

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
 - Orients 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]

- Three Campbell Collaboration Meta-Analyses provided the studies in this analysis: Mitchell, Wilson, & MacKenzie (2005), Wilson, MacKenzie, & Mitchell (2005), and Lipsey, Landenberger & Wilson (2007)
- 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)

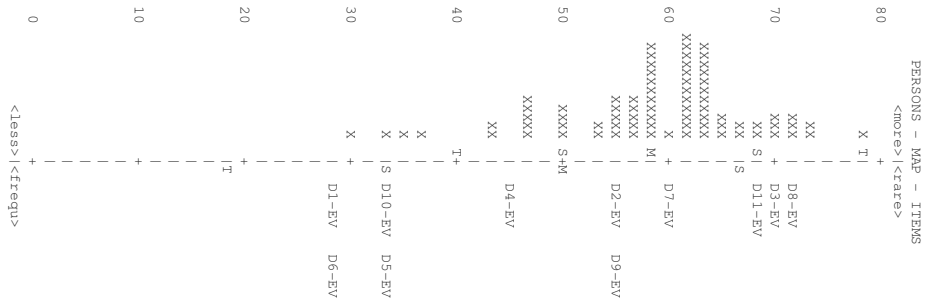
[IRT Model]

- 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)

Item Map



Results Cont'd

Table 1
Item Fit Statistics for Original Study DIAD Rating Scale Structure

Item	Average Measure	Infit MnSq	Infit Z std	Item/Measure Correlation
D11 Variation in Implementation	68.9	1.49	2.0	.31
D3 Variation in Target Setting	69.4	.79	-1.1	.47
D7 Tested Participant Subgroups	59.7	.98	-.1	.58
D5 Appropriate Time of Measurement	33.9	1.23	.6	.47
D1 Appropriate Participants	29.1	1.23	.5	.51
D6 Appropriate Time Period	29.1	1.23	.5	.51
D2 Variation in Participants	55.8	1.11	.8	.39
D8 Tested Variation in Setting	71.5	.70	-1.5	.49
D9 Tested Important Outcomes	55.3	.85	-.8	.71
D10 Tested Time of Measurement for Effects	32.9	.82	-.2	.46
D4 Included Important Outcomes	44.5	.78	-1.6	.68

[Results Cont'd]

Table 1
Item Fit Statistics for Original Study DIAD Rating Scale Structure

Item	Average Measure	Infit MnSq	Infit Z std	Item/Measure Correlation
D11 Variation in Implementation	68.9	1.49	2.0	.31
D3 Variation in Target Setting	69.4	.79	-1.1	.47
D7 Tested Participant Subgroups	59.7	.98	-.1	.58
D5 Appropriate Time of Measurement	33.9	1.23	.6	.47
D1 Appropriate Participants	29.1	1.23	.5	.51
D6 Appropriate Time Period	29.1	1.23	.5	.51
D2 Variation in Participants	55.8	1.11	.8	.39
D8 Tested Variation in Setting	71.5	.70	-1.5	.49
D9 Tested Important Outcomes	55.3	.85	-.8	.71
D10 Tested Time of Measurement for Effects	32.9	.82	-.2	.46
D4 Included Important Outcomes	44.5	.78	-1.6	.68

[Results Cont'd]

- 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 infit = 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

[Results Cont'd]

Table 2

Item Fit Statistics for Revised Study DIAD Rating Scale Structure

Item	Average Measure	Infit MnSq	Infit Z std	Item/Measure Correlation
D11 Variation in Implementation	67.9	1.40	1.6	.35
D3 Variation in Target Setting	70.3	.82	4.5	.40
D7 Tested Participant Subgroups	56.9	1.16	1.2	.48
D5 Appropriate Time of Measurement	24.0	.88	.0	.62
D2 Variation in Participants	50.2	1.01	.2	.40
D8 Tested Variation in Setting	73.4	.81	-.9	.40
D9 Tested Important Outcomes	52.1	.64	-2.0	.72
D10 Tested Time of Measurement for Effects	22.8	.80	-.1	.58
D4 Included Important Outcomes	32.5	.85	-1.3	.62

[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

- Farrington, D. P., Gottfredson, D. C., Sherman, L. W., & Welsh, B. C. (2002). The Maryland Scientific Methods Scale. In Sherman, L. W., Farrington, D. P., Welsh, B. C., & MacKenzie (eds.) *Evidence-Based Crime Prevention* (pp. 13-21). London: Routledge.
- Jadad, A. R., Moore, A., Carroll, D., Jenkinson, C., Reynolds, D. J., Gavaghan, D. J., & McQuay, H. J. (1996). Assessing the quality of reports in randomized clinical trials: Is blinding necessary? *Controlled Clinical Trials*, 17, 1-12.
- Jüni, P., Witschi, A., Bloch, R., & Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. *Journal of the American Medical Association*, 282, 1054-1060.
- Lipsey, M.W., Landeberger, N. A., & Wilson, S. J. (2007). *Effects of Cognitive-Behavioral Programs for Criminal Offenders*. A Campbell Collaboration meta-analysis, available at: <http://www.aic.gov.au/campbellcj/reviews/titles.html>.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 60, 523-547.
- Mitchell, O., Wilson, D. B., & MacKenzie, D.L. (2005). *The Effectiveness of Incarceration-Based Drug Treatment on Criminal Behavior*. A Campbell Collaboration meta-analysis, available at: <http://www.aic.gov.au/campbellcj/reviews/titles.html>.
- Mosteller, F., Gilbert, J. P., & McPeck, B. (1980). Reporting standards and research strategies for controlled trials. *Controlled Clinical Trial*, 1, 37-58.
- Valentine, J. C., & Cooper, H. (2008). A systematic and transparent approach for assessing the methodological quality of intervention effectiveness research: The Study Design and Implementation Assessment Device (Study DIAD). *Psychological Methods*, 13(2), 130-149.
- Wilson, D. B., MacKenzie, D. L., & Mitchell, F. N. (2005). *Effects of Correctional Boot Camps on Offending*. A Campbell Collaboration Meta-analysis, Retrieved on July 20, 2008, from <http://www.aic.gov.au/campbellcj/reviews/titles.html>.