As Assessment of Instructor/Evaluator's Inter-rater Reliability: Judgments of Crew LOE Performance

Timothy E. Goldsmith and Peder J. Johnson
University of New Mexico

Fred Kirijan and Vince Mancuso
Delta Airlines

The FAA established the Advanced Qualification Program (AQP) with a primary goal of allowing airlines to develop true proficiency-based training programs. Here proficiency is defined by an explicit set of performance objectives that are systematically developed and then continuously validated throughout the collection and evaluation of empirical data. Data management (collection and analysis) is, therefore, an integral part of AQP. To succeed, there must also be a process for demonstrating the reliability of these data. The present study was specifically conducted to assess the reliability of methods used by a particular fleet within a major carrier. More generally, the study may serve as an example, demonstrating a process or set of procedures that may be used for assessing the reliability of Instructor/Evaluators (I/E) scoring of grade sheets used for evaluating a flight crew's performance during an LOE.

Method

A group of twenty-eight I/E's at a major carrier were presented a 25 minute video of a crew flying an LOE and asked to complete the grade sheet developed for this particular flight scenario. The flight scenario was divided into three event sets. At the beginning of each event set the chief trainer for this fleet set the context by explaining to the I/E's where in the flight scenario the video was beginning and what would be occurring in the upcoming event set. As will be discussed later, most of the I/Es were quite familiar with the particular LOE used in the present video.

The first event set was marked by an engine failure after V1, the second involved an engine out Cat II missed approach, and the third involved a retry with a visual approach with limited fuel and a tailwind. Across the three event sets there were a total of 29 decision/actions (referred to as success criteria) that were scored as performed or not performed. A more detailed analysis of the composition of the LOE is presented in Table 1.

We should emphasize that this was the first experience that the I/Es had with the new grade sheet and the 2-point scoring procedure (i.e., observed versus non-observed). Previously the fleet had used a 4-point scale (4 = outstanding; 3 = average or acceptable level of performance; 2 = a weak pass; 1 = failure). The I/Es were instructed that in using the 2-point scale they should score what in the past was a 2, 3, or 4 as a plus (observed), and what was previously scored as a 1, was now a minus (non-observed). The I/Es were instructed to make these entries on the Grade Sheet at the time they normally completed this task when evaluating an LOE (i.e., during or at the completion of the LOE).

Table 1
Composition of LOE

	No. of Criteria	No. Missing	Criticality Scores	Global Rating
Event Set 1	13	3	3 3 3	3
Event Set 2	9	2	3 2	2
Event Set 3	7	1	3	3
Total	29	6		

Note. No. of Criteria refers to the number of success criteria in each event set.
No. Missing refers to the number of success criteria that were not performed by the crew flying the LOE.
Criticality Scores refer to the criticality of the missing success criteria where:
1: indicates that failing this success criterion could result in the loss of life or injury to passenger or crew; 2: indicates that failure could result in damage to the craft and danger to the crew and passengers; 3: indicates that failure to exhibit this success criterion would result in unnecessary difficulty or inefficiency, but would not result in damage to passengers, crew, or craft); 4: reflects proper performance of the task.
Global Rating is the overall rating on a 4-point scale (1 = unsatisfactory; 2 = acceptable; 3 = standard performance; 4 = above average) for each event set.

In addition to completing the grade sheet, each I/E provided basic demographic information on a separate form. This information, along with the grade sheets, was collected at end of the session. The session took approximately one hour to complete.

Results

The demographic information revealed that 54% of the I/E's were captains and the remaining 46% were first officers. Participants also indicated whether they completed LOE work sheets during or after completion of the simulated flight. A slight majority, 57%, reported normally completing the grade sheet during the LOE, and the remaining 43% reported completing the form after the simulated flight. The means and standard deviations for the remaining demographic information collected is summarized in Table 2.

Table 2
I/E Demographic Characteristics

	Mean	SD
Flying Hours Commercial Air	8523	4435
Years with current Carrier	16.3	7.2
Years with current Fleet	5.2	2.4
No. months as I/E	38.8	40.6
Familiarity with LOE (5=high)	4.1	1.3

Table 2 shows that most of the I/Es were very experienced pilots, plus having flown with the current carrier for a considerable time. Although it appears that these pilots were also experienced I/E's (averaging approximately 40 months), the standard deviations (40.6) show a considerable variance on this factor. In fact, eleven pilots had been an I/E for five or fewer months. Finally, as alluded to earlier, the I/E's, as a group, were familiar with the LOE used in the present study. On a 5-point scale of familiarity (5 = high), the mean was 4.1.

Turning to the grade sheet, we first assessed inter-rater reliabilities by computing the correlation between all pairwise combinations of the 28 I/E's ratings on the success criteria. From this 28 X 28 matrix of Pearson correlations we computed each I/E's mean inter-rater reliability by simply averaging across the remaining 27 correlations (i.e., that I/E's correlations with each of the other 27 I/E's). Figure 1 presents a frequency distribution of the inter-rater reliabilities. Reliabilities ranged from .00 to .52, with a mean of .38. Although the overall mean inter-rater reliability was quite low, this was in part related to four I/E's with reliabilities below .30. When these four I/E's were removed from the analysis the mean increased to .46. Finally, we re-computed the reliabilities by correlating each I/E's ratings with a referent (the correct judgment as determined by the developers of the success criteria used on the grade sheet). The mean correlation with the referent was .47, suggesting that the I/E's generally agreed with a referent more than they did with one another. This is not surprising given that the sample contained several I/E's whose ratings were apparently inconsistent with most of the other raters. These individuals tended to lower the inter-correlations for the better raters.

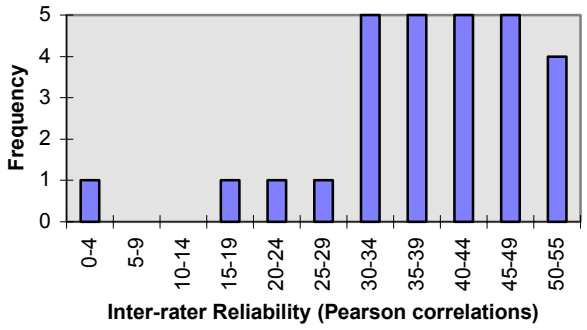

Figure 1. Frequency of Mean I/E Inter-rater Reliabilities

It is generally recognized that Pearson correlations tend to be rather unstable when computed on dichotomous data. The problem is magnified when the distribution is heavily skewed (i.e., towards all 0's or all 1's). This raises a question regarding the interpretation of the above inter-rater reliability scores. Specifically, how much disagreement does a Pearson correlation of .5 imply? Because of these concerns we explored some alternative methods of estimating inter-rater agreement. After comparing different statistics we concluded that percent agreement with the referent ratings most clearly communicated I/E's performance on the present

grade sheet[1]. Using this referent as defining the correct judgment the percent correct was computed for each of the 29 success criteria.

Figure 2 presents the frequency distribution for percent correct. The overall mean percent correct was 79.6%. If the three outliers at or below 62% correct are excluded from the analysis the mean percent correct increases to 82.2%. Recall that the mean Pearson correlation with the referent was .47, illustrating that a reasonably high percent agreement results in a relatively low correlation statistic with the present data.

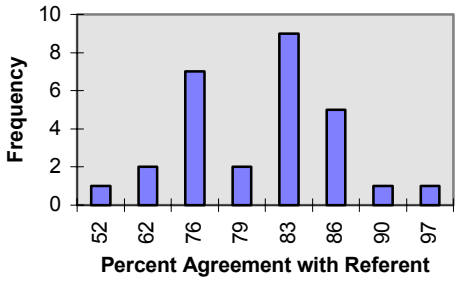

Figure 2. Frequency Distribution of I/E's Agreement with Referent

These results presented in Figure 2 were next organized, first in terms of whether the correct judgment was a plus (observed, Table 3) or a minus (non-observed, Table 4). Within each of these tables the success criteria were then rank ordered in terms of level of performance, from lowest to highest.

The overall mean performance for the observed criteria was 83% correct. Thus, on 17% of their judgments, I/E's incorrectly concluded that a success criterion did not occur when it fact it had occurred. The mean percent correct for criticality scores of 1, 2, and 3 were 83%, 85%, and 81%, respectively, indicating that performance was not influenced by the importance of the consequences of a success criterion. Note also that many of the errors were centered around a few of the criteria. If we drop the first three criteria in Table 3 (Engine out..., Plans for possible..., Engine failure ...) and recompute the overall mean percent correct it increases to 89%.

Table 4 contains the comparable values for success criteria that did not occur. The overall mean percent correct for these negative criteria was 68%. On 32% of these criteria I/E's incorrectly judged the criterion as having occurred, when in fact it had not occurred. Compared

[1] In future work, we plan to explore the nature of different measures of agreement (e.g., Kendall's tau, phi coefficient) for dichotomous data, including percent agreement, and for different types of distributions.

to the performance on the positive criteria, the I/E's were more likely (32%) to incorrectly observe a success criterion that had not occurred, than to fail to observe a success criterion that had occurred (17%).

Because most of the criticality values for the 6 negative criteria were 3's (5 of the 6 had criticality values of 3 and the remaining one was a 2), we did not analyze percent correct by criticality. Mean percent correct increases to 74% if we drop the lowest negative criterion (e.g., Utilizes automation ...), still significantly less than the performance observed on the positive criterion.

Table 3
Mean Percent Correct and Standard Deviation for Positive Success Criteria.

Mean	SD	Criticality Score	Positive Success Criteria
0.36	0.49	2	Engine out missed approach procedures
0.39	0.50	3	Plans for possible single engine go-around
0.54	0.51	1	Engine Failure after V1
0.68	0.48	2	Identifies and utilizes correct POM procedures
0.68	0.48	3	Selects automation level that reduces workload
0.71	0.46	3	Considers straight-out missed approach
0.75	0.44	3	Completes all published missed approach procedures
0.79	0.42	1	All landing with engine out operations performed within standards
0.86	0.36	3	Maintains fuel balance between tanks
0.86	0.36	3	Evaluates and determines best course of action for return/divert
0.86	0.36	3	Coordinates crew approach tasks during visual approach
0.89	0.31	2	Crew completes with 1000 foot cleanup procedures
0.89	0.31	2	Considers CAT II equipment requirements
0.93	0.26	3	Request clearance to takeoff alternate
0.93	0.26	3	Coordinates communication during rollout kept PF informed
0.96	0.19	2	Captain makes a legal return/divert decision
0.96	0.19	3	Notifies company of decision and requests assistance
0.96	0.19	3	Automation tasks do not detract from crew performance
1.00	0.00	2	Determines if an inflight start is appropriate
1.00	0.00	2	Considers weather at departure airport
1.00	0.00	1	Considers CAT III equipment requirements for decision
1.00	0.00	1	Engine out precision approach
1.00	0.00	2	CAT II approach procedures

Table 4
Mean Percent Correct and Standard Deviation for Negative Success Criteria.

Mean	SD	Criticality Score	Negative Success Criteria
0.36	0.49	3	Utilizes automation to reduce workload during missed approach
0.57	0.50	2	Missed approach automation performed according to procedures
0.64	0.49	3	Anticipates possibility of low fuel
0.79	0.42	3	Communicates with flight attendants when workload permits
0.82	0.39	3	Considers max tailwind when selecting runway for landing
0.89	0.31	3	Uses S/E climb page for an enroute altitude & airspeed

We next examined the relation between performance of individual I/E's and the demographic information reported above. First we computed a percent correct score for each I/E based on all 29 success criteria. The mean percent correct across the 28 I/E's was 79.6 (SD=9.01) with a minimum score of 52% and a maximum score of 97%. Figure 2 shows the frequency of these scores.

Performance was then analyzed in terms of position (Captains = 79.9%, First Officers = 79.4%) and when the I/E's made their entries in the grade sheet (During = 79.3%, After = 80.2%). Although, neither of these factors appeared to have any influence on the level of I/E's performance, a word of caution may be called for in the case of when the entries in the grade sheets were made. The conditions under which the LOE was observed in the present study were far less resource demanding than what might be expected to occur in an actual simulator. The I/E's in the present situation only had to observe the video, whereas if they were running a simulated flight they would be responsible for controlling numerous parameters that influenced aspects of the simulator flight. Consequently the I/E's in the present study were able to direct all of their attention to observing the video and to take notes if they so choose. For these reasons it would be premature to conclude that there are no risks associated with delaying competition of the grade sheet.

Performance on the other demographic variables was examined by dividing the I/E's into upper and lower 50th percentiles and computing percent correct separately for the I/E's in each of the two groups. The results, presented in Table 5, show once again that rating performance is not influenced by any of the demographic variables examined in this study.

Table 5
Overall Percent Correct on Grade Sheet for I/E's
Below and Above Median on Selected Demographics

	Below Median	Above Median
Flying Hours Commercial Air	80.4%	79.9%
Years with Delta	78.4%	80.9%
Years with current Fleet	77.2%	82.1%
No. months as I/E	80.1%	79.1%
Familiarity with LOE (5=high)	78.9%	80.4%

Finally, we examined the relationship between I/E's ratings of event sets made on a global versus local basis. The global ratings of each event set were made on the 4-point scale described earlier. What we are calling local ratings, were computed on the basis of I/E's judgments of observed/non-observed to each success criterion within an event set. An event set was scored as a "4" if all of the success criteria received a "+" (i.e., were judged to have occurred); a "3" if one or more -'s occurred only at level 3 criticality; a "2" if a - occurred on a single 2; and a "1" if there occurred a - on a criticality score of 1 or on two or more "2"'s. Table 6 shows the means and standard deviations for the global and local ratings along with the correlations between the two types of ratings, for each of the three event sets.

From Table 4 it can be seen that the mean values for global and local measures were in close agreement across all three events sets. Moreover, for event sets 1 and 3 the correlations between the global and local ratings was reasonably high (both p values < .01). The correlation between global and local measures for event set 2, while considerably lower, was statistically significant, $p < .05$. Despite the higher correlations for Event Sets 1 and 3 there was more variability in both the global and local ratings of these event sets (e.g., global ratings ranged from 1 to 4 on these event sets) than what was found with Event Set 2.

Table 6
Ratings of Events Sets at Global and Local Levels

	Global					Local				
	Mean	**SD**	**Min**	**Max**	**Referent**	**Mean**	**SD**	**Min**	**Mix**	**Corr**
Event Set 1	2.0	.90	1	4	3	1.9	.97	1	3	.74**
Event Set 2	2.5	.58	2	4	2	2.3	.44	2	3	.40*
Event Set 3	2.6	.83	1	4	3	2.6	.91	1	4	.76**

Note. SD = standard deviation; Min = minimum rating given by any I/E; Max = maximum rating given by any I/E; Ref = local Referent score; Corr = Pearson correlation between global and local scores; * = .05; ** = .01 significance levels.

Summary

Although the mean inter-rater reliabilites were lower than what may be considered acceptable, there were three important factors that very likely operated to lower the reliabilities. First, inter-rater reliabilities computed by a Pearson correlation coefficient are possibly based on an inappropriate statistic given the nature of the data generated from the current grade sheet. It appeared that percent correct, relative to an expert referent, communicated a more accurate picture of I/E's performance. Second, the I/E's were using a new form of the grade sheet. The I/E's had to shift from a 4-point to a 2-point scale. Based on I/E comments, some of the I/E's were confused on how the 4-point scale was to be transformed to a 2-point scale. A future calibration study is planned to determine if this difficulty can be corrected with additional training on the use of the 2-point scale, or if the grade sheet should be changed to a 3- or 4-point scale. A third factor contributing to the lower inter-rater reliabilities was that the descriptions of each of the success criteria had not been calibrated. Again, I/E's comments at the conclusion of the task that indicated that some of the success criteria were misinterpreted. As noted earlier, exclusion of only a few of these problem criteria resulted in a substantial improvement in inter-rater reliabilities. Future calibration sessions with I/E's should allow us to determine the precise source of ambiguity in the verbal descriptions of these problem criteria and through careful rewording eliminate the ambiguity.

Future calibration sessions should also pay particular attention to the negative criteria where I/E's had a higher percentage of disagreements. Finding that I/E's tended to judge non-occurring success criteria as observed suggests that there was a positive bias influencing their judgments. This may be related to the fact that these I/E's were accustom to only making entries on the grade sheet when a crew member performed below qualification standards.

Turning to the findings on the demographics information it can be summarized as indicating that none of the demographic factors were related to I/E rating performance. Failing to find an experience effect, either in terms of flying time or time as I/E may be taken as positive news in that it suggests that it may be possible train I/Es relatively quickly. Of course we can come to no firm conclusion on this until the mean inter-rater reliability correlations are improved to an acceptable level.

Finally, the findings with the global ratings raise some interesting questions. Despite a rather high degree of variability across I/E's on these ratings they correlated quite highly with the derived local ratings for two of the three event sets. This indicates a certain level of consistency within an I/E in how they made both their local and global ratings. In other words, while I/E's disagreed in what they saw in the video, their local ratings were consistent with the overall impressions of the quality of the performance on an event set. This may suggest that the low inter- rater reliabilities were not the so much a result of unfamiliarity with the 2-point scale, as they were the result of real disagreements regarding the quality of crew performance observed in the video.