Meta-Analysis

Goals

- Same as traditional narrative review but more quantitative
- Do Narrative review first
- Integrative Review
  - Uses statistical analyses to combine results of previous studies
  - Less likely to allow researcher bias to enter into conclusions
  - Can compute mean effect sizes for IV
  - Can compute significance of mean effect size, and of difference between mean effect sizes in different conditions of a moderator
  - For testing mediational hypotheses – (Shadish, 1996)

Cooper & Rosenthal (1980)

- Professors and Graduate students reviewed 7 studies: Sex and persistence at tasks
- A) traditional narrative review
  - Perceived larger difference between males and females, who were more persistent
- B) Statistical review

Brief History

- 1904 – 1st application
  - Pearson – 11 studies of vaccine against typhoid
    - Averaged measures of treatment’s effect across two groups of studies
    - On basis of average correlations, concluded that all other vaccines were more effective
- 1932 – Fisher
  - Statistical Methods for Research Workers
    - Test for combining p values from independent tests of same hypothesis
- Techniques not widely implemented until 60s
- 1976 – phrase coined by Gene Glass
Cooper (1982) five-stage model

- Threats to inferential validity
  - Later users of data must be as accountable for the validity of their methods as the original data gatherers
- Check Validity
  - Internal
  - Theoretical
    - Are conditions met?
  - Ecological

Mullen et al. (1991) Validity Check

1. Exclude studies highly flawed in internal or construct validity
   - E.g., use of measure later deemed invalid
   - Construct design flaw analysis
     - Matrix where rows = studies and columns = validity threats
2. Establish explicit set of criteria for judging validity
   - E.g., random assignment?
3. Classify studies as to their degree of validity and factor into analysis

Procedures

- Literature Search
  - Published AND unpublished sources
    - Why?
    - Must include estimates of effect size
    - Problems?
  - 10–15 studies minimum
    - 10–15 studies per condition of moderator
- Level of analysis
  - “Mixing apples and oranges” – e.g., combine effect sizes across different types of therapy
  - Mixing across DVs even more problematic
Operationally Defining Study Outcomes

- 1. support/not support hypothesis
  - Vote-counting
- 2. multiple outcome categories
  - 1. sig. and supported H1
  - 2. not sig. but supported H1
  - 3. IV had no effect
  - 4. not sig and contradicted H1
  - 5. sig and contradicted H1
- 3. effect size
  - $d$ and/or $r$

Example

- Remedial education and self-esteem
  - $H_0 = $ adults receiving and not receiving education do not differ in SE
  - Extract from Methods and Results, information on each of the relevant study characteristics
    - E.g., age, measures, sex etc.
    - Reliability from a sample of those studies

Procedures

- Vote Counting
  - Divide reports into piles:
    - Statistically significant, no differences, null hypothesis
    - Side with larger pile
  - Problems with this method?

Procedures

- Vote Counting
  - If null is true, 1/20 (5%) studies will suggest significance by chance alone
  - The “largest pile wins” strategy requires that 7/20 (34%) of the studies must be significant before that conclusion is accepted
  - (fewest # in a pile to be considered largest when 20/3)
  - But what if five studies showed significant relationship between self-esteem and remedial education?
  - Two studies can have same effect size (e.g., $r = .25$), but larger sample ($N = 100$) be sig. and smaller sample ($N=50$) NS
Procedures

› **Vote Counting**
  - Susceptible to Type II errors
  - Strategy does not weight reports differently based on sample size!
  - Effect sizes from larger samples should be given more weight
  - Also does not weight large and small mean differences differently

› **Combining Probabilities**
  - Extract $p$ associated with each test of the null hypothesis
  - Generate a single probability that relates to the likelihood of obtaining a run of studies with these results given that null is true
  - E.g., what is the combined probability of finding that education has no effect on self-esteem with 20 studies?

Procedures

› **Combining Probabilities**
  - E.g., Remedial education and self-esteem
  - What should researcher conclude if:
    - combined probability was $p < .03$?
    - Combined probability was $p < .19$?

  - Overcomes improper weighting problems BUT is very powerful
    - Very high likelihood of rejecting null if treatments have generated a large N of studies
    - Also, tells you effect exists but not its size

 › **Effect size estimation**
   - Reframe – *how much does* remedial education influence self-esteem?
   - Positive values indicate that effect size is consistent with hypothesis
   - Negative values indicate opposite hypothesis
Effect size estimation
If examining relationship between two continuous variables (e.g., GPA and self-esteem) – use Pearson’s product moment correlation.

Effect size estimation
If comparing treatment to control group
Cohen’s d – standardized mean difference
Scale-free measure of the number of SDs between two group means

\[ D = \frac{x_1 - x_2}{SD_1 - SD_2} \]

Effect size estimation
To determine how big of a difference exists between education and control conditions for all studies in the sample on average:
- Calculate d for each outcome in each study
- Weight them by sample size
- Average all d indexes
- This average effect size ignores characteristics of the studies

- Comprehensive Meta-Analysis
Influences on Effect Sizes

- Calculate average $d$ indexes for subsets of studies with common characteristics
- Homogeneity analysis
  - Test whether these factors are reliably associated with different magnitudes of effect (different average $d$ indexes)
  - Group studies according to potentially important characteristics and test for between-group differences
  - If significant, differences in effect size are not due to sampling error alone
  - *Results do NOT allow causal statements*

Sensitivity Analysis

- What happens if some aspect of the data or the analysis is changed?
- Funnel plot
  - Depicts sample size of studies versus estimated effect size for the group of studies
  - Should approximate shape of normal distribution
  - But publication bias will restrict range of distribution – overrepresentation at one tail

- Trim and fill method (Duval & Tweedie, 2000)
  - Through iterative process ‘fills-in’ effect sizes from studies that were not represented in data set
  - Nonparametric method that estimates missing effect sizes based on normal distribution

- Could also prepare stem-and-leaf and box plots to examine distribution of standardized mean differences
- Remove any outlying effect size and compare result to total effect with all studies included.
Problems

- Missing information
- Coding ambiguities
- Correlated data points
- Problems with original data collection
- Timeliness
- Be mindful that moderators are correlational

Useful Site: Meta Analysis Calculator
http://www.lyonsmorris.com/ma1/index.cfm

To Ponder

- A. What were the conceptual variables of interest?
- B. What inclusion criteria were used in selecting research for the meta-analysis?
- C. How many different measures of each of the conceptual variables were found in the literature review?
- D. What method was used to determine the average effect size?
- E. Was the statistical significance of the effect size estimate calculated? If so, how?
- F. Was the file drawer problem addressed?
- G. What problems did the authors encounter in conducting the meta-analysis? How did the authors attempt to solve these problems?
- H. What was the authors' conclusion about the relation between the variables of interest?