Chapter 8

Replication and Meta-Analysis

All truth passes through three stages. First, it is ridiculed. Second, it is violently opposed. Third, it is accepted as being self-evident.

—Arthur Schopenhauer

Overview

- Principles of replication
- Traditional methods for research synthesis
- Meta-analysis as a method for research synthesis
- Steps of meta-analysis
- Meta-analytic models and tests
- Threats to the validity of a meta-analysis

Principles of replication

- Replication is a critical scientific activity, but it is underappreciated in the behavioral sciences

- The use of statistical tests as if they directly estimated the likelihood of replication and editorial bias against studies without H_0 rejections may contribute to this problem

Principles of replication

- Behavioral data may not have less inherent empirical cumulativeness than data in the natural sciences (e.g., Hedges, 1987)
- *Empirical cumulativeness* is the observed degree of agreement of results across different studies
- Replication also requires *theoretical cumulativeness*, the degree to which empirical and theoretical structures build on one another in a way that permits results of current research to extend earlier work

Principles of replication

- B. Thompson (1997) distinguishes between internal and external replication
- *Internal replication* is conducted by the original researcher(s)
- *External replication* is conducted by others and involves new samples collected at different times or places

Principles of replication

- Internal replication includes statistical resampling and cross-validation
- *Statistical resampling* is a computer-based method that combines the cases in the original data set in different ways to estimate the effect on the results of idiosyncrasies in the sample (e.g., bootstrapping)
- The total sample in *cross-validation* is divided at random into a *derivation sample* and a *cross-validation sample*, and the same analyses are conducted in each one

Principles of replication

- Types of external replication of experimental studies (Carlsmith, Ellsworth, & Aronson, 1976; Lykken, 1968):

  - *Exact* or *literal replication*:

    All major aspects of the original study are copied as closely as possible, but true exact replication exists more in theory than in practice

  - *Operational replication*:

    Just the sampling and experimental methods of the original study are duplicated—deals with whether the effect is robust enough to be found by another researcher who uses the same methods

Principles of replication

- Types of external replication of experimental studies:
  - *Balanced replication*:

    Operational replications are used as control conditions for new conditions that may represent the manipulation of additional substantive variables to test new hypotheses

  - *Construct replication*:

    Avoids close imitation of the specific methods of the original study—the new researcher specifies the design, measures, and data analysis methods deemed appropriate to test the generality of the original finding

Principles of replication

- Most replication studies in the behavioral sciences are probably construct replications
- A problem is that variation in measures, samples, or methods across construct replications may be associated with actual changes in the phenomenon of interest
- Without a systematic cataloging of how construct replications differ from each other, it may be difficult to synthesize their results (e.g., understand cross-study variability in results)
- This is one of the main goals of a meta-analysis

Traditional methods for research synthesis

- Integrative summaries of the status of a research area are among the most cited articles in the research literature (Cooper & Hedges, 1994)
- The traditional method for research synthesis is the *qualitative* (*narrative*) *literature review*, which is a rational exercise
- Traditional literature reviews in the behavioral sciences were often combined with *voting counting (box-score method)*, in which numbers of statistically significant results were counted across studies

Traditional methods for research synthesis

- Problems with vote counting:
  - If power is .50 (which is typical), the box score will be “tied,” which makes the review look inconclusive
  - If restricted to published studies, results may overestimate the size of the population effect
  - Subject to all the general limitations of statistical tests
  - Not generally a scientific way to synthesize research results

Meta-analysis as a method for research synthesis

- Meta-analysis is a type of quantitative literature review
- It analyzes results from primary studies (i.e., study is the unit of analysis)
- Derivation of effect size indexes from statistics reported in primary studies is a crucial part of most meta-analyses
- Meta-analysis usually seeks to explain variability in observed effect sizes across studies
- It uses many of the same kinds of statistical methods used in primary studies, such as multiple regression and ANOVA, to analyze effect sizes

Meta-analysis as a method for research synthesis

- Quantitative methods for research synthesis have actually been around for decades (e.g., see Bangert-Drowns, 1986)
- Meta-analysis in its present form is usually attributed to Glass (1976) and R. Rosenthal (1976)
- The first published modern meta-analysis was the study of psychotherapy outcomes by Smith and Glass (1977)
- There are now hundreds of meta-analytic summaries in many different areas, such as medicine and environmental sciences
- See the review "Meta-Analysis at 25" by Glass (2000) at

  http://glass.ed.asu.edu/gene/papers/meta25.html

Meta-analysis as a method for research synthesis

- A meta-analysis attempts to identify and measure characteristics of primary studies that may underlie variability in effect sizes
- Some of these characteristics concern attributes of samples, settings where cases are tested, or the type of treatment administered
- Other characteristics concern properties of the outcome measures, quality of the research design, source of funding, professional background(s) of the author(s), or date of issuance (see Table 8.1)

Meta-analysis as a method for research synthesis

- Lipsey's (1994) typology of study characteristics:

  - *Substantive factors*:

    Presumed relevant for understanding an effect—includes features of treatments, samples, or settings

  - *Method factors*:

    Procedural aspects about how studies are conducted—may be seen as artifacts that confound cross-study comparisons of effect sizes

  - *Extrinsic factors*:

    Include things such as author attributes, the form or outlet of publication, and funding source

Meta-analysis as a method for research synthesis

- The definition of a study factor as substantive, method, or extrinsic depends on the research context
- Substantive, method, or extrinsic factors are usually seen as meta-analytic predictors
- The outcome variable (criterion) is study outcome measured with the same standardized effect size index

Meta-analysis as a method for research synthesis

- Each meta-analytic predictor in actually a moderator variable, which implies interaction

- This is because the criterion usually represents the association between the independent and dependent variables in each study

- If observed variation in effect sizes is explained by a meta-analytic predictor, then the relation between the independent and dependent variables changes across the levels of that predictor

- For example, the magnitude of a treatment effect may depend on illness chronicity (e.g., Table 8.2)

Steps of meta-analysis

- The steps of meta-analysis are basically the same ones as in a primary study
- These steps are iterative because it is necessary to return to an earlier stage when problems are discovered at later stages
- These steps are:

1. *Formulate the research question*: Hypotheses and basic operational definitions should help distinguish between relevant and irrelevant studies

2. *Collect the data* (find the studies): Characterized by searches in multiple sources, including traditional kinds of published works (journals, books), reports from public agencies, and unpublished works—see M. Rosenthal (1994) for search tips

Steps of meta-analysis

- These steps are:

   3. *Evaluate data quality*: Avoid the garbage-in, garbage-out problem (i.e., studies with serious design or method flaws may need to be eliminated from further consideration—there are some standard systems for coding research quality (e.g., Wortman, 1994)

   4. *Measure the predictors and criterion*: Code the studies according to meta-analytic predictors and standardized effect sizes, but the results still may not be directly comparable due to use of very different outcome measures (i.e., the apples-and-oranges problem)

Steps of meta-analysis

- These steps are:

5. *Analyze the data*: Lau, Ioannidis, and Schmid (1997) outline the following four iterative phases:

   a. Decide whether and what results to combine across studies
   b. Estimate a common (average) effect size
   c. Estimate the heterogeneity in results and try to explain it
   d. Assess the potential for bias

6. *Describe, interpret, and report the results*: See R. Rosenthal (1995) for suggestions about writing and understanding meta-analytic articles

Meta-analytic models and tests

- Study effects sizes are usually weighted by a function of their sample sizes, which gives greater weight to results from larger studies

- The general form of the weighted average of k observed effect sizes is:

$$M_{\text{ES}} = \frac{\sum_{i=1}^{k} w_i \; ES_i}{\sum_{i=1}^{k} w_i}$$

ES_i is the ith observed effect size index and w_i is its weight

Meta-analytic models and tests

- A weight that minimizes the variance of M_{ES} is:

$$w_i = \frac{1}{s^2_{ES_i}}$$

$s^2_{ES_i}$ is the *conditional variance* (squared standard error) of the ith effect size, also called the *within-study variance*

Recall that standard error reflects sample size

Meta-analytic models and tests

- The conditional variance of M_{ES} is affected by the number of effect sizes and their weights:

$$s^2_{M_{\text{ES}}} = \frac{1}{\sum_{i=1}^{k} w_i}$$

The square root of $s^2_{M_{\text{ES}}}$ is the standard error of M_{ES}

Meta-analytic models and tests

- Assuming normality, the form of an approximate confidence interval for μ_{ES}, the population average weighted effect size, is:

$$M_{ES} \pm s_{M_{ES}} \left(z_{2\text{-tail},\, \alpha} \right)$$

- If the confidence interval for μ_{ES} includes zero and $z_{2\text{-tail},\, \alpha} = 1.96$, then the nil hypothesis H_0: $\mu_{ES} = 0$ cannot be rejected at the .05 level
- This is an example of a meta-analytic statistical test, one conducted with study effect sizes

Meta-analytic models and tests

- There is a meta-analytic statistical test that evaluates whether the variability in study effect sizes is great enough to reject the hypothesis that they estimate a common population effect size

- It is based on the following homogeneity test statistic:

$$Q = \sum_{i=1}^{k} \frac{(ES_i - M_{\text{ES}})^2}{s^2_{ES_i}}$$

which is distributed as a central chi-square statistic with $df = k - 1$ for a large number of effect sizes and assuming a true null hypothesis

Meta-analytic models and tests

- Rejection of the homogeneity hypothesis leaves two basic options:

  1. Continue to assume a *fixed-effects model* (i.e., there is one true effect size) but disaggregate studies by the levels of one or more meta-analytic predictors—continue until the homogeneity hypothesis is not rejected within each category (e.g., Tables 8.2 and 8.3)

  2. Specify a *random-effects* or a *mixed-effects model* instead of a fixed-effects model (i.e., there are multiple true effect sizes)

Meta-analytic models and tests

- Random- or mixed-effects models are more complicated compared with fixed-effects models and may require special statistical methods
- Recall a similar distinction in ANOVA
- The choice of a model for explained vs. unexplained variability is not entirely statistical, but should also be guided by the researcher's domain knowledge (Shadish & Haddock, 1994)

Threats to the validity of a meta-analysis

- Although it is useful to know average effect size magnitudes, effect size by itself says little about substantive significance (chap. 4)

- Explaining a relatively high proportion of the observed variability in effect sizes with a set of meta-analytic predictors does not imply that these factors are actually the ones involved in the underlying processes

- The results of a meta-analysis are subject not only to the particular studies found in the literature search but also to many decisions about things such as

  - how to code study characteristics
  - how to screen studies
  - the choice of a statistical model for effect size variability

Threats to the validity of a meta-analysis

- This is where a *sensitivity analysis* can be useful: the same data are analyzed under different assumptions or methods

- If the basic results of the meta-analysis do not change under different assumptions, the results are robust against those assumptions

- See Cooper and Hedges (1994), Eysenck (1995), Hunt (1997), and Sohn (1995) for more information about strengths and weaknesses of meta-analysis

References

Bangert-Drowns, R. L. (1986). Review of developments in meta-analytic method. *Psychological Bulletin*, *99*, 388-399.

Carlsmith, J., M., Ellsworth, P. C., & Aronson, E. (1976). *Methods of research in social psychology*. Reading, MA: Addison-Wesley.

Cooper, H., M., & Hedges, L. V. (1994). Research synthesis as a scientific enterprise. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 3-14). New York: Russell Sage Foundation.

Eysenck, H. J. (1995). Meta-analysis squared—Does it make sense? *American Psychologist*, *50*, 110-111.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *10*, 3-8.

Glass, G. V. (2000). Meta-analysis at 25. Retrieved October 3, 2001, from http://glass.ed.asu.edu/gene/papers/meta25.html

Hedges, L. V. (1987). How hard is hard science, how soft is soft science? *American Psychologist*, *42*, 443-455.

Hunt, M. (1997). *How science takes stock*. New York: Russell Sage Foundation.

Lau, J., Ioannidis, J. P. A., & Schmid, C. H. (1997). Quantitative synthesis in systematic reviews. *Annals of Internal Medicine*, *127*, 820-826.

Lipsey, M. W. (1994). Identifying potentially interesting variables and analysis opportunities. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 111-124). New York: Russell Sage Foundation.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin*, *70*, 151-159.

Rosenthal, M. C. (1994). The fugitive literature. In H. Cooper and L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 85-94). New York: Russell Sage Foundation.

Rosenthal, R. (1976). *Experimenter effects in behavioral research*. New York: Halstead Press.

Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin*, *118*, 183-192.

Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 261-281). New York: Russell Sage Foundation.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, *32*, 752-760.

Sohn, D. (1995). Meta-analysis as a means of discovery. *American Psychologist*, *50*, 108-110.

Thompson, B. (1997). Editorial policies regarding statistical significance tests: Further comments. *Educational Researcher*, *26*, 29-32.

Wortman, P. M. (1994). Judging research quality. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 97-123). New York: Russell Sage Foundation.