

APPLICATION OF PSYCHOMETRICS TO THE CALIBRATION OF AIR CARRIER CHECK AIRMEN

Robert W. Holt
George Mason University
Fairfax, Virginia

Peder Johnson
Timothy Goldsmith
New Mexico State University
Albuquerque, New Mexico

The FAA's Advanced Qualification Program (AQP) encourages airlines to implement proficiency-based training programs and requires collection of reliable and valid performance assessment data. We present applications of traditional and innovative psychometric methods to this domain.

A primary goal of the FAA in establishing the Advanced Qualification Program (AQP) was to encourage airlines to implement proficiency-based training programs. Such programs focus on the collection of empirical data that will allow the proficiency of crews to be validly assessed and continuously monitored. Crew proficiency is defined by an explicit and systematic set of performance objectives. The collection and analysis of quality data is integral to the success of these programs.

Psychometric methods have historically been used to assess and ensure the quality of subjective measurements. The present paper describes our efforts to apply traditional and innovative psychometric methods to assess AQP data quality. We focus specifically on assessing the reliability, sensitivity, and validity of evaluator's judgments of crew performance in high fidelity simulations.

Approaches to assessing Reliability

Reliability is the amount of systematic variance in a measure. We have elaborated traditional approaches to assessing and training inter-rater reliability, and developed an innovative approach to reliability assessment using an external referent.

Inter-Rater Reliability

Inter-rater reliability (IRR) was developed from traditional approaches to ensuring rater reliability which use the set of group judgments as the standard for assessing and training each evaluator (Holt et. al. HF, 1996). Traditional approaches have emphasized either inter-rater consistency, often estimated by inter-rater correlations across items, or inter-rater agreement, often estimated by $r(wg)$ (ref). These approaches were combined and augmented with a systematic analysis of the distribution of a rater's judgments to give more complete information about rater reliability.

The distribution of a rater's judgments is partly important because it can limit the maximum possible values of inter-rater correlations and agreement. If the rater has a positively skewed distribution while the group distribution is negatively skewed, the maximum possible value of the average inter-rater correlation for this evaluator is decreased. Similarly, such distinct judgment distributions will necessary decrease the possible values of $r(wg)$. Therefore, shape of each rater's judgment distribution is relevant for IRR. A rater's judgment distribution can be compared to the group distribution in two conceptually distinct and meaningful ways.

Conceptually, when rating an equivalent set of stimuli a rater should not have a systematically different average evaluation than the group. The rater's distribution can be compared to the group average distribution with a t-test. If significantly high or low, raters must understand why they have lenient or harsh ratings and adjust the mean tendency of their ratings. A general preliminary test of systematic mean differences in a group of raters is available from an analysis of variance of ratings using the "rater" as one independent variable.

Since comparing the mean only compares one aspect of a rater's distribution to the group distribution, this logic can be extended by comparing the variance, skewness, and so forth of the rater's distribution to the group distribution. To simplify this process, a congruency index was defined which includes all such aspects of the distribution and has a range of values from 1.0 (perfect congruency) to 0.0 (random congruency) and -1.0 (completely contrasting judgment distributions). This index is one minus the sum of the absolute values of the discrepancy in judgment probabilities of the rater vs. the group, where the sum is across all scale categories. The average inter-rater correlation, the systematic differences t-test, and the congruency index for each person as well as the agreement indexes on each item are used for rater feedback and training.

Reference Reliability

- Referent normed

- Construction of referent or "Gold Standard"

- Comparison of evaluator judgments to referent:

 - Dichotomous judgments

 - Percent Agreement

 - Signal Detection Theory d'

 - Scale judgments

Approaches to Assessing Sensitivity

Conceptually, sensitivity is extent to which real variability in performance is reflected in variability in the evaluations of each rater. That is, sensitivity is the ability of each rater to discern fine gradations in performance and appropriately assign distinct ratings to each level of performance. In AQP, sensitivity of discerning and assigning

different ratings to unsafe vs. safe levels of crew performance is critical for detection and remediation of unsafe crew performance. Furthermore, sensitivity in discerning and rating different gradations of safe performance is important for detecting subtle trends or shifts in performance over time that have training implications. Within reliability and validity constraints, the sensitivity of a multi-point rating scales can be higher than a dichotomous rating and enable more precise delineation of gradations or shifts in performance.

Assessing sensitivity of judgment requires first establishing known differences in evaluated performance on videotaped flight segments. Subject matter experts (SMEs) evaluate overall performance levels of each videotaped segment. Representative samples of High, Medium, and Low performance are presented to the group of raters for evaluation.

To create a meaningful index of sensitivity for rater feedback and training, each rater's evaluations of different performance levels are analyzed with an Analysis of Variance (ANOVA). Based on the results of the ANOVA, Hays' (Ref) omega-squared strength-of-effect index is computed based on the expected mean squares for the ANOVA. Values for this index range from essentially zero if the rater's judgments show no discrimination of the different performance levels to one if the rater's judgments perfectly discriminate the different performance levels with almost no error.

Approaches to Assessing Validity:

Validity is the extent to which the variability of the measure reflects variability in the targeted construct as opposed to extraneous or random variability. Traditional validity concepts emphasizing the relationship of a measure to other variables can be augmented with the use of internal evidence concerning the judgment process.

Internal evidence of validity

If a theory or systematic set of expectations can be developed for the judgment process, evidence that the structure of relationships among the judgments fits the specified pattern is evidence for validity. Conceptually, this process is similar to confirmatory factor analysis (Mulaik, ref). The stages or flow of the judgment process can be mapped with structural equation modeling (SEM). Alternatively, a specified pattern of relationships among sets of variables can be confirmed or disconfirmed with normal statistical techniques such as multiple regression.

For example, suppose the raters have been trained to use a specific judgment sequence or process which progresses from behavioral observations to judgment dimensions of performance and finally to overall evaluations for each person. Path analysis or SEM can be used to track the predicted judgment sequence. Strong relationships should occur from each stage of judgment to the next and support validity. Conversely, not finding the predicted structure of relationships or finding extraneous, non-mediated relationships among ratings from very different stages of inference is evidence against validity. Furthermore, these analyses have also been used for additional feedback and training of the decision process of aircrew evaluators.

External evidence of validity

External evidence of validity requires specifying the theory upon which crew performance assessments are based. The basic theory underlying an LOE is that the LOE measures general and stable skills/abilities that underlie individual and crew performance. As Nunnally and Berstein (Ref) discuss, several types of external validity are relevant: content, predictive and construct validity.

Content Validity. An LOE is initially developed to have appropriate content. That is, SMEs develop the content of the LOE and the content of the assessment instruments such as worksheets to be applicable to actual flight operations. The domain that LOE is attempting to assess is a rather large both in the scope of situations comprising the task (e.g. phases of flight, types of operations) and in assessing both the technical and crew resource management skills (CRM) aspects of the crew performing the task. For such a large and fuzzy domain, there are likely to be a multitude of measures of the domain, some of which will not correlated very highly with one another, which would reduce internal consistency reliability.

Content validity of an LOE should be evaluated by the extent to which the LOE content adequately samples the performance domain. The large, fuzzy performance domain precludes an exhaustive delineation of domain content and empirical assessment of content validity. Using the more general perspective that the airlines' domain of interest is the safe and efficient operation of the aircraft, expert judgments can be used to ensure that the LOE's sample of required behavior is highly similar to behavior required for safe and efficient operation of the aircraft in normal and abnormal situations.

Predictive Validity. One of the most direct means of demonstrating validity is to show that the measure predicts an appropriate external criterion. The LOE is specifically designed and assumed to measure CRM and technical skills under abnormal operating conditions. Therefore, an appropriate external criterion should pertain directly to CRM and technical skills used under abnormal (e.g., high workload) flight conditions.

Maneuver validations are high workload but emphasize technical proficiency and do not have a strong CRM component. Line-check evaluations (where an evaluator observes the crew fly an actual flight from beginning to end) involve technical and CRM skills, but are almost always assessed under normal rather than abnormal flight conditions. If flying under normal and abnormal conditions requires some different pilot abilities, the correlation between LOE and line-check performance will be low. Thus, neither maneuver validations nor line-checks by themselves are acceptable external criterion for LOE performance.

Construct Validity. Since no single external criterion can completely validate LOE performance, we contend that a broad construct validity approach is necessary. The LOE is intended to assess multiple facets of performance that may have a wide variety of manifestations. Thus, the manifestations of each facet may be expected to have only a moderate to low correlation with LOE performance. However, the total pattern of relationships of the measures we propose below can help establish the construct validity of LOE performance.

In an LOE, specific CRM skills (e.g., workload management, situational awareness, decision making, etc.) and technical skills may be evaluated by multiple items

across event sets. The magnitude of the intercorrelations of items measuring the same construct across different event sets is evidence for convergent validity. Since some sets of skills may be relatively independent (e.g. CRM and technical skills), scores from items measuring these skills can be compared in a multi-trait multi-method matrix or equivalent confirmatory factor analysis technique to establish divergent as well as convergent validity.

If the LOE measure is valid, differences in levels of averaged performance across CRM elements should correspond to the incidence of certain types of problems as reflected by other measures (e.g., line check, reported incidents, etc.). In other words, CRM or technical problem areas identified with the LOE data should correspond to problem areas observed with other measures.

Since CRM performance depends to some extent on both procedural/skill knowledge and declarative knowledge, there should be a correlation between LOE CRM performance and declarative knowledge of CRM. CRM knowledge could be assessed by a separate oral or written test. LOE CRM performance should significantly correlate with this knowledge test.

Maneuver validations are intended to assess pilot's ability to perform specific technical maneuvers. Since these maneuvers are executed under abnormal or emergency situations (e.g., executing a VI cut) this performance should moderately predict the technical skills assessed on the LOE. These scores should also predict the overall LOE score to the extent that the overall LOE score depends on assessed technical skills.

Summary

Assessing safety-critical performance requires high levels of reliability, sensitivity, and validity. Both traditional psychometric methods must be applied wherever possible and innovative psychometric methods developed for unique requirements of each domain. The FAA's AQP program has fostered the development of new approaches to traditional psychometric methods and innovative methods which could be used in other domains.

References