Validity Evidence Using Generalizability Theory for an Objective Structured Clinical Examination (OSCE): a multiple occasions rationale

Michael J. Peeters, PharmD, MED, FCCP, BCPS*; M. Kenneth Cor, PhD, MED, BScEng#; Bryan M. Bishop, PharmD, BCPS*; Sarah E. Petite, PharmD, BCPS*; Michelle N. Schroeder, PharmD, BCACP, CDE*

*University of Toledo College of Pharmacy & Pharmaceutical Sciences, Toledo OH, USA
#University of Alberta, Faculty of Pharmacy & Pharmaceutical Sciences, Edmonton AB, Canada

Purpose

- To evaluate how our reliability of OSCE grading changed depending on number of stations over multiple weeks/occasions

Key Findings [Implications]

- This OSCE was successfully modeled with multivariate G-theory
- Students accounted for only some variation in OSCE scores
  - Variation in scores from other influences as well (station, rater, week)
- Our reliability improved by increasing the number of stations each week and/or number of weeks

Why did we do this study? [Background]

- OSCEs are gold-standard for performance assessment (skills)¹
  - Can be used to help evaluate PharmD students’ practice-readiness
- Pharmacy colleges/schools should be generating validation evidence for assessments used to make important decisions; including OSCEs²
- Kane’s Framework for validation¹
  - Scoring ➔ Generalization ➔ Extrapolation ➔ Implications
- Generalizability Theory (G-theory) is one method for reliability⁴
- Prior studies in pharmacy education rarely report the use of G-theory to produce reliability evidence as part of the validation process³

What did we do? [Methods]

- IRB approved
- Assessment of OSCE scores from 3rd-year PharmD students
  - In Spring Semester before their final-year rotations
  - Assessed skills with: counselling & communication, case presentation, prescription verifying, medication reconciliation, non-adherence, drug information, pharmacy calculations and general drug knowledge
- OSCE included 14 stations over 3 weeks with 4 or 5 stations per week
  - 1 or 2 stations per week were scored by faculty-raters
  - 3 stations per week required students’ written responses
  - mGENOVA software was used for Multivariate G-theory
  - \( p \times w \) model used
  (Students crossed with stations nested in 3 weeks/occasions)

What did we find? [Results]

- Ninety-seven students completed this OSCE
- Stations were scored independently
- G-study: Our estimated g-coefficient (reliability) for the total score based on three weeks and approximately five stations per week was estimated at 0.74

<table>
<thead>
<tr>
<th>Week#</th>
<th>Stations/week</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Week 1</td>
<td>0.19</td>
</tr>
<tr>
<td>Week 2*</td>
<td>0.06</td>
</tr>
<tr>
<td>Week 3</td>
<td>0.33</td>
</tr>
<tr>
<td>Total-Score</td>
<td>0.37</td>
</tr>
</tbody>
</table>

Further analysis (decision studies) revealed how our reliability changed depending on combinations of stations per week

Decision studies (G-Coefficients Estimates)

- Ninety-seven students completed this OSCE
- Stations were scored independently
- G-study: Our estimated g-coefficient (reliability) for the total score based on three weeks and approximately five stations per week was estimated at 0.74

<table>
<thead>
<tr>
<th>Our Variation Sources (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>OSCE Week1</td>
</tr>
<tr>
<td>student</td>
</tr>
<tr>
<td>station</td>
</tr>
<tr>
<td>( p \times w ) (and other error)</td>
</tr>
</tbody>
</table>

Further analysis (decision studies) revealed how our reliability changed depending on combinations of stations per week

<table>
<thead>
<tr>
<th>Week#</th>
<th>Stations/week</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Week 1</td>
<td>0.19</td>
</tr>
<tr>
<td>Week 2*</td>
<td>0.06</td>
</tr>
<tr>
<td>Week 3</td>
<td>0.33</td>
</tr>
<tr>
<td>Total-Score</td>
<td>0.37</td>
</tr>
</tbody>
</table>

*Improving stations especially in Week #2 should also help improve our total-score reliability

Expected Reliability [Graph of purple table above]

References