Scientific Evaluation of Aircrew Performance

George Mason University Research Team

Historically, pilots have been evaluated for their technical proficiency. More recently, pilots and crews have also been assessed on the quality of their CRM performance at some carriers. Interestingly, the aviation community has been sensitive to the difficulties in fairly and accurately measuring CRM performance, despite the fact that the same issues exist in the assessment of technical performance.

In general, scientifically evaluating performance requires that a measurement system demonstrate *reliability* and *validity*. *Reliability* refers to a consistency in measurement or a lack of measurement error. As an example, a bathroom scale should always report the same weight of the person being weighed. If you were to weigh yourself twice one morning and the weights reported were very different, the scale would be considered unreliable and unfit for measuring your weight. *Validity* refers to the accuracy of measurement; that is, you need to be measuring what you intend to measure. The weight given by your bathroom scale should correspond to the weight measured on the same day by a doctor's professional scale. If the bathroom scale reports a different weight from the criterion of the doctor's scale, it is invalid. Since validity is often harder to establish, the reliability of a measure is usually established first followed by validity.

Scientists are concerned that measures of both CRM and technical performance meet high standards of reliability and validity. Reliability and validity can be estimated in several ways. The various methods may result in slightly different values, hence it is important to combine information gleaned from different methods and samples of aircrews. Across these methods, there are some standards for the reliability and validity of measures. One standard for reliability for measures used in research is a value of about 0.70 where the minimum value is 0.0 (no reliability) and the maximum value +1.0 (perfect reliability).

Over the last decade, there have been a variety of research projects that have measured CRM and technical performance of aircrews. The reliability and validity estimates for evaluating CRM and technical performance from these research projects can be collected, summarized, and compared. This information should clarify the extent of scientific justification for measuring each area of aircrew performance. Additionally, careful inspection of the methods used in the studies with the best results should point to the best measurement methods. One way to easily compare these results is to put them in tables.

The table on the next page comes from research conducted by the George Mason University research team over the last three years. This research involved both a phase/task specific form of CRM that was implemented at a regional carrier and a global (phase independent) form of CRM training at a major domestic carrier. In both cases, CRM was primarily evaluated at a detailed level in Line Oriented Evaluation (LOE) event sets. Technical performance was evaluated both in maneuver validation training, and,

separately, as part of the LOE. Results using the same type of measures and estimates on the same sample of pilots are most directly comparable.

The basic point that the table makes is that some values of the reliability of the measures of CRM performance are as good as the reliability values for measuring technical performance. For example, the agreement index on overall CRM for 5 event sets is .86 for the regional carrier, which is as good as the agreement index obtained for the overall Technical performance for the same 5 event sets of .85. Similarly, the internal consistency estimate for CRM for 11 event sets (.75) is slightly higher than for the consistency of overall technical ratings on those event sets (.66). ***These data suggest that we can measure CRM performance with at least the same degree of consistency as we can measure technical performance.***

)

Reliability Estimates

Performance Type: CRM/Technical	Aviation Domain:	Type of Evaluator	Type of Evaluation	Type of Reliability estimate	Reliability Value
CRM	Civilian/Regional	Instructor/Evaluators	Inter-Rater Reliability training For LOE evaluations	Agreement Index : r_{wg} on Overall CRM for 5 event sets	.86
CRM	Civilian/Regional	Instructor/Evaluators	Line Operational Evaluation: overall CRM evaluation for each event set	Internal Consistency of 11 Overall CRM ratings	.75
CRM	Civilian/Major	Standards Captains	Line Operational Evaluation: Topic level evaluations for each event set	Internal Consistency of 20 topic ratings	.88
Technical	Civilian/Regional	Instructor/Evaluators	Inter-Rater Reliability training For LOE evaluations	Agreement Index : r_{wg} on Overall Tech for 5 event sets	.85
Technical	Civilian/Regional	Instructor/Evaluators	Line Operational Evaluation: 2 overall Technical evaluations for each event set	Internal Consistency of 22 Overall Technical ratings	.66
Technical	Civilian/Regional	Instructor/Evaluators	Maneuver Validation: Maneuver Performance as Pilot Flying	Internal Consistency of 4 First Look and 6 Fixed maneuvers	.78

)