Evaluating Quality Assessments in Meta-Analysis

Ryan T. Williams
& Terri D. Pigott
Loyola University Chicago

Brief History

- Quality reporting standards have aggressively emerged over the past three decades (e.g. Mostellar, Gilbert, & McPeak, 1980)
- Quality assessment instruments have also emerged as a product of the evidence-based practice movement
Quality Assessment

- Intended to capture and measure quality as a scientific construct
- The Jadad Scale (Jadad et al., 1996) in the medical sciences; the Maryland Scientific Methods Scale (Farrington, Gottfredson, Sherman, and Welsh, 2002) in crime and justice research

The Logic

- What does the “best” evidence say?
- The information yielded from higher quality studies will be more informative than lower quality studies
- Will ultimately lead to better policy decisions
The Continuum

- Variations in quality control
  - Quality reporting standards
  - Quality assessment as part of moderator analysis in meta-analysis
  - Weighting effect-size coefficients by quality indicators

The Controversy

- Two approaches:
  - Inclusiveness
    - Examine the big picture including methodologically poor studies and perform relevant subgroup analyses
  - Exclusiveness
    - Exclude substandard methodologically implemented studies or charge extra for admission to the show
Some Empirical Insight

- Using the Jadad Scale, Moher et al. (1998) found that low-quality studies may inflate treatment effects by 30-50%.
- Jüni et al. (1999), however, found divergent results based on the quality instrument that was used in the analysis.

Purpose of This Work

- Demonstrate the use of Item Response Theory (IRT) methods in evaluating study quality assessment instruments in meta-analysis.
- Evaluate the measurement properties of one subscale of the Study Design and Implementation Device (DIAD): the external validity scale.
Why DIAD?

- Comprehensiveness
- Why the external validity subscale?
  - Most number of items (11) and uses both ordinal and dichotomous rating scales

Study DIAD

- The DIAD (Valentine & Cooper, 2008) proposed as a unique quality assessment device
  - Departure from single-score reliance
  - Operationalization
  - Transparency
Structure of the DIAD

- 16 contextual items
  - Orient the rater and forces operationalization
- Four subscales
  - Internal Validity
  - External Validity
  - Construct Validity
  - Statistical Conclusion Validity

Structure of the DIAD Cont’d

- Three item tiers
  - 32-34 Study design and implementation items
  - Eight composite items based upon the results of the design and implementation items
  - Four global items based on the results from the eight
- External subscale
  - 6 ordinal items (four point scale)
  - 5 dichotomous items
Methods

- Rate 78 crime and justice experimental and quasi-experimental studies on the external validity subscale of the DIAD
- Each study represents on “respondent”
  - Quality being a construct belonging to the designed study, not raters (psychometric)

Studies

- A total of 78 studies have been analyzed for the presented results
Studies Cont’d

- 51% of the studies were publications from refereed journals
- 47% were program evaluation reports that were not published in refereed journals
- 1% were dissertations or theses, and 1% were edited book chapters

Inter-Rater Reliability

- Establishment of Inter-rater reliability (IRR)
  - 20% of the study sample was randomly assigned to expert rater for scoring
- IRR was sufficient (.83)
IRT Model

- Examined the DIAD external validity scale using a partial credit model (PCM; Masters, 1982)
- Depends on unidimensionality of construct
- Depends on absence of local item dependence
- Assumes each item to have its own rating scale (allowing both dichotomous and polytomous items to be entered)

Results

- First Iteration – Original Structure
  - Study (person) reliability (.69)
  - Sample ability variance
  - Item reliability (.97)
  - Item difficulty variance
  - Internal consistency (.72)
  - Local item dependence between two items (appropriate participants and appropriate time period)
Table 1
Item Fit Statistics for Original Study DIAD Rating Scale Structure

<table>
<thead>
<tr>
<th>Item/Measure</th>
<th>Average Measure</th>
<th>Infit MaSq</th>
<th>Infit Z std</th>
<th>Item/Measure Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>D11 Variation in Implementation</td>
<td>68.9</td>
<td>1.49</td>
<td>2.0</td>
<td>.31</td>
</tr>
<tr>
<td>D3 Variation in Target Setting</td>
<td>69.4</td>
<td>.79</td>
<td>-1.1</td>
<td>.47</td>
</tr>
<tr>
<td>D7 Tested Participant Subgroups</td>
<td>59.7</td>
<td>.08</td>
<td>-1</td>
<td>.58</td>
</tr>
<tr>
<td>D5 Appropriate Time of Measurement</td>
<td>33.9</td>
<td>1.23</td>
<td>.6</td>
<td>.47</td>
</tr>
<tr>
<td>D1 Appropriate Participants</td>
<td>29.1</td>
<td>1.23</td>
<td>5</td>
<td>.51</td>
</tr>
<tr>
<td>D6 Appropriate Time Period</td>
<td>29.1</td>
<td>1.23</td>
<td>5</td>
<td>.51</td>
</tr>
<tr>
<td>D2 Variation in Participants</td>
<td>55.8</td>
<td>1.11</td>
<td>.8</td>
<td>.39</td>
</tr>
<tr>
<td>D8 Tested Variation in Setting</td>
<td>71.5</td>
<td>.76</td>
<td>-1.5</td>
<td>.49</td>
</tr>
<tr>
<td>D9 Tested Important Outcomes</td>
<td>53.3</td>
<td>.85</td>
<td>-8</td>
<td>.71</td>
</tr>
<tr>
<td>D10 Tested Time of Measurement for Effects</td>
<td>32.9</td>
<td>.82</td>
<td>-2</td>
<td>.46</td>
</tr>
<tr>
<td>D4 Included Important Outcomes</td>
<td>44.5</td>
<td>.78</td>
<td>-1.6</td>
<td>.68</td>
</tr>
</tbody>
</table>
Results Cont’d

<table>
<thead>
<tr>
<th>Item Fit Statistics for Original Study DIAD Rating Scale Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>D11 Variation in Implementation</strong></td>
</tr>
<tr>
<td><strong>D9 Tested Important Outcomes</strong></td>
</tr>
<tr>
<td><strong>D8 Tested Variation in Setting</strong></td>
</tr>
<tr>
<td><strong>D7 Tested Participant Subgroups</strong></td>
</tr>
<tr>
<td><strong>D6 Appropriate Time Period</strong></td>
</tr>
<tr>
<td><strong>D5 Appropriate Time of Measurement</strong></td>
</tr>
<tr>
<td><strong>D4 Included Important Outcomes</strong></td>
</tr>
<tr>
<td><strong>D3 Variation in Target Setting</strong></td>
</tr>
<tr>
<td><strong>D2 Variation in Participants</strong></td>
</tr>
<tr>
<td><strong>D1 Appropriate Participants</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Item</th>
<th>Average Measure</th>
<th>Infit Mnsq</th>
<th>Infit Z std</th>
<th>Item/Measure Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>D11</td>
<td>68.9</td>
<td>1.49</td>
<td>2.0</td>
<td>0.31</td>
</tr>
<tr>
<td>D9</td>
<td>69.4</td>
<td>0.79</td>
<td>-1.1</td>
<td>0.47</td>
</tr>
<tr>
<td>D7</td>
<td>50.7</td>
<td>0.98</td>
<td>-1.1</td>
<td>0.38</td>
</tr>
<tr>
<td>D8</td>
<td>33.9</td>
<td>1.23</td>
<td>6.0</td>
<td>0.47</td>
</tr>
<tr>
<td>D4</td>
<td>59.1</td>
<td>1.23</td>
<td>5.0</td>
<td>0.51</td>
</tr>
<tr>
<td>D6</td>
<td>29.1</td>
<td>1.23</td>
<td>5.0</td>
<td>0.51</td>
</tr>
<tr>
<td>D2</td>
<td>58.8</td>
<td>1.11</td>
<td>8.0</td>
<td>0.39</td>
</tr>
<tr>
<td>D1</td>
<td>71.5</td>
<td>0.76</td>
<td>-1.5</td>
<td>0.49</td>
</tr>
<tr>
<td>D3</td>
<td>55.3</td>
<td>0.85</td>
<td>-8.0</td>
<td>0.71</td>
</tr>
<tr>
<td>D10</td>
<td>32.9</td>
<td>0.82</td>
<td>-2.0</td>
<td>0.46</td>
</tr>
<tr>
<td>D4</td>
<td>44.5</td>
<td>0.78</td>
<td>-1.6</td>
<td>0.68</td>
</tr>
</tbody>
</table>

- Misordered step thresholds for two items:
  - (inclusion of important characteristics of target setting; and testing for effectiveness across subgroups of participants)
- All item-measure correlations >.30
- One item slightly misfit the model (mnsq inft = 1.49); measured testing varied methods of treatment implementation
- No floor or ceiling effects
Results Cont’d

- Second Iteration – Revised Structure
  - Collapsed ordinal items
  - Removed locally dependent items
  - Study reliability (.66)
  - Item reliability (.98)
  - Internal consistency (.60)
  - Item-measure correlations >.35

Table 2
Item Fit Statistics for Revised Study DIAD Rating Scale Structure

<table>
<thead>
<tr>
<th>Item</th>
<th>Average Measure</th>
<th>Infit MaSq</th>
<th>Infit Z std</th>
<th>Item/Measure Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1 Variation in Implementation</td>
<td>67.9</td>
<td>1.40</td>
<td>1.6</td>
<td>.35</td>
</tr>
<tr>
<td>D3 Variation in Target Setting</td>
<td>70.3</td>
<td>.82</td>
<td>.5</td>
<td>.40</td>
</tr>
<tr>
<td>D7 Tested Participant Subgroups</td>
<td>56.9</td>
<td>1.16</td>
<td>1.2</td>
<td>.48</td>
</tr>
<tr>
<td>D5 Appropriate Time of Measurement</td>
<td>24.0</td>
<td>.88</td>
<td>.9</td>
<td>.62</td>
</tr>
<tr>
<td>D2 Variation in Participants</td>
<td>50.2</td>
<td>1.01</td>
<td>.2</td>
<td>.40</td>
</tr>
<tr>
<td>D8 Tested Variation in Setting</td>
<td>73.4</td>
<td>.81</td>
<td>-.9</td>
<td>.40</td>
</tr>
<tr>
<td>D9 Tested Important Outcomes</td>
<td>52.1</td>
<td>.64</td>
<td>-2.0</td>
<td>.72</td>
</tr>
<tr>
<td>D10 Tested Time of Measurement for Effects</td>
<td>22.8</td>
<td>.80</td>
<td>-.1</td>
<td>.58</td>
</tr>
<tr>
<td>D4 Included Important Outcomes</td>
<td>32.5</td>
<td>.85</td>
<td>-1.3</td>
<td>.62</td>
</tr>
</tbody>
</table>
Results Cont’d

- All items fit the measurement model
- No misordered step thresholds
- No evidence of prominent local item dependence

Limitations

- Have not yet rated all studies from the three reviews used in this study
- Rated reported methods and not necessarily implemented methods
Discussion

- The measurement properties of the external validity subscale of the DIAD were mixed
- A simplified rating scale may help minimize subjectivity and increase reliability
  - More ordinal items in EV subscale than any other
- Study separation reliability and internal consistency remained low

Conclusion

- Insufficient reliability to use quality assessment scores as weights in meta-analysis
- Screening studies out based on quality also unjustified
- Moderator analysis provides an appropriate venue for quality assessment
Where to from Here?

- Future applications of IRT methods to the DIAD are needed
  - Facets model to examine judgment severity
- Future applications of IRT methods to other quality assessment scales are needed
- Additional analyses on the DIAD are also needed across the full scale and disciplines

Questions?
References