Addressing Rater Error in Performance Assessments
Bringing the Many Facet Rasch Model (MFRM) to Assessment

Myrah R. Stockdale, M.S.
Objectives

Participants will be able to:

• describe the impact of rater variability on score validity and fairness

• recognize benefits of Many Facet Rasch Model (MFRM) in performance assessment contexts

• explain ways to improve performance assessments and scores
Presentation Outline

- Introduce Performance Assessments (PA)
- Issues Presented by Rater / Judge Variability
- Influence of Norm- vs. Criterion-Referenced Assessment on PAs
- Assessing Score Reliability with the Many-Facet Rasch Model (MFRM)
Performance Assessments

A Brief Introduction
Performance Assessment

• What is it?
  • What are some examples?

• Why is it important?
Performance Assessment

Performance Assessment is

• Evaluation of student performance on a given formative assessment task.
• Comprised of a task and a rubric
• Students showing judges, raters, instructors what they know
• Often used in conjunction with other assessments and tests

... and is not

• Looking at how to teach (looks at how to assess learning)
• Usually summative assessment
Performance Assessment

Test Formats:

- Constructed Response
- Essay
- Demonstration
- Oral Presentation
- Role-Playing
- Exhibits
- Portfolios
- Direct Observations
**Performance Assessment**

**Benefits**

• Examinees can demonstrate skills that are not easily assessed with multiple-choice (MC) items (Messick, 1996).
  • Such as higher order thinking and/or problem solving skills (Muraki, Hombo, & Lee, 2000)

• Considered more authentic and relevant to real-life situations (Haertel & Linn, 1996; Sax, 1997).
Performance Assessment

Trade-offs

- Typical PAs have less items/tasks for examinees to complete; statistically this negatively affects reliability.
  - However, there are ways to introduce ancillary exam information from other sources to increase reliability (Ackerman, 1991; Luecht & Miller, 1992)

- PAs can be more resource intensive than traditional exams
Variability and Validity

And A Little About Context
Performance Assessments and Rater Error

- Raters regularly provide inconsistent ratings and judgements.

- Variation in ratings/judgement can have significant impacts on:
  - Examinees’ grades
  - Validity

- NOTE: Validity is not a property of the assessment, but rather of the meaning of the scores.
Assessment Contexts

Norm-Referenced Assessment

- Performance is relative to peers
- Arguably *inappropriate* for performance assessment in higher education assessment contexts

Criterion-Referenced Assessment

- Performance is compared relative to a particular standard
# Rater Error

<table>
<thead>
<tr>
<th>Type of Error</th>
<th>Tendency for Rater to</th>
</tr>
</thead>
<tbody>
<tr>
<td>Halo Effect</td>
<td>Give student the benefit of the doubt and provides high score</td>
</tr>
<tr>
<td>Gestalt Phenomenon</td>
<td>Provide ratings that are based on the overall impressions because the rater is unable to differentiate various aspects of the student’s performance</td>
</tr>
<tr>
<td>Leniency</td>
<td>Consistently provide inflated ratings</td>
</tr>
<tr>
<td>Scale Interpretation</td>
<td>Provide inappropriate ratings because the rater misunderstands the scale and perceives a rating to mean something different</td>
</tr>
<tr>
<td>Skimming</td>
<td>Evaluate only a limited portion of a performance and assign a score</td>
</tr>
<tr>
<td>Severity</td>
<td>Consistently provide overly critical ratings</td>
</tr>
</tbody>
</table>

(Royal & Hecker, 2015)
Example: Scale Interpretation

On a scale of 1 to 7, how well did the student perform on the following tasks:

<table>
<thead>
<tr>
<th>Task</th>
<th>Substantially exceeds expectations</th>
<th>Far exceeds expectations</th>
<th>Exceeds expectations</th>
<th>Equals expectations</th>
<th>Short of expectations</th>
<th>Far short of expectations</th>
<th>Substantially short of expectations</th>
<th>Not Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interpreting physician orders</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
</tr>
<tr>
<td>Labeling of pharmaceuticals</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
</tr>
<tr>
<td>Compliance with state and federal drug laws</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
</tr>
</tbody>
</table>
Improving Validity

• Make sure objectives are clearly defined and operationalized.

• Match assessment measures to objectives.

• Get students involved; have the students look over the assessment for troublesome wording, interpretation issues, or other difficulties.

• If possible, compare your measure with other measures, or data that may be available.
The Many Facet Rasch Model

A Measurement Model
Measurement vs. Statistics

Measurement

• Focuses on:
  • Meaningfulness of numbers (Falmange & Narens, 1983)
  • Individual processes and relations within the variables (Duncan, 1992)
  • Imposes strong requirements on data quality in order to achieve a common metric (Cohen, 1994)

Statistics

• Focuses on:
  • group processes and relationships among variables (Fisher, 2010)
  • Makes assumption about factors beyond its control
Rasch Family of Models

• In the Rasch model, the probability of a specified response (e.g. right/wrong answer) is modeled as a function of person and item parameters.

• The models represent the structure which data should exhibit in order to obtain measurements from the data.
Many-Facet Rasch Model Formula

\[
\log \left( \frac{P_{nijk}}{P_{nijk-1}} \right) = B_n - D_i - C_j - F_k
\]

where

- \( P_{nijk} \) is the probability of examinee n being awarded on item I by judge j a rating of k
- \( P_{nijk-1} \) is the probability of examinee n being awarded on item I by judge j a rating of k-1
- \( B_n \) is the ability of examinee n
- \( D_i \) is the difficulty of item I
- \( C_j \) is the severity of judge j
- \( F_k \) is the difficulty of step up from category k-1 to category k
Many-Facet Rasch Model

- MFRM can be used to evaluate a number of specific hypotheses related to potential sources of rater error that represent threats to the validity of ratings.

- It is recommended that the MFRM be considered a priori rather than post hoc

  - **Why?** Identifying and labeling the elements being measured will help with data analysis, data cleaning, and interpretation
Many-Facet Rasch Model

Logit-to-Probability Conversion Table

<table>
<thead>
<tr>
<th>Logit Diff.</th>
<th>% Success</th>
<th>Logit Diff.</th>
<th>% Success</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.0</td>
<td>99%</td>
<td>-5.0</td>
<td>1%</td>
</tr>
<tr>
<td>4.0</td>
<td>98%</td>
<td>-4.0</td>
<td>2%</td>
</tr>
<tr>
<td>3.0</td>
<td>95%</td>
<td>-3.0</td>
<td>5%</td>
</tr>
<tr>
<td>2.2</td>
<td>90%</td>
<td>-2.2</td>
<td>10%</td>
</tr>
<tr>
<td>2.0</td>
<td>88%</td>
<td>-2.0</td>
<td>12%</td>
</tr>
<tr>
<td>1.0</td>
<td>73%</td>
<td>-1.0</td>
<td>27%</td>
</tr>
<tr>
<td>0.0</td>
<td>50%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Person 10 has a logistic measure of 1.0.
Person 9 and 17 have a logistic measure of 0.0.
Essay 3 has a logistic measure of 0.0.
Person 29 has a logistic measure of -1.0.
Presentation and Measurement Perspectives on a Representative Rating Scale

The scale as it is printed:

<table>
<thead>
<tr>
<th>Deficient</th>
<th>Poor</th>
<th>Medium</th>
<th>Good</th>
<th>Excellent</th>
<th>(Word Label)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>(Numerical Label)</td>
</tr>
</tbody>
</table>

The scale as it functions:

<table>
<thead>
<tr>
<th>Deficient</th>
<th>Poor</th>
<th>Medium</th>
<th>Good</th>
<th>Excellent</th>
<th>(Word Label)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(Estimated Location)</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>(Numerical Label)</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>(Qualitative Steps)</td>
</tr>
</tbody>
</table>

(Linacre, 1993)
Many-Facet Rasch Model

- Examinee A has a higher measure but a lower average rating than Examinee B, because A was rated by a more severe judge.

(Linacre, 1993)
Test Reliability Analysis

MFRM

• Provides detailed information about what caused the inconsistency or where the disagreement is.

• The analysis of rating data with the MFRM does not require all judges to rate all items to arrive at an estimate of interrater reliability.

Classical Test Theory

• Cronbach’s alpha is used to determine how consistent elements (e.g., judges) are.

• Generates a sketchy picture of the internal consistency of the test and inter-rater agreement.

Final Remarks

- Validity is of the utmost importance when designing and implementing performance assessments
  - Use a criterion-references context
  - Train raters and judges

- The Many-Facet Rasch Model (MFRM), is a powerful tool that can help in pinpointing the location of inconsistencies and/or discrepancies among a number of elements
References


- Cohen, J. (1994). The earth is round (p < 0.05). American Psychologist, 49, 997-1003


Resources

• Program for running MFRM : FACETS (Trial Version: MiniFac)

• Tutorials on How to Run MFRM by Michael Linacre
  • Tutorial 1 Software operation and basic concepts
  • Tutorial 2 Fit analysis and measurement models
  • Tutorial 3 Estimation and interactions
  • Tutorial 4 Anchoring