Test-retest reliability for common tasks in vision science
Kait Clark¹, Charlotte R. Pennington¹, Craig Hedge ², Joshua Lee¹, & Austin Petrie¹
¹University of the West of England, ²Cardiff University

Reliability
- Consistency in results produced by a measure
- Test-retest reliability: Correlation of scores taken at 2+ points in time
- How well can a measure consistently distinguish between individuals who have high/low scores?

Repeatability = reliability
- Highly robust tasks are often unreliable (e.g., Stroop)
- Test-retest reliability is required to assess individual differences

Intraclass correlation coefficient (ICC)
\[ ICC = \frac{\text{Variance between individuals}}{\text{Variance between individuals} + \text{error variance}} \]

- Low ICC could represent high error or the fact that individuals are very similar
  - *cognitive tasks often designed to minimize variance

Test-retest reliability is required to assess individual differences

Participants
- 165 undergraduate psychology students
- Testing sessions:
  - Two sessions, separated by 1-3 weeks
  - Each session two hours in length:
    - 1 hour social cognition tasks (Pennington et al., in prep)
    - 1 hour perceptual tasks (current paper)

Tasks
- Motion coherence (MoCo)
  - Indicate direction of motion
  - 400 dots, random motion
  - Starting trial: 24% coherent
  - 3 staircases, 1-up/1-down (%)

- Useful field of view (UFOV)
  - Fixate in the centre
  - Central number / peripheral dot flash for 90 ms
  - Report number and dot location

- Multiple-object tracking (MOT)
  - 3-6 items flash (targets)
  - All items move for 6.5 seconds
  - Track the locations of targets
  - Items stop moving
  - Click on all of the targets

- Visual working memory (VWM)
  - Two sets of squares presented
  - One square might change (50%) indicate change or no change
  - Set sizes range from 2-7

Test-retest reliabilities

<table>
<thead>
<tr>
<th>Task/measure</th>
<th>ICC</th>
<th>Rho</th>
</tr>
</thead>
<tbody>
<tr>
<td>MoCo Threshold</td>
<td>0.60</td>
<td>0.57</td>
</tr>
<tr>
<td>UFOV Number Accuracy</td>
<td>0.47</td>
<td>0.48</td>
</tr>
<tr>
<td>Inner Accuracy</td>
<td>0.43</td>
<td>0.48</td>
</tr>
<tr>
<td>Middle Accuracy</td>
<td>0.60</td>
<td>0.65</td>
</tr>
<tr>
<td>Outer Accuracy</td>
<td>0.74</td>
<td>0.75</td>
</tr>
<tr>
<td>MOT Max Items</td>
<td>0.41</td>
<td>0.36</td>
</tr>
<tr>
<td>Threshold</td>
<td>0.36</td>
<td>0.31</td>
</tr>
<tr>
<td>VWM Capacity</td>
<td>0.75</td>
<td>0.77</td>
</tr>
</tbody>
</table>

*similar to Xu et al., 2018

Spearman’s Rho
- Accounts for outliers
- Tends to be a more lenient measure
- Aligns with ICC

Relative size of variance components:

Consider test-retest reliability before assessing individual differences

- Many cognitive tasks not designed to discriminate between individuals
- Unlikely to find IDs for tasks with low ICCs
- If a measure does not correlate with itself, it’s unlikely to correlate with anything else

References

Participants and testing
- 165 undergraduate psychology students
- Testing sessions:
  - Two sessions, separated by 1-3 weeks
  - Each session two hours in length:
    - 1 hour social cognition tasks (Pennington et al., in prep)
    - 1 hour perceptual tasks (current paper)

References