DISCUSSION: ISSUES IN EVALUATING CREW PERFORMANCE IN LINE ORIENTED EVALUATION

R. Key Dismukes
NASA Ames Research Center

In this discussion I will try to put the papers in this panel session in a larger context of issues raised by Line Oriented Evaluation (LOE). I will give you my personal opinion of what questions should be addressed and what we know so far about the answers.

The FAA's Advanced Qualification Program (AQP) is forcing a new and deeper look at CRM (FAA, 1991; Longridge, 1998). Heretofore, crews have practiced CRM in LOFTs that were non-jeopardy. However AQP requires participating airlines to formally evaluate CRM in Line Oriented Evaluations (LOE). Many conceptual and methodological issues are raised whenever we attempt to measure any aspect of human behavior. We can no longer get by with fuzzy concepts and uncritical assumptions about the relationships among classroom discussion of CRM, simulation practice of CRM, and real-world performance in the cockpit.

The framework for LOE was initially set out by innovative and thoughtful work by several individuals collaborating from their respective organizations (George Mason University, 1996; Seamster, Edens, McDougall, & Hamman, 1994). This framework consists of simulations designed around a series of event sets in which crews encounter situations designed to test their skills along specific CRM and technical dimensions. I am very impressed by this work. At the same time I feel that many questions remain to be answered and I hope that the implementation of AQP will not freeze the methodology before these questions can be answered.

LOE must be meaningful, valid, and reliable, and research is required on all three aspects. Meaningful: does LOE deeply probe crew performance across the range of real-world aviation operations? Valid: does LOE actually measure what we intend it to measure? This is an especially challenging question because we are in virgin territory with LOE. Normally a new instrument to measure behavior is validated by comparing it to existing measurement instruments but to what should we compare our results with LOE?

Reliability has been studied to a fair degree. Much of the research on LOE has focused on interrater reliability: Will two raters give the same rating for a given example of crew performance? However, there are other important aspects of reliability. For example, will we get the same result from two different scenarios designed to measure some particular aspect of crew performance, such as workload management? We would also like to know whether we would get the same result if we retested the crew at some later date.

With these three principles in mind -- validity, meaningfulness, and reliability -- I have framed what I consider the eight most critical questions about LOE. The first question has already been addressed fairly substantially in the initial development of LOE, however the remaining seven questions require considerably more research. Some of them have not been addressed at all.

To what extent does a given event-set scenario, designed to test a specific aspect of CRM, probe crew performance under the range of conditions which the crew might encounter in the cockpit? The original developers of LOE set out detailed methods by which scenario developers could design LOE scenarios that probe dimensions of CRM and technical performance critical to managing the kinds of situation that arise in line operations (Seamster et al., 1994).

If we designed several different LOE scenarios to probe, say, decision-making, to what extent would we see the same level of decision-making performance by a given crew on each of the several scenarios? In fact, we must develop new scenarios fairly frequently so crews do not know what to expect during recurrent LOEs. Underlying this question are two subordinate questions: (i) To what extent does crew performance on a particular dimension generalize? Does performance hinge more on the general design of the event set or on the specific details of the event set? (ii) To what extent does the evaluator's observation and evaluation vary with the details of the event set scenario? This question also raises the issue of whether evaluators will have to be trained and calibrated specifically on every LOE scenario they use.

To what extent do evaluators' skills on a given LOE scenario generalize across crews? Although each crew encounters the same events in the scenario, individual differences among crews generate many subtle and complex variations in how they respond. One crew's performance may be much harder to evaluate than another's even though the two crews are equally effective.

How stable is crew performance on a given dimension over time? Are there random variations in crew performance on a given dimension? Does performance systematically decay over time? How often do we need to test crew performance? Unfortunately we cannot test this directly because we cannot run the crew in a given scenario more than once because the effectiveness of a simulation scenario depends on the crew not knowing exactly what will happen. However if we could first show that different scenarios can consistently measure performance on a particular dimension then we could address this question of stability of performance over time.

How stable are evaluators' observation and evaluation skills over time? We must address this question in order to know how often to retrain or recalibrate evaluators.

To what extent is crew performance influenced by the unique interaction of two particular individual pilots? For example, can we assume that a first officer who is appropriately assertive with one captain will be appropriately assertive with other captains?

How predictive is LOE performance of real-world performance? If a particular crew does well or not-so-well along particular CRM dimensions in the simulation can we assume they will do much the same in the real cockpit? One hears a wide range of speculation on this topic but little data exists to answer the question. It would be very interesting to do a study in which crews were given both an LOE and a line audit of the sort conducted by Helmreich and his colleagues (Klinect, J.R., Wilhelm, J.A., & Helmreich, R.L., in press) and measure the degree of correlation.

How good is interrater reliability in LOE? Several studies have addressed this question, and I will talk about this in more detail.

David Baker and Casey Mulqueen (these proceedings) have just presented what I regard as a very sensible set of guidelines for training evaluators to rate crew performance accurately. They have drawn upon evaluator training studies in fields outside of aviation that have been studied more extensively than has LOE. The guidelines, though sensible, are also necessarily rather general. To implement these guidelines we will need to make detailed decisions and these decisions will require answers to at least some of the questions I have raised. For example, to provide frame of reference training we need to decide how many different crews and how many different versions of a scenario evaluators need to see. How often do the evaluators need to be recalibrated?

To date several preliminary reports have been published of studies indicating that it is possible to achieve a fairly high level of agreement among evaluators (George Mason University, 1996; Law & Sherman, 1995; Williams, Holt, & Boehm-Davis, 1997). This is very encouraging, however we will have to wait until the full details of the methods are published to know exactly how to interpret the results.

One question that the authors of these studies have themselves raised is on what basis are evaluators making their overall assessments of CRM and technical performance. In the most commonly used LOE grading system, evaluators give an overall assessment of CRM performance on a particular event set and an overall assessment of technical performance. The evaluators also judge whether the crews exhibited specific CRM and technical behaviors listed on the grade sheets. The specific observable behaviors are chosen by the LOE designers to represent behaviors crucial to success in the scenario. However, the data published so far reveal only very slight correlation of the observable behaviors with the overall CRM or with the overall technical ratings. Thus the evaluators appear to agree but it is not clear what is driving their ratings of the crews.

Perhaps the most critical question in the interrater reliability area is this: If a particular crew performs at a less than adequate level on an event set, what is the probability that a randomly assigned evaluator will rate that performance as inadequate? The converse question is: What is the probability that a randomly assigned evaluator will rate as inadequate a crew that really performed adequately, if the truth were known? It would be useful if investigators would analyze and report their data in these terms.

I will conclude with some general observations. The first point is what I call "the curse of aviation psychology". We all make our 15 minute presentations at this symposium and publish the short proceedings articles, which is fine, but these formats do not provide enough detail about methods for other scientists to evaluate research design and to understand the extent to which a study's conclusions generalize. In reviewing LOE research reports I found it difficult to assess the state of the field because so little of the work has been published yet in full form. This problem is in no way unique to LOE research. We need to encourage timely publication of full details of research in all areas of aviation psychology, especially since policy decisions are frequently made on the basis of preliminary reports.

To properly design research to answer the questions I have raised here places a heavy demand for research subjects, in this case airline instructors and pilots. Several airlines have made important contributions by providing access to their instructors and pilots.

Nevertheless there are practical limits. For example, it would be very hard for an airline to provide 30 hours of time from each of 50 instructors, yet in some cases that is roughly what it would take to design a study adequately to answer the research question. Instead of running a large number of studies to answer questions one at a time, perhaps we need large-scale collaborations to conduct a large multi-factorial study that would address several of these questions simultaneously. Also we might think about seeking additional sources of research subjects. For example, Navy training researchers have done quite a bit of valuable work paralleling LOE research using Navy personnel (see, for example, Brannick, Prince, Salas, & Stout, 1995).

I am very positive about LOE because I feel it can lead us to a deeper understanding of CRM issues and more powerful approaches to training CRM and practicing CRM on the line. However, given the very incomplete state of our knowledge we do need to exercise some caution in interpreting the performance of crews in LOE, especially when making policy decisions.

REFERENCES

Brannick, M., Prince, C., Salas, E., & Stout, R. (1995). Assessing aircrew coordination skills in TH-57 pilots. In R. S. Jensen & L.A. Rakovan (Eds.), *Proceedings of the Eighth International Symposium on Aviation Psychology* (pp. 1069-1071). Columbus: The Ohio State University.

FAA (1991). Advisory Circular 120-54: Advanced Qualification Program. Washington, D.C.: Federal Aviation Administration.

George Mason University (1996). Developing and Evaluating CRM Procedures For A Regional Air Carrier, Phase I Report. Washington, D.C.: Federal Aviation Administration.

Klinect, J.R., Wilhelm, J.A., & Helmreich, R.L. (in press). Threat and error management: Data from line operations safety audits. In R. S. Jensen (Ed.), *Proceedings of the Tenth International Symposium on Aviation Psychology.* Columbus: The Ohio State University.

Law, J. R. & Sherman, P.J. (1995). Do raters agree? Assessing inter-rater agreement in the evaluation of air crew resource management skills. In R. S. Jensen and L.A. Rakovan (Eds.), *Proceedings of the Eighth International Symposium on Aviation Psychology* (608-612). Columbus: The Ohio State University.

Longridge, T. M. (1998). Overview of the advanced qualification program (on-line). Federal Aviation Administration, Washington, D.C. http://www.faa.gov/avr/afs/tlpaper.htm.

Seamster, T.L., Edens, E.S., McDougall, W.A., & Hamman, W.R. (1994). Observable Crew Behaviors in the Development and Assessment of Line Operational Evaluations (LOE's). Washington, D.C.: Federal Aviation Administration.

Williams, D., Holt, R., & Boehm-Davis, D. (1997). Training for inter-rater reliability: Baselines and benchmarks. In R. S. Jensen (Ed.), *Proceedings of the Ninth International Symposium on Aviation Psychology* (pp.514-519). Columbus: The Ohio State University.