Running Head: CONSTRUCT VALIDITY OF LOS RATINGS

Improving the Construct Validity of Line Operational Simulation (LOS) Ratings: Lessons Learned from the Assessment Center

J. Matthew Beaubien

David P. Baker

Amy Nicole Salvaggio

American Institutes for Research

Washington, DC

June 4, 2003

in press, *International Journal of Aviation Psychology*

Abstract

Line Operational Simulations (LOS) are commonly used for training and evaluating pilot crews under realistic conditions. Despite their widespread use, the construct validity of LOS ratings remains largely unexplored. Preliminary evidence suggests that LOS ratings cluster by phase of flight, rather than by the technical and Crew Resource Management (CRM) skills that they are intended to measure. These results are consistent with findings from the assessment center literature. After comparing and contrasting the LOS and assessment center techniques, we provide research-based guidelines for improving the construct validity of LOS ratings.

Improving the Construct Validity of Line Operational Simulation (LOS) Ratings: Lessons Learned from the Assessment Center

Today, most pilot crew training takes place in a simulator. Much of this training is conducted in the form of Line Operational Simulations (LOS) that mimic gate-to-gate operations. For measurement purposes, LOS are organized into a series of "event sets." Each event set represents a distinct phase of flight during which the crew must demonstrate their technical and Crew Resource Management (CRM) skills. During the LOS, an instructor manipulates the simulator, interacts with the crew by role playing the air traffic controller, and evaluates their performance using a standardized rating form. Following the LOS, the instructor de-briefs the crew (Federal Aviation Administration, 1990).

Despite their widespread use, few studies have assessed the construct validity of LOS ratings. On the surface, this may seem like a purely academic issue. However, because Line Operational Simulations are used to train pilot crews (e.g., Line Oriented Flight Training; LOFT) and certify their airworthiness (e.g., Line Operational Evaluation; LOE), it is imperative that they measure what they purport. To the extent that LOS ratings do not validly measure crewmembers' technical and CRM skills, the feedback that instructors provide may be inappropriate (Lievens & Conway, 2001; Howard, 1997).

The Construct Validity of LOS Ratings

As noted earlier, few studies have explored the construct validity of LOS ratings. Fortunately, recent years have witnessed a growing culture of trust among the FAA, the airline industry, and the research community that has allowed issues such as the present one to be openly discussed.

In a seminal article, Trumpower and colleagues (Trumpower, Johnson, & Goldsmith, 1999) tested three competing hypotheses regarding the construct validity of LOS ratings. Their first hypothesis (the sub-skill hypothesis) was that LOS ratings would cluster across event sets by the technical and CRM skills that were being evaluated. Their second hypothesis (the context specificity hypothesis) was that LOS ratings would cluster by the event sets in which they were measured. Their third hypothesis (the general skill hypothesis) was that LOS ratings would cluster as a single skill (Trumpower et al., 1999).

Using de-identified ratings from eight separate Line Operational Evaluations (LOEs)[1], Trumpower and colleagues found strong support for the context specificity hypothesis. Specifically, they found that the mean correlation between skills within an event set (regardless of what technical or CRM skills were being measured) ranged between .64 and .70. By way of comparison, the mean correlation between identical skills that were measured during different event sets ranged between .17 and .33. Trumpower and colleagues' research suggest that the LOEs were not accurately measuring the specific technical and CRM skills for which they were designed. If they were, there would have been greater convergent validity between identical skills that were measured during different event sets. There would also have been greater discriminant validity between different skills that were measured during the same event set (Campbell & Fiske, 1959).

Beaubien and colleagues observed similar results in two separate samples (Beaubien, Holt, & Hamman, 1999). Unlike the Trumpower et al. (1999) study which used multitrait-multimethod (MTMM) matrices to assess the LOEs' construct validity, Beaubien et al. used principal components analysis to examine the pattern of correlations among the technical and CRM skill ratings. Their first sample included 636 Boeing 757 crews from an international air carrier. All crews completed the same recurrent LOE. The LOE included six different event sets, and measured a total of 12 technical and CRM skills. Three principal components emerged, accounting for approximately 57% of the total item variance. With one exception, all skills loaded on a single component. The first component included the technical and CRM skills from the "Cruise" and "Descent" event sets. The second component included the technical and CRM skills from the "Pre-Departure" and "Taxi-Out" event sets. The final component included the technical and CRM skills for the "Climb" and "Approach" event sets (see Table 1).

Table 1.
Rotated Component Matrix (Sample 1).

	Component		
Event Set and Type of Skill Measured	1	2	3
Cruise (CRM)	.768		
Cruise (TECH)	.731		
Descent (CRM)	.727		
Descent (TECH)	.655		
Pre-Departure (CRM)		.728	
Pre-Departure (TECH)		.706	
Taxi-Out (TECH)		.691	
Taxi-Out(CRM)		.668	
Climb (TECH)			.738
Climb (CRM)			.706
Approach (TECH)			.686
Approach (CRM)	.376		.633

Note: Varimax rotation converged in 5 iterations.

Other than the event sets in which they were measured, no discernable pattern emerged to describe the pattern of skill ratings (Beaubien, Holt, & Hamman, 1999).

The second sample included 837 Boeing 757 crews who were assessed one year later. As before, all participants completed a single recurrent LOE that included six different event sets and measured a total of 12 technical and CRM skills. This time, four components emerged, accounting for approximately 73% of the total item variance. Only two skills had cross-loadings greater than .30. The first component included the technical and CRM skills from the “Top of Descent to Final Approach” and “Final Approach to Taxi-In” event sets. The second component included the technical and CRM skills from the “Pre-Departure to Taxi-Out” and

"Takeoff to Top of Climb" event sets. The third component included the technical and CRM skills for the "Reaching Top of Climb" event set. The fourth and final component included the technical and CRM skills for "Takeoff to Top of Climb" and "Cruise" event sets (see Table 2).

Table 2.
Rotated Component Matrix (Sample 2).

	Component			
Event Set and Type of Skill Measured	1	2	3	4
Final Approach to Taxi-In (TECH)	.867			
Final Approach to Taxi-In (CRM)	.860			
Top of Descent to Final Approach (TECH)	.693			
Top of Descent to Final Approach (CRM)	.688			
Pre-Departure to Taxi-Out (TECH)		.798		
Pre-Departure to Taxi-Out (CRM)		.790		
Take-Off to Top of Climb (TECH)		.688		.395
Take-Off to Top of Climb (CRM)		.685		.368
Reaching Top of Climb (TECH)			.956	
Reaching Top of Climb (CRM)			.951	
Cruise (TECH)				.848
Cruise (CRM)				.810

Note: Varimax rotation converged in 5 iterations.

As before, other than the event sets in which they were measured, no discernable pattern emerged to describe the pattern of skill ratings (Beaubien et al., 1999). In summary, Beaubien and colleagues' research suggest that the LOEs were not accurately measuring the specific skills for which they were designed. If they were, the ratings would have clustered by the skills that were being evaluated, rather than by the event sets in which they were measured.

Collectively, the Trumpower et al. (1999) and Beaubien et al. (1999) studies assessed a total of 10 Line Operational Simulations, all of which had questionable construct validity. The data suggest that LOS ratings may cluster by event set, rather than by the technical and CRM skills that they are designed to measure. These preliminary findings need to be replicated in a greater sample of carriers, fleets, and LOS formats (e.g., Line Oriented Flight Training; LOFT) to assess their generalizability. Because the findings from the Trumpower et al. (1999) and Beaubien et al. (1999) studies are similar to those from the assessment center literature, we looked to the assessment center literature to identify guidelines for improving the construct validity of LOS ratings.

The Assessment Center

The assessment center technique provides a holistic view of an individual’s job-related knowledge, skills, and abilities (Howard, 1997). Initially developed during World War II to select recruits for espionage missions in occupied Europe, the assessment center has since been applied to the selection of police officers, business executives, and a host of other occupations (Moses & Byham, 1977).

All assessment centers have three defining characteristics. First, assessment centers use multiple exercises to assess the participants’ job-related skills and abilities. Multiple exercises are used because they provide participants with the opportunity to display their skills in a range of situations (Moses & Byham, 1977). Second, assessment centers use trained raters to evaluate the participants' performance. Multiple raters are used to ensure that key behaviors are not missed, thereby providing a more accurate assessment of the participants' qualifications. Typically, a separate group of raters evaluates the participants' performance in each exercise (e.g., role-play, leaderless group discussion, etc.). Third, assessment centers use judgment pooling to create an overall, composite score for each participant. This typically occurs at the assessment center's conclusion, when the raters gather to discuss their individual evaluations. Either clinical or statistical integration can be used to create composite scores for each participant (Howard, 1997).

The Validity of Assessment Center Ratings

Over the years, there has been a substantial amount of research concerning the validity of assessment center ratings. Research concerning their predictive validity – the extent to which assessment center ratings predict future job performance – has generally been positive. For example, assessment center ratings have been validated against numerous criteria, including performance evaluations, training performance, career progress, salary, and ratings of managerial potential (Howard, 1997). A meta-analysis by Gaugler and colleagues (1987) suggests that the uncorrected mean validity of assessment centers for predicting subsequent job performance is approximately .32. These results demonstrate that people who perform well in the assessment center also perform well on the job.

Evidence concerning the construct validity of assessment center ratings – the extent to which assessment centers actually measure the skills and abilities that they were designed to measure – has been somewhat less positive (Howard, 1997). Research on the construct validity of assessment centers typically focuses on the convergent and discriminant validity of assessment center ratings. In an assessment center, each skill is assessed during multiple exercises. Convergent validity is said to exist when ratings of the same skill correlate highly across exercises. For example, ratings of oral communication assessed during the role-play exercise should be highly correlated with ratings of oral communication assessed during the leaderless group discussion. Discriminant validity is said to exist when different skills that are measured during the same exercise are not correlated. For example, ratings of oral communication assessed during the role-play exercise should be only weakly correlated with ratings of other skills (e.g., decision-making skills) that are also assessed during that exercise.

Like the LOS technique, assessment centers have problems with convergent and discriminant validity (Howard, 1997). However, recent research suggests that these results are much less problematic than previously thought. For example, Lievens and Conway (2001) suggest that the skills which are measured during an assessment center explain as much of the total variance (approximately 34%) as the exercises that are used to measure them.

Comparing and Contrasting the LOS and Assessment Center Techniques

The LOS and assessment center techniques share a number of features. For example, both provide multiple opportunities to assess the participants' job-relevant skills under realistic conditions. In the assessment center, this is accomplished by requiring the participant to complete a series of "exercises." In the LOS, this is accomplished by requiring the crewmembers to complete a series of "event sets."

Both techniques also use trained evaluators to synthesize a large amount of information when forming an overall score for each participant. In the assessment center, this is achieved by "convergence sessions," during which the raters discuss their individual exercise ratings for each participant and decide upon overall performance scores after weighing all the evidence. In the LOS, this is typically achieved by decision rules that specify how to aggregate the skill ratings to create overall event set and overall LOS evaluations for each crew.

Finally, both techniques provide developmental feedback to each participant. In the assessment center, this is achieved by providing each participant with a diagnostic report that describes his or her job-related strengths and weaknesses, and which provides individually tailored development plans. In the LOS, crews are verbally debriefed about their performance immediately following the simulation. The debrief sessions typically use videotaped examples of the crew's own performance to highlight the instructor's diagnosis (Federal Aviation Administration, 1990). Except in extreme circumstances, written development plans are rarely provided.

Nevertheless, there are several key differences between the LOS and assessment center techniques. First, unlike the assessment center, Line Operational Simulations are designed to assess both individual and crew performance. This raises some interesting measurement issues (Prince, Brannick, Prince, & Salas, 1997). For example, in addition to assessing and debriefing the crew's performance, the instructor must also assess the technical and CRM skills of each pilot, and must provide tailored feedback to each crewmember. This is necessary because focusing exclusively on the crew's performance may not help an individual pilot improve his or her skills. This is most clearly illustrated when a pilot's performance deficiency (e.g., lack of assertiveness) is trait-based, thereby requiring individual skills practice to overcome. At the same time, focusing exclusively on the skills of individual pilots does not completely address how the crew performed as a unit. This is most clearly illustrated when a crew fails a LOS. Although the failure may be due largely to the actions or inactions of a single crewmember, both crewmembers are jointly responsible for the safety of the passengers, the cabin crew, and the aircraft. As a result, they must share the blame for the crew’s failure.

A second critical difference involves the complexity of the tasks that are preformed. Wood (1986) characterizes task complexity along 3 primary dimensions: component complexity (i.e., the number of distinct behaviors required to perform the task), coordinative complexity (i.e., the extent to which these behaviors must be precisely timed and sequenced), and dynamic complexity (i.e., the extent to which the behavior-performance relationship changes over time). Using this framework, it is clear that the tasks performed during a LOS are considerably more complex than those performed in a typical assessment center. For example, unlike the in-basket or leaderless group discussion exercises, event sets typically measure a larger number of skills, require greater coordination of these skills to successfully complete the task, and rely more heavily on situational moderators (i.e., normal vs. emergency conditions) to determine the appropriate course of action.

Third, the number of people rating (and being rated) varies between the two assessment techniques. In a LOS, each crew is evaluated in isolation from other crews. Moreover, a single instructor assesses each crew, and that instructor has to coordinate multiple roles, such as interacting with the crew (e.g., role playing the air traffic controller), evaluating their performance, and manipulating the simulator parameters. In an assessment center, the participants are evaluated both separately (e.g., the in-basket exercise) and in a group setting (e.g., the leaderless group exercise) by multiple raters whose sole task is to assess their performance.

Finally, the LOS and assessment center techniques differ in the duration of their use. Unlike a LOS scenario that may be used for an entire recurrent training cycle (i.e., up to one year), assessment centers are typically used for only a few days at a time. As will be noted later, the extended lifespan of LOS scenarios provides an advantage over the assessment center, because many of the recommendations for improving their construct validity require a substantial time to take effect.

Guidelines for Improving the Construct Validity of LOS Ratings

Over the years, a number of guidelines have been proposed for improving the construct validity of assessment center ratings. Several of these mirror recommendations for improving the validity of LOS ratings (Lauber & Foushee, 1981; Prince, et al., 1997; Prince, Oser, Salas, & Woodruff, 1993). However, many are unique enough to deserve special attention here. In the paragraphs below, we propose a list of guidelines for improving the construct validity of LOS ratings. Although many of these guidelines have received empirical support for improving the validity of assessment center ratings, few have been directly tested in commercial aviation. Therefore, until their generalizability from the assessment center to LOS can be empirically verified, they must be considered practice-based hypotheses.

The guidelines are organized along two major dimensions: 1) reducing the cognitive demands of the rating task, and 2) selecting, training, and retaining qualified pilot instructors. Our experience suggests that no single intervention can substantially improve the construct validity of LOS ratings. Therefore, we recommend that multiple changes be implemented simultaneously to complement one another (Murphy & Cleveland, 1995). Given the recent economic downturn in commercial aviation, we recognize that some of these guidelines may not be immediately feasible. However, over the long term, these guidelines have tremendous potential for improving the construct validity of LOS ratings, and by extension, the quality of pilot crew training and evaluation.

Reducing Instructor Workload

Previous research suggests that raters’ limited information processing capacity can degrade the quality of their LOS and assessment center ratings (Reilly, Henry, & Smither, 1990). In this section, we outline several strategies for reducing instructor workload.

Evaluate Fewer Skills per Event Set. LOS designers should keep the number of skills evaluated per event set to a minimum

(Lievens & Conway, 2001; Smith-Jentsch, Johnson, & Payne, 1998). This can be achieved by including only those skills that are identified as "mission critical." Criticality ratings can be identified via a team task analysis, which is required under the Advanced Qualification Program (AQP). In addition to reducing instructor workload, rating fewer skills requires the instructor to spend less time "heads down" while taking notes and completing the assessment form. The extra time can be used to observe and evaluate the crew's performance.

Increase the Length of Each Event Set. Increasing the length of each event set can also reduce instructor workload. The additional time may allow the instructors to complete their ratings, make additional notes regarding the crew's performance, compare the crew's performance on the current event set to previous event sets, and prepare for the next event set. However, increasing the length of an event set can have unintended side effects. For example, if the goal is to create a stressful scenario that tests the crew's ability to react under time pressure, increasing the length of the event set may counteract the stress manipulation (Orasanu & Backer, 1996). Therefore, LOS developers should carefully consider the instructional objectives of each event set before determining whether or not to increase its length.

Design a "User-Friendly" Evaluation Form. The evaluation worksheet should be designed to accommodate the instructors' typical working conditions (Seamster, Boehm-Davis, Holt, & Schultz, 1998). For example, to offset low-light conditions, the forms should be developed using large print and brightly colored paper to increase the contrast between the text and the background. To offset the cramped workspaces, the worksheets should incorporate a spiral-bound booklet format that can be easily folded when necessary. To minimize the instructors' need to flip pages during an event set, the evaluation form should be designed such that background information (e.g., background information, simulator manipulations, "ATC" requests, skill definitions, etc.) is located on the left hand page, and the rating form is located on the right hand page. Finally, the evaluation form should be designed using a simple "check in the box" format rather than a more cumbersome "fill in the bubble" format.

Automate the Simulator as Much as Possible. When possible, the simulator manipulations (i.e., the event set "trigger," weather characteristics, background chatter, etc.) should be automated. For example, if the simulator is capable of reproducing "background chatter," this option should be used rather than having the instructor simulate the chatter him/herself. Although this may require some initial programming effort by the simulator maintenance staff, automating this task will free up a significant amount of the instructors' time, thereby allowing them to focus more on observing and evaluating the crew (Burki-Cohen, Kendra, Kanki, & Lee, 2000). Simulated background chatter may also appear more realistic than instructor-generated chatter, because the crewmembers may become attuned to the instructor's voice during the LOS.

Use a Behavioral Checklist Instead of Likert-type Rating Scales. There are many techniques for assessing crew performance. Perhaps the most common is to use a 4- or 5-point Likert scale. Although scale anchors vary from carrier to carrier, they usually cover the range of performance from "Repeat Required" to "Excellent." Unfortunately, given the many demands placed on instructors, evaluating the crew's performance using a Likert scale can be a complex task. One alternative is to evaluate the crew's performance using a behavioral checklist. Behavioral checklists include a list of task-relevant behaviors for each event set. Because behavioral checklists only require the instructor to indicate whether or not the behavior was performed successfully (not how well it was performed), behavioral checklists have the potential for reducing cognitive workload and the validity of performance ratings(Donahue, Truxillo, Cornwell, & Gerrity, 1997; Reilly et al., 1990).

Clearly Specify Skill Definitions and Example Behaviors. Regardless of what type of scale is used, the skill definitions and example behaviors must be clearly defined and concisely worded (Seamster et al., 1998; Lovler, Rose, & Wesley, 2002). Because vague terminology such as "Captain exhibits leadership" can be interpreted in a number of ways, behaviorally based statements should be used instead. A better example might read "Captain delegates Pilot Flying (PF) and Pilot Not Flying (PNF) duties." Similarly, unobservable phenomena, such as "Captain maintains situational awareness," should be abandoned in favor of more readily observable behaviors. A better example might read "Captain monitors radar and radio communications for potential traffic."

Provide Multiple Opportunities for Crews to Demonstrate their Skills. In order to obtain reliable measures of the crews' performance, it is necessary to provide the crewmembers with multiple opportunities to demonstrate their skills. This ensures that their overall performance ratings are not unduly influenced by any single event set. The most common way is to build multiple event sets into the simulation (Prince et al., 1997; Smith-Jentsch et al., 1998). However, Brannick and colleagues (1995) caution LOS developers to carefully design and pre-test each event set to ensure that they truly assess the same skills at roughly the same level of difficultly.

Videotape Crew Performance to Confirm Initial Expectations. We are all subject to a number of decision-making errors (Reilly et al., 1990). These include "halo error" (e.g., when an example of particularly effective or ineffective performance biases subsequent ratings), "central tendency error" (e.g., the tendency to rate all crews as "average"), and the "recency effect" (e.g., when the most recent behavior biases our recall of previously observed behaviors). To reduce the effect of these decision-making errors, pilot instructors should be encouraged to videotape examples of particularly effective or ineffective crew performance during the LOS (Federal Aviation Administration, 1990). These videotaped examples can later be consulted when making overall ratings of the crew's performance.

Videotaped examples have a number of other benefits. For example, if the crew does not recognize that they made a particular error, they are less likely to dispute the instructor when faced with incontrovertible proof of their own behavior. Our personal experience suggests that videotape is best suited for demonstrating CRM behaviors that have a strong verbal component such as command, leadership, communication, assertiveness, and decision-making.

Provide Decision Tools to Help Instructors Make their Final Ratings. At the completion of each event set, the pilot instructor is required to create a summary score for the crew. Such overall ratings are typically determined by pre-established decision rules. For example, at one carrier, an overall event set technical score of "Repeat Required," is assigned if two or more technical skills are rated less than standard or any technical skill requires repeating. Similar judgments must be typically made to determine whether or not the crew passed the LOS. In an absolute sense, these decision rules are not difficult to apply. However, given the time- and space-limited environment in which pilot instructors work, these decision rules can be difficult to apply in practice. Therefore, we recommend that carriers provide their instructors with tools for automating their overall event set and overall LOS decisions. This may involve, for example, a spreadsheet that calculates a crew's overall event set rating based upon their skill ratings.

Document All Skill Ratings. Our experience suggests that several carriers inadvertently reward their instructors for rating most crews as "average," even though their performance might technically warrant a different rating. This often occurs when the carrier requires the pilot instructor to document the crew's performance using reason codes or handwritten notes, but only when the crew performs above or below average. Because such documentation is essential for diagnosing all levels of performance, we recommend that documentation accompany all ratings, including ratings of average performance. This may remove the instructors' incentive to rate a large percentage of crews as "average" on all skills, thereby resulting in a less skewed distribution of ratings.

Recruiting, Selecting, Training, and Retaining Qualified Pilot Instructors

In this section, we outline several strategies for recruiting, selecting, training, and retaining qualified pilot instructors. These Human Resources (HR) functions are complementary. For example, to the extent that a carrier can select individuals with the relevant observational and communication skills required to be a pilot instructor, the amount of time required to train such individuals may be lessened (Cascio, 1991).

Make the Position of Pilot Instructor More Attractive to Potential Recruits. The position of pilot instructor is an extremely prestigious one. However, the position's status may not be sufficient to attract a large number of qualified candidates. This can occur for many reasons. For example, because much of their time is spent at the carrier's training facility, pilot instructors may have difficulty meeting their currency requirements. Pilots may also be discouraged from taking an instructor position because some instructors are perceived as being "out of touch" with typical line operations. Therefore, carriers should survey their pilot instructors to identify their reasons for and against becoming an instructor. Similarly, carriers should survey their line pilots regarding their opinions of their instructors, as well as their perceptions of the duties involved as an instructor. Armed with this information, the carrier can re-design the instructor position to make it more attractive to potential instructors. The carrier must also conduct a thorough recruitment effort, for example, by publicizing these changes via newsletter articles and presentations during the pilots' recurrent training. Taken together, these steps may encourage a larger number of line pilots to consider applying for the instructor position.

Select Pilots Instructors based on their Ability to Perform the Tasks Required of a Pilot Instructor. At many carriers, pilot instructor selection is a haphazard affair. Many times, the carrier desperately needs to fill a certain number of instructor positions. As a result, they may be willing to accept almost anyone who volunteers for the position. Although this may fulfill their short-term need, it certainly does not meet their long-term needs. Therefore, carriers should screen candidates based on their ability to perform the skills required of pilot instructors. These skills can easily be identified by conducting a critical incident analysis (Flanagan, 1954) of the pilot instructor position.

A critical incident analysis involves identifying examples of previously good (and bad) instructor performance. These examples will help the carrier to identify the major tasks required of instructors (e.g., observing the crew’s performance, taking effective notes, evaluating their performance, debriefing them), as well as the knowledges, skills, and abilities required to perform each task. Once complete, the carrier can develop their own tests to select candidates who possess these skills, or can purchase commercial-off-the-shelf (COTS) tests that have previously been validated.

In practice, recruitment and selection are complementary activities. To the extent that few pilots apply for the position of instructor, selection becomes moot. Therefore, carriers should begin by assessing, and if necessary, redesigning their instructor role. Once this is complete, they should actively recruit potential instructors. During the recruitment process, they should begin to develop their selection systems, so that once the recruitment drive is complete, the selection system can be immediately implemented. As a general rule, selection systems work best when many applicants are applying for a limited number of positions.

Provide Instructors with Behavioral Observation Training (BOT). Even though well-developed selection programs minimize the need for instructor training, they can never completely eliminate it. We recommend that all pilot instructors receive behavioral observation training (BOT) to enhance their observational skills. BOT teaches raters to accurately detect, perceive, recall, and recognize specific behavioral events (Thornton & Zorich, 1980). Instructors should also receive training in note-taking skills. These notes are essential for debriefing the crew after the LOS has been completed.

Provide Instructors with Frame-of-Reference (FOR) training. Previous research suggests that inadequate rater training contributes to the poor construct validity of assessment center scores (Klimoski & Brickner, 1987). Therefore, instructors should receive some form of training that emphasizes the skills to be assessed during the LOS. This training should provide explicit behavioral definitions for each skill, use behavioral examples of good and poor performance on each skill, and provide practice and feedback using the actual rating form. Frame-of-reference (FOR) training is one such program (Baker, Mulqueen, & Dismukes, 2001; Bernardin & Buckley, 1981). Previous research has identified FOR training as the most effective technique for minimizing rater errors (Woehr & Huffcutt, 1994).

Provide Frequent Training. Regardless of what type of training the instructors receive, they must receive frequent refresher training. Like all skills, the skills involved in observing and evaluating crew performance degrade over time. Unfortunately, little research has been conducted regarding the most appropriate re-training interval. Therefore, carriers will need to experiment by varying the time intervals between refresher training sessions. We suggest that, at least initially, recurrent pilot instructor training take place every three to six months. During each training session, the pilot instructors should be calibrated to a “gold standard” (Baker, Swezey, & Dismukes, 1998). If, after this initial period, the pilot instructors remain calibrated, the interval can be safely extended. However, if the instructors do not maintain calibration, the recurrent training intervals may need to be shortened.

Provide Incentives for Reducing Unwanted Turnover. At many carriers, instructor turnover is extremely common. Unwanted turnover is a problem for two major reasons. First, high turnover can lead to poor LOS construct validity. This typically occurs when a sizeable percentage of the pilot instructor population is composed of relative novices who do not share the same "mental model" with the more veteran instructors. Although training can offset the effect of turnover, it can take a long time to replace a seasoned instructor.

Second, if the instructor returns to line flight duties too soon, the company may not realize a profit on their training investment. Practically speaking, training a line pilot to serve as an effective instructor requires a substantial investment of the carrier's time and money. If carriers experience a high level of unwanted turnover, they should systematically examine their HR practices, such as their work schedules, currency requirements, and compensation plans. It may be that the carrier's HR practices are inadvertently resulting in high turnover. Changing these practices may help alleviate the problems associated with unwanted turnover.

Conclusions and Applications for Practice

As one reviewer pointed out, many people in the industry "know" that LOS ratings exhibit poor convergent and discriminant validity. However, during our literature review, we were unable to identify any published studies that directly addressed the construct validity of LOS ratings. Therefore, the primary purpose of this paper was to publicly identify an issue that some in the community has privately discussed for years. A secondary purpose was to show that this problem is not unique, and that there is an established body of research on assessment centers that can be drawn upon to improve the validity of LOS-type performance ratings.

As we have shown throughout this paper, the LOS and assessment center techniques share a number of similarities. For example, both involve assessing job-related performance on a number of skills using a series of structured exercises. Both also involve trained assessors who use judgment pooling to create overall performance ratings. Finally, both reveal similar problems with convergent and discriminant validity.

These similarities suggest a number of implications for practice. First, although the predictive validity of LOS ratings has yet to be explicitly tested, they may mirror the relatively high validities of assessment center ratings in predicting subsequent job performance (Gaugler, et al., 1987). As a result, LOS may be an effective tool in the organization's selection system. Second, while both the LOS and assessment center ratings have problems with construct validity, recent research suggests that the problem may not be insurmountable (Lievens & Conway, 2001). Third, the assessment center literature suggests a number of practical and cost-effective guidelines for improving the validity of LOS ratings – and by extension, the value of LOS as a training and evaluation tool. Although some of these guidelines – such as developing new recruitment and selection systems – may have large start-up costs, they can provide valuable long-term benefits to the organization.

Previous research suggests that a variety of factors can affect the construct validity of LOS ratings. These include LOS design issues, the usability of the rating form, instructor training programs, the carrier’s Human Resources systems, and the use of technology. However, no single intervention can be expected to exhibit any meaningful change. Rather, multiple interventions must be adopted that complement one another (Murphy & Cleveland, 1995). Poor LOS construct validity is a multifaceted problem that develops over time. As a result, improvements will likely occur only over the long term. We recognize that given the industry’s current financial crisis, some of these guidelines may not be immediately feasible. However, once the industry emerges from this crisis, we believe that they will again come to view pilot training and evaluation not as a cost to be minimized, but as a long-term investment in maintaining safety and profitability.

Because poor LOS construct validity can have real-world effects on pilot training and performance, assessing and improving the construct validity of Line Operational Simulations is more than just an academic or scientific issue. It is also a practical and political issue that it involves multiple stakeholders who may have competing concerns. These include safety, justice/fairness, technical feasibility, and cost-effectiveness (Austin, Klimoski, & Hunt, 1996). Therefore, we recommend that all potential stakeholder groups be involved in identifying and improving the construct validity of Line Operational Simulations. These groups may include pilot unions, training staff, flight standards staff, and officials from the regional FAA offices. Moreover, all groups must be prepared to compromise some of their own goals/needs to achieve a balanced solution. In the end, only by working together can industry address the issue of LOS construct validity, and by extension, the quality of pilot crew training and evaluation.

As with any study, this one has its limitations. First, the available data concerning the construct validity of LOS ratings is limited to a handful of samples, all of which were conducted as LOEs. Second, most of the guidelines for improvement have only been tested with assessment centers, not LOSs. As a result, these preliminary results will need to be verified with other forms of LOS, such as Line-Oriented Flight Training (LOFT). Nevertheless, we hope that at a minimum, this paper will prompt the industry to further explore this important topic. As we noted earlier, the construct validity of LOS ratings is not merely an academic or scientific issue; it has important implications for pilot crew training and evaluation.

References

Austin, J. T., Klimoski, R. J., & Hunt, S. T. (1996). Dilemmatics in public sector assessment: A framework for developing and evaluating selection systems. *Human Performance, 93*, 177-198.

Baker, D. P., Mulqueen, C., & Dismukes, R. K. (2001). Training raters to assess resource management skills. In E. Salas, C. Bowers & E. Edens (Eds.) *Improving teamwork in organizations: Applications of resource management training* (pp. 131-145). Mahwah, NJ: Erlbaum.

Baker, D. P., Swezey, R. W., & Dismukes, R. K. (1998). *A methodology for developing gold standards for rater training video tapes*. Unpublished technical report. Prepared under contract to the NASA-Ames Research Center (NASA Cooperative Agreement No. NCC 2-911), Moffet Field, CA.

Beaubien, J. M., Holt, R. W., & Hamman, W. R. (1999). *An evaluation of the rating process used by instructor/evaluators in a line-operational simulation: Preliminary evidence of internal structure validity*. Technical Report #98-002. Fairfax, VA: George Mason University.

Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review, 6*, 205-212.

Burki-Cohen, J., Kendra, A. J., Kanki, B. G., & Lee, A. T. (2000). *Realistic Radio Communication in Pilot Simulator Training*. FAA Technical Report #DOT-VNTSC-FAA-00-13. Cambridge, MA: Volpe National Transportation Systems Center.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Cascio, W. F. (1991). *Applied psychology in personnel management* (4th ed.). Englewood Cliffs, NJ: Prentice Hall.

Donahue, L. M., Truxillo, D. M., Cornwell, J. M., & Gerrity, M. J. (1997). Assessment center construct validity and behavioral checklists: Some additional findings. *Journal of Social Behavior & Personality, 12*(5), 85-108.

Federal Aviation Administration. (1990). *Line operational simulations: Line oriented flight training, special purpose operational training, line operational evaluation*. Advisory Circular 120-35B. Washington, DC: Author.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 41*, 327-358.

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72*, 493-511.

Howard, A. (1997). A reassessment of assessment centers: Challenges for the 21st century. *Journal of Social Behavior and Personality, 12*, 13-52.

Klimoski, R. J., & Brickner, M. (1987). Why do assessment centers work? The puzzle of assessment center validity. *Personnel Psychology, 40*(2), 243-260.

Lauber, J. K., & Foushee, H. C. (1981). *Guidelines for line-oriented flight training (Volume 1)*. National Aeronautics and Space Administration. Conference Publication #2184. Moffett Field, CA: NASA-Ames Research Center.

Lievens, F., & Conway, J. M. (2001). Dimension and exercise variance in assessment center scores: A large-scale evaluation of multitrait-multimethod studies. *Journal of Applied Psychology, 86*, 1202-1222.

Lovler, B., Rose, M., & Wesley, S. (2002, April). *Finding assessment center construct validity: try behaviors instead of dimensions*. Paper presented at the 17th Annual Meeting of the Society for Industrial and Organizational Psychology, Toronto, Canada.

Moses, J. L. & Byham, W.C. (1977). *Applying the assessment center method*. New York: Pergammon Press.

Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives*. Thousand Oaks, CA: Sage.

Orasanu, J. M., & Backer, P. (1996). Stress and military performance. In J. E. Driskell and E. Salas (Eds.), *Stress and human performance* (pp. 89-125). Mahwah, NJ: Erlbaum.

Prince, A., Brannick, M. T., Prince, C., & Salas, E. (1997). The measurement of team process behaviors in the cockpit: Lessons learned. In M. T. Brannick, E. Salas, & C. Prince (Eds.), *Team performance assessment and measurement* (pp. 289-310). Mahwah, NJ: Erlbaum.

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.

Reilly, R. R., Henry, S., & Smither, J. W. (1990). An examination of the effects of using behavior checklists on the construct validity of assessment center dimensions. *Personnel Psychology, 43*, 71-84.

Smith-Jentsch, K. A., Johnson, J. H., & Payne, S. C. (1998). Measuring team-related expertise in complex environments. In J. A. Cannon-Bowers & E. Salas (Eds.) *Making decisions under stress: Implications for individual and team training* (pp. 61-87). Washington, DC: American Psychological Association.

Seamster, T. L., Boehm-Davis, D. A., Holt, R. W., & Schultz, K. (1998). *Developing Advanced Crew Resource Management*

(ACRM) Training: A Training Manual. Washington, DC: Federal Aviation Administration.

Thornton, G. C., & Zorich, S. (1980). Training to improve observer accuracy. *Journal of Applied Psychology, 65*, 351-354.

Trumpower, D. L., Johnson, P. J., & Goldsmith, T. E. (1999). Structural analysis of line-oriented evaluation data. In R. S. Jensen (Ed.), *Proceedings of the 10th International Symposium on Aviation Psychology* (pp. 1220-1223). Columbus, OH: The Ohio State University Press.

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A meta-analytic review. *Journal of Occupational and Organizational Psychology, 67*, 189-205.

Wood, R. E. (1986). Task complexity: Definition of a construct. *Organizational Behavior and Human Decision Processes, 37*, 60-82.

Author Notes

This research was supported through grants from the NASA-Ames Research Center (NCC 2-1084) and the FAA's Office of the Chief Scientific and Technical Advisor for Human Factors (99-G-0048). Dr. David P. Baker is the Principal Investigator on both projects.

The views expressed in this paper are solely those of the authors, and do not necessarily reflect those of their sponsoring agencies.

Correspondence concerning this article should be addressed to J. Matthew Beaubien, American Institutes for Research, 1000 Thomas Jefferson Street NW, Washington DC, 20007. E-mail: jbeaubien@air.org.

Footnotes

[1] The sample size varied across LOEs, but ranged between 15 - 64 crews each.