

FOR REFERENCE ONLY

Application of Psychometrics to the Calibration
of Air Carrier Evaluators

Robert W. Holt
George Mason University
Fairfax, Virginia

Peder J. Johnson
Timothy E. Goldsmith
University of New Mexico
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) is 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 proportion 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 (IRR)

IRR components. 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, Meiman, and Seamster, 1996). Traditional approaches have emphasized either inter-rater consistency, often estimated by inter-rater correlations across items, or inter-rater agreement, often estimated by $r(wg)$ (Law and Sherman, 1995). These indexes are fundamentally different since the inter-rater correlation is the Pearson correlation calculated for pairs of raters across items while the $r(wg)$ index compares the sample variance of each item to the variance of a uniform distribution across the same number of scale points. These two 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 relative to the case where both distributions have the same shape. Similarly, such distinct judgment distributions will necessarily 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 distinct and meaningful ways.

First, when rating the same stimuli or an equivalent set of stimuli, a rater should have the same average evaluation as the other raters. That is, a rater should not be systematically harsher or more lenient than the group. Each rater's average can be compared to the group average with a t-test. If significantly high or low, raters must understand the source of their 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.

IRR training. Rater training based on the above components plus sensitivity (discussed below) was used for calibrating air carrier check airmen at two air carriers (Williams, Holt, and Boehm-Davis, 1997). Individual rater feedback included congruency, systematic differences, average inter-rater correlation, and sensitivity. Group level feedback included group averages for these four individual indexes and group agreement for each item. Group discussion and resolution of different rating standards and processes was used to increase reliability.

Referent Reliability

A second method of examining evaluator reliability involves first defining a standard of performance, which we call a referent, for each crew behavior and then comparing an evaluator's judgment of these behaviors to this standard (Goldsmith & Johnson, 1997). Referent reliability is fundamentally distinguishable from inter-rater reliability in relying on the comparison of each rater's judgments to an external referent rather than the group norm. This external referent is sometimes called a "Gold Standard" for judgments.

This approach to examining evaluator reliability is possible within the piloting domain because qualification standards have been defined for all relevant air crew

FOR REFERENCE ONLY

behaviors as part of the AQP development process. Indeed, it is these performance standards that crews are trained to achieve and these same standards that evaluators are trained to recognize and use to evaluate crew behaviors.

Using a referent to assess evaluator reliability has three important implications. First, high referent-rater reliability results in high inter-rater reliability. If evaluators are judging crew behavior in accordance with the same set of standards, then they will necessarily agree with another. Second, the distribution of referent scores defines the appropriate distributional characteristics of evaluator scores. Hence, by training to match the referent judgments, evaluators can be trained to match the mean, variance, and skewness of the referent distribution. Third, a major advantage of referent reliability over inter-rater reliability is that the problem of an incorrect group norm for judgments is avoided. That is, in the case where the majority of raters are rating incorrectly, the other raters will not be trained to this incorrect standard.

The major disadvantage of referent reliability is the extra effort required to construct such a referent. Since each rater judgment is compared to the corresponding judgment in the referent, the construction process must ensure that every judgment in the referent is precisely correct. Construction of the referent relies partly on selecting appropriate judgment materials or stimuli for which a small group of very expert evaluators can have very congruent and consistent judgments. The degree of agreement and consistency of the expert evaluators must be very high to justify the use of the referent in rater training. The resulting referent is conceptually similar to an external criterion for establishing criterion validity except that (1) the referent is established to a very high level of precision, and (2) the referent is established on a limited subset of cases or stimuli which can be judged unambiguously rather than covering the entire range and domain of performance.

To ensure calibration of air carrier check airmen, the check airmen judgments must be systematically compared to the referent. To the degree that noticeable discrepancies are found, the discrepant rater must be trained. Different approaches to comparing rater judgments with the referent are possible.

Traditionally, the Pearson correlation coefficient has been used to measure rater reliability. Although the correlation coefficient has several properties that make it a desirable measure, the fact that it measures relative rather than absolute consistency can sometimes be a disadvantage. For example, an evaluator who always grades air crew behaviors one score value less than the referent will show perfect relative consistency. A measure that assesses absolute consistency, such as mean absolute difference, corrects for this problem. If mean absolute difference from the referent is low (i.e., good match to the referent), the correlation coefficient will necessarily be high. Hence, mean absolute difference is a more stringent measure of consistency with the referent and so should have advantages over the correlation coefficient in calibrating evaluator judgments.

One thorny issue that arises in assessing evaluator reliability is the base-rate problem. The base-rate problem occurs when the frequency of some grades is very different from others. For example, on a 4-point grading scale where a "4" is excellent and a "3" is standard, it is possible for upwards of 90% of crew behaviors to be graded a "3". Of course, evaluators are cognizant of this fact and by simply selecting the modal grade can achieve relatively high agreement. Clearly these distributions can influence simple agreement measures such as percent agreement. However, such non-normal

distributions could also make it difficult to interpret the real meaning of a correlation of say $r=0.80$ between two evaluators.

Goldsmith & Johnson (1997) carried out a series of Monte Carlo studies to analyze the extent of the base-rate problem under conditions similar to realistic evaluator data. They compared a signal detection theory analysis using d' to a correlation coefficient for various types of skewed distributions. They also examined the expected agreement under chance alone between sets of scores with various types of skewed distributions and reported the resulting agreement values as measured by a correlation coefficient, mean absolute difference, and d' . In general, the results indicated that highly skewed distributions may result in unstable measures of agreement particularly with dichotomous score values.

In addition to mean absolute difference and signal detection methods, the IRR methods discussed above can be used to compare each rater's judgments to the referent. That is, the analyses discussed in the IRR section can be used to check if the rater (1) has a congruent distribution of judgments with the referent, (2) has harsher or more lenient average judgments than the referent, (3) is consistently shifting up and down with the referent, and (4) is sensitively distinguishing the different levels of performance specified by the referent.

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. Although conceptually similar to the process for establishing a referent, these judgments focus on a global assessment of the overall level of performance for large performance segments rather than the precise correct judgment for each scale. Using these global judgments, 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' (1981) 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 only chance-level discrimination of the different performance levels to 1.0 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 logic of this approach is similar to confirmatory factor analysis (Mulaik, 1972). 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 basic 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. This pattern of results supports validity (Holt, Meiman, and Seamster, 1996). 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.

Rater training. Since the internal validity analyses should reflect the judgment process, the results of these analyses can be used for feedback and training of the decision process of air carrier check airmen. For example, at one carrier the rater must first observe specific behaviors and then combine these observations into technical or CRM dimension judgments. This step of the judgment process was analyzed with multiple regression to empirically establish the relative weight of each observable behavior for the dimension judgments. In training, path diagrams based on the multiple regression results were used to inform raters about the relative weights for combining behavioral information into dimension judgments. Results inconsistent with group expectations were the focus of group discussion with the goal of training a common information-processing schema.

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 (1978) discusses, 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, situation 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,

FOR REFERENCE ONLY

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 V1 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. To ensure high quality performance evaluations, traditional psychometric methods must be applied wherever possible and innovative psychometric methods developed for the unique requirements of each domain. The FAA's AQP program has fostered the development of new approaches to traditional psychometric methods and innovative methods for establishing the calibration of evaluators of air crew performance. These methods can be used in any other domain which emphasizes the quality or process of complex team performance and which relies on expert raters for performance evaluation.

References

Goldsmith, T. E. & Johnson, P. J. (1997). Assessing Instructor/Evaluator Reliability under AQP. Paper presented at the meeting of the International Symposium on Aviation Psychology, Columbus, OH

Hays, W.L. (1981) Statistics. New York: Holt Rinehart and Winston.

Holt, R.W., Meiman, E., and Seamster, T.L. (1996) Evaluation of aircraft pilot team performance. In Proceedings of the 40th Annual Meeting of the Human Factors Society, Vol. 1 (pp. 44-48). Santa Monica, CA: The Human Factors Society.

Law, J.R., and Sherman, P.J. (1995) Do raters agree? Assessing inter-rater agreement in the evaluation of air Crew Resource Management skills. In Proceedings of the Eighth International Symposium on Aviation Psychology. Columbus, Ohio: Ohio State University

Mulaik, S.A. (1972) The foundations of factor analysis. New York: McGraw-Hill.

Nunnally, J. (1978) Psychometric Theory (2nd ed.). New York: McGraw-Hill

Williams, D.M., Holt, R.W., and Boehm-Davis, D.A. (1997) Training for inter-rater reliability: baselines and benchmarks. In Proceedings of the Ninth International Symposium on Aviation Psychology. Columbus, Ohio: Ohio State University.