Factors that Impact Analytic Skill Ratings

Jessica Heineman-Pieper
Mary E. Lunz
Measurement Research Associates, Inc.

Holistic ratings lack sufficient information to measure candidates with the accuracy required for high-stakes certification examinations. When examiners make only one holistic rating of candidate performance, decisions about candidate ability are consumed with measurement error. Holistic ratings also make it impossible to determine the basis for the examiners' ratings, and to separate examiner severity from candidate ability. If another examiner gives a holistic rating to the same candidate, they often differ significantly.

In an effort to gather more information about the candidate, the pertinent clinical skills encompassed in the holistic rating were broken out, and examiners were asked to give separate analytic ratings, one for each skill. The problem is how to collect enough information to make pass\fail decisions about candidates that have minimal measurement error and reasonable confidence in their accuracy, while not asking examiners for redundant ratings.

The medical skills tested in an oral certification examination, diagnosis, treatment, and technical skill are conceptually related by the nature of the clinical situation. This is why they are selected for use in the examination. Can these skills be evaluated independently by examiners in the examination environment. Is it possible to evaluate the choice of treatment independently from the diagnosis?

Candidates have an ability to perform the clinical skills. This ability is expected to be reasonably stable across time, skills and applications. The goal of the examination is to certify candidates as safe, competent physicians. If candidate performance on the examination across skills or across cases were extremely volatile, this would challenge the expectation that candidate competence represents a single meaningful construct.

It seems impossible for a candidate who received a low mark for the pivotal skill, diagnosis, to receive a high mark for treatment, since it would be highly unlikely that the candidate's inaccurate diagnosis would happen to have the same treatment as does the correct diagnosis. In fact, when skills are arranged in their clinical sequence, it should be unlikely that a higher grade would ever follow a

Jessica Heineman-Pieper

Jessica Heineman-Pieper is working towards a Ph.D. in Psychology and The Conceptual Foundations of Science at the University of Chicago. She is also training to be a licenced clinical psychologist, and is a research associate at Measurement Research Associates. Jessica majored in American History and Literature in college and graduated Phi Beta Kappa (junior year) and with high honors from Harvard University. She was subsequently awarded a Rhodes Scholarship and graduated with honors in Philosophy and Psychology from Oxford University. Jessica hopes to teach and practice the new psychology Intrapsychic Humanism, popularly articulated in the parenting book "Smart Love: The Compassionate Alternative to Discipline That Will Make You a Better Parent and Your Child a Better Person."

Mary E. Lunz

Mary E. Lunz earned a Ph.D. from Northwestern University. After teaching and consulting for several years, Mary worked as Director and Psychometrician for the Board of Registry of the American Society of Clinical Pathologists for 17 years. During this time, she began working with Ben Wright and Michael Linacre on issues relating to performance examinations, and computerized adaptive testing. Research is still ongoing. Mary is currently Director and Senior Associate at Measurement Research Associates, Inc.

lower grade. Therefore, skills conceptually arranged in clinical sequence, should show the same or consistently decreasing scores.

However, conceptual relation and lack of rating independence do not consider the relative difficulties of the skills. Relative skill difficulty levels result from the unique demands each skill requires. Skill difficulties are established independently of candidate abilities or examiner severities, with the Rasch multi-facet model (Linacre, 1989). Generally, candidates receive lower scores on more difficult skills and higher scores on easier skills, regardless of the clinical sequencing of the skills. When an easier skill is followed by a harder skill, candidates' scores are likely to decrease more often than not. Likewise, when a harder skill is followed by an easier skill, we expect candidates' scores to increase more often than not.

Data are from two different medical oral certification examinations. Skill ratings were given to candidates on a four point scale (EX1 scale = 1,2,3,4 and EX2 scale = 0,1,2,3). Both examinations were analyzed with the FACETS program (Linacre, 1990).

In the first examination, EX1, oral examiners rated candidates on three skills on each of four standardized cases. The skills were: 1) data /interpretation; 2) diagnosis; and 3) management. In this examination, examiners informed candidates of errors to insure that candidates continued through the standardized case as established. This examination is structured to minimize the effects of conceptual dependence and foster independent skills assessments. The second medical examination EX2, examined each candidate on cases from the candidate's actual practice. Candidates were rated on six skills: (1) data gathering; (2) diagnosis; (3) treatment; (4) technical skills (of surgery); (5) outcomes; and (6) ethics.

The FACETS program establishes a fair average score for each candidate on each skill. The fair average score is the score expectation of the logit measure and accounts for the severity of the examiner and difficulty of the standardized case. The fair average score is used in this analysis to make it easier to relate the scores to the rating scale. When fair average scores are the same for two skills, the ratings may not be independent, or the candidate may have the same level of ability on both skills. When the fair average scores differ, this suggests that examiners were able to distinguish candidate performance or that candidates demonstrated different levels of ability on each skill.

Diagnosis, a pivotal skill, is used for comparison to the other skills. Diagnosis is also a relatively easy skill for both EX1 and EX2, as shown in Tables 1 and 2. Therefore candidates should earn 1) the same or lower fair average scores on

Table 1. Skill Difficulty Measures for EX1

Conceptual Order	Difficulty (in logits)
Data Gathering	0.00
Diagnosis	-0.18
Treatment	0.18

Graph 1. Comparison of Performance on Two Skills

Data/Interpretation was more difficult
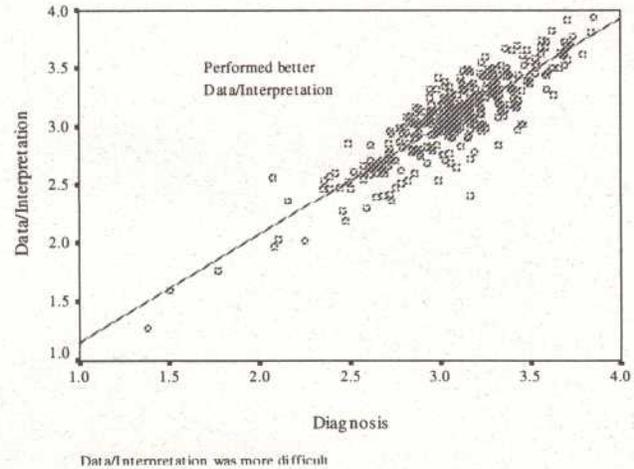
Graph 2. Comparison of Performance on Two Skills

Management was more difficult
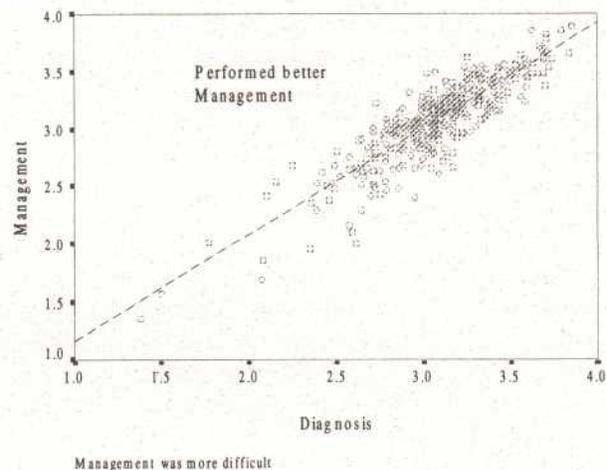
subsequent skills according to the conceptual relations; 2) the same fair average scores among skills if the ratings are dependent or the candidate is consistent; or 3) varying fair average scores according to the calibrated difficulty, and independent assessment of candidate ability.

Table 2. Skill Difficulty Measures for EX2

Conceptual Order	Difficulty (in Logits)
Data Gathering	.09
Diagnosis	−.21
Treatment	.16
Technical Skill	.08
Technical Skill	.05
Ethics	−.52

Table 1 shows the Rasch calibrated skill difficulties for EX1. Diagnosis is the easiest skill. Graphs 1 and 2 show the comparison of the fair average scores for diagnosis (easier) with data gathering (harder) or management (harder) respectively. Most candidates earned comparable fair average scores among skills, supporting the consistency of candidate ability among skills. However,

Graph 3.

Comparison of Performance on Two Skills

Performed better
Data Gathering

Diagnosis

Data Gathering was more difficult

some candidates earned higher or lower fair average scores on data/interpretation or management. This provides some evidence that examiners rate the skills independently based on their observation of the candidate and the difficulty of the skill.

Table 2 shows the calibrated difficulties of the skills for EX2. Diagnosis is one of the easiest skills on which to earn a high score. Graphs 3 - 6 show the comparisons of fair average scores when diagnosis is compared to data gathering, treatment, technical skills, and outcomes respectively. Many of the candidates earn comparable fair average scores among skills. This is commensurate with the premise that candidates have a stable ability that can be measured. However, some candidates earn higher fair average scores on the clinically subsequent skills, showing that examiners can evaluate candidate performance, independent of the underlying conceptual relationships. These results show that the calibrated difficulty of the skill is not driven by conceptual relations among skills. While the functional relationship among the skills is critical to the coherence of the overall examination, the functional relationship does not control examiners' ratings. Rather, examiners seem to be able to rate candidates on each skill independently. This pattern holds true when examiners rate candidates on cases from their actual medical practices, or on standardized cases developed by the Board. The use of analytic ratings may not be foolproof, but examiners' analytic ratings appear to be independent, even when skills are conceptually related. In addition, the use of analytic rather than holistic ratings,

Graph 4.

Comparison of Performance on Two Skills

Performed better
Treatment

Diagnosis

Treatment was more difficult

Graph 5.

Comparison of Performance on Two Skills

Performed better
Technical Skill

Diagnosis

Technical Skill was more difficult

Graph 6.

Comparison of Performance on Two Skills

Performed better
Outcome

Diagnosis

Outcomes was more difficult

has the advantage of collecting a sufficient amount of information about each candidate to make pass and fail decisions with minimal measurement error and a high level of confidence.

References

Linacre, J.M.. (1989). Many-facet Rasch measurement. Chicago: MESA Press.

Facets: A computer program. Chicago: MESA Press.