THE INTERNATIONAL JOURNAL OF AVIATION PSYCHOLOGY, *12*(3), 287–303

Pilot Instructor Rater Training: The Utility of the Multifacet Item Response Theory Model

Casey Mulqueen and David P. Baker
American Institutes for Research
Washington, DC

R. Key Dismukes
National Aeronautics and Space Administration Ames Research Center
Moffett Field, CA

A multifacet 1-parameter item response theory (i.e., Rasch) model was used to examine interrater reliability training for pilot instructors. This model provides a means for examining individual instructor leniency or severity in ratings, difficulty of grade-sheet items, skill levels of flight crews, and interactions among these components. It was found that pilot instructor trainees differed in their levels of rating severity, and that higher crew resource management scores were easier to achieve than technical scores. Interaction analyses identified several pilot instructors who were evaluating crews in an unexpected manner, which is useful when providing feedback during training.

The introduction of the Advanced Qualification Program (AQP) has led to a significant amount of research on the process by which pilots are evaluated on their technical and crew resource management (CRM) skills. Under AQP, pilots are trained and assessed during line operational simulation (LOS) scenarios, line-oriented flight training (LOFT), line operational evaluation (LOE), and special-purpose operational training (SPOT) during initial or recurrent training. LOFT and SPOT are used for training, whereas LOE involves actual evaluation of the flight deck crew's technical and CRM skills. All LOS scenarios involve a

Requests for reprints should be sent to Casey Mulqueen, American Institute for Research, 1000 Thomas Jefferson Street NW, Washington, DC 20007. E-mail: cmulqueen@air.org

complete cockpit crew—captain, first officer, and flight engineer (depending on aircraft type)—flying a scenario in a realistic high-fidelity flight simulator. These scenarios usually begin at the departure gate and include specific scenario events that are introduced as the flight progresses to the destination airport. Each scenario event set is designed to elicit technical and CRM behaviors by the crew (Hamman, Seamster, Smith, & Lofaro, 1991). A pilot instructor, seated in the back of the simulator, observes the crew's response to each event set and rates the performance of the crew and each crewmember regarding their technical and CRM skills.

A critical element in LOS is the pilot instructor.[1] These individuals observe how each aircrew performs on the LOS scenario event sets and assign technical and CRM performance ratings (Birnbach & Longridge, 1993). In LOE, the resulting ratings are used to determine whether or not each pilot in the crew should be certified to fly the line or requires additional training before certification. Therefore, the extent to which pilot instructors make accurate judgments about crew and crewmember performance is critical to the effectiveness of AQP training and airline operations.

A reliable and accurate assessment of a crew's CRM skills cannot be made during an LOS scenario if pilot instructors do not agree on the CRM behaviors observed and the level of performance demonstrated for each skill. When pilot instructors do not agree, performance ratings are a function of the particular instructor conducting the LOS as opposed to performance of the crew. To safeguard against this problem, Longridge and others have suggested that pilot instructors should receive formal rater training (Birnbach & Longridge, 1993; Williams, Holt, & Boehm-Davis, 1997). Under AQP, pilot instructor training is required, and the proficiency and standardization of instructors and evaluators must be verified on a recurrent basis (Federal Aviation Administration, 1990a, 1990b). This training, known as interrater reliability (IRR) training, has now been implemented at several AQP airlines.

In one common format, IRR training consists of a single-day workshop during which pilot instructors receive information about the LOS scenario rating process and practice rating the performance of several videotaped crews. During the first part of the day, videotapes of one or two crews flying a specific LOS scenario are viewed, and pilot instructors independently rate each crew's technical and CRM performance using a grade sheet specifically designed for the LOS. During a class break, ratings are analyzed to determine the level of agreement that exists across class participants and areas in which significant rating discrepancies occur. When the class reconvenes, the results of these analyses are fed back to the instructors

[1]The term *pilot instructor* is used throughout this article. It encompasses any qualified individual involved in training and evaluating aircrew performance: instructors, check airmen, and standards captains.

and discrepancies are discussed. A videotape of a different crew flying the same LOS scenario is then viewed and rated to determine the level of agreement achieved within the class (Williams et al., 1997).

In traditional IRR-training classes, several statistical indexes are used to provide feedback to participants including average interrater agreement, systematic differences, congruency, and sensitivity (Holt, Johnson, & Goldsmith, 1998). The latter three indexes represent analyses of the distributions of pilot instructors' ratings of the videotapes used for practice. Analysis of systematic mean differences attempts to discover pilot instructor levels of severity or leniency by comparing each individual instructor's mean crew performance rating with the group mean. The congruency index is conceptually similar to systematic differences in that it compares other aspects of an individual pilot instructor's crew performance ratings with those of the group (e.g., variance and shape of the distribution of ratings). The rationale behind these measures is that individual instructor's judgments of performance should not deviate too much from the rest of the group undergoing rater training (Holt et al., 1998).

Sensitivity is the extent to which variability in ratings matches real variability in performance. In other words, it is the ability of pilot instructors to judge gradations in a crew's technical and CRM skills and to assign accurate ratings to different levels of performance (Holt et al., 1998). Assessing sensitivity requires that subject-matter experts (SMEs) first establish "true scores" for the videotapes used in training. SMEs review each tape, assign technical and CRM performance ratings to each crew depicted on the video, and discuss these ratings to establish consensus ratings or true scores. True scores serve as referents against which instructor ratings are compared during rater training (Johnson & Goldsmith, 1998).

In regard to rater agreement, measures such as Pearson's *r,* the within-group interrater agreement coefficient (r_{wg}*;* James, Demaree, & Wolf, 1984), and the intraclass correlation coefficient (ICC; Shrout & Fleiss, 1979) are often used for IRR training. Pearson's *r* is a measure of interrater consistency and indicates the extent to which the videotapes of flight crews, used in training, are ranked similarly by pilot instructors. As such, Pearson's *r* is a measure of relative agreement of instructor performance ratings. The r_{wg} statistic is an indicator of interrater consensus. It measures the absolute agreement among pilot instructors (i.e., the extent to which pilot instructors assign the same technical and CRM performance ratings to the same pilots and crews). The ICC is a special case of the one-facet generalizability study. It measures the correlation among different pilot instructors who rated the training videotapes (Shrout & Fleiss, 1979) and provides estimates of the variance associated with instructors, videotapes (i.e., crews), and the interaction among these facets. These measures, particularly the Pearson's *r,* are widely used and familiar to most researchers. However, because Pearson's *r* is only a measure of the consistency of ratings, it is limited as an indicator of rater

performance. The r_{wg} statistic by itself provides a good measure of interrater agreement, an important aspect of the outcome of pilot instructor rater training. The ICC is useful for providing an index of both the consistency and agreement among pilot instructors, as well as the amount of variance in ratings that is accounted for by instructors, crews, and their interactions. Generalizability theory, in its full form, is discussed later.

These statistics provide information about factors crucial to flight-training operations. If instructors differ substantially in their ratings, decisions about pilots' skills will be driven largely by the happenstance of which instructor does an evaluation. Inappropriate leniency in evaluations may jeopardize safety, and inappropriate severity will increase training costs. Inconsistency by individual pilot instructors produces all of these negative effects and makes it difficult for air carriers to track the effectiveness of pilot training. Some aspects of reliability may be easier to achieve than others are. For example, it may be easier to train instructors to be consistent and to avoid leniency or severity than it is to develop high levels of agreement among instructors. Many rater-training programs in other domains have failed to achieve high levels of interrater agreement (e.g., see Cason & Cason, 1984; Lunz & Stahl, 1993).

Although useful, the analyses described previously focus solely on the quality of the ratings alone and therefore are limited as indicators of overall training effectiveness. More multifacet investigations into the performance-rating process have been recommended (Baker & Salas, 1997). Facets that are of particular interest in this context include not only the pilot instructors themselves but also the rating forms (i.e., LOS grade sheets) and the training materials (i.e., videotapes of crews performing LOS scenarios or scenario event sets). Each can contribute to error in the LOS assessment process and each is briefly expanded on next.

The grade sheets that are used for data collection should be examined to determine whether the rating scale is broad enough to capture the range of crew performance inherent in the event sets that are being rated. The difficulty of rating specific technical and CRM items should also be assessed. Some grade-sheet items may be more difficult to evaluate than others are and some items may be ambiguous or misleading. Alternatively, event sets may not probe all the technical and CRM behaviors covered on a rating form. Such analysis is necessary to improve LOS grade sheets and scenario design. It can also identify event sets for which instructors need extra practice in learning to evaluate.

The range of performance levels depicted by the videotapes used to train instructors is important for statistical analysis and training. Statistical measures will be misleading if the full range of possible performance is not used. Instructors will not learn to make appropriate distinctions if not exposed to the full range of performance exhibited by pilots for the various technical and CRM skills.

MULTIFACET RASCH MODEL

In this article we use a multifacet measurement technique, the multifacet Rasch model, to analyze the results of an IRR-training program. Our approach is an alternative to the procedures—congruency, consistency, agreement (r_{wg}), sensitivity, and systematic differences—currently used during IRR training within the airline industry. We believe that this multifacet procedure can improve the quality of pilot instructor training by providing pilot instructors with important information that is not available with other techniques. We used the multifacet Rasch method instead of generalizability (G) theory, another multifacet technique. Similar to multifacet Rasch analysis, G-theory provides information about facets—pilot instructors, videotapes of aircrews used in IRR training, and LOS grade sheets—and their interactions with one another. However, G-theory partitions the variance attributable to each of these facets using an analysis of variance (ANOVA) framework and thus focuses on groups as the unit of analysis (i.e., whether or not pilot instructors as a group are reliable or unreliable as opposed to the performance of a particular instructor undergoing IRR training). Unlike the multifacet Rasch model, G-theory is a classical test theory model. Within classical test theory, an aircrew's performance ratings on a LOS are a function of their true score (i.e., true performance) and error (i.e., instructor variability and grade sheet variability). The larger the error component, the less reliable the LOS grades. Generalizability studies partition this error term into identifiable components. By identifying potential sources of error and the magnitude of each source's contribution to the error component, steps can be taken to reduce error and improve reliability (Stahl & Lunz, 1992). Inasmuch as high reliability is the goal, reducing variability caused by instructors, grade sheets, and crews are important components in achieving that goal. Within the air-carrier industry, IRR training has been the primary strategy used for error reduction.

The multifacet Rasch technique is an item response theory (IRT) model that focuses on individual elements of the LOS assessment process (Stahl & Lunz, 1992). This model is useful for pilot instructor rater training because it provides individual-level, as opposed to group-level, information that can be directly fed back to individual pilot instructors. Information about the LOS grade sheet and the videotapes used for practice and feedback can also be gleaned from this analysis.

Table 1 lists the information provided by some of the more common statistical methods used in IRR training, and rater training in general, and by the multifacet Rasch analysis. In our study, we demonstrate the importance of the multifacet Rasch analysis for providing individualized feedback to pilot instructors, as well as information about the grade sheets and the videotapes used in IRR training.

TABLE 1
Information Provided by Different Statistical Methods

Facet	*Pearson's r*	r_{wg}	*G-Theory*	*Multifacet Rasch*
Pilot instructors	X	X	X	X
Grade sheets			X	X
Videotapes			X	X
Interactions			X	X
Measurement focus				
Group	X	X	X	
Individual				X

Note: r_{wg} = within-group interrater agreement coefficient; G-theory = generalizability theory.

The Rasch model, a one-parameter IRT model, has traditionally been used for analysis of multiple-choice examinations in which the difficulty of the test items and the ability of the examinees are parameters. The model provides estimates of each examinee's ability and each item's difficulty and conveys them on a common log-linear scale. The probability of a correct response to an item is a function of an examinee's ability and an item's difficulty (Wright & Stone, 1979).

Multifacet Rasch measurement provides the capability to model additional facets making it particularly useful for analysis of subjectively rated performance tasks such as LOS scenarios. With this method, the chances of success on such tasks are related to a number of aspects of the performance setting itself. In the case of LOS, these include the ability of the aircrew performing the LOS, difficulty of the event set (as reflected by items on the LOS grade sheet), and characteristics of the pilot instructor conducting the LOS (e.g., pilot instructor severity/leniency). These facets are related to each other as increasing or decreasing the likelihood of a particular crew of given ability achieving a given score on an LOS event set.

Interactions among facets can be modeled, allowing detection of unusual interactions between pilot instructors and LOS event sets or between pilot instructors and particular aircrews. This information is useful when evaluating IRR training because systematic patterns in pilot instructor behavior can be identified. Pilot instructors may display particular patterns of severity or leniency in relation to only one crew and not others or in relation to particular scenario events. In multifacet Rasch analysis, these types of interactions are referred to as bias. Thus, individual instructors rating inconsistently in relation to specific crews can be identified and provided feedback regarding this pattern. For a more detailed explanation of the multifacet Rasch model, see Linacre (1994).

METHOD

Participants

The participants were 33 pilot instructors (i.e., 6 check airmen and 27 pilot instructors) at a major U.S. commercial airline. These individuals are responsible for observing and evaluating aircrews during LOS, specifically LOFT or LOE. In both cases, instructors observe a crew's performance during a scenario and rate the crew's technical and CRM performance on each event embedded in the LOFT or LOE. Instructors also provide an overall performance rating for each crewmember (i.e., pilot-in-command [PIC] and second-in-command [SIC]) during this evaluation.

Pilot instructor trainees were divided into four separate classes that received training on separate days. The class sizes were 7, 7, 11, and 8. The same trainer, an experienced commercial airline pilot, facilitated each of the training sessions.

IRR-Training Program

IRR training focused on LOE and consisted of three major components: overview of the LOE grading process, review of the LOE grade sheets, and practice with the LOE rating task. Review of the LOE grading process and the LOE grade sheets was accomplished through in-class lecture, discussion, and demonstrations, whereas practice involved rating videotapes of aircrews flying specific LOE event sets and then providing feedback to instructors about their performance. The nature of this practice and the method by which feedback was provided is described in more detail next.

First, videotapes of two different aircrews (Crews 1 and 2) flying the same LOE scenario events (i.e., Event Sets A, B, and C) were shown. After observing each event, pilot instructors independently rated each crew's technical and CRM performance. In addition, pilot instructors rated the overall performance of each crewmember (i.e., PIC and SIC) on each event. That is, Crew 1, Event Set A, was shown and rated, and then Crew 2, Event Set A, was shown and rated. Next, during a class break, ratings were analyzed to determine the level of interrater agreement (using r_{wg}) that existed. In addition, measures of congruency and systematic differences between each pilot instructor and the class as a whole were computed. When the class reconvened, results of these analyses were fed back to the instructors and rating discrepancies were discussed. Finally, a videotape of a third crew (Crew 3) flying the same three scenario events was shown and rated by the pilot instructors to determine the level of postfeedback agreement. The videotapes were always shown and rated in the same order in each of the IRR-training classes.

Crew performance on the IRR-training videos varied across the tapes in such a fashion that Crew 1 demonstrated average performance, Crew 2 demonstrated

low performance, and Crew 3 demonstrated high performance. The videotapes were of actual crews performing an LOE scenario. These videotapes were recorded during development of the LOE scenario. Therefore, the videos depicted crews flying somewhat different versions of the same LOE, because the scenario went through several revisions. The videos were not scripted in any way; however, they were purposely edited to create the different performance levels.

LOE Grade Sheet

The LOE grade sheet used by the pilot instructors to evaluate each crew consisted of several graded parts: CRM behaviors, technical behaviors, and overall grades for CRM, technical, PIC, and SIC (for a detailed discussion of the LOE grading process and an example grade sheet, see Baker & Dismukes, this issue). CRM behaviors were graded on a 3-point scale (i.e., *observed, partially observed,* or *not observed*), whereas the remaining items were graded on a 4-point scale (i.e., 1 [*repeat*], 2 [*debrief*], 3 [*standard*], and 4 [*excellent*]). Only the overall CRM, technical, PIC, and SIC grades were used in this analysis. Grades for the technical and CRM behaviors were not analyzed for two reasons. First, the scale used to rate CRM behaviors (3-point scale) was different from the other scales (4-point scale). This difference would have necessitated performing separate analyses for each scale. Second, because the videos were collected during the development of the LOE, some of the technical behaviors for each event set differed among crews. For instance, Event Set A may have had two more technical behaviors for Crew 3 than for Crews 1 and 2. The use of overall ratings allowed for the grouping of items into technical, CRM, PIC, and SIC components for each of the three videotaped crews across the three event sets. Thus, each pilot instructor provided technical, CRM, PIC, and SIC grades; on Event Sets A, B, and C for Crews 1, 2, and 3; a total of 36 ratings per instructor.

RESULTS

The computer program FACETS (Linacre, 1988) was used to analyze the data. Figure 1 provides a graphical map that contains measures for each facet (i.e., pilot instructors, LOE grades, and aircrews). The measures in Figure 1 are pilot instructor *severity/leniency,* crew *ability,* and LOE grade-sheet item *difficulty.* The pilot instructors, crews, and grade-sheet items have been measured on one common linear scale, represented by the logit (log odds units) measures in the right-hand column, which is labeled “Linear Measure.” The far left-hand column contains the ratings from the LOE grade sheet that would be expected according to the Rasch model. The discussion of results is organized according to each facet of measurement.

Expected Rating	P/I Severity	Crew Ability	Item Difficulty	Linear Measure
(High)	(Lenient)	(More Able)	(Easy)	
4				+2
	*			
	*			
	*			
	*			
	**			
	*			
	****			
3	***	3		
	**		4 (CRM)	+1
	**		7 (PIC)	
	***		12 (SIC)	
	*			
	***			
	**		10 (SIC)	
	***		5 (CRM)	
			6 (CRM)	0
	*	1	1 (Tech)	
			11 (SIC)	
	*		8 (PIC)	
			3 (Tech)	
	*		9 (PIC)	
			2 (Tech)	
		2		–1
2				
1				–2
(Low)	(Severe)	(Less Able)	(Difficult)	

Note: * = 1 rater.

FIGURE 1 Estimated measures for pilot instructors, aircrews, and grade-sheet items.

Pilot Instructors

The pilot instructors are well spread out on the severity continuum and have a separation reliability of .77 ($\chi^2 = 150.8$, $p < .01$). This indicates that on the whole the pilot instructors are significantly different from one another in their level of rating severity, although as can be seen by comparing their distribution to the expected rating column, the majority tend to rate at the middle to high end of the scale. The mean rating is 2.9 (*SD* = .72) across the LOE 4-point rating scales. The measure of severity ranges from a low of –.67 (*more severe rater*) to a high of 1.78 (*more lenient rater*).

Two fit statistics are calculated with the multifacet Rasch analysis that identifies any pilot instructor that deviates from the mathematical model generated by the Rasch analysis. This model reflects expected rating performance on the part of the pilot instructors and is based on the following assumptions: (a) more skilled crews will score higher on the grade-sheet items than less skilled crews, (b) more difficult grade-sheet items will receive lower scores than easier items, and (c) more severe pilot instructors will give lower scores than more lenient pilot instructors. Two fit statistics (infit mean square and outfit mean square) assess for different types of rating inconsistencies on the part of a pilot instructor. The infit statistics are sensitive to inconsistencies at points close to the center of a rating scale. The outfit statistic provides information for detecting inconsistencies among pilot instructors at the high and low ends of a rating scale.

In reference to Table 2, the fit statistics, logit severity measures, and frequency of ratings are provided for 6 pilot instructors who have the greatest amount of misfit with the mathematical model. It can be readily seen that the 3 instructors (4, 6, and 22) identified as having low fit statistics have very limited variance in their ratings, with the majority occurring in the middle of the

TABLE 2
Fit Statistics, Severity Measures, and Rating Category Frequencies of Raters Identified as Misfitting

				Frequencies of Ratings (Percentages)			
Instructor	*Infit Mean Square*	*Outfit Mean Square*	*Severity Measure*	*1*	*2*	*3*	*4*
4	0.5	0.5	.08	5	28	67	0
6	0.5	0.5	.29	0	36	58	6
22	0.5	0.5	1.07	0	17	69	14
33	1.6	1.6	.15	17	33	19	31
8	1.6	1.7	.67	6	22	56	17
11	2.0	1.9	1.07	11	14	42	33

rating scale, particularly for response category 3. Those pilot instructors (33, 8, and 11), identified with high values of the infit and outfit statistics, are distinguished by their use of the extreme categories of the scale (i.e., high numbers of 1 and 4 ratings). The misfit analysis provides a quick and simple means for identifying instructors who are engaging in certain unexpected rating patterns, making it useful for providing feedback to specific pilot instructors about the variability of crew and pilot performance when conducting LOFT or LOE ratings. More detailed information concerning specific pilot instructors can be gained through an interaction analysis.

Crew Videotapes

The estimates for crew ability are provided in Table 3. Crew measures of ability range from –1.01 for Crew 2 (*low performing*) to 1.08 for Crew 3 (*high performing*). Crew 1 is estimated to be average in ability, with a logit measure of –.07. The separation reliability is .99 ($\chi^2 = 297.2$, $p < .01$), indicating an excellent degree of ability differentiation among these aircrews.

Grade-Sheet Items

Item difficulty is well spread out for the 12 items, with a separation reliability of .90 ($\chi^2 = 117.8$, $p < .01$) and a difficulty range from –.80 logits (harder item) to .99 logits (easier item). An examination of the item difficulties, provided in Table 4, indicates that there is some degree of difference in difficulty among the types of items rated on the LOS grade sheet (i.e., CRM, technical, PIC, or SIC ratings). The item estimated to be least difficult, with a mean rating of 3.2, is one of the overall CRM ratings, whereas the most difficult item (*M* rating = 2.6) is one of the overall technical ratings. FACETS was used to group the items that comprise overall technical and CRM performance. The mean difficulty estimate for CRM is .36 logits, and the mean estimate for the technical ratings is –.46 logits. A paired *t* test between difficulty estimate means for technical and CRM ratings indicated a significant difference between difficulty estimates of the

TABLE 3
Crew Ability Estimates, Standard Errors, and Mean Ratings

Crew	*Ability Measure*	*Standard Error*	*Mean Composite Rating*
1	–.07	.09	2.9
2	–1.01	.08	2.5
3	1.08	.09	3.2

Note: Reliability of separation index = .99 ($\chi^2 = 297.2$, $p < .01$).

TABLE 4
Item Difficulty Estimates, Standard Errors, Mean Ratings, and Item Dimensions Arranged Easiest to Most Difficult

Item No.	*Difficulty Measure*	*Standard Error*	*Mean Rating*	*Item Dimension*
4	.99	.18	3.2	CRM
7	.76	.18	3.1	PIC
12	.56	.18	3.1	SIC
10	.25	.17	3.0	SIC
5	.08	.17	2.9	CRM
6	.02	.17	2.9	CRM
1	–.07	.17	2.8	Technical
11	–.24	.17	2.8	SIC
8	–.35	.17	2.7	PIC
3	–.51	.16	2.7	Technical
9	–.69	.16	2.6	PIC
2	–.80	.16	2.6	Technical

Note: Reliability of separation index = .90 ($\chi^2 = 117.8$, $p < .01$); CRM = crew resource management; PIC = pilot-in-command; SIC = second-in-command.

technical and CRM items ($p < .05$). Thus it appears that it may be somewhat easier for crews to achieve better CRM scores than technical flight skill ratings. Because the standard error terms for the item difficulty estimates are fairly large, an equivocal statement concerning the difference in difficulty between CRM and technical items is tentative.

Interaction Analysis

One of the more interesting features of multifacet Rasch measurement is the ability to examine interactions among elements of facets. In this case, the interactions between pilot instructors and particular crews were examined. In such an analysis, bias measures, in logits, and their corresponding standardized *z* scores are reported. Table 5 provides the results for pilot instructors who were displaying the highest degree of bias in measurement. Once again, the term bias has a specific meaning in multifacet Rasch measurement, and it is not the same as the more common use of the term in traditional measurement. In Table 5, for each pilot instructor–crew interaction, the bias measure and corresponding *z* score are given. In addition, for each pilot instructor and crew interaction, the observed score and expected score are given. The *observed* score is the sum total of rating points awarded to the crew by the pilot instructor on the 12 graded items, whereas the *expected* score is the sum of ratings that are mathematically expected on the basis of the ability of the crew, the difficulty of the rating items, and the severity of the instructor.

The 3 pilot instructors with highly negative *z* scores are interacting with specific crews in an unexpectedly lenient manner. For example, Pilot Instructor 32 awarded Crew 2 with a sum of 41 points across all ratings, whereas the expectation was that this crew deserved a total of 32 points from this particular instructor. Once again, this estimate is based on the ability of the crew, the severity of the instructor, and the difficulty of the rating items. The pilot instructors with extreme positive *z* scores are rating specific crews more severely than is expected by the model (i.e., awarding ratings lower than expected).

This analysis readily identifies 2 instructors who are rating in an inconsistent manner, Pilot Instructors 32 and 33. These pilot instructors have radically different perceptions of the performance of Crews 2 and 3, as can be seen from Table 6. Pilot Instructor 32 has an unexpectedly high opinion of Crew 2, whereas Pilot Instructor 33 saw this crew as performing even worse than the other raters saw them. These same 2 raters also interact with Crew 3 but this time in opposite directions. Instructor 33 is unexpectedly lenient and Pilot Instructor 32 is unexpectedly severe.

TABLE 5
Pilot Instructor/Crew Bias Measures, *z* Scores, and Observed and Expected Scores Arranged by *z* Score

Pilot Instructor	*Crew*	*Bias Measure*	*z Score*	*Observed Score*	*Expected Score*
29	2	–2.11	–3.89	41	33.2
33	3	–2.67	–3.99	45	36.4
32	2	–2.37	–4.36	41	32.0
33	2	1.70	3.46	18	26.9
11	2	1.61	3.79	23	31.6
32	3	1.87	3.90	33	40.0

TABLE 6
Frequencies of Ratings for Raters 32 and 33 With Crews 2 and 3

		Frequencies of Ratings (Percentages)			
Crew	*Instructor*	*1*	*2*	*3*	*4*
2	32	0	0	58	42
2	33	50	50	0	0
3	32	0	25	75	0
3	33	0	0	25	75

DISCUSSION

This article illustrates ways in which the multifacet Rasch technique can be used to analyze IRR-training data. This analysis provides significantly more detailed and individualized information than traditional IRR-training methods such as r_{wg}. By way of example, we computed r_{wg} for raters for each item used in this analysis. The average r_{wg} values for Crews 1, 2, and 3 were .70, .58, and .78, respectively. Other than providing some general information about the level of pilot instructor agreement, r_{wg} provides little information to pilot instructors about their severity and consistency of rating. Furthermore, except in the case of G-theory, statistics such as r_{wg} provide little if any information about the characteristics of the LOS grade sheets and the videotapes used in training.

With respect to providing feedback to pilot instructors, one of the benefits of the Rasch analysis is its ability to identify discrepant and unexpected interactions between pilot instructors and the aircrews they evaluate. Feedback can be given to, and just as importantly sought from, pilot instructors concerning their perceptions of particular crews with whom they have unexpected interactions. It is this individual level of interaction analysis that makes the multifacet Rasch approach useful for the purpose of IRR training. Interactions cannot be identified with any existing IRR measures, and although interactions can be modeled using G-theory, information about the interactions of individual pilot instructors and individual aircrews is not possible.

If it were acceptable to the air carrier involved, an adjustment to pilot instructors' total scores for specified crews could be made based on the results of these analyses. For example, Table 5 provides the scores for Crews 2 and 3 that would be expected from Pilot Instructors 32 and 33 on the basis of their modeled severity, the difficulty of the rating items, and each crew's ability. These expected scores could be substituted for the observed scores. From the standpoint of the actual evaluations that are given to aircrews following training, such corrections could be made on the basis of each pilot instructor's estimated severity. In multifacet Rasch parlance, this would result in a more objective assessment. Corrected data could be compared with actual data to identify any cases in which a different decision would have been made about a crew. Furthermore, such corrections could be applied to LOE data stored in an airline's program proficiency database and used in examining AQP training effectiveness.

From the perspective of ongoing development of an IRR-training program, specific information was provided on the ability levels of each of the videotaped crews used in training. On the basis of this information, it would be possible for additional videotapes of crew performance to be calibrated to this sample, increasing the precision of the estimates of crew ability and rater severity. As additional crews are videotaped for use as training tapes, they can be calibrated on the same scale as previous crews. Crews with a varying range of abilities can

be gathered, adding to the cadre of tapes available for use in training. This level of detailed knowledge about training materials is not possible with the other approaches for examining rater training. Ongoing analysis of this sort can help to modify and improve the overall training program.

It was found that CRM scores were easier to achieve than technical scores. This information is useful to trainers in that it may indicate that the pilot instructors are more comfortable with rating the technical skills of aircrews as opposed to their CRM skills. The vast majority of training that airline pilots receive is technically oriented, and therefore they may be more able to discriminate among levels of performance. In regard to the components of CRM behavior, these instructors may have difficulty in recognizing and discriminating among certain behaviors and therefore most often rate CRM as *standard* (rating 3). Alternatively, it may be the case that CRM tasks are simply easier to perform than technical tasks (Bowers, Morgan, Salas, & Prince, 1993). Because individual CRM and technical items from the rating form were not included in this analysis, no information concerning which particular items were relatively more or less difficult is available.

Although the information that is provided by the use of the multifacet Rasch technique is rich, there are certain drawbacks to this procedure. First, the data setup and programming for the FACETS program are cumbersome and time-consuming, particularly when first being learned. Once a particular analysis is decided on, the program can be kept and new data files run with minimal alterations; however, actually putting the data into the correct format can be somewhat tedious and time-consuming. This aspect of the procedure may limit its real-time usefulness during a training class. Second, the IRT framework is not as well known or understood among practitioners and researchers or training recipients as the more traditional methods for assessing IRR training. Specialized education is required to have a full understanding and working knowledge of IRT and, particularly, the multifacet Rasch model. The feedback itself could be somewhat difficult to explain to pilot instructors attending IRR training. If scores for crews were to be changed on the basis of the expected scores generated from the procedure, resistance from both the crews being evaluated and the instructors would be a real possibility.

The research presented here could be expanded on in several ways. First, a larger sample of pilot instructor trainees would be helpful for obtaining more stable parameter estimates. The sample presented here is somewhat small for this procedure. Second, as noted previously, as additional videotapes of crews performing LOS become available, they can be calibrated and their performance estimated. This would provide value-added information to the training instructors who use the tapes. Third, additional analyses can be performed on IRR-training data through the use of this model. For example, the four IRR-training classes examined in this study could be analyzed as a unique facet within the model to

examine whether the trainees in the different classes established different group-rating standards. This would be valuable information in determining the generalizability of the training program. Finally, a direct comparison between the multifacet Rasch model technique and generalizability theory would be interesting. Analyzing the data using a generalizability approach could serve to validate results obtained using the Rasch method and would also be valuable for explaining more clearly the distinctions and similarities between the two models and different circumstances under which each model might be appropriate.

ACKNOWLEDGMENTS

An earlier version of this article was presented at the 15th Annual Meeting of the Society for Industrial and Organizational Psychology in New Orleans, Louisiana in April 2000. The authors thank John Stahl for many helpful suggestions regarding the analysis used in this article. This research was supported by grant NCC–2–1084 from the National Aeronautics and Space Administration (NASA) Ames Research Center. The views presented in this article are those of the authors and should not be construed as an official NASA position, policy, or decision unless so designated by other official document.

REFERENCES

Baker, D. P., & Salas, E. (1997). Principles for measuring teamwork: A summary and look towards the future. In M. T. Brannick, E. Salas, & C. Prince (Eds.), *Assessment and measurement of team performance: Theory, methods, and applications* (pp. 331–355). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Birnbach, R. A., & Longridge, T. M. (1993). The regulatory perspective. In E. L. Wiener, B. G. Kanki, & R. L. Helmreich (Eds.), *Cockpit resource management* (pp. 263–281). New York: Academic.

Bowers, C. A., Morgan, B. B., Jr., Salas, E., & Prince, C. (1993). Assessment of coordination demand for aircrew coordination training. *Military Psychology, 5,* 95–112.

Cason, G. J., & Cason, C. L. (1984). A deterministic theory of clinical performance rating. *Evaluation and the Health Professions, 7,* 221–247.

Federal Aviation Administration. (1990a). *Advanced Qualification Program* (Advisory Circular 120–54). Washington, DC: Department of Transportation.

Federal Aviation Administration. (1990b). *Special Federal Aviation Regulation 58—Advanced Qualification Program* (Federal Register, Vol. 55, No. 91, Rules and Regulations, pp. 40262–40278). Washington, DC: National Archives and Records Administration.

Hamman, W. R., Seamster, T. L., Smith, K. M., & Lofaro, R. J. (1991). The future of LOFT scenario design and validation. *Proceedings of the 6th International Symposium on Aviation Psychology, 1,* 589–594.

Holt, R. W., Johnson, P. J., & Goldsmith, T. E. (1998). *Application of psychometrics to the calibration of air carrier evaluators.* Federal Aviation Administration. Retrieved April 2001 from http://www.faa.gov/avr/afs/aqphome.htm

James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology, 69,* 85–98.

Johnson, P. J., & Goldsmith, T. E. (1998). *The importance of quality data in evaluating aircrew performance.* Federal Aviation Administration. Retrieved April 2001 from http://www.faa.gov/avr/afs/aqphome.htm

Linacre, J. M. (1988). *FACETS.* Chicago: MESA.

Linacre, J. M. (1994). *Many-faceted Rasch measurement.* Chicago: MESA.

Lunz, M. E., & Stahl, J. A. (1993). Impact of examiners on candidate scores: An introduction to the use of multifacet Rasch model analysis for oral examinations. *Teaching and Learning in Medicine, 5,* 174–181.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–428.

Stahl, J. A., & Lunz, M. E. (1992, May). *A comparison of generalizability theory and multifacet Rasch measurement.* Paper presented at the Midwest Objective Measurement Seminar, Chicago.

Williams, D., Holt, R., & Boehm-Davis, D. (1997). Training for inter-rater reliability: Baselines and benchmarks. *Proceedings of the 9th International Symposium on Aviation Psychology, 1,* 514–520.

Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago: MESA.

Manuscript first received June 2001