Systematic Reviewing and Meta-Analysis

Jeffrey C. Valentine

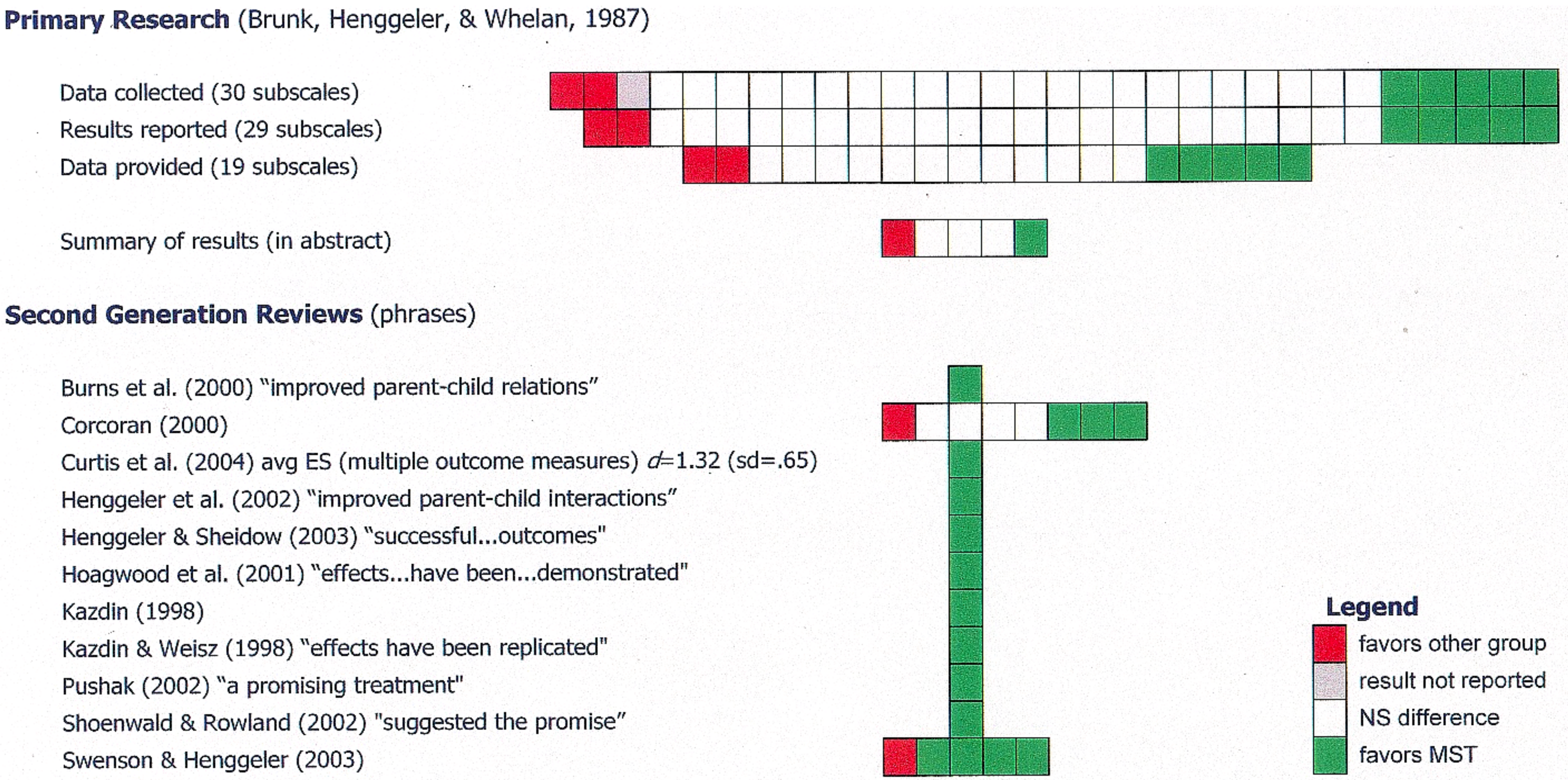
University of Louisville

February 6, 2015

Plan for Today

- Definition of and rationale for systematic reviewing / meta-analysis
  - Brief history
- Basic principles
- Conducting a simple meta-analysis
- Five questions to help evaluate the quality of a systematic review
- Future directions
- Your goals??

Why Do We Need a Systematic Approach to Reviewing Studies?



Primary Research (Brunk, Henggeler, & Whelan, 1987)

Data collected (30 subscales)

Results reported (29 subscales)

Data provided (19 subscales)

Summary of results (in abstract)

Second Generation Reviews (phrases)

Burns et al. (2000) "improved parent-child relations"

Corcoran (2000)

Curtis et al. (2004) avg ES (multiple outcome measures) *d*=1.32 (sd=.65)

Henggeler et al. (2002) "improved parent-child interactions"

Henggeler & Sheidow (2003) "successful...outcomes"

Hoagwood et al. (2001) "effects...have been...demonstrated"

Kazdin (1998)

Kazdin & Weisz (1998) "effects have been replicated"

Pushak (2002) "a promising treatment"

Shoenwald & Rowland (2002) "suggested the promise"

Swenson & Henggeler (2003)

Legend

🟥 favors other group

▨ result not reported

□ NS difference

🟩 favors MST

Definitions

- Systematic review
  - A summary of the research literature that uses explicit, reproducible methods to identify relevant studies, and then uses objective techniques to analyze those studies.
  - The goal of a systematic review is to limit bias in the identification, evaluation, and synthesis of the body of relevant studies that address a specific research question.
- Meta-analysis
  - Statistical analysis of the results of multiple studies
  - Often conflated with systematic review

Not all meta-analyses are based in a systematic review, and not all systematic reviews result in a meta-analysis

What are Some of the Options for Conducting a Review?

- Narrative review: impressionistic determination of what a literature “says” (“I carefully read and evaluated 15 studies and it seems clear that…”)
  - Historically, the only method that was used
  - Still relatively common today

- Vote count: Comparing the statistical significance of the results across studies
  - Sometimes used in conjunction with a narrative review process

- Lists of “effective” programs: Require that at least two “good” studies with statistically significant results exist
  - Lists of “evidence based” practices

- Systematic review and meta-analysis

Why Did We Move Toward Systematic Reviewing and Away From Narrative Reviewing?

- Recognition that reviews should be held to the same standards of transparency and rigor as primary studies
- Methodological improvements needed to help address
  - Reporting biases (publication bias, outcome reporting bias)
  - Need for transparency and replicability
  - Study quality assessments
- Statistical improvements needed to help address
  - Reporting biases
  - Efficient processing of lots of information
  - Moderator effects
    - Study quality
    - Study context variables (sample, setting, etc.)

Publication Bias

- Tendency for studies lacking statistical significance on their primary findings to go unreported
- Difficult problem because it is hard to…

  …detect

  …estimate how much of an impact it is having if believed to be present

Outcome Reporting Bias

- Emerging evidence base on ORB
  - Medicine (Chan & Altman, 2005; Vedula et al., 2009)
  - Education
    - Pigott et al.
      - Examined dissertations in education that were later published
      - Statistically significant dissertation findings were about 30% more likely to appear in subsequent journal publication than nonsignificant (OR = 2.4)

Scientists Have Been Thinking About How to Integrate Studies for a Long Time

- James Lind, English naval surgeon (18th Century):
  - "...it became requisite to exhibit a full and impartial view of what had hitherto been published on the scurvy ... by which the sources of these mistakes may be detected."
- 1904: K. Pearson. Report on certain enteric fever inoculation statistics. *British Medical Journal, 3*, 1243-1246.
- 1932: R. A. Fisher. *Statistical Methods for Research Workers*. London: Oliver & Boyd.
  - "…it sometimes happens that although few or [no statistical tests] can be claimed individually as significant, *yet the aggregate* gives an impression that the probabilities are lower than would have been obtained by chance." (p.99, emphasis added)
- 1932: R. T. Birge. The calculation of errors by the method of least squares. *Physical Review, 40*, 207-227.

Resurgence in the 1970’s

- Explosion of research since the 1960’s
  - About 100 randomized experiments in medicine per year in the 1960’s
  - About 20,000 randomized experiments in medicine per year today

- 1978: R. Rosenthal & D. Rubin. Interpersonal expectancy effects: The first **345** studies. *Behavioral and Brain Sciences*, *3*, 377-415.

- 1979: G. V. Glass & M. L. Smith. Meta-analysis of research on class size and achievement. *Educational Evaluation and Policy Analysis*, *1*, 2-16.
  - Over **700** estimates

- 1979: J. Hunter, F. Schmidt & R. Hunter. Differential validity of employment tests by race: A comprehensive review and analysis. *Psychological Bulletin*, *86*, 721-735.
  - Over **800** estimates

Resurgence in the 1970’s (cont’d)

- Given the potential for reporting biases
  - e.g., publication bias
- And the likely presence of moderator effects
  - e.g., some people are more likely to elicit interpersonal expectancy effects than others
- And the likely variation in study quality across studies
- And the sheer number of studies

- A narrative reviewer simply has no hope of arriving at a systematic, replicable, and valid estimate of the main effect and the conditions under which it varies

Steps in the Systematic Review Process

SRs are a form of research

Follow basic steps in the research process:

- Problem formulation
- Sampling
  - Like a survey, except surveying studies not people
    - Studies are the sampling unit
    - Sampling frame = all relevant studies
    - Sample = studies available for analysis
- Data collection
  - Data derived (extracted) from studies (usually two trained coders working independently)
    - Study quality assessments
- Analysis
  - Qualitative (descriptive, study quality assessment)
  - Quantitative (effect sizes, meta-analysis)
- Reporting

Basic Principles of Systematic Reviewing and Meta-Analysis

- Conduct a thorough literature search looking for all relevant studies
- Systematically code (survey) studies for information
  - Study context
  - Participants
  - Quality indicators
- Combine effects using a justifiable weighting scheme (like inverse variance weights)

Study Quality Assessments

- Study quality will likely vary (sometimes quite a bit) from study to study
- Study quality often covaries with effect size
- Therefore, meta-analytic effects may vary as a function of study quality
- Therefore, every good systematic review needs to deal with study quality thoughtfully

Methods People use to Address Study Quality (All Bad!)

- Ignore it
- Treat publication status as a proxy for study quality
  - If published then good else bad
- Use a study quality scale

Quality Scales Do Not Work as Intended

- Jüni et al. (1999)
  - Found an existing meta-analysis on the effects of a new drug on post-operative DVT
  - Found 25 quality scales
    - Most (24 of 25) published in peer-reviewed medical journals
  - Conducted 25 separate meta-analyses
    - Each one used a different quality scale
    - Were interested in what the "high quality" studies said about the drug relative to the "low quality" studies
    - What did they find?

Jüni et al. (1999) Results

- In about 50% of the meta-analyses, the high and low quality studies agreed about the effectiveness of the drug
  - In about 25% of the meta-analyses
    - *high quality* studies suggested the *new drug was more beneficial* than the old
    - low quality studies said it was no better then the old drug
  - In about 25% of the meta-analyses
    - high quality studies suggested the new drug was no better than the old drug
    - *low quality* studies suggested the *new drug was more beneficial* than the old
- The conclusion about the effectiveness of the new drug depended (in part) on the quality scale chosen

Implications of Jüni et al.

- Study quality doesn’t matter in this area

And/or

- The quality scales were so bad that they masked the effects of study quality

Weaknesses of Study Quality Scales

- Different numbers of items
  - suggests different criteria in use (some are more comprehensive than others)
- Different weightings of same criteria
  - Even with the same number of criteria, the same aspect of study quality can have a different weight across scales
- Possibility that biases may act in opposite directions is ignored
- Difference between study quality and reporting quality is ignored

- Reliance on single scores to represent quality
  - Study A: high internal validity, low external validity = 80
  - Study B: low internal validity, high external validity = 80

Why Are Study Quality Assessments So Hard to Develop?

- Study quality is almost certainly context dependent
  - A threat to the validity of one study may not be a threat to the validity of another study
    - Valentine & McHugh (2007) on attrition in randomized experiments in education
    - Means that it is hard to build an evidence base
- Study quality is multidimensional
- Study quality indicators may not “add up” in expected ways

Study Quality Assessment: Best Practices

- Know your research context
  - What are the likely markers of quality?
- Markers that likely matter a lot
  - Consider excluding studies that do not have the trait
    - e.g., in a drug treatment study I would be very worried about the effects of attrition
- Markers that might matter less
  - Code studies for these, and explore how the are associated with variations in effect size

Study Quality Assessment Best Practices: Example

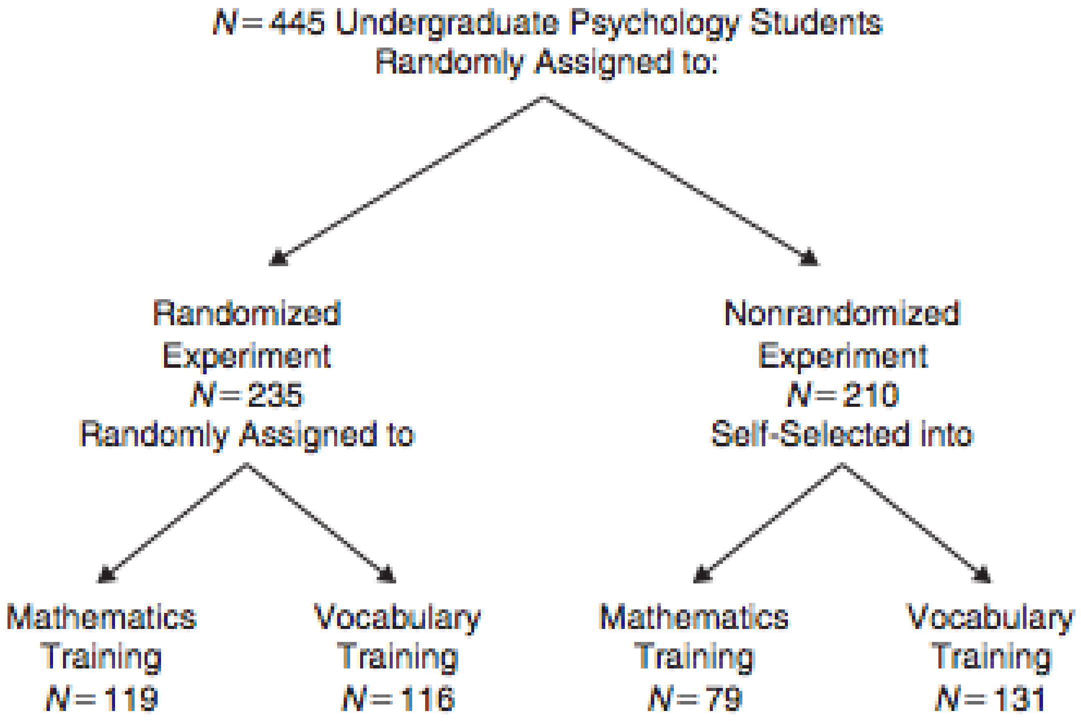
- Internal validity is a prime concern in nonrandomized experiments
- Controlling for important and well-measured covariates can help, but what’s an important covariate?
- Answer depends on the context of the research question
  - What was the selection mechanism?
    - Specific answer will vary from question to question

Side Note: Comparability in Nonrandomized Experiments

- A randomized experiment involves forming study groups randomly (e.g., a coin flip)
  - Main benefit is that we can assume that groups are equivalent on all measureable and unmeasurable variables
  - This facilitates interpretation of study results
- A nonrandomized experiment is a study in which participants were not randomly assigned to conditions
  - e.g., someone else chose the condition for them; participants selected their conditions, etc.
- Nonrandomized experiments are problematic because we don’t know if participants in different groups are comparable

Conditions Under Which Nonrandomized Studies May Approximate RCTs

- Scholars have used “within study comparisons” to investigate this
- Within study comparison = randomly assign participants to an RCT or nonrandomized experiment
- If RCT, randomly assign to treatment or control
- If nonrandomized, use another method to form groups (e.g., let participants choose)
  - Important to know a lot about the selection mechanism


N = 445 Undergraduate Psychology Students
Randomly Assigned to:
Randomized
Experiment
N = 235
Randomly Assigned to
Nonrandomized
Experiment
N = 210
Self-Selected into
Mathematics
Training
N = 119
Vocabulary
Training
N = 116
Mathematics
Training
N = 79
Vocabulary
Training
N = 131

Selection Mechanism (Example)

- Adapted from Shadish et al. (2008)
- Q: What determines whether an individual will choose to participate in a math vs. vocabulary intervention?
  - Math ability
  - Verbal ability
  - Math anxiety
  - Etc.
    - These were derived from a thorough review of both theoretical and empirical literatures
- Quality judgments are best made by individuals with a deep understanding of the research context, and developing this understanding might require research!

What Do Within Study Comparisons Tell Us?

- Develop a rich understanding of the selection process
  - e.g., why participants chose to participate in one condition vs. another
- Good pretest measures help a lot
- Measure the relevant variables well
- Do not rely on easily measured data that just happens to be collected (e.g., basic demographics) as these are unlikely to help

Quality Assessments (the better way)

- Adopt relatively few, highly defensible exclusion criteria that are related to study quality
  - Hopefully, you will have empirical evidence suggesting that your criteria reduce the bias of the results
    - Keep in mind that a bias in an individual study does not necessarily mean that the *set* of studies will be biased
- Treat other potential sources of bias empirically
  - e.g., develop a good model of the selection process and compare effects from studies that model the process well vs. those that do not
- Avoid confusing study quality and reporting quality
  - Contact authors for information
    - Missing or ambiguous

Meta-Analysis

- Meta-analysis is the statistical combination of the results of multiple studies
- As most commonly practices, meta-analysis involves weighting studies by their precision (more precise studies = relatively more weight)
- In other words, we weight by a function of sample size
  - Specifically, we (almost always) use inverse variance weights

Meta-Analysis (cont'd)

- Often, study results have to be transformed for meta-analysis
- Most usually because studies do not present results on the same scale
  - e.g., some studies use the ACT, others the SAT
- Common ways of expressing effects across studies are:
  - Standardized mean difference *d* (Cohen's *d*)
  - Correlation coefficient
  - Odds ratio

Common Standardized Effect Sizes

- The correlation coefficient is defined as usual
- The standardized mean difference (*d*) is defined as:

$$d = \frac{\overline{X}_T - \overline{X}_C}{s_p}$$

Subtract the mean of the treatment group from the mean of the control group, and divide by the pooled ("average") standard deviation

- The odds ratio is defined as: $OR = \frac{a/b}{c/d}$

Where a, b, c, and d refer to cells in a 2x2 table:

	Graduated	Did Not Graduate
Treatment	a	b
Control	c	d

Meta-Analysis Using *r* and OR

- Correlation coefficients and odds ratios have undesirable distributional properties
- Therefore they are transformed for meta-analysis
  - *r* is transformed using Fisher's *z* transformation
    - *z* has a mean of 0 and a very convenient variance (1/n-3) – since we use inverse variance weights, the weight for a correlation coefficient is just n-3!
  - Odds ratios are log transformed for analysis
    - The weight for a logged odds ratio is the square root of

(1/a + 1/b + 1/c + 1/d)

Choosing a Meta-Analytic Model

- Need to decide whether to use a fixed effect or random effects model
  - Fixed effect (aka common effect)
    - Assumes all studies are estimating the same population parameter
      - i.e., if all studies were infinitely large they would all yield the same effect size
  - Random effects
    - Studies arise from a distribution of effect sizes
      - And are therefore expected to vary. This variability is taken into account in point and interval estimation

Fixed vs. Random Effects

- Functionally, use FE when
  - You believe studies are very close replications of one another and/or
  - You only want to generalize to studies highly like the ones you have and/or
  - You have a small number of studies (e.g., < 5)
- Use RE when
  - You believe studies are not very close replications
  - And/or you want a broader universe of generalizability
- FE models almost always have greater statistical power that RE models
  - RE CIs will never be smaller than FE, and are almost always larger

Conducting a Basic Meta-Analysis

- Assume we have three studies on the effects of a summer bridge program for at risk students
  - Randomly assigned to bridge program or usual orientation
- DV is score on a math placement test
- The data are:

	Treatment			Control			
	Mean	SD	n	Mean	SD	n	*r*
Rodgers (2013)	6.061	6	55	3.0	6	48	+.250
Dalglish et al. (2012)	3.113	2	30	1.0	2	33	+.462
Benitez (2004)	36.35	10	7	11.0	10	6	+.762

Inverse Variance Weighting

- Recall that for meta-analysis, we transform correlation coefficients to Fisher's z, and the weight for Fisher's z is n-3

Study	*r*	Fisher z (=fisher(r))	n	Weight (n-3)
Rodgers (2013)	+.25	+.25	103	100
Dalglish et al. (2012)	+.462	+.50	63	60
Benitez (2004)	+.762	+1.00	13	10

Meta-Analysis

- Because we only have three studies, fixed effect meta-analysis is probably best
- The meta-analytic average is a weighted average, and is computed in the usual way. Expressed generically:

$$\overline{ES} = \frac{\Sigma w_j (ES_j - \overline{ES})^2}{\Sigma w_j}$$

In words, subtract each effect size from the mean ES, square that value, and multiply the result by the ES's weight. Do this for all ES's, and sum the weighted squares. Divide this quantity by the sum of the weights.

Inverse Variance Weighting

Study	*r*	Fisher’s z (=fisher(r))	n	Weight (n-3)	Weight x Fisher z
Rodgers (2013)	+.25	+.25	103	100	
Dalglish et al. (2012)	+.462	+.50	63	60	
Benitez (2004)	+.762	+1.00	13	10	
Sums					

Inverse Variance Weighting

Study	*r*	Fisher’s z (=fisher(r))	n	Weight (n-3)	Weight x Fisher z
Rodgers (2013)	+.25	+.25	103	100	25
Dalglish et al. (2012)	+.462	+.50	63	60	
Benitez (2004)	+.762	+1.00	13	10	
Sums					

Inverse Variance Weighting

Study	*r*	Fisher's z (=fisher(r))	n	Weight (n-3)	Weight x Fisher z
Rodgers (2013)	+.25	+.25	103	100	25
Dalglish et al. (2012)	+.462	+.50	63	60	30
Benitez (2004)	+.762	+1.00	13	10	
Sums					

Inverse Variance Weighting

Study	*r*	Fisher's z (=fisher(r))	n	Weight (n-3)	Weight x Fisher z
Rodgers (2013)	+.25	+.25	103	100	25
Dalglish et al. (2012)	+.462	+.50	63	60	30
Benitez (2004)	+.762	+1.00	13	10	10
Sums					

Inverse Variance Weighting

Study	*r*	Fisher’s z (=fisher(r))	n	Weight (n-3)	Weight x Fisher z
Rodgers (2013)	+.25	+.25	103	100	25
Dalglish et al. (2012)	+.462	+.50	63	60	30
Benitez (2004)	+.762	+1.00	13	10	10
Sums				170	

Inverse Variance Weighting

Study	*r*	Fisher's z (=fisher(r))	n	Weight (n-3)	Weight x Fisher z
Rodgers (2013)	+.25	+.25	103	100	25
Dalglish et al. (2012)	+.462	+.50	63	60	30
Benitez (2004)	+.762	+1.00	13	10	10
Sums				170	65

- Now we know: the sum of the weights x ES (numerator) and the sum of the weights (denominator)

Carrying out the Meta-Analysis

- Recall that the generic formula for the weighted mean effect size is:

$$\overline{ES} = \frac{\Sigma w_j (ES_j - \overline{ES})^2}{\Sigma w_j}$$

- So here, $\overline{z}_r = \frac{65}{170} = .382$

- Because this is a Fisher's *z*, we back transform (using exponentiation, or =fisherinv(z)) in a spreadsheet) to a correlation coefficient. Here, *r* = +.365
- The point is: This is easy!

Basic Principles of Systematic Reviewing and Meta-Analysis

- Conduct a thorough literature search looking for all relevant studies
- Systematically code (survey) studies for information
  - Study context
  - Participants
  - Quality indicators
- Combine effects using a justifiable weighting scheme (like inverse variance weights)

“Is this a good review?”: Five Questions

- Does the background do a good job of setting out the research problem?
- Did the researchers look under every rock for potentially eligible studies?
- Did the researchers take study quality into account in a convincing way?
- Was a reasonable model chosen for the meta-analysis?
- Do the conclusions follow from the analyses?

Also Very Nice: Lots of Study Detail

- Also a really good idea for reviewers to create a table that lays out important study characteristics and effect sizes

- For example, see Stice et al. (2009), who provided two different types of tables. I like these because the provide **a lot** of information and make the review easier to replicate

Table 2
Descriptions of the Sample, Intervention Content, and Findings From Depression Prevention Trials

Study	Sample	Intervention	Findings
Barrett et al., 2006	669 girls and boys	Efficacy trial of a universal school-based CBT intervention designed to prevent child anxiety by teaching children coping and problem-solving skills.	No significant effects for depressive symptoms (CDI) compared with an assessment-only control group.
Beardslee et al., 2003	121 girls and boys	Efficacy trial of selective psychoeducational intervention targeting children of depressed parents that presented information on mood disorders, risk, and resilience, and how to facilitate relationships.	No significant effects for depressive symptoms (SADS–L) at 1-, 2-, and 4.5-year follow-ups compared with an attention control group.
Bearman et al., 2003	74 girls	Efficacy trial of selective CBT intervention targeting adolescent girls with elevated body dissatisfaction.	Significant effects for depressive symptoms (BDI) at posttest but not 6-month follow-up compared with a waitlist control group.
Burton et al., 2007	145 young women	Efficacy trial of selective CBT intervention targeting women with elevated depressive symptoms.	Significant effects for depressive symptoms (BDI) at posttest, 3-month, and 6-month follow-ups compared with control group.
Cardemil et al., 2007	168 girls and boys	Two-year follow-up of efficacy trial of a universal school-based CBT intervention that taught cognitive and social problem-solving skills.	Significant effects for depressive symptoms (CDI) compared with an assessment-only control group.

Table 4

Moderator Values and Effect Sizes for Depression Prevention Trials

	Participant				Content											Effect size	
Study	Risk status	Gender (% female)	Ethnicity (% White)	Age	Cognitive change	Behavioral activation	Problem solving	Social skills	Intervention duration (hr)	Home-work	Professional interventionist	Interview assessment	Published	Incorrect unit of analysis	Random	Post-test	Follow-up
Barrett et al., 2006	0	54	—	12.0	1	0	1	0	11.7	0	0	0	1	1	1	.05	.06
Beardslee et al., 2006	1	43	94	11.6	1	0	0	1	8.5	1	1	1	1	1	1	—	.00
Bearman et al., 2003	1	100	47	18.9	1	1	0	0	4	1	1	0	1	0	1	.29	.07
Burton et al., 2007	1	100	52	18.6	1	1	0	0	4	1	1	0	1	0	1	.58	.24
Cardemil et al., 2007	0	50	0	11.3	1	0	1	0	18	1	1	0	1	0	1	.27	.19
Chaplin et al., 2006																	
Girls only vs. control	0	100	89	12.2	1	0	1	0	18	1	1	0	1	0	1	−.47	—
Co-ed vs. control	0	44	89	12.2	1	0	1	0	18	1	1	0	1	0	1	−.29	—
Clarke et al., 1993																	
Study 1	0	42	—	15.4	0	0	0	0	2.5	0	0	0	1	0	1	.02	−.03
Study 2	0	46	—	15.1	0	1	0	0	4.2	0	0	0	1	0	1	.02	.05
Clarke et al., 1995	1	70	93	15.3	1	0	0	0	11.2	0	0	1	1	0	1	.18	−.02
Clarke et al., 2001	1	28	90	14.6	1	0	0	0	15	1	1	1	1	0	1	.22	.16
Forsyth, 2000	1	97	93	19.4	1	0	1	1		1	1	0	0	0	1	.68	.76
Garber et al., 2008	1	59	75	14.6	1	0	0	0	15	1	1	1	0	0	1	.14	—
Gillham, 1994, Study 2																	
Child & parent vs. control	0	47	—	14.6	1	0	1	1	16	1	1	0	0	0	1	0.03	−.07
Child only vs. control	0	47	—		1	0	1	1	24	1	1	0	0	0	1	0.25	.21

Where Is This Enterprise Heading?

- Multivariate meta-analysis
  - Lots of recent developments in meta-analysis and structural equation modeling and factor analysis
- All Bayes all the time
  - Bayesian meta-analysis becoming more popular
  - Solves some problems
    - Meta-analysis when number of studies is small
    - User interpretation of output (*p*-values)
    - More helpful output (probability that effect is > 0)
  - Introduces others
    - Defensible priors

Thank You!! ☺

- Your questions?

Journal of Educational and Behavioral Statistics
Month 2009, Vol. XXX, No. XX, pp. 1–33
DOI: 10.3102/1076998609346961
© 2009 AERA. http://jebs.aera.net

How Many Studies Do You Need? A Primer on Statistical Power for Meta-Analysis

Jeffrey C. Valentine
University of Louisville

Therese D. Pigott
Loyola University-Chicago

Hannah R. Rothstein
Baruch College

In this article, the authors outline methods for using fixed and random effects power analysis in the context of meta-analysis. Like statistical power analysis for primary studies, power analysis for meta-analysis can be done either prospectively or retrospectively and requires assumptions about parameters that are unknown. The authors provide some suggestions for thinking about these parameters, in particular for the random effects variance component. The authors also show how the typically uninformative retrospective power analysis can be made more informative. The authors then discuss the value of confidence intervals, show how they could be used in addition to or instead of retrospective power analysis, and also demonstrate that confidence intervals can convey information more effectively in some situations than power analyses alone. Finally, the authors take up the question ''How many studies do you need to do a meta-analysis?'' and show that, given the need for a conclusion, the answer is ''two studies,'' because all other synthesis techniques are less transparent and/or are less likely to be valid. For systematic reviewers who choose not to conduct a quantitative synthesis, the authors provide suggestions for both highlighting the current limitations in the research base and for displaying the characteristics and results of studies that were found to meet inclusion criteria.

Keywords: *meta-analysis; research methodology; statistics*

Scholars undertaking a literature review can have a variety of goals. Some reviews, for example, are meant to provide insight into important theories or themes in a literature. Other reviews have more specific aims. For example,

Please address correspondence to Jeff Valentine, 309 CEHD, Louisville, KY 40290; e-mail: jeff.valentine@louisville.edu