

THE EFFECT OF EVALUATION PURPOSE ON CREW PERFORMANCE RATINGS ACROSS COMPARABLE TRAINING EVENTS

J. Matthew Beaubien
American Institutes for
Research
Washington, DC
Dr. Robert W. Holt
George Mason University
Fairfax, Virginia

Capt. William R. Hamman
United Air Lines
Denver, Colorado

Dr. Deborah A. Boehm-Davis
George Mason University
Fairfax, Virginia

ABSTRACT

Previous research suggests that performance evaluations conducted for personnel decisions tend to be substantially more lenient than performance evaluations conducted for research purposes. Because LOE is a “jeopardy” event, we hypothesized that LOE ratings would be substantially more lenient than comparable LOFT ratings. The results failed to support this hypothesis. However, path analyses suggest that the instructors were using different rating strategies when evaluating overall PIC and SIC performance in LOFT than in LOE. Specifically, PIC and SIC ratings in the LOE tended to emphasize specific behavioral examples (i.e., TECH topics) to a much greater extent than in LOFT.

INTRODUCTION

Previous research suggests that performance evaluations conducted for personnel decisions such as promotion or compensation tend to be substantially more lenient than comparable evaluations conducted for research purposes (Dobbins, Cardy, & Truxillo, 1988; Kozlowski, Chao, & Morrison, 1998; Murphy & Cleveland, 1995). Moreover, this effect tends to be magnified when both sets of evaluations use rating scales that are relatively insensitive to small differences in actual performance (e.g., typical 4-point or 5-point rating scales).

Although the specific reasons for this phenomenon remain unclear, it has been suggested that the raters’ implicit and explicit goals play a key role (Kozlowski, Chao, & Morrison, 1998). For example, because performance evaluations that are conducted for research purposes have, by definition, no negative impact on the ratees’ careers, raters may be more motivated to provide highly accurate and diagnostic evaluations. Because performance evaluations that are conducted for personnel decisions undoubtedly

influence the ratees’ careers, raters may consciously or unconsciously distort their performance evaluations to achieve other goals (e.g., maintaining equity among all employees within a department). This curious phenomenon has been observed in diverse contexts ranging from the military to manufacturing and service organizations. To date, however, no research has explored its generalizability to the aviation domain.

LOE and LOFT, two hallmarks of commercial aviation training, share a number of similarities. For example, both involve training and evaluating complete crews in a full-motion, simulated flight from take-off to landing. The underlying purpose is to make the training as realistic as possible, thereby increasing the probability that trained behaviors will transfer to the line.

LOFT and LOE also differ in several important ways. Specifically, LOFT ratings are largely collected for training purposes. Although LOFT contains an evaluation component, it is emphasized much less than in LOE. For example, the consequences of failing a LOFT tend to be relatively minor. Typically, they involve no more than an additional day of training, after which time the pilot is immediately returned to flight duties. To a lesser extent, LOFT ratings are also used for research purposes, such as developing follow-up training programs.

Unlike LOFT, LOE places a much higher emphasis on evaluation. By extension, LOE ratings are much more likely to be used in personnel decisions such as compensation and promotion. For example, failing an LOE typically results in removal from flight duties until substantial remedial training and evaluation has been completed. Depending on the logistical constraints involved with scheduling the additional training and subsequent LOE, this could take up to several weeks. As a result, LOE has a much stronger

emphasis on evaluation than LOFT.

Due in part to its “jeopardy” nature, we hypothesized that all things being equal, mean LOE ratings would be substantially more lenient (i.e., higher) than mean LOFT ratings. Moreover, because LOFT and LOE are typically assessed with rating scales that are somewhat insensitive to small differences in performance, we hypothesized that LOE ratings would exhibit substantially less variability (i.e., smaller standard deviations) than comparable LOFT ratings. By extension, the high mean and decreased variability associated with LOE ratings should result in attenuated correlations among LOE grades, but not among LOFT grades.

METHOD

Data were collected from two training events (i.e., one LOFT, one LOE) which were designed according to standard industry and regulatory practices (Federal Aviation Administration, 1990; Prince et al., 1993). Both events were designed to be of roughly equivalent difficulty. For example, both involved windshear during takeoff and re-routing to alternate airports during descent. However, the two events differed in content to reduce testing effects.

During each event, a highly-trained instructor manipulated the simulator settings, role-played the ATC, and evaluated the crew using a standardized evaluation form. Upon completion, the instructor debriefed the crews regarding their performance. LOFT sessions were conducted by Pilot Instructors, while LOE sessions were conducted by Standards Captains. No information was available regarding differences in their levels of experience or seniority.

To increase measurement reliability, each training event was decomposed into multiple event sets (Seamster, Edens, & Holt, 1995). Each event set represented a distinct phase of flight that included an environmental trigger, specific behaviors that the crew were expected to perform, and a set of pre-defined rating criteria.

Participants

Data were collected from pilots in the Boeing 757/767 fleet at United Air Lines who were undergoing recurrent training and evaluation in 1999. All pilots had been trained in Crew Resource Management (CRM) principles, and were expected to exert maximal effort during both LOFT and LOE

(Dubois et al., 1993; Sackett et al., 1988).

Although several LOEs were conducted during the course of the year, there was only one LOFT. This LOFT provided usable data from 643 crews. To ensure that the results between the LOFT and LOE were as stable as possible, we chose the LOE with the largest sample size. This LOE provided usable data from 273 crews. Unfortunately, because all data was de-identified, we were unable to describe the pilots’ background characteristics and expertise. However, because all 273 crews from the LOE also participated in the LOFT, and because crews were quasi-randomly assigned to different LOEs, we assume that the “true” level of proficiency (Nunnally, 1978) across the two events was roughly equivalent.

As noted earlier, both events were designed to be of roughly equal difficulty, and all pilots were expected to exert maximal performance. Therefore, to the extent that systematic differences in means, variances, and covariances were observed, this would provide indirect support for our hypothesis that the observed differences were due to the type of evaluation (research purposes vs. personnel decisions).

The Rating Process

For each event set, the pilot instructors rated the crews along several dimensions, including: technical (TECH) topics, CRM topics, overall TECH performance, overall CRM performance, pilot-in-command (PIC) performance, and second-in-command (SIC) performance. With the exception of the CRM topics, all variables were measured using a 4-point scale with the following anchors: “Repeat Required” (1), “Debrief” (2), “Standard” (3), and “Excellent” (4). CRM topics were rated using a 3-point scale with the following anchors: “Not Performed” (1), “Partially Performed” (2), and “Performed” (3).

The pilot instructors were instructed to first rate the crews on the TECH and CRM topics (i.e., behavioral markers). Next, they were to use the topic ratings when rating overall TECH and CRM performance. Finally, they were to use the overall TECH and CRM ratings, along with their independent judgment, to make PIC and SIC ratings. The rating process is described graphically in Figure 1.

RESULTS

Mean Differences

We hypothesized that because LOE is a “jeopardy” evaluation, mean LOE ratings would be significantly more lenient than mean LOFT ratings. Mean differences between were assessed using independent samples t-tests with pooled variance terms. An independent samples test was chosen because the de-identified data precluded more sophisticated, repeated-measures analyses. The pooled variance term was used to compensate for the different sample sizes. All tests were performed using one-tailed tests ($p = .025$).

Because the LOFT and LOE contain slightly different content, we averaged across event sets (within each training event) to generate overall means and standard deviations for TECH topics, CRM topics, overall TECH performance, overall CRM performance, PIC ratings, and SIC ratings. For example, to generate the mean PIC rating for the LOFT, we averaged the PIC ratings from all 8 event sets.

The data did not support the hypothesis that mean LOE ratings were significantly more lenient than mean LOFT ratings. While several mean differences were observed, none exceeded an absolute value of .20 (approximately 5% of the total 4-point rating scale). Only one comparison (Average PIC) was significant and in the hypothesized direction ($t_{(914)} = 2.929$). Three others were significant (Average SIC, Average TECH, Average CLR Topics), but in the opposite direction ($t_{(914)} = -4.022, -4.555, \text{ and } -7.012$, respectively). These statistically significant, but practically non-existent differences were most likely due to the high level of statistical power.

Differences in Variability

As noted earlier, when using relatively insensitive measurement scales, leniency is also associated with decreased variability (i.e., smaller standard deviations). Differences in variability between the LOE and LOFT were compared by calculating coefficients of variation. The coefficient of variation (CV), which is calculated by dividing the standard deviation by the mean, is used to compare the variability of items that are measured using different scales (Howell, 1997). This technique was chosen because the CRM topics were assessed using a 3-point scale, whereas the remaining measures were assessed using a 4-point scale.

A visual comparison of like terms across the two training events revealed virtually no differences in

CVs between like terms on the LOFT and LOE. The largest difference (-.044) occurred between overall TECH performance (See Table 1). Because this represents approximately 1% of the 4-point scale, we concluded that evaluation purpose had no effect on the variability of the performance ratings.

However, the reader should note that for both events, the coefficients of variation for CRM topics were substantially smaller than the remaining measures. By way of comparison, the coefficient of variation for TECH topics on the LOFT and the LOE are 1.49 and 2.22 times larger than their corresponding CRM topics, respectively. This lack of variability in the CRM topics will become increasingly important in the tests of structural validity that follow.

Differences in Structural Validity

Despite the fact that there were no significant differences in either means or variabilities, we were still interested in examining the structural validity of ratings on the two training events. We estimated the structural validity of the two training events using path analysis (Pedhazur, 1982). For each event set, overall TECH and CRM performance were (separately) regressed onto the TECH and CRM topics. In all cases, CRM topics were entered in the first block; the TECH topics were then entered in the second block.

Next, Pilot in Command (PIC) and Second in Command (SIC) ratings were (separately) regressed onto overall TECH and CRM performance, TECH topics, and CRM topics. In all cases, overall CRM and TECH performance were entered in the first block; CRM topics were entered in the second block; and TECH topics were entered in the final block.

Separate analyses were performed for each event set in each of the two training events (i.e., LOE and LOFT). This resulted in sixty-eight separate regression analyses (4 criterion variables in the LOFT x 8 event sets; 4 criterion variables in the LOE x 9 event sets). At each stage of each analysis, the overall amount of variance accounted for was assessed by R^2 measures of effect size. Due to space constraints, only summary tables will be presented. Additional data are available from the first author upon request.

A consistent pattern of results emerged for overall TECH and CRM performance across the two training events. In general, CRM topics (entered in the first block) explained between 4 to 7 percent of the criterion variance. However, TECH topics (entered in

the second block) explained an additional 30 to 50 percent variance in the criterion variables (See Table 2).

A similar pattern of results emerged for overall PIC and SIC ratings across the two training events. In general, overall TECH and CRM performance (entered in the first block) explained between 50 and 64 percent of the criterion variance. CRM topics (entered in the second block), exhibited virtually no incremental validity. TECH topics (entered in the third block), explained an additional 15 and 18 percent of the criterion variance, but only for LOE ratings.

We believe that the observed results may be due, in part, to the differential predictive validity of the overall CRM and TECH performance across the two training events. Specifically, in the LOFT, overall CRM and TECH performance explained about 60 to 64 percent of the PIC and SIC criterion variance. In the LOE, however, overall CRM and TECH ratings explained only about 50 percent of the PIC and SIC criterion variance.

This pattern of differential validities suggest that the instructors were using different aggregation criteria when calculating overall PIC and SIC ratings in LOFT and LOE. In LOFT, it appears that the instructors relied heavily on overall TECH and CRM performance, as described in the rating guidelines. In LOE, however, it appears that the instructors relied to a lesser extent on overall TECH and CRM performance, but also considered specific examples of the crewmembers behaviors (TECH topics).

We must temper our conclusions, however, by emphasizing the indirect nature of these tests, and the potential (but unverified) differences in experience among the Pilot Instructors and Standards Captains. To truly understand what the pilot instructors were thinking as they made their ratings, more direct evidence is required.

DISCUSSION

Because LOFT and LOE are designed for somewhat different purposes, we hypothesized that differences in means, variabilities, and covariances would vary as a function of training event (LOFT vs. LOE). The data did not support these hypotheses. Differences in means and variabilities across the two training events were trivial. None of the mean differences exceeded an absolute value of .20 on the 4-point rating scale. Similarly, differences in variability rarely exceeded .05

on the 4-point rating scale.

Path analyses suggest that regardless of the training event, behavioral markers (CRM and TECH topics) were strongly predictive of overall CRM and TECH performance. In addition, overall CRM and TECH performance were strongly predictive of PIC and SIC performance. However, it appears that instructors were using different strategies when generating their overall PIC and SIC grades in the LOE than in the LOFT. Specifically, PIC and SIC ratings in the LOE tended to emphasize specific behavioral examples (i.e., TECH topics) to a much greater extent than in LOFT. While we recognize that this difference may be due differences in the instructors' experience or seniority, we suggest that future research employ more direct techniques (e.g., retrospective interviews) for assessing instructors' decision processes.

Given the relatively poor predictive validities for the CRM topics, we suggest dropping the current 3-point scale in favor of the 4-point scale used for the remaining measures. We believe that altering the scale may result in greater variance, thereby allowing the CRM topics to exhibit greater covariance with the other event set grades. However, because many pilots feel that CRM behaviors cannot be evaluated with the same precision as TECH behaviors, this recommendation may be difficult to implement. We suggest using a small group tryout, for example, with data collected for research purposes on approximately 100 or so crews, to determine whether changing the rating scale makes a difference.

Finally, because both the LOFT and LOE ratings exhibited relatively desirable psychometric properties, we suggest that researchers begin to explore the utility these performance ratings in their own research. For example, by linking training grades with questionnaire measures (e.g., crew cohesion, leadership skills), it may be possible to test practically- and theoretically-meaningful models of crew behavior. Such local validation studies could provide detailed information regarding the antecedents of effective crew performance. At the same time, they may also identify areas where additional training is needed. Given the relatively large sample sizes available and low cost of using such additional measures, we believe that this would be an effective use of an otherwise underutilized resource.

However, we must caution the reader that the LOFT and LOE ratings' psychometric properties have not come without a substantial dollar investment. United

Airlines has been a pioneer in CRM and AQP research for nearly two decades, and the psychometric properties of their performance ratings has continually improved over time. As a result, given sufficient support from upper management, it may be possible for other carriers to achieve similar results. Although we recognize that this may take some time, we believe that LOFT and LOE performance ratings can become an integral part of the aviation psychologists' research "toolbox" in the not too distant future.

REFERENCES

- Birnbach, R. A., & Longridge, T. M. (1993). The regulatory perspective. In E. L. Weiner, B. J. Kanki, & R. L. Helmreich (Eds.), *Cockpit resource management* (pp. 263-281). New York, Academic Press.
- Dobbins, G. H., Cardy, R. L., & Truxillo, D. M. (1988). The effects of individual differences in stereotypes of men and women and purpose of appraisal on sex differences in performance ratings: A laboratory and field study. *Journal of Applied Psychology, 73*, 551-558.
- Dubois, C. L. Z., Sackett, P. R., Zedeck, S., & Fogli, L. (1993). Further exploration of typical and maximal performance criteria: Definitional issues, prediction, and White-Black differences. *Journal of Applied Psychology, 78*, 205-211.
- Howell, D. C. (1997). *Statistical methods for psychology* (4th edition). Belmont, CA: Duxbury Press.
- Kozlowski, S. W. J., Chao, G. T., & Morrison, R. F. (1998). Games raters play: Politics, strategies, and impression management in performance appraisal. In J. W. Smither (Ed.), *Performance appraisal: State of the art in practice* (pp. 163-205). San Francisco: Jossey-Bass.
- Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational and goal-based perspectives*. Thousand Oaks, CA: Sage.
- Nunnally, J. (1978). *Psychometric theory* (2nd edition). New York: McGraw-Hill.
- Pedhazur, E. J. (1982). *Multiple regression in behavioral research* (2nd edition). Fort Worth: Harcourt Brace.
- Prince, C., Oser, R., Salas, E., & Woodruff, W. (1993). Increasing hits and reducing misses in CRM/LOS scenarios: Guidelines for simulator scenario development. *International Journal of Aviation Psychology, 3*, 69-82.
- Sackett, P. R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and maximal performance. *Journal of Applied Psychology, 73*, 482-486.
- Seamster, T. L., Edens, E. S., & Holt, R. W. (1995). *Scenario event sets and the reliability of CRM assessment*. Paper presented at the 8th International Symposium on Aviation Psychology, Columbus, OH.

Figure 1
Graphic Representation of the Rating Process Used in LOFT and LOE

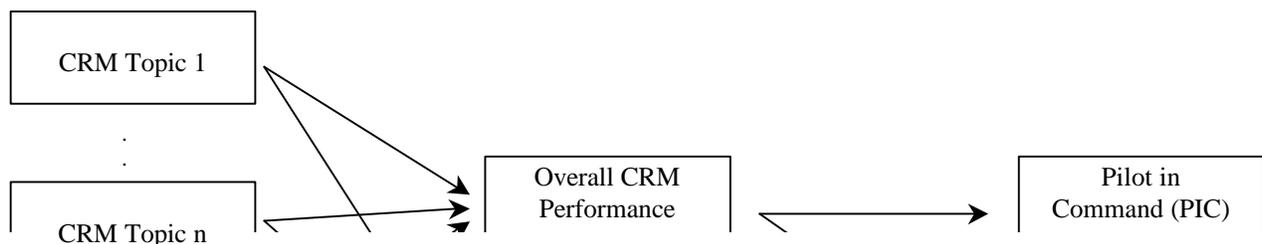


Table 1

Means, Standard Deviations, and Coefficients of Variation for LOFT and LOE

	<u>LOFT</u>				<u>LOE</u>		
	Mean	SD	CV		Mean	SD	CV
Average PIC	3.153	0.458	0.145	Average PIC	2.985	0.468	0.157
Average SIC	3.159	0.435	0.138	Average SIC	3.003	0.449	0.149
Average CLR	3.178	0.444	0.140	Average CLR	3.099	0.500	0.161
Average TECH	3.081	0.470	0.153	Average TECH	2.980	0.585	0.196
Average CLR Topics	2.917	0.270	0.092	Average CLR Topics	2.957	0.189	0.064
Average TECH Topics	3.136	0.430	0.137	Average TECH Topics	3.012	0.428	0.142

Note 1: Cell values have been averaged across multiple event sets.

Note 2: LOFT contained 8 event sets; LOE contained 9 event sets.

Table 2

Multiple Regression Results Depicting the LOFT/LOE Ratings

	<u>LOFT</u>	<u>LOE</u>
<u>DV = Overall TECH ratings</u>		
Block 1: CRM Topics	0.037	0.071
Block 2: TECH Topics	0.497	0.411
<u>DV = Overall CRM ratings</u>		
Block 1: CRM Topics	0.033	0.066
Block 2: TECH Topics	0.409	0.322
<u>DV = PIC ratings</u>		
Block 1: Overall CRM & TECH performance	0.642	0.490
Block 2: CRM Topics	0.004	0.009
Block 3: TECH Topics	0.018	0.175
<u>DV = SIC ratings</u>		
Block 1: Overall CRM & TECH performance	0.614	0.488
Block 2: CRM Topics	0.003	0.010
Block 3: TECH Topics	0.015	0.154

Note 1: Cell values = $\bar{A}R^2$ estimates of effect size that have been averaged across multiple event sets.

Note 2: LOFT contained 8 event sets; LOE contained 9 event sets.