Identifying Problematic Examiners in an Oral Examination

Jessica Heineman-Pieper, M.A. and Mary E. Lunz, Ph.D.

Measurement Research Associates, Inc.

Introduction

Oral examinations are most valid and reliable when the analyses identify and account for stable differences in how examiners assess and rate examinees (Heineman-Pieper & Lunz, 2000). This can be achieved using the multi-facet model. When oral examinations are analyzed using the multi-facet model, the multi-facet analysis calculates and adjusts for each examiner's unique level of 'severity' in rating examinees. As a result, examiners need no longer strive for the impossible goal of perfect agreement on candidate raw scores. Instead, examiners need only agree on the basic meaning of the rating scale categories and implement that understanding consistently within their own grading behavior. The multi-facet model provides fit statistics to indicate how well an examiner's rating behavior fits the expectations of the model. These fit statistics are frequently used to indicate how well an examiner understands and implements the rating scale categories. These statistics will also help determine which examiners will be allowed to participate in subsequent examinations.

Examiner fit statistics can signal when examiners are inconsistent, or when they fail to distinguish among candidates of different abilities. However, only the latter examiner behavior, *non-discrimination*, can be identified unambiguously from model outputs. In contrast, examiner *inconsistency* cannot be definitively diagnosed without careful review of specific raw data that may or may not have been collected. This paper demonstrates: (1) how non-discrimination can accurately be diagnosed using facets outputs; (2) why examiner inconsistency cannot be accurately diagnosed exclusively from standard model outputs, and (3) what additional information can accurately identify the quality of apparently inconsistent examiners.

Data

These data are from an oral certification examination. Candidates were rated on several skills across several practice areas by two independent but co-present examiners. Accordingly, the data were structured to allow a four facet analysis with the following facets: candidate, examiner, skill, and practice area.

Methods

The data were analyzed using the multi-facet model (FACETS), which adjusts the raw ratings that candidates earn to account for variations in examiner severity, skill difficulty, and difficulty of the practice area in which the ratings were earned. The multi-facet model generates measures of candidate ability, examiner severity, etc., in logits. Logit scores reflect the log odds that a candidate will receive a rating of a particular value, when rated by an examiner of a particular severity, in a practice area of a particular difficulty, on a skill of a particular difficulty.

In addition to constructing measures, the multi-facet model generates quality control statistics to indicate where the data do not provide a good fit to the model. The most important of these are the mean squared fit statistics (infit and outfit). The mean squared fit statistics construct a ratio that compares observed to expected responses. When an examiner's observed and expected responses perfectly coincide, the fit statistics for that examiner will be 1.0. When the examiner's observed responses are inconsistent with the expectations of the model, the mean squared fit statistics will be "high." Numerically, high has been found to correspond to approximately 1.5 or greater. When the variation in observed responses is less than would be expected by the model, the mean squared fit statistics will be low (generally, approximately 0.5 or lower).

When an examiner's fit statistics are too high or too low, the examiner is usually singled out as a potentially incompetent examiner. In the case of low fit statistics, this conclusion can indeed be verified using additional model outputs (specifically, the pattern of ratings). However, in the case of high fit statistics, it is far more difficult (and sometimes impossible) to distinguish a capricious examiner from an examiner faced with several slightly inconsistent candidates. A case example will show why (see Results section).

Results

1. When low fit statistics reveal that examiners are not differentiating among candidates of different abilities

Occasionally, examiners have extremely low fit statistics (<0.5). When this occurs, the most likely reason is that the examiner is not distinguishing among relevantly different levels of candidate ability. This hypothesis can be confirmed by examining the distribution of the examiner's ratings among the various rating scale categories. If the overwhelming majority of the examiner's ratings are concentrated in a single rating category, the examiner is not adequately distinguishing among candidate ability (see Table 1). If this pattern is pervasive among examiners, then the problem may lie with the definition of the skills and rating scale categories, which may not allow differentiation at the level necessary for the tested population. However, if this pattern is unusual among examiners, then the problem can be attributed to the examiner's understanding and implementation of those definitions. Table 1 shows a non-discriminating rating pattern and a well-distributed rating pattern. Figures 1 and 2 portray these contrasting patterns graphically.

2. Why high fit statistics are ambiguous with regard to examiner quality

To see why high fit statistics are ambiguous with regard to examiner quality, we turn to a case example. Examiner A was of thoroughly average severity (0.00 logits), and had a very high outfit statistic (1.9). This examiner gave some unexpectedly low ratings to some relatively able candidates. Overall, Examiner A gave 44 highly misfitting ratings (standard residual >=3.0), out of a total of 2329 ratings (2% of ratings). Based on these indicators, Examiner A appears inconsistent.

However, closer inspection of the rating patterns from Examiner A reveals that many of these unexpectedly low ratings were *corroborated* by Examiner B. In other words, Examiner B independently gave the same candidate the same low rating on the same skill and practice area — a response that similarly appeared as a misfit. Thus, the lack of fit may not reflect any problems in Examiner A's rating behavior. Examiner A accurately discriminated and documented fluctua-

tions in candidate ability. Of the 44 misfitting responses by Examiner A, half (22) were corroborated by the examiner's various partners.

Still, this may still be an inflated figure. Skill #6 was so easy that the model expects everyone to score well. When an examiner of average or lower severity gives a middling to excellent candidate a low mark on an easy skill, misfits may result even when the low mark is well deserved. In these instances, the misfit can actually indicate laudable examiner behavior, namely, that the examiner was able to report accurately the levels of performance on a skill where other examiners were non-discriminating. When we further eliminate misfitting responses for Examiner A that occurred on Skill #6, only 10 (0.4%) misfitting responses remain.

Conclusions

As these data and analyses reveal, extreme caution should be exercised before concluding that a misfitting examiner is internally inconsistent and should be excused. Whereas non-discriminating examiners can be accurately identified from model statistics alone, inconsistent examiners can only be identified definitively when additional information can be brought to bear. Specifically, if the examination structure enables examiner ratings to be compared with the ratings of a partner, evidence should be weighed for signs that candidate performances significantly differed from expectation. These data are necessary to determine when significantly elevated fit statistics indicate genuine instances of examiner inconsistency, and when they are produced artifactually from unusual candidate performance patterns.

In contrast, standard outputs can definitively identify examiners who fail to distinguish among candidates of widely varying ability. This pattern of non-discrimination is possible whenever an examiner has extremely low fit statistics (<0.5). The pattern can be definitively confirmed whenever the examiner's ratings overwhelmingly favor a middle category (Table 1). When this pattern of non-discrimination occurs pervasively, the responsibility for the problem may lie with vague or overly inclusive skill definitions and rating scale category definitions. Contrarily, when this pattern occurs in a minority of examiners, the problem may lie in examiners' individual interpretations and implementations of the rating scale. When non-discrimination is a problem only for a minority of examiners, those examiners should be instructed in the types of distinctions that must be recognized among candidates. Sometimes examiners do not adequately discriminate among candidates because the examiners are insecure about their rating abilities. At other times, examiner non-discrimination may reflect a problem in scaling; the examiner may have in mind a much wider ability range than is manifest in the actual population being tested. In either case, if training does not fix the problem, the examiner can justifiably be excused from future examinations.

Table 1. Examiner Use of the Rating Scale*

Rating Category	Percent of Ratings at each level	Mean Candidate Ability at each level
Well-Distributed Rating Pattern: Examiner Fit = 1.00 (Ex. 387)		
Unsatisfactory	4%	-0.76
Marginal	27%	-0.18
Satisfactory	53%	0.74
Excellent	15%	2.52
Non-Discriminating Rating Pattern: Examiner Fit = 0.4 (Ex. 431)		
Unsatisfactory	1%	-1.64
Marginal	2%	-.76
Satisfactory	90%	1.17
Excellent	7%	5.20

* For each examiner, the sample of candidates examined has a distribution of candidate abilities that is consistent with the overall candidate pool. Accordingly, rating patterns do not result from sampling errors.

Figure 1. Probability of Ratings at Each Level (0-3) Graphed against Candidate Ability: Well Distributed Rating Pattern

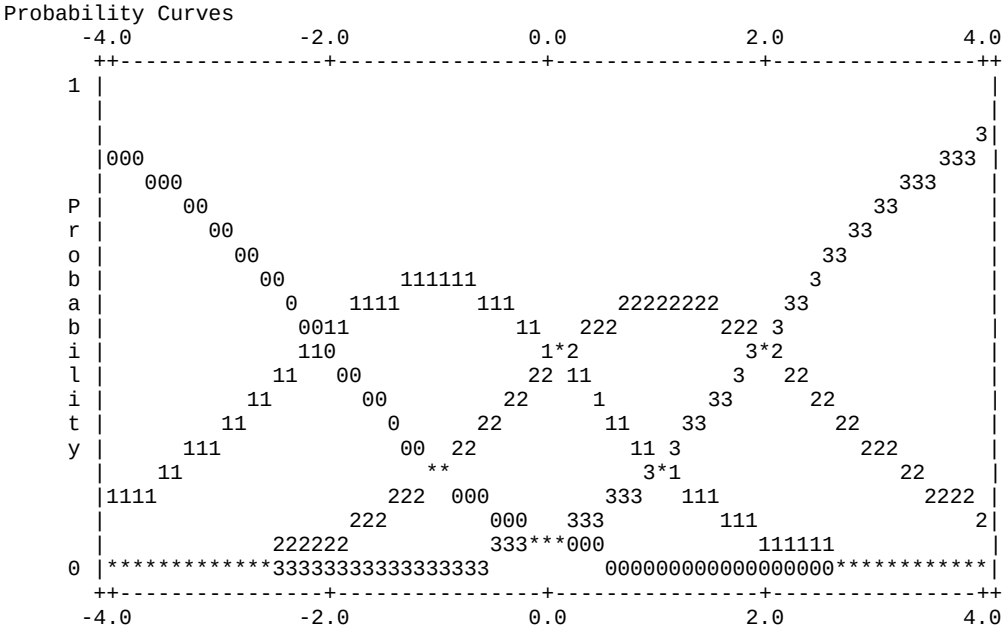

Figure 2. Probability of Ratings at Each Level (0-3) Graphed against Candidate Ability: Non-discriminating Rating Pattern