Running head: MULTIFACET RASCH ANALYSIS OF RATER TRAINING PROGRAM

Using Multifacet Rasch Analysis to Examine the Effectiveness of Rater Training

Casey Mulqueen

David Baker

American Institutes for Research

R. Key Dismukes

NASA Ames Research Center

Poster paper presented at the 15th Annual Meeting of the Society for Industrial/Organizational Psychology (SIOP) in New Orleans, LA. Requests for information should be sent to Casey Mulqueen (cmulqueen@air.org). The authors would like to thank John Stahl for many helpful suggestions regarding the analysis used in this paper.

April 2000

Abstract

Multifacet Rasch (e.g., one-parameter IRT) analysis was used to examine the effectiveness of rater training for individuals that are required to conduct end-of-training work performance evaluations. The results are presented with emphasis on the additional information provided by this technique, and the relative advantages and disadvantages of this approach vis-a-vis other methods of analysis.

Introduction

For years, organizations have relied on the subjective judgments of raters to provide measures of work performance among employees. More recently, the use of rater training has been advocated as a method for reducing the amount of distortion that is inherent in many of these ratings due to their subjective nature (Smith, 1986; Woehr, 1994). Not surprisingly, most of the studies that are conducted to assess the effectiveness of rater training focus on dependent measures that assess rater tendencies (e.g., halo, leniency) and rater accuracy (i.e., the extent to which raters assign the correct rating to the particular level of performance that was observed) (Woehr & Huffcutt, 1994). Although useful, analyses that focus solely on the quality of the ratings alone are limited as indicators of overall rater training effectiveness. More multifaceted investigations into the performance rating process have been recommended (Baker & Salas, 1997).

Facets that are of particular interest in the rater training context include the raters, the rating forms, and the training materials (e.g., videotapes of workers performing tasks). In traditional rater training analysis, an index of interrater reliability is often computed. However, other aspects of rater performance are important, such as raters' consistency in the use of their own rating standards across ratees, as well as their individual levels of rating severity. Ideally, raters should remain consistent in their judgments and ratings between target ratees whose performance will necessarily vary. Indeed, many have claimed that rater consistency is the best outcome that can be hoped for through rater training, since the attainment of high levels of rater agreement has been an infrequent outcome of many rater training programs (Cason & Cason, 1984; Lunz & Stahl, 1993).

Regarding the rating forms that are used to collect ratings, an analysis of these forms would be useful for determining whether the rating scale is broad enough to capture the range of performance inherent in the tasks that are being rated. Additionally, information concerning the difficulty of the rating items would be useful. Certain items may be found to be ambiguous in either wording or content, or the work tasks simply don't lend themselves to a display of all the content domains that are covered on a rating form. This may cause certain items to be more difficult for raters to respond to, or for ratees to actually provide the behaviors necessary to be rated.

Regarding the training materials that are used, information about the performance level of the ratees could be useful to trainers. For instance, if videotapes of workers performing tasks are used, the trainers would want to know if the performance of each of these workers was high, moderate, or low on the performance continuum. This would particularly be the case if the videotapes were not scripted beforehand to represent varying levels of performance, thus providing some objective verification of the accuracy of the ratings that are provided by raters.

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 to assess rater training effectiveness. Pearson's *r* is a measure of interrater *consistency*, and indicates the extent to which target ratees are ranked similarly. As such the Pearson's *r* is a measure of *relative* agreement in ratings. The r_{wg} statistic is an indicator of interrater *consensus*. It measures the *absolute* agreement between raters (i.e., the extent to which raters assign the same ratings to the same target ratees). The ICC is a special case of the one-facet generalizability study, and measures the correlation between ratings on a target (Shrout & Fleiss, 1979). In particular, it provides estimates of the variance associated with raters, ratees, and the interaction between these facets. These measures, particularly the Pearson's *r*, are widely used and familiar to most researchers. However, since the 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 rater training. The ICC is useful for providing an index of both the consistency and agreement between raters, as well as the amount of variance in ratings that is accounted for by raters, ratees, and their interactions. Generalizability theory, in its full form, is discussed below.

The purpose of this paper is to analyze the results of a rater training program using a multifaceted measurement technique, the multifacet Rasch model. This method was used as an alternative to generalizability (G-) theory, another multifaceted technique. Like multifacet Rasch analysis, G-theory provides information about facets 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. The multifacet Rasch technique is an item response theory (IRT) model that measures latent traits, and thus focuses on *individual* elements of each variable (Stahl & Lunz, 1992). The use of this model may be beneficial in the rater training context since it provides individual-level information that can be used for the purposes of direct feedback to individual raters concerning their performance, as well as specific information concerning the difficulty and performance of individual rating form items and scales. In addition, information about the training materials themselves can be gleaned from such an analysis. Table 1 provides a listing of the information that is provided by some of the more common statistical methods that are used to evaluate rater training, as well as multifacet Rasch analysis. The results of this study will be presented with emphasis on the information that is provided for individualized feedback to rater trainees, as well as information concerning the rating forms and training materials that is relevant to the ongoing development of a rater training program.

Multifacet Rasch Model

The Rasch model, a one-parameter item response theory (IRT) model, has traditionally been used for analysis of multiple choice-examinations, where the parameters involved are the difficulty of the test items and the ability of the examinees. 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 simply a function of the difference between examinee ability and item difficulty (Wright & Stone, 1979).

Multifacet Rasch measurement is an extension of the general model that provides the capability to model additional facets of interest, making it particularly useful for analysis of subjectively rated performance tasks. With this method, the chances of success on a performance task are related to a number of aspects of the performance setting itself. These aspects (i.e., facets) include the ability of the target ratee , the difficulty of the performance task, and characteristics of the raters themselves (i.e., rater severity/leniency). These facets are related to each other as increasing or decreasing the likelihood of a ratee of given ability achieving a given score on a particular task.

Interactions among facets can be modeled, allowing for the detection of unusual interactions between raters and tasks/items, or raters and particular target ratees. This is particularly useful when evaluating rater training because systematic patterns in rater behavior can be identified. Raters may display particular patterns of severity or leniency in relation to only one ratee and not others, or in relation to particular tasks. In multifacet Rasch analysis these types of interactions are referred to as *bias*. Thus, individual raters that are rating inconsistently in relation to specific ratees can be identified and provided feedback regarding this pattern. For a more detailed explanation of the multifacet Rasch model, see Linacre, 1994.

The training program that was analyzed in this study involved training raters who are responsible for evaluating individual airline pilots and aircrews in critical flight scenarios. The outcome of the evaluations that these raters provide determines whether airline pilots are certified to fly or are in need of additional training. Because of the high stakes involved in this setting, we wanted to conduct an analysis that would provide us with as much information as possible concerning the performance of the individual raters, as well as the quality of the rating form and training materials. It was felt that the multifacet Rasch technique might provide a useful framework for gathering this information in a usable form. To the best of our knowledge, this model has not been utilized for evaluating rater training. The airline pilot rater training course provides an ideal context for testing the utility of the multifacet Rasch model. It contains all the facets of interest, and is conducted in a realistic high stakes scenario.

Method

Participants

The participants were 33 airline pilot instructors at a major commercial airline. These instructors are responsible for observing and evaluating aircrews during an end-of-training flight scenario. Essentially, this scenario is a job simulation. It includes identifiable events that are designed to elicit specific technical and teamwork responses by the crew (ATA, 1994). An instructor observes a crew's performance during the scenario and rates the crew's technical and team performance on each event embedded in the scenario. They also provide an overall performance rating for each crew member (i.e., pilot-in-command [PIC] and second-in-command [SIC]) during this evaluation.

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

Rater Training Program

The rater training that was studied in the current investigation consisted primarily of practice and feedback with the rating task. First, videotapes of two different aircrews flying the same three scenario events were shown. After viewing each event, the instructor trainees independently observed and rated each crew's technical and team performance. In addition, instructor trainees rated the overall performance of each crew member (i.e., PIC and SIC) on each event. Next, during a class break, ratings were analyzed to determine the level of interrater agreement (using r_{wg}) that existed among the instructor trainees in the class and the areas where significant rating discrepancies existed. Upon reconvening the class, the results of these analyses were fed back to the instructor trainees and rating discrepancies were discussed. Finally, a videotape of a third crew flying the same three scenario events was shown and rated by the trainees to determine the level of post-feedback agreement.

Regarding the videotaped flight crews, performance varied across the videotapes and the component events in such a fashion that, on average, one crew demonstrated low performance, one crew demonstrated average performance, and one crew demonstrated high performance. The videotapes were rated in the same order by each class: average followed by low performing crew prior to the break, and the high performing crew at the end of the day.

Instrument

The rating form used by the instructors to evaluate overall aircrew performance consisted of a 4-point Likert scale. The scale consisted of the following anchors: repeat (1), debrief (2), standard (3), and excellent (4). The same scale was used to evaluate each crew's technical and team performance as well as the overall performance of each crew member. Thus, each instructor provided a total of 36 ratings for use in this analysis. Each crew received 12 ratings, three each for technical, teamwork, PIC and SIC performance.

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., raters, rating form items, and aircrew videotapes). The measures in Figure 1 are rater *severity/leniency*, aircrew *ability*, and rating item *difficulty*. The raters, crews, and rating items have been measured on one common linear scale, represented by the logit (log odds units) measures in the left hand column. Discussion of results is organized according to each facet of measurement.

Raters

The raters 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 raters are significantly *different* from one another in their level of rating severity, although the majority tend to rate at the mid to high end of the scale. The overall mean rating is 2.9 (SD = .72). The logit measure of severity ranges from a low of -.67 (more severe rater) to a high of 1.78 (more lenient rater). Examination of the infit and oufit statistics identifies those raters who are misfitting the expectations of the model. In Table 2, the fit statistics, logit severity measures, and frequency of ratings is provided for six raters who have the greatest amount of misfit (2 or more standard deviations from the expected mean of 1). It can be readily seen that the three raters identified as having low fit have very muted variance in their ratings, with the majority occurring in the middle of the range, particularly for response category "3." Those raters identified as misfitting at the high end are distinguished by their use of the extreme categories of the scale. The misfit analysis provides a quick and simple means for identifying raters who are engaging in certain unexpected rating patterns, making it useful for providing feedback to specific raters about the variability of target performance when conducting performance ratings. More detailed information concerning specific raters can be gained through an interaction analysis, covered below.

Training Materials

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 between crews is .99 (χ^2 = 297.2, p < .01), indicating an excellent degree of ability differentiation between these aircrews. This result validates the judgments of the rater training developers, who had chosen these three crews because of their differentiation of overall performance.

Rating Form

Item difficulty is well spread out for the 12 items, with a separation reliability of .90 (χ^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 between the types of items (i.e., teamwork, technical, PIC or SIC ratings). The item estimated to be least difficult, with a mean rating of 3.2, is one of the three event ratings that comprise overall crew teamwork, while the most difficult item (mean rating =

2.6) is one of the events comprising overall crew technical performance. FACETS was used to group the items that comprise overall technical and teamwork performance. The mean difficulty estimate for teamwork 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 teamwork ratings indicated a significant difference between difficulty estimates of the technical and teamwork items ($p < .05$). Thus it appears that it is somewhat easier for crews to achieve better teamwork scores than technical flight skill ratings.

Interaction Analysis

One of the more interesting features of multifacet Rasch measurement is the ability to examine interactions between elements of facets. In this case, the interactions between raters and particular aircrews was examined. In such an analysis, bias measures, in logits, and their corresponding standardized Z-scores are reported. Table 5 provides the results for raters who were displaying the highest degree of bias in measurement. Once again, the term *bias* has a specific meaning in multifacet Rasch measurement, and is not the same as the more common use of the term in traditional measurement. In Table 5, for each rater/crew interaction, the bias measure and corresponding Z-score are given. In addition, for each rater 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 rater on the 12 items, while the expected score is the sum of ratings that are mathematically expected based upon the ability of the crew, the difficulty of the rating items, and the severity of the rater.

The three raters with highly *negative* Z-scores are interacting with specific crews in an unexpectedly lenient manner. For example, rater 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 rater. Once again, this estimate is based on the ability of the crew, the severity of the rater, and the difficulty of the rating items. The raters with extreme *positive* Z-scores are rating specific crews more severely than is expected by the model.

This analysis readily identifies two instructor trainees who are rating in an inconsistent manner, raters 32 and 33. These two trainees have radically different perceptions of the performance of crews 2 and 3, as can be seen from Table 6. Rater 32 has an unexpectedly high opinion of crew 2, while rater 33 saw this crew as performing even worse than the other raters saw them. These same two raters also interact with crew 3, but this time in opposite directions. Rater 33 is unexpectedly lenient and rater 32 is unexpectedly severe.

Discussion

This paper sought to provide a comprehensive analysis of a rater training program through the use of multifacet Rasch measurement. The purpose was to display how such an analysis can provide specific information on raters that is useful for feedback, and also important information concerning the performance of the rating form and training materials. This information is particularly useful for the ongoing development of a rater training program.

The interaction analysis indicated that several rater trainees were engaging in inconsistent rating patterns with specific crews. This provides a particularly valuable piece of information for the training facilitator. It allows the facilitator to provide this feedback to these raters and to investigate their reasons for the ratings they provided. It also begs the question of how consistent raters will remain following a training program. If follow-up training were to be provided, the consistency of raters over time could be analyzed using the multifacet Rasch bias analysis.

One of the benefits of this type of bias analysis is in its ability to identify discrepant and unexpected interactions between raters and ratees. Feedback can be given to, and just as importantly sought, from raters concerning their perceptions of crews with whom they have unexpected interactions. It is this individual-level of interaction analysis that makes the multifacet Rasch approach useful for the evaluation of rater training. Although interactions can be modeled using G-theory, information about the interactions of individual raters and ratees is not possible. If it were acceptable to the parties involved, an adjustment to raters' total scores for specified crews could be made, based upon the results of the bias analysis. Table 5 provides the scores for crews 2 and 3 that would be expected from raters 32 and 33, based upon 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 based upon a rater's estimated severity. In multifacet Rasch parlance, this would result in a more "objective" assessment.

From the perspective of ongoing development of the rater training program, specific information was provided on the ability levels of each of the crews used in training. It is also possible for additional videotaped crews and rater trainees 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 and their abilities estimated. In potential, 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 teamwork scores were significantly easier to achieve than technical scores. This information is useful to trainers in that it may indicate that the instructor trainees are more comfortable with rating the technical skills of aircrews as opposed to their teamwork skills. The vast majority of training that airline pilots receive is technically oriented, and therefore they are most comfortable with judging the technical performance of crews and are able to discriminate between levels of performance. Regarding the components of teamwork behavior, these raters may have difficulty in recognizing and discriminating among certain behaviors, and therefore most often rate teamwork as “standard” (rating 3). Alternatively, it may be the case that teamwork tasks are simply easier to perform than technical tasks (Bowers, Morgan, Salas, & Prince, 1993).

The analysis presented here is a limited example of a full investigation. Additional facets could be modeled through the analysis. For example, the four rater training classes could be analyzed as a facet in order 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.

Although the information provided by the use of the multifacet Rasch technique is rich, there are certain drawbacks to this procedure. The data set up and programming for the FACETS program are cumbersome and time consuming when first being learned. Also, the IRT framework is not as well known as the more traditional methods for assessing rater training effectiveness, and requires specialized education.

References

ATA, AQP Subcommittee (1994). Line operational simulation: LOFT scenario design and validation . Washington, DC: Author.

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 . New Jersey: Lawrence Erlbaum Associates.

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.

James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. Journal of Applied Psychology, 69(1), 85-98.

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

Linacre, J. M. (1994). Many-Facet Rasch Measurement. Chicago: MESA Press.

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(3), 174-181.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86(2), 420-428.

Smith, D. E. (1986). Training programs for performance appraisal: A review. Academy of Management Journal, 11(1), 22-40.

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

Woehr, D. J. (1994). Understanding frame-of-reference training: The impact of training on the recall of performance information. Journal of Applied Psychology, 79(4), 525-534.

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

Wright, B. D., & Stone, M. H. (1979). Best Test Design. Chicago: MESA Press.

Table 1

Information provided by different statistical methods

Facet	Pearson's *r*	r_{wg}	G-Theory	Multifacet Rasch
Raters	X	X	X	X
Rating Forms			X	X
Training Materials			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.

Table 2

Fit statistics, severity measures, and rating category frequencies of raters identified as misfitting

Rater	Infit MnSq	Outfit MnSq	Severity Measure	Frequencies of Ratings (percents) 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

Table 3

Crew ability estimates, standard errors, and mean ratings

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

Reliability of separation index = .99 (χ^2 = 297.2, *p* < .01).

Table 4

Item difficulty estimates, standard errors, mean ratings, and item dimensions, arranged easiest to most difficult

Item #	Difficulty Measure	Standard Error	Mean Rating	Item Dimension
4	.99	.18	3.2	Teamwork
7	.76	.18	3.1	PIC
12	.56	.18	3.1	SIC
10	.25	.17	3.0	SIC
5	.08	.17	2.9	Teamwork
6	.02	.17	2.9	Teamwork
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

Reliability of separation index = .90 ($\chi^2 = 117.8$, $p < .01$).

Table 5

Rater/crew bias measures, Z-scores, observed and expected scores, arranged by Z-score

Rater	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 (percents)			
Crew	Rater	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

Figure 1. Estimated measures for raters, aircrews, and rating items.

```
------------------------------------------------------
Linear     Rater     Crew         Item       Expected
Measure   Severity   Ability    Difficulty    Rating
------------------------------------------------------

          (Lenient)  (More Able)    (Easy)    (High)

+   2   +          +            +            + 4     +
|       |          |            |            |       |
|       | *        |            |            |       |
|       | *        |            |            |       |
|       | *        |            |            |       |
|       | *        |            |            |       |
|       | **       |            |            |       |
|       | *        |            |            |       |
|       | ****     |            |            |       |
|       | ***      | 3          |            | 3     |
+   1   + **       +            + 4          +       +
|       |          |            |            |       |
|       | **       |            | 7          |       |
|       | ***      |            |            |       |
|       |          |            | 12         |       |
|       | *        |            |            |       |
|       | ***      |            |            |       |
|       | **       |            | 10         |       |
|       |          |            |            |       |
|       | ***      |            | 5          |       |
*   0   *          *            * 6          *       *
|       | *        | 1          | 1          |       |
|       |          |            | 11         |       |
|       | *        |            | 8          | ---   |
|       |          |            |            |       |
|       |          |            | 3          |       |
|       |          |            |            |       |
|       | *        |            | 9          |       |
|       |          |            | 2          |       |
|       |          |            |            |       |
+  -1   +          + 2          +            +       +
|       |          |            |            |       |
|       |          |            |            |       |
|       |          |            |            |       |
|       |          |            |            | 2     |
|       |          |            |            |       |
|       |          |            |            |       |
|       |          |            |            |       |
|       |          |            |            |       |
|       |          |            |            |       |
+  -2   +          +            +            + 1     +
------------------------------------------------------

          (Severe)   (Less Able)  (Difficult)  (Low)
```

note: * = 1 rater.