THE PSYCHOMETRICS OF PERFORMANCE ASSESSMENTS – A PHYSICIAN’S PERSPECTIVE

Jeanne M. Sandella, DO
Objectives

• Discuss the reliability of performance assessments – why are they used, what are the threats to having a “good assessment?”
• Identify the impact of human raters on the psychometrics of performance assessments
• Define the examinee – centered approach for setting standards for performance assessments
Why a physician’s perspective?

\[ y = \log_e \left( \frac{x}{m} - sa \right) \]

\[ yr^2 = \log_e \left( \frac{x}{m} - sa \right) \]

\[ e^{yr^2} = \frac{x}{m} - sa \]

\[ m e^{yr^2} - x - sma \]

\[ me^{rry} = x - mas \]
What makes a good assessment? (Kane’s framework)

1. Scoring
2. Generalization
3. Extrapolation
4. Interpretation/decision
1. Scoring

- Fair administration
- Individuals were evaluated accurately
- Scoring rules consistently applied
Fair administration
Accurate evaluations

Scoring

- **Checklists**
  - measure an explicit process (H and P)
  - Evidence based development – by experts

- **Rubrics**
  - measure implicit processes;
  - Holistic
  - Require significant expertise/training
Human Raters – are they consistently applying rules?

- Training of raters (physician and SP)
- Score equating/calibration
- Quality assurance of raters
Human raters
2. Generalization

Generalization

- Good sampling of observations
- Enough samples
Generalization factors contributing to errors.

Heterogeneity of candidates
Generalization factors contributing to errors.

Number of cases
## Generalize results – adequate sampling

<table>
<thead>
<tr>
<th>Case 1</th>
<th>Judge 1</th>
<th>Judge 2</th>
<th>Judge 3</th>
<th>Judge 4</th>
<th>Judge 5</th>
<th>Judge Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>Case 2</td>
<td>B</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>Case 3</td>
<td>B</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>Case 4</td>
<td>B</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>Case 5</td>
<td>B</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>Case X</td>
<td>B</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
</tbody>
</table>
3. Extrapolation

- Observations are relevant to the construct of interest
- Scores are not unduly influenced by sources of variance not related to the construct being measured
FOR A FAIR SELECTION EVERYBODY HAS TO TAKE THE SAME EXAM! PLEASE CLIMB THAT TREE
• Framework for score interpretation can be supported
• Categorization is supported
Standard setting

• What do the scores mean?
  – E.g. Pass/Fail
Standard Setting Anatomy

A • Standard setting method

B • Defining a performance standard

C • Derive the Cut-point

D • Finalization of Cut score
Panelists make independent judgments of qualified or not qualified performance on the clinical skill of interest by reviewing actual or proxy performance on the examination.
Defining a performance standard

Not qualified

Qualified
Deriving a cut score

$P = Q$
Finalization of Cut score

Triangulation

NBOME Executive Committee

Stakeholder Surveys

Expert Panelists
Factors contributing to classification errors.

The cut score

Higher cutscore = more false negatives.
Lower cutscore = more false positives.
Objectives

• Discuss the reliability of performance assessments – why are they used, what are the threats to having a “good assessment?”

• Identify the impact of human raters on the psychometrics of performance assessments

• Define the examinee – centered approach for setting standards for performance assessments
Selected references


THANK YOU.