

## Chapter 6

### Latent Trait Models for Performance Examinations

*Mary E. Lunz and Benjamin D. Wright*

American Society of Clinical Pathologists, Chicago, and  
University of Chicago

For planning, developing, implementing, and scoring a performance examination, the assessment design is the basis for the construction of meaningful ability estimates. The purpose of a performance examination is to infer candidate abilities that go beyond the particular sample of tasks, items, and judges encountered. Whether the goal is to make reproducible pass/fail decisions or to position candidates according to demonstrated ability, the performance examination must measure candidate ability consistently. This is most efficiently accomplished by using a latent trait model to give examinees an ability estimate which is independent of the current value of individual facet elements, such as judges, tasks and items. Thus, candidate ability estimates are comparable regardless of the particular judge or examination encountered.

This paper applies a multi-facet Rasch model to the analysis of two medical performance examinations. Comments concerning the facets typically included in performance examination designs are followed by a description of the multi-facet Rasch latent trait model used. The data analyses show the facets of each examination design and the variance of the elements within each facet.

#### 1. Facets of a Performance Examination

There are typically five separate facets in a performance examination that must be accounted for in latent trait analysis. The first facet is *candidate ability* which encompasses the knowledge and skill possessed by the candidate with regard to the problem, task or product measured by the performance examination. It is expected that candidates will vary in their ability.

The second facet is the *task* and may take many forms. Some performance tasks have detailed specifications that are comparable across candidate performances. Examples may be medical cases, essay prompts, or science projects. The task requirements are described to candidates who then perform them to the best of their ability. Other performance examinations allow candidates to select a sample of their own work. In medicine, candidates may select cases from their medical practice to present in an oral examination. Portfolios may be developed in art or writing. The performances usually cover specific content specifications so that general areas of knowledge and skill are represented.

The third facet is the *judge*. Judges are critical to performance examinations; however, judges have unique physical and mental characteristics, as well as unique reactions to the assessment, all of which affect the judges' ratings. Training focuses and directs a judge's attention, but is usually unable to alter permanently the knowledge and skill that has developed over a lifetime (Stahl and Lunz, 1994).

The fourth facet is the *items* rated for each task. Considerations for this facet include: (1) the number of items rated, (2) the extent of detail in the definition of the items, (3) the relevance of the items to the task and (4) whether or not the items are measurable.

The fifth facet is the *definition of the rating scale*. Rating scales may have only two ordered categories (0/1) or an infinity (0/∞). Usually each category on the scale has a specific definition. The definitions of the rating scale categories are important, because the categories influence how judges use the scale. The measurement distance between rating categories also influences the ratings given (e.g., 0/1 or 1, 2, 3, 4, 5).

The treatment of these five facets in the performance examination produces the assessment design. Each facet must be considered for its contribution to the observed ratings. Because two facets are the easiest to visualize, planners often emphasize some combination of two facets and neglect the others in planning, implementation, and scoring.

The motivation for measurement presumes that a candidate has a measurable ability to perform the tasks required; therefore, the decisions concerning a candidate should be relatively consistent regardless of which tasks, items or judges are encountered. The reality, however, is that judges differ in their severity, tasks and items, differ in their difficulty. Latent trait models provide the statistical methods for estimating the differences in task difficulty, judge severity, item difficulty and rating scale category usage. Considering candidates, tasks, or judges to be invariant (see Fagot, 1991) is unrealistic and a primary cause of the unreliability associated with performance examinations (Koretz, 1992). Interjudge correlations have been used to argue that candidates have comparable opportunities for success when assigned randomly to judges. If judges could meet the theoretical criteria for perfect agreement, giving identical ratings to the same performance consistently, and differing only by an additive or multiplicative constant (Fagot, 1991), then the interjudge correlations would be perfect. However, even with perfect correlations, the actual candidate scores can vary due to the severities of the judges and/or difficulties of tasks.

## 2. The Multi-Facet Rasch Model

For analysis of performance examinations the basic Rasch model (Rasch 1960/1980) is extended to the multi-facet Rasch model (Linacre, 1989), so that facets for task and item difficulty, judge severity, and rating scale usage can be added to the equation. When a candidate is rated, the log odds of a candidate being rated in category  $x$  is modelled to be governed by:

$$\log\left(\frac{P(x_{vmji} = x)}{P(x_{vmji} = x - 1)}\right) = \theta_v - \alpha_m - \beta_j - \sigma_i - \tau_x \quad (1)$$

where:  $P(x_{vmji} = x)$  is the probability of being rated in category  $x$   
 $P(x_{vmji} = x - 1)$  is the probability of being rated in category  $x-1$   
 $\theta_v$  is the ability of candidate  $n$   
 $\alpha_m$  is the difficulty of task  $m$   
 $\beta_j$  is the severity of judge  $j$   
 $\sigma_i$  is the difficulty of item  $i$   
 $\tau_x$  is difficulty of being rated in category  $x$  rather than category  $x-1$

It is the linear (additive) formulation of this model which enables the separation of parameter estimates (see chapter 1.2). Each facet is calibrated conjointly from the observed ratings. Candidate ability is estimated from all ratings given by all judges on all tasks and/or items. Judge severity is estimated from all ratings given across all candidates, tasks and/or items; and so on. The positioning of candidates, tasks, judges, and items on a common linear scale provides the frame of reference for understanding the relationship of the facets of the performance examination. The probability of a satisfactory performance is a function of the difference between candidate ability and the combined impact of the judge(s) severity and task(s) and/or item(s) difficulty encountered by the candidate.

Performance examination designs usually direct candidates to encounter different combinations of facet elements, so that raw score variations among candidates express far more than candidate differences in ability. The multi-facet Rasch model corrects for the incidental variations when estimating candidate ability, then candidate separation reliability,  $Rel(\theta)$ , is calculated as

$$Rel(\theta) = \frac{\text{Var}(\hat{\theta}) - \text{MSE}}{\text{Var}(\hat{\theta})} \quad (2)$$

$$\text{where } \text{MSE} = \frac{\sum se^2(\theta_v)}{N}$$

This confirms variation among candidate ability after the effect of variation among the other facet elements is accounted for.

The multi-facet Rasch model also provides estimates of the consistency of the observed rating patterns. Fit is a quantitative measure of the discrepancy between the statistical model and the observed data. The expectation is that, more able candidates achieve consistently higher scores than low able candidates, and easy items or tasks precipitate higher ratings. The residual difference between expected and observed scores is the basis of fit analysis. Unexpected high ratings for low performers are improbable successes on difficult tasks, often due to special knowledge. Unexpected low ratings for high performers are improbable failures, often due to carelessness or misunderstanding.

Two fit statistics are reported. The Outfit mean-square is a conventional chi-square statistic divided by its degrees of freedom. Thus, under Rasch model conditions, the modelled variance of an observation around its expectation is,

$$\text{Var}(x_{vij}) = \sum_{x=0}^{m_i} (x - E(x_{vij}))^2 p(x_{vij}) \quad (3)$$

where  $\text{Var}(x_{vij})$  is the variance of the observation,  $x_{vij}$  awarded to person  $v$ , on item  $i$  by judge  $j$

$m_i$  is the highest numbered category for observations on item

$E(x_{vij})$  is the expected value of the observation

$p(x_{vij})$  is the probability that person  $v$  will be observed on item  $i$  by judge  $j$  in category  $x$ .

The Outfit mean-square,  $U_i$ , for item  $i$  (and similarly for person  $v$  and judge  $j$ ) is

$$U_i = \frac{\sum_{v=1}^N \sum_{j=1}^J (x_{vij} - E(x_{vij}))^2 / \text{Var}(x_{vij})}{NJ} \quad (4)$$

where  $N$  is the number of persons and  $J$  is the number of judges.

Similarly, the Infit mean-square is an information-weighted chi-square statistic divided by its modelled degrees of freedom. The Infit mean-square,  $V_i$ , for item  $i$  (and similarly for person  $v$  and judge  $j$ ) is (see Rasch, (1960/1980) p. 193-194).

$$V_i = \frac{\sum_{v=1}^N \sum_{j=1}^J (x_{vij} - E(x_{vij}))^2}{\sum_{v=1}^N \sum_{j=1}^J \text{Var}(x_{vij})} \quad (5)$$

The Infit compares the sum of squared rating residuals with their expectation. It is sensitive to an accumulation of on-target deviations that are less or more consistent than expected. The Outfit compares the sum of standardized residuals with their count and is sensitive to off-target responses due to carelessness or misunderstanding.

The range of these fit statistics is 0 to infinity with a modelled expectation of 1.00 and a variance inversely proportional to the number of independent replications in the statistic referred to.

Because the interpretation of fit is situationally dependent, there are no fixed levels for fit statistic acceptance or rejection. Situational considerations include the type of test or rating scale, type of observation, examination design, and/or expectations for judge agreement (see Wright and Linacre, 1994 and Wright, 1995). The criteria for acceptable fit used in this study are mean-squared residuals between .5 and 1.5. Judges and tasks with fit statistics beyond these criteria are reviewed for inconsistency or overconsistency in their rating pattern.

The fit statistics for *judges* indicate the degree to which each judge is internally self-consistent across candidates, tasks, and items (intrajudge consistency). Judges who award what are for them unexpectedly high or low ratings to a particular candidate on a particular

task or item are identified and the effect of the unexpected ratings on the candidate ability estimates reviewed. The fit statistic for each *task* indicates inter-judge consistency when grading that task. Misfit indicates that some judges deviated significantly from others when grading some tasks or candidates. The fit statistics for each *item* indicate inter-judge item consistency when rating items within tasks across candidates. *Candidate* misfit indicates inter-judge disagreement on the quality of that candidate's performance.

### 3. Data

Two performance examinations for medical certification provide data. Judges rated candidate performances using a defined rating scale, and a unique assessment design.

The *histology certification examination* required candidates to prepare 15 laboratory slides (tasks) to detailed specifications. Six judges rated the slides prepared by each of the 93 candidates on three items. Two items required a 0/1 rating and the third required a 0, 1, 2, 3 rating in which 0 was unacceptable, 1 was marginal, 2 was satisfactory and 3 was excellent. Candidates' slides were allocated to judges in a rotational pattern. All examinations involved the same tasks, items and rating scales. No slide was rated more than once. The allocation of different judges to different candidate performances introduced the need to adjust for judge severity.

The rotational system directed each judge to grade some examples of each of the 15 slide tasks sometime during the grading session on all the items. Random combinations of three different judges rated subsets of five slides for each candidate. Thus, even though no slide was graded more than once, three judges had input into each candidate's overall performance. This constructed the linking necessary to calibrate judge severity and slide difficulty. Because each slide was graded only once, there were many missing data, but judges had all slide tasks and items and some candidate performances in common (for more complete information, see Lunz, Wright, and Linacre, 1991).

The *oral examination* was part of a medical specialty board certification. This is an example of a loosely linked examination design. Twelve structured cases (tasks) were used. Each structured case described the nature of a patient's problem. During an interview with one judge, the candidate acquired additional information from the judge until a diagnosis could be made, a treatment determined, or a complication controlled. Candidates were rated holistically on each case, using a four category scale. Specific items were not defined. It was impossible for all judges to assess all candidates on all structured cases (312 candidates x 12 cases = 3744 performances), so cases were grouped by grading session and two judges were randomly assigned to interview each candidate.

Each candidate participated in two 20 minute oral examinations each of which included three structured cases, and a different judge. Thus, candidates interacted with different random pairs of judges on different structured cases depending on the particular grading session in which they were tested. This examination design is loosely linked. The judges graded all 12 structured cases, graded in all four grading sessions, and assessed some candidates in common. The oral examination offers the most opportunity for the case difficulty and the judge severity to affect candidate scores. Ratings were holistic, so each candidate had only six ratings (3 cases x 2 judges = 6 ratings). The statistical error of measurement for candidate ability estimates is thus necessarily large.

Data for these performance examinations were analyzed using FACETS (Linacre, 1988), a computer program for the Rasch analysis of performance assessment examinations.

#### 4. Results

*Histology Practical:* Table 1 shows the results of the histology examination. The severities of the judges range from .32 to -.29 logits with a mean standard error of .07 logits and the difficulties of the slides (tasks) range from 1.21 to -.88 logits with a mean error of .11 logits. The three items range in difficulty from -.62 to 1.05 logits. All judges are consistent within themselves. But there is more variance than expected in their assessment of slide #6, the easiest slide (infit = 1.6 and outfit = 2.9). This suggests that some judges gave lower than expected ratings on this easy task to some candidates. Candidate ability estimates ( $N = 93$ ) range from .09 to 4.25 logits with a mean of 1.51, mean error of .29,  $MSE = .096$  and  $V(\hat{\theta})$  of .56 logits. The reliability of candidate separation is .83.

Candidates performed the same tasks, were graded on the same items and had some judges in common due to the rotational judge linking design. Because candidates were rated by differing trios of judges, the varying severity of the judges who rated them impacts candidate scores. By accounting for judge severity, candidates are measured against a common criterion, regardless of the particular judges they encounter. Two candidates were selected for illustration. Both candidates earned a raw score of 108; however, candidate #68, who had moderate judges earned a logit measure of 1.70 logits, while candidate #87, who had severe judges, earned a higher logit measure of 2.00 logits. The outfit for these candidates were 1.7 and 1.9 respectively, indicating that each candidate received at least one unexpected rating from one of their three judges. Since these were relatively able candidates, it is likely they received one lower than expected rating.

TASKS IN DIFFICULTY ORDER						
Slide Number	Score	Count	Difficulty Calib	Error	Infit MnSq	Outfit MnSq
Least Difficult						
6	608	465	-0.88	0.16	1.6	2.9
12	575	465	-0.35	0.12	1.0	0.4
1	575	465	-0.32	0.12	0.8	0.6
5	575	465	-0.32	0.12	0.8	0.8
4	565	465	-0.18	0.12	1.0	1.3
11	562	465	-0.17	0.12	1.3	0.7
8	566	465	-0.12	0.12	0.9	0.4
10	559	465	-0.02	0.11	1.0	0.6
13	550	465	-0.01	0.11	1.1	0.9
3	537	465	0.17	0.11	1.0	1.0
2	534	465	0.20	0.10	1.0	1.2
7	539	465	0.22	0.11	1.3	0.5
15	530	465	0.22	0.10	1.1	1.0
14	516	465	0.36	0.10	0.9	0.7
9	430	465	1.21	0.09	1.1	0.9
Most Difficult						
Mean	548.1	465.0	0.00	0.11	1.1	0.9
S.D.	38.6	0.0	0.44	0.01	0.2	0.6
MSE 0.014 Separation Reliability 0.93						

JUDGES IN SEVERITY ORDER						
Judge Number	Score	Count	Severity Calib	Error	Infit MnSq	Outfit MnSq
Least Severe						
5	1141	825	-0.29	0.07	1.1	0.9
13	1067	795	-0.17	0.07	1.0	0.9
50	900	675	-0.02	0.07	1.1	1.1
34	880	690	0.03	0.07	1.1	1.0
6	760	585	0.13	0.08	0.9	1.0
3	780	615	0.32	0.07	0.9	1.0
Most Severe						
Mean	921.3	697.5	0.00	0.07	1.0	1.0
SD	140.0	87.4	0.18	0.00	0.1	0.1
MSE 0.005 Separation Reliability 0.84						
ITEMS IN DIFFICULTY ORDER						
Item Number	Score	Count	Difficulty Calib	Error	Infit MnSq	Outfit MnSq
Least Difficult						
2	1220	1395	-0.62	0.08	1.0	1.0
3	1097	1395	0.08	0.07	1.0	1.0
1	3210	1395	1.05	0.04	1.1	1.1
Most Difficult						
Mean	1842.3	1395.0	0.17	0.06	1.0	1.0
SD	968.4	0.0	0.69	0.02	0.0	0.1
MSE 0.005 Separation Reliability 0.99						

**Table 1:** Histology Practical Performance Examination

*Oral Examination:* Judges ranged in severity from 1.39 to -1.07 logits with a mean error .22, MSE .05 logits and cases ranged in difficulty from .43 to -.65 logits with a mean error .11, MSE .012 logits. Both facets show significant separation reliability (.84 and .87) indicating that the particular combination of judges and cases encountered by a candidate would impact his raw scores. Three judges showed inconsistency. Judge 18 (outfit = 1.6), a relatively severe judge, gave two ratings that were lower than expected to two candidates and one that was higher than expected to another. Judges 45 (outfit = 1.8), a moderate judge, and Judge 8 (outfit = 1.7), a lenient judge, gave lower than expected ratings to one candidate on one case. The patient cases varied in difficulty but were rated consistently across all judges and candidates.

Candidate estimated measures (N = 312) ranged from 4.02 to -1.40 logits. The mean was .77 ( $V(\theta) = 1.10$  and mean error = .60, MSE = .40) logits. The error of measurement associated with each candidate ability estimate was high because candidates were rated holistically on only six cases. Candidate ability estimates were adjusted for the severity of judges and the difficulty of the cases encountered, because the particular combination of facet elements influenced the probability of a satisfactory score. Two candidates will be used as examples. Candidate #45 encountered Judge #29 (severity = 1.33) and Judge #63 (severity = .75), and earned an ability estimate of .51 logits from a score of 7, because the adjustment accounted for severe judges. Candidate #252, in contrast, encountered Judges #96 (severity = -.43) and #84 (severity = -.15), and earned a total score of 12, but an estimated measure of

only .24 logits, because the adjustment accounted for the more lenient judges. The fit statistics for Candidate #252 were out of range (infit = 2.4 and outfit = 2.2). This shows that the judges did not agree on the ability of this candidate.

## 5. Discussion

In the histology performance examination, all candidates performed the same tasks (slides), while in the oral examination, candidates interacted with different tasks (structured cases). But both examinations showed different levels of measured difficulty for the tasks, suggesting that some tasks are easier for candidates to perform than others. In both examinations judges demonstrated measurably different levels of severity. Since this affects the probability of success for individual candidates, it must be accounted for to construct objective ability estimates. The multi-facet Rasch model objectifies the scoring.

The more items rated, the larger the number of observations, and the smaller the error of measurement associated with each item, task, judge, and candidate estimate. Of course, the rating scale must be constructed to apply to the examination, and the judges must be trained to use the rating scale consistently. While measurement error is never eliminated, it can be reduced by including the opportunity for more ratings. The holistic ratings given on the oral examination entail more measurement error because they are fewer ratings per candidate performance. Also the more abstract the holistic judgment, the more likely judges will vary in severity.

A reliance on judges to evaluate candidate performance must be coupled with statistical analysis to achieve "fairness". The use of a multi-facet Rasch model does not diminish the importance of training judges, but helps redefine their role in the examination process. Instead of being the final, individual arbiters on the quality of candidate performance, they are an element in the examination design. The elements of each facet are calibrated independently so that statistical adjustments can be made for the unique combinations of judges and tasks encountered by a candidate. This enables a uniform pass point to be set for all candidates. Passing or failing becomes independent of which judges, tasks, and items are encountered.

Implementing the multi-facet Rasch model for performance examinations requires some rethinking. The facets of the examination must be linked properly. Judges are relieved of the responsibility of making pass/fail decisions; however, they are also deprived of the time-honored mystique that a qualified judge can, somehow, individually, correctly, and absolutely determine the quality of a candidate's performance. Similar patterns were observable in both sets of data presented in this study and in writing assessment data presented by others (Englehard, 1992; Burger and Burger, 1994). Latent-trait modelling is accepted for multiple choice examinations. Evidence for its usefulness for performance examinations is demonstrated in this study.

## References

- Burger, S.E. and Burger D.L. (1994). Determining the validity of performance-based assessment. *Educational Measurement Issues and Practice*, 13, 1, 9-15.
- Englehard, G. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education*, 5, 3, 171-191.
- Fagot, R.F. (1991). Reliability of grades for multiple judges: Intraclass correlation and metric scales. *Applied Psychological Measurement*, 15, 1-11.
- Koretz, D. (1992). New report on Vermont portfolio project documents challenges. *National Council on Measurement in Education Quarterly Newsletter*, 1, 4, 1-2.
- Linacre, J.M. (1988). *FACETS*, a computer program for analysis of examinations with multiple facets. Chicago, IL: MESA Press.
- Linacre, J.M. (1989). *Many-Facet Rasch Measurement*. Chicago, IL: MESA Press.
- Lunz, M.E., Wright, B.D. and Linacre, J.M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, 3, 331-345.
- Lunz, M.E. and Stahl, J.A. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Professions*, 13, 4, 425-444.
- Lunz, M.E. and 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.
- Lunz, M.E., Stahl, J.A. and Wright, B.D. (1993). Interjudge Reliability and Decision Reproducibility. *Educational and Psychological Measurement*.
- Rasch, G. (1969/1980). *Probability models for some intelligence and achievement tests*. Chicago, IL: University of Chicago Press.
- Wright, B., and Stone, M. (1979). *Best Test Design*. MESA Press.
- Wright, B.D. (1995). Diagnosing person misfit. *Rasch Measurement Transactions*, 9:2, 430-431.
- Wright, B.D. and Linacre, J.M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8:30, 370.