Systematic Review Methods: The Science and Practice of Research Synthesis

Julia H. Littell
Bryn Mawr College

Brandy R. Maynard
Saint Louis University

Society for Social Work and Research
New Orleans, LA
January 15, 2015

Agenda

I. Introductions
II. Overview: Principles, methods, and evidence for research synthesis
III. How to conduct a systematic review (an introduction)
   A. Problem formulation
   B. Information retrieval
   C. Data extraction: screening, coding, and evaluation
   D. Data analysis, synthesis, and interpretation
   E. Reporting results
IV. Summary and conclusions
V. Additional resources
I. INTRODUCTIONS

- Your name, organization
- Experience with systematic reviews or meta-analysis
- What you hope to learn in this workshop

II. OVERVIEW: PRINCIPLES, METHODS, AND EVIDENCE FOR RESEARCH SYNTHESIS
Terminology

- *Research synthesis*: combining or summarizing results of multiple studies
- *Systematic review*: attempts to minimize bias and error at each step in the review process
- *Meta-analysis*: statistical analysis and synthesis of quantitative results of multiple studies

Purposes of Research Reviews

- Summarize existing empirical research to:
  - Take stock of a body of research
  - Identify gaps in knowledge
  - Organize knowledge (master the information tsunami)
  - Provide directions for further research
  - Inform policy and practice

Thanks to Brandy Maynard for slide & graphic
Review Questions

- What do we know and how do we know it?
- Possible topics include
  - Rates and trends (e.g., incidence/prevalence, differences over time/place/subgroups)
  - Correlates and causes (e.g., risk and protective factors)
  - Prevention and treatment (e.g., outcomes, impacts, cost effectiveness, comparative effectiveness)
  - Diagnosis (e.g., accuracy of various dx categories/tests)
  - Prognosis (e.g., predictively validity of categories/tests)
  - Methods and measures (e.g., reliability, validity)

Questions and Methods

- Different review questions call for
  - Different types of evidence
  - Different synthesis methods
- “Evidence hierarchies” do not work across questions
Objectives

- Reviews have various objectives
  - Descriptive
    - Document what’s been done
    - No synthesis
    - Scoping reviews & systematic maps
  - Synthesis
    - “Summing up”
    - Describing patterns (averages, trends, variations) across studies

Research Synthesis

Combining results of multiple studies

1. Provides more compelling evidence than results of any single study
   - Single studies can have undue influence on practice & policy
   - We don’t use single subject (N=1) designs to assess public opinion, shouldn’t rely on single studies to answer important questions

2. Provides new opportunities to investigate
   - What works for whom under what conditions
   - Why results may vary across studies
   - Using analyses that capitalize on natural variations across studies
"Science is suppose to be cumulative, but scientists only rarely accumulate evidence scientifically" (Chalmers, Hedges, & Cooper, 2002, p. 12)

Scientific methods of research synthesis are
- Available
- Rapidly advancing

The Problem: Studies Pile Up

- "What can you build with thousands of bricks?" (Lipsey, 1997)
- Many studies are conducted on the same topic
- Which one(s) do we use? How do we use them?
Research, Reports, and Reviews: Ideal

Types of Reviews

- Reviews vary in amount of planning, transparency, and rigor
- Different approaches to research synthesis:
  - Traditional, narrative reviews (may include “vote counting”)
  - Systematic reviews (aim to minimize bias & error)
  - Meta-analysis (quantitative synthesis)
  - Rapid evidence assessment (AKA rapid reviews) - hybrids
Traditional Narrative Reviews

- Convenience samples of published studies
- Narrative description of studies
- Cognitive algebra or “vote counting” to synthesize results
  - How many studies had positive, null, negative, or mixed results
  - Relies on statistical significance in primary studies
    - Counting the wrong thing. Significance depends, in part, on sample size (studies may be too small to detect meaningful effects, large studies can detect differences that are meaningless)
- Decision rules are not transparent
- Vulnerable to many sources of bias and error …

Research, Reports, and Reviews: Ideal

![Diagram showing the hierarchy of research, reports, and reviews with positive and negative results]

Reviews
Publications
Research reports
Studies (all data collected)

Negative results  Positive results
Empirical Evidence of Bias

- Dissemination of research results is a biased process (Song et al., 2009, 2010)
  - Selective reporting, publication, citation, selection of evidence
  - Confirms favored theories
  - Overestimates benefits and underestimates harms of favored treatments
Outcome Reporting Bias

- Reporting of results is influenced by their direction and/or statistical significance
- “Cherry picking”
Evidence of ORB - 1

- Statistically significant and positive results are more likely to be
  - reported (mentioned at all)
  - fully reported (data provided)

- These reporting biases occur within studies
  (Chan et al., 2004a, 2004b; Chan & Altman, 2005; Dwan et al., 2008, 2014; Hahn et al., 2002; Pigott et al., 2011; Williamson et al., 2006)

- Unrelated to study or outcome “quality”  (Chan et al., 2004, 2005; Pigott et al., 2011; Williamson et al., 2006)

Evidence of ORB - 2

Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias

Kerry Dwan1, Douglas G. Altman2, Juan A. Arnaiz3, Jill Bloom4, An-Wen Chan5, Eugenia Cronin6, Evelyne Decullier7, Philippa J. Easterbrook8, Erik Von Elm9,10, Carrol Gamble11, Davina Gheusi11, John P. A. Ioannidis12,13, John Simes14, Paula R. Williamson1

- Statistically significant outcomes are more likely to be reported than nonsignificant outcomes
- Odds ratios 2.2 to 4.7  (Dwan et al., 2008)
Evidence of ORB - 3

**Frequency and reasons for outcome reporting bias in clinical trials: interviews with trialists**

R M D Smyth, research associate, J J Kirkham, research associate, A Jacoby, professor of medical sociology, D G Altman, professor of statistics in medicine, C Gamble, senior lecturer, P R Williamson, professor of medical statistics

- BMJ (2010)
- “The prevalence of incomplete reporting is high. Trialists seem generally unaware of the implications for the evidence base of not reporting all outcomes...”

Evidence of ORB - 4

**The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews**

Jamie J Kirkham, Kerry M Dwan, Douglas G Altman, Carrol Gamble, Susanna Dodd, Rebecca Smyth, Paula R Williamson

- BMJ (2010)
- 19/42 (45%) of meta-analyses had substantial errors due to ORB
  - 8 (19%) became non-significant after adjusting for ORB
  - 11 (26%) overestimated treatment effect by 20% or more
50% of completed studies are published (Dwan et al., 2008; Jones et al., 2013)

Publication rates may be lower in social sciences, observational studies, and low/middle income countries

31% publication rate in psychology
Publication Status

- Publication status is not a proxy for methodological quality (McLeon & Weitz, 2004; Moyer et al., 2010)
- Should never be used as an inclusion criteria in reviews (Chandler et al., 2013; Higgins & Green, 2011; Institute of Medicine, 2011)

Publication Bias

- Studies with statistically significant, positive results are 2-3 times more likely to be published than similar studies with null or negative results (Song et al., 2009, 2010)
  - likelihood of publication is related to direction and significance of results -- net of influence of other variables
  - (Begg, 1994; Cooper et al., 1997; Coursol & Wagner, 1986; Dickersin, 1987, 2005; Dwan et al., 2008; Easterbrook et al., 1991; Hopewell et al., 2007, 2009; Scherer et al., 2007; Song et al., 2000, 2009, 2010; Torgerson, 2006; Vecchi et al., 2009)
Publication Bias May be Worsening

Negative results are disappearing from most disciplines and countries

Daniele Fanelli

Received 1 May 2011; revision received 11 September 2011
© Akadémiai Kiadó, Budapest, Hungary 2011

Abstract Concerns that the growing competition for funding and citations might distort science are frequently discussed, but have not been verified directly. Of the hypothesized problems, perhaps the most worrying is a worsening of positive-outcome bias. A system that disfavors negative results not only distorts the scientific literature directly, but might also discourage high-risk projects and pressure scientists to fabricate and falsify their data. This study analyzed over 4,600 papers published in all disciplines between 1990 and 2007, measuring the frequency of papers that, having declared to have “tested” a hypothesis, reported a positive support for it. The overall frequency of positive supports has grown by over 22% between 1990 and 2007, with significant differences between disciplines and countries. The increase was stronger in the social and some biomedical disciplines. The United States had published, over the years, significantly fewer positive results than Asian countries (and particularly Japan) but more than European countries (and in particular the United Kingdom). Methodological artefacts cannot explain away these patterns, which support the hypothesis that research is becoming less pioneering and/or that the objectivity with which results are produced and published is decreasing.

Sources of publication bias are complex

- Investigators
  - don’t think null/negative results are worthwhile and/or don’t expect these results to be accepted/published
  - are less likely to submit null results for conference presentations (Song et al., 2009) and publication (Dickersin, 2005; Song et al., 2009)
- Peer reviewers & editors may be less likely to accept/publish null results? (Mahoney, 1977 vs. Song et al., 2009)

“Publication bias appears to occur early, mainly before the presentation of findings at conferences or submission of manuscripts to journals” (Song et al., 2009).
Dissemination Bias

- Studies with significant results are
  - Published faster (Hopewell et al., 2001)
  - Cited and reprinted more often (Egger & Smith)
- Easier to locate (esp. in English)

Reporting, Publication, Dissemination

Reporting, publication, dissemination biases
- Are ubiquitous
- Are cumulative
- Inflate effect size estimates
- (Altman, 2006; Hopewell et al., 2005, 2007, 2009; Song et al., 2009)
Biases in Haphazard Reviews

Haphazard reviews → Reviews → Publications → Research reports → Studies (all data collected)

Negative results → Positive results
Bias and Error in the Review Process

- Can occur at several stages, including:
  - Searching for studies
  - Selection of studies
  - Data extraction
  - Data analysis
  - Synthesis of results across studies

- Some examples...

Searching

- Bibliographic databases
  - Largely limited to published studies
  - Search results are likely to be affected by publication bias
Selection Bias

- Trivial properties of studies or reports affect recall and evaluation of information

- Memorable titles (Bushman & Wells, 2001)

Data Extraction

- Extracting data from studies is difficult
- Errors are common (Gøtzsche et al., 2007)
- Initial agreement is low (Tendal et al., 2009)
- Experimental evidence shows that duplicate extraction reduces errors (Buscemi et al., 2006)
Synthesis

- Narrative synthesis is
  - Unduly influenced by trivial properties of studies (Bushman & Wells, 2001)
  - Less accurate than meta-analysis (Bushman & Wells, 2001; Cooper & Rosenthal, 1980; Mann, 1994)

- Vote counting is not a good alternative
  - Does not consider sample size or heterogeneity
  - E.g., 10 studies: 6 positive, 2 null, 2 negative
    - Overall results depend on N and SE
    - Overall effect could be positive, null, or negative

Traditional Reviews and Well-Meaning Experts can be Misleading

- Scholars are human
- Rely on “natural” methods to filter and synthesize data
- The human brain is
  - Good at detecting patterns, maintaining homeostasis, defending territory
  - Bad at complex math, revising beliefs (Runciman, 2007)
- Research synthesis is too complex for informal methods, “cognitive algebra”
- Vulnerable to many sources of bias.
Summary

- Bias and error are common at every stage
  - Reporting
  - Publication
  - Dissemination
  - Reviews

Research, Reports, and Reviews: Reality

- Studies (all data collected)
- Research reports
- Publications
- Reviews

Positive results
Negative results
A concrete example

One study published in 1987
- How did investigators make use of data?
- How did reviewers make use of data?

Littell, 2008

Parent Training vs Multisystemic Therapy (Brunk et al., 1987)

- 43 families of abused/neglected children randomly assigned to
- Parent training (PT) groups or Multisystemic Therapy (MST)
- 33/43 families completed treatment and provided data on outcomes immediately after treatment
- 30 outcomes (scales and subscales)
Results Expected by Chance (30 outcomes)

Expected results:
- 1 favors PT
- 1 favors MST
- 28 no difference

Results Obtained (Brunk et al., 1987)

Actual results:
- 2 favor PT
- 5 favor MST
- 22 no difference
- 1 not reported
**Results Obtained (Brunk et al., 1987)**

Actual results:
- 2 favor PT
- 5 favor MST
- 22 no difference
- 1 not reported

**Outcome data**

- Data provided
- All (7) statistically significant results
- 12/22 non-significant results
- Outcome reporting bias

**Data Reported (Brunk et al., 1987)**

- Data provided on
Investigators’ Summary

- Both groups showed decreased psychiatric symptoms, reduced stress, and reduced severity of identified problems.
- MST was more effective than PT at restructuring parent-child relations.
- PT was more effective than MST at reducing identified social problems.
Reviewers’ Summaries

- Most reviewers used a single phrase to characterize results of this study, highlighting advantages of one approach (MST)
- Ignoring valuable information on relative advantages, disadvantages, and equivalent results of different approaches (Littell, 2008)

Systematic Reviews

- Aim to sum up the best available evidence in a way that minimizes errors and biases
- Use explicit, replicable research methods to identify relevant studies and objective techniques to analyze those studies
  - Develop and follow pre-determined plan (protocol)
  - Secondary analysis of existing data
- Treat review process as a form of survey research and follows basic steps in research process
  - Research reports, rather than people, are surveyed
  - Each research report is “interviewed” by a coder who codes information and quantitative findings
Systematic Reviews (SRs)

Steps to reduce bias and error:
- Set explicit inclusion/exclusion criteria
- Develop and document strategies for locating all relevant studies (regardless of publication status)
- Inter-rater agreement (reliability) on key decisions, data extraction, coding
- Formal study quality assessment (risk of bias)
- Meta-analysis (when possible) to synthesize results across studies

Meta-Analysis (MA)

Set of statistical procedures used to assess
- Averages across studies
- Variations across studies
- Potential sources of variation (moderators)
- Risk of bias (e.g., tests for publication & small sample bias)
- Systematic reviews don’t always include meta-analysis
  - Might include narrative synthesis (or no synthesis)
  - Can include multiple meta-analyses

- Meta-analyses are not always based on systematic reviews
  - Many use convenience sample of published studies
  - Vulnerable to publication and dissemination biases

**Meta-Analysis Alone is Insufficient**

- Meta-analysis
- Publications
- Research reports
- Studies (all data collected)
## Systematic Reviews Minimize Bias

![Diagram showing the hierarchy of studies, research reports, publications, and systematic reviews]

## Organizations Devoted to SRs

- Cochrane Collaboration – health care
  - [www.cochrane.org](http://www.cochrane.org)
- Campbell Collaboration – social care
  - [www.campbellcollaboration.org](http://www.campbellcollaboration.org)
  - Social Welfare
  - Education
  - Crime & Justice
  - International Development
  - Methods
  - Knowledge Translation & Implementation
- Society for Research Synthesis Methodology (SRSM)
  - Journals: *Research Synthesis Methods, Systematic Reviews*
Some “Systematic Reviews” Aren’t

- Evidence-based standards for SRs & MA
  - Based on methodological research (Cochrane Library)

- Standards for conduct of SRs
  - Developed by Cochrane and Campbell Collaborations
    (Higgins & Green, 2011; Chandler et al., 2013)

- Standards for reporting SRs & MA
  - PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses; Moher et al., 2009)
  - MECIR and MEC2IR (Methodological Expectations of Cochrane/Campbell Intervention Reviews)

- Standards not followed by US Evidence-based Practice Centers, most peer-reviewed journals, etc.

III. HOW TO CONDUCT A SYSTEMATIC REVIEW

one step at a time
Stages of a Research Synthesis - 1

A. Problem formulation
   - Clarify key questions and scope
   - Set explicit inclusion/exclusion criteria
   - Develop a detailed protocol (plan) for the review

B. Information retrieval
   - Literature search
   - Begin gathering information on studies

C. Data extraction and evaluation
   - Eligibility decisions: apply inclusion/exclusion criteria
   - Assess quality of included studies

Adapted from Cooper (1982, 2009)

Stages of a Research Synthesis - 2

D. Data analysis, synthesis, and interpretation
   - Integrate results of included studies
   - Interpret results

E. Report preparation
   - Narrative, statistics, graphics, tables

➢ Each stage presents challenges
➢ Most of the work is not spent in statistical analysis
➢ The scientific contribution of the final product is dependent on all stages of the review.
What’s required?

- A team with
  - Substantive expertise
  - Methodological expertise
  - Statistical expertise
  - Information retrieval expertise
- Time and money
  - Systematic reviews are labor intensive
  - $50-$150K depending on scope, complexity, and number of studies
  - But can be done with no external funding
- Research question (hypothesis) and clear purpose

A. Problem Formulation

“To be or not to be, that is the question. Next question?”

This section adapted from Wilson & Lipsey; Littell & Pigott; Valentine; and Shlonsky
Problem Formulation

- Establishing the research question
  - What is the specific hypothesis of interest in this synthesis?
  - What evidence is relevant to this hypothesis?
  - A well formulated problem will define the variables of interest so that relevant studies can be distinguished from irrelevant studies

- How broad do we go? Issues related to scope

- Strategy for establishing inclusion/exclusion criteria

- Writing a protocol

Conceptual Considerations

- Conceptual issues arise in attempts to combine results of studies that vary (to some degree) in their methods, treatments, samples, outcome measures
  - Apples, oranges, and other fruits
  - Parallel problems in studies of individuals (no 2 people are identical and there is no “average” person)

- Given variation in primary research
  - What should be included in a synthesis?
  - Depends on the question(s) we are asking
The Research Question and Objective

- What makes a good research synthesis question?
- What problem/intervention(s), outcomes, populations and contexts are of interest?
- Is there sufficient research on the topic to merit a synthesis?
- What purposes could be served by synthesizing knowledge in this area (objectives of the review)?
- Has the question already been asked and answered?
  - Search for and assess prior reviews before you start
  - If prior reviews have been done, is another one needed? If so, how will yours be different and improve upon prior reviews?

What Kind of Questions Can We Ask?

1. **Effectiveness**: Is one tx more effective than another or no tx at ameliorating or diminishing a problem or condition experienced by a client?
2. **Prevention**: Is one tx aimed at preventing or stopping the initial occurrence of a problem or condition more effective than another or no treatment?
3. **Diagnostic / Prognostic**: The likelihood or probability that a client will experience undesirable consequences within a given interval of time; which test is a better predictor of Y?
4. **Assessment**: Whether a client has a problem/condition/strength.
5. **Association / Description**: How does x₁ relate to x₂? What are the dynamics of a given population (e.g., population trends), satisfaction with services, needs assessment, other descriptive-type of questions.
Scope: How Narrow or Broad?

- **Specific, narrow questions**
  - Useful for testing effects of specific treatments
    - e.g., What are the effects of MST on delinquency?
    - Pros and cons of a narrow scope?
- **Broad, global questions**
  - Useful for generating new knowledge
    - Identify common elements of effective programs (Lipsey, 2008)
    - Build better intervention theories to guide program development and evaluation design (Lipsey, 1997)
    - e.g., Are truancy interventions effective in improving attendance?
  - Pros and cons of broad scope?
- **Differences in scope can lead to different conclusions**

Steps in Problem Formulation

- Determine the *conceptual* definitions relevant to the research
- Determine the *operational* definitions relevant to the research
- Set the review parameters in terms of **PICOS**
  - **Population/Participants** (problems/conditions)
  - **Interventions** (if applicable)
  - **Comparison group** (e.g., absolute vs. relative effects, counterfactual conditions)
  - **Outcomes** (primary and secondary outcomes, acceptable outcome measures)
  - **Study Design** (should be fit for purpose)
    - Additional criteria: geographic area, time, language, other relevant criteria
Participants:
Who is included in the sample of subjects?

- How is the problem/condition defined?
- What are the most important characteristics that describe the population/participants?
- Are there any relevant demographic factors (e.g. age, sex, ethnicity, risk status)?
- What is the setting (e.g. hospital, community, etc.)?
- Who should make the diagnosis/identify the problem?
- Are there other types of people who should be excluded from the review (because they are likely to react to the intervention in a different way)?
- How will studies involving only a subset of relevant participants be handled?

Participants: Examples

- Youth (ages 10-17) with social, emotional, and/or behavioral problems (MST review; Littell et al., 2005).
  - Why 10-17 only?
  - Exclude special populations?
  - Exclude studies with male or females?
  - Include families of these children?
  - What is the rationale for the choice?

- At-risk primary or secondary students (after-school program review)
  - Who are “at risk” students? How will we define “at-risk”?
  - Why primary or secondary students? Why not specific age or grade?
Types of Interventions

- What are the experimental and control (comparison) interventions of interest?
- Does the intervention have variations?
- Are all variations to be included?
- How will trials including only part of the intervention be handled?
- How will trials including the intervention of interest combined with another intervention (co-intervention) be handled?

MST Interventions: Conceptual and Operational Considerations?

- What kinds of programs are considered MST?
  - Are there certain components that must be present to identify the intervention of interest?
  - How will we deal with variation?
    - Exclude?
    - Code and examine in analysis?

- What types of treatments are the “right” comparisons?
  - Other interventions, “business as usual”, no treatment?
  - Inclusion criteria?
  - Code and examine in analysis?
Types of Comparisons

- Treatment vs. no treatment (absolute effects)
- Treatment A vs. Treatment B (relative effects)
- Treatment A vs. TAU (treatment as usual; relative effects)

- Absolute and relative effects should not be combined
- Different control conditions should be examined separately

Types of Outcomes

- Primary outcomes
  - Those specified by the intervention logic model
  - Needed to reach a conclusion about effects of the intervention(s)
  - Essential for decision-making
  - Emphasize outcomes of importance to clients
- Secondary outcomes
  - Additional outcomes useful for explaining effects
  - May not be a primary outcome specified by the logic model, but being measured by primary researchers and of interest to you

- Outcomes should cover potential and actual adverse effects
- Consider all relevant outcomes (including economic)
- Type and timing of outcome measurements
What are the Outcomes of Interest?

- MST research question states:
  - Placement in out-of-home living arrangements
  - Crime and delinquency
  - Behavioral outcomes
  - Psychosocial outcomes

- Conceptual, operational, and measurement considerations
  - Standardized vs. home-grown instruments
  - Collateral reports vs. self-reports
  - Post-treatment, short term, long term follow-up

Types of Study Designs

- Set a minimal threshold for acceptable study design(s)
  - Depends on the purpose of your review
  - Synthesis of invalid studies produces invalid conclusions (garbage in, garbage out)

- Certain designs are superior for certain questions
  - Need to include designs that can generate credible answers to the specific research question
  - For intervention effects: RCT, QED, other designs?

- May depend on/be limited by the types of designs being used in primary research

- Studies should not be excluded based on:
  - Sample size (or statistical power considerations)
  - Publication status
Additional Inclusion/Exclusion Criteria

- Geographic/political boundaries
- Language restrictions
- Time period

- These should be justified, potential biases weighed, and resulting limitations discussed in the review

Protocols for Systematic Reviews (SRs)

- A detailed protocol (plan) for the SR should be developed and made available to readers (Higgins & Green, 2011; Moher et al., 2009)
  - Protocols increase transparency, limit ad hoc decisions (Stewart et al., 2012)
- The review process is iterative and plans may change during the process
  - The final report should document and explain changes made (deviations from the protocol)
Small group exercise: Problem formulation

- Form a working group of 2 – 5 people and develop a preliminary research question for a systematic review using the PICOS framework
- Elect a recorder for the group who will report back to the whole
- Each group will share thoughts/challenges they faced and next steps

B. Information Retrieval

“This section adapted from Wilson & Lipsey, Hammerstrom, Lilell & Pigott"
Searching the Literature

The goal of a good literature search:

- Uncover all relevant studies (i.e., studies that meet inclusion criteria)
  - Published and unpublished studies

The search should be:

- Systematic -- how the search is conducted
- Transparent -- how well the search is documented and reported
- Comprehensive -- consistent with the ambitions and scope of the review

Sensitive Vs. Specific Searches?

- Sensitive searches: Finding any potentially relevant studies
- Specific searches: Finding studies that clearly meet criteria
- Advice: Aim for more sensitivity rather than specificity

\[\text{Sensitive: Find many studies that may be important} \quad \text{Specific: Find only relevant studies} \]
Important Steps in Locating Studies

1. Consult with information specialists (librarian, trial search coordinator) before searching
   - We cannot stress this enough!
2. Identify sources for your search
   - Multiple sources are required
   - Published and gray literature
3. Create search strategies for each source
   - Sensitive and specific searches are needed (more later)
4. Review results and revise search strategies as necessary
5. Process/manage references and reports
6. Log and report the search

Where will you Search for Studies?

- Electronic Databases
- Primary studies and prior reviews (citation searches)
- Contacting experts/research centers
- Internet searching (?)
- Hand searching
- Searching for Grey Literature
Searching Electronic Databases

- Part of every systematic review
- Mostly focused on published literature, although there are databases that include grey literature (e.g. ERIC, OpenGrey)
- To reach goal of more sensitivity, must search several electronic databases in related fields and on the web
- Aim for a wide variety of search terms combined with OR (more later)

Selection of Databases

- Consult a trial search specialist or librarian
- List of potentially important databases for social sciences in the C2 Information Retrieval Guide available on C2 website
- Be sure to search not only your discipline but other related disciplines as well
Examples of Databases

- **Education**: ERIC, British Education Index, Australian Education Index, CBCA Education, Education: A SAGE Full-text Collection; Education Abstracts.
- **Psychology**: PsycINFO, PubMed (Medline), Ageline, Psychology: A SAGE Full-Text Collection, Criminology: A SAGE Full-Text Collection
- **Sociology**: Sociological Abstracts, Contemporary Women’s Issues. Sociology: A SAGE Full-text Collection
- **Multidisciplinary**: Academic Search Premier, ProQuest Dissertations and Theses, FRANCIS, Social Sciences Index, SCOPUS, Web of Science

Citation Searches

Identify potentially relevant studies by scanning other citations and reference lists

- Reference lists of eligible studies
- Reference lists in prior reviews (included AND excluded studies)
- **ISI Web of Knowledge and Google Scholar**
  - Works that cite a particular reference
  - Works by a particular author
  - For SSCI, the journal must be indexed by Thompson Scientific (about 4,000 or roughly 20% of all journals in all academic disciplines)
  - Example on next slide looks up Henggeler
Example with Google Scholar

Generic vs. Specialized Search Engines

- Generic search engines (e.g., Google) can sometimes be a good resource for unpublished material
  - Far less efficient than specialized databases (db)

<table>
<thead>
<tr>
<th></th>
<th>Primary Benefit</th>
<th>Primary Drawback</th>
</tr>
</thead>
<tbody>
<tr>
<td>Specialized db</td>
<td>✷ retrieval of</td>
<td>✷ retrieval of unpublished work</td>
</tr>
<tr>
<td>(Specific)</td>
<td>unpublished work</td>
<td></td>
</tr>
<tr>
<td></td>
<td>✷ false hits</td>
<td></td>
</tr>
<tr>
<td>Generic search</td>
<td>✷ retrieval of</td>
<td>✷ false hits (dramatically)</td>
</tr>
<tr>
<td>engine</td>
<td>unpublished work</td>
<td></td>
</tr>
<tr>
<td>(Sensitive)</td>
<td>✷ false hits</td>
<td></td>
</tr>
</tbody>
</table>
Contacting Experts

- Email correspondence with active researchers
  - Snowball sample: Who else should I contact for information on relevant studies that may be unpublished, in progress, or in press?
- Email correspondence and website searches of relevant research centers, advocacy/professional organizations, government agencies
- Postings (calls for studies) on relevant listservs and newsletters
  - Reviewers must be aware of these sources to use them. Again, substantive expertise is essential.

Grey Literature

Definition of grey literature: “That which is produced on all levels of government, academics, business and industry in print and electronic formats, but which is not controlled by commercial publishers.” Fourth International Conference on Grey Literature

- Grey literature search required for Campbell/Cochrane reviews
- Important due to potential for publication bias if review are limited to published studies
- Some of the strategies identified above can be used to find grey literature (e.g. contact with experts)
- OpenGrey (www.opengrey.eu/)
- Google, Google Scholar, specific websites, some databases
Developing the Search Strategy: Keywords

- What are the key concepts? (include as search terms)
- How are these concepts presented in relevant disciplines?
- What are the related terms (synonyms)?
- How are these key concepts represented in the controlled vocabulary within each database/source to be searched?

Using a Thesaurus

1. From the research question, determine the main concepts to be searched. For example:
   - Multisystemic therapy or treatment
   - Youth or adolescent
   - Research or evaluation or outcome or impact
2. Consult the main databases to be searched
3. Look up each concept in the thesaurus for this database.
   - A thesaurus is an alphabetical listing of the controlled vocabulary (or descriptors) used within a subject database
   - A hierarchical arrangement is used so that Broader, Narrower and Related headings may be discovered
   - The user will be sent from invalid headings to valid headings
Construction of the Search Statement

- Construct your search statement by combining search terms relevant to PICO
- Use of operators (AND; OR; NOT)
- Important to use correctly as to not unintentionally limit or expand your search
  - Can be your friend or enemy!
- Text word variation, spelling and truncation
  - Ab*sen* to capture absence, absences, absenteeism, absentee
  - Be sure words are spelled correctly; English or foreign language nuances
- Use of database limiters? Consult librarians
Boolean Operators

**AND:** Both terms must be present in order for a record to be retrieved. Used to combine different concepts.

- parent participation AND achievement

**OR:** Either term may be present in order for a record to be retrieved. Used to search for related terms or synonyms.

- parent OR family

**NOT:** Used between two terms to ensure that the second term will not appear in any of the results.

- literacy NOT adult

(Parent* involvement OR parent* participation) AND academic achievement AND (elementary OR primary)

---

**Example Using PICOS**

```
<table>
<thead>
<tr>
<th>P</th>
<th>I</th>
<th>C</th>
<th>O</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>OR</td>
<td>Adolescents Youth</td>
<td>Multisystemic therapy Multisystemic treatment</td>
<td>Delinquency Crime Arrest Depression</td>
<td>Center Family</td>
</tr>
</tbody>
</table>
```

AND
Caution: Inconsistent Labeling of Methodology

- Reviewers only interested in experiments will sometimes use search strategies aimed at capturing design
  
  (After-school or afterschool)
  
  AND
  
  (random or evaluation or RCT)

- Several studies have shown that this strategy will miss relevant studies (25-67%)!
1. Multisystemic Therapy for Adolescents with Poorly Controlled Type 1 Diabetes: Stability of Treatment Effects in a Randomized Controlled Trial.


Abstract Only

2. Outcomes from Wraparound and Multisystemic Therapy in a Center for Mental Health Services System-of-Care Demonstration Site.

By: Stambough, Leyla Faw; Mustillo, Sarah A.; Burns, Barbara J. Journal of Emotional and Behavioral Disorders, v15 n3 p143-155 Fall 2007. (EJ779174)

ETF Full Text

ETF to Java
Managing Your Results

- Export the results
  - Save as a Text file
- Import into a bibliographic management software:
  - RefWorks,
  - Reference Manager,
  - EndNote
  - Zotero (freeware available at www.zotero.com)
- Edit your in-house database
  - Add Source code for each database searched (e.g., ERIC1, PsycINFO1)
  - Add notes to the records (e.g., included vs. excluded)
- Compile a Search History document listing the original search strategies

Documenting the Search Process

It is critical that you document your search in sufficient detail

- So that it could be replicated
- For adequate reporting

1. Document all information sources, dates covered, who performed the search, and date of the search
2. Document full electronic search strategy, including any limits used, so that the strategy can be replicated
3. Provide a flow chart of information through the search and screening process

(Moher et al., 2009)
Documentation of Database Search Strategy

<table>
<thead>
<tr>
<th>Database</th>
<th>Date Searched</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academic search premier</td>
<td>Jan 1980 to Nov 2013</td>
<td>US</td>
</tr>
<tr>
<td>Strategy</td>
<td>(anxiety OR &quot;school refus*&quot; OR &quot;school phobia&quot;) AND (attendance OR absent*) AND (evaluation OR intervention OR treatment OR outcome OR program) AND (student* OR school* OR child* OR adolescent*)</td>
<td></td>
</tr>
<tr>
<td>Limits: humans and children (0-18)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Database</th>
<th>Date Searched</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australian Education Index</td>
<td>Jan 1980 to June 2013</td>
<td>Australia</td>
</tr>
<tr>
<td>Strategy</td>
<td>(preschool* OR Lower school* OR primary school* OR kindergarten* OR middle school* OR upper school* OR high school* OR grammar school*) AND (absent* OR refus* OR phobia* OR anx* OR distress* OR attrition*) AND (evaluation* OR intervention* OR psychosocial OR treatment* OR outcome* OR program* OR what works*) AND (student* OR child* OR adolescent* OR youth OR young OR teenage*)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Database</th>
<th>Date Searched</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medline</td>
<td>Jan 1980 to Nov 2013</td>
<td>US</td>
</tr>
<tr>
<td>Strategy</td>
<td>(anxiety OR &quot;school refus*&quot; OR &quot;school phobia&quot;) AND (attendance OR absent*) AND (evaluation OR intervention OR treatment OR outcome OR program) AND (student* OR school* OR child* OR adolescent*)</td>
<td></td>
</tr>
</tbody>
</table>

**Information Retrieval: A Continuous Process**

- **Preliminary Searches**
  - Supports beginning steps: Definition of key concepts & research question
  - Use of standard reference tools and broad searches for review articles and key primary studies

- **Main Searches**
  - Identification of primary studies through searches of online databases, printed indices, Internet, branching, hand-searches
  - Most difficult given a number of challenges

- **Final Searches**
  - Occurs towards the end of the Review Process
  - Refine search terms and update original searches
Information Retrieval: Wrap Up

“Shoestring-budget information retrieval is likely to introduce bias, and should be avoided” (Hammerstrøm et al., 2004)

- Importance of information retrieval process
  - Not a “one-shot” deal
  - Requires expertise in the planning and implementation of searches
  - Consulting with a Librarian or Information Specialist is highly recommended (some Cochrane/Campbell groups have Search Coordinators who develop and execute search strategies)

- Use of bibliographic management software
  - Store, manage and organize results

- Ability to replicate review
  - Documentation of entire process, including search strategies used for each database, decisions taken, etc.

Campbell Systematic Reviews
2010: Supplement 1
First published: 7 September, 2010
Last updated: 19 August, 2010

Searching for studies:
A guide to information retrieval for Campbell Systematic Reviews
Kariinne Hammerstrøm, Anne Wade, Anne-Marie Klint Jørgensen

This guide is based on chapter 6 of The Cochrane Handbook (Lefebvre C, 2006 and Highe JPN, 2008)
C. Data Extraction: Screening, Coding, and Evaluation

Why Code Studies?

- Document decisions made in the review
- Provide an accounting of the research included in the review
  - Also helps identify what's missing
- Identify characteristics of participants, interventions, methods, and measures in the research
- Assuming that results may vary across studies, coding allows you to identify variables that might explain those differences (moderators)
Steps in the Process

1. Screen titles and abstracts
   - Retrieve full texts as needed
2. Review full texts
   - Identify eligible studies
   - Document reasons for exclusion of other studies
3. Extract data on included studies
   - Hierarchical structure: studies, reports, measures (more later)

Abstract Level Screening

- Does the document look like it might be relevant?
  - Based on title and abstract
    - If yes, retrieve full text
  - Exclude obviously irrelevant studies, but don’t assume the title and abstract are going to give reliable information on study design, outcome, subject characteristics or even interventions
- When in doubt, retrieve the full text
Key Decisions

- Important to have inter-rater agreement on key decisions:
  1. Full text retrieval (if either screener thinks it necessary)
  2. Study eligibility for the review (based on information in the full text of study reports, inter-rater agreement)
  3. Extraction and coding of key study elements
- When in doubt, double code
  - At least two trained raters working independently (more later)

Study Eligibility Form

- Develop a form using PICOS eligibility criteria
- Complete form for all studies retrieved as potentially eligible
- Modify criteria after examining sample of studies (controversial)
- Double-code eligibility decisions
- Maintain database on eligibility decisions for each study screened
  - Code included studies
  - Document reasons for exclusion for other studies
Study Coding Process

- Example from MST review:
  - Level 1 = initial screening (from titles and abstracts)
  - Level 2 = eligibility decisions (based on full text, multiple reports if applicable)
  - Level 3 = study characteristics (included studies only)
  - Level 4 = outcome measures and data (included studies only)
  - Level 5 = risk of bias (included studies only)

Screening Form: MST example (paper)

Level 1: Initial Screening (document level; from titles and abstracts)

1.1. Is this document about MST (perhaps in addition to other topics)?
   - Yes
   - No [STOP, code as unrelated]
   - Can’t tell [RETRIEVE FULL TEXT]

1.2. What is this?
   - Study of effects of MST for social, emotional or behavioral problems [GO TO LEVEL 2]
   - Study of effects of MST for medical condition(s) [STOP]
   - Review of MST outcome studies [HARVEST REFERENCES]
   - Descriptive, correlational, single-group, or case study [STOP]
   - Theoretical or position paper, editorial, or book review [STOP]
   - Practice guidelines or treatment manual [STOP]
   - Can’t tell [RETRIEVE FULL TEXT]
Eligibility Form: MST example (paper)

Level 2: Eligibility Decision (study level; from full text)
2.1. Does this study include two or more parallel cohorts (groups that received different treatments and were assessed at the same points in time)?
   • Yes
   • No [STOP, excluded]
   • Can’t tell
2.2. Is it a randomized experiment?
   • Yes
   • No
   • Can’t tell
2.3. Does this study include a licensed MST program?
   • Yes
   • No [STOP, excluded]
   • Can’t tell
2.4. Does it include youth (ages 10-17) with social, emotional, or behavioral problems?
   • Yes
   • No [STOP, excluded]
   • Can’t tell

Data Extraction Forms: MST example (Excel)

Excel sheet with multiple tabs
Shifts in units of analysis
• Level 1 – documents (screening)
• Compile documents in groups by studies
• Level 2 – study (eligibility)
• Level 3 - study (characteristics)
• Level 4 – outcome measures and outcome data
• Level 5 – study (risk of bias)
Once screening of all, relevant, full-text reports is complete, you will have:

- An accounting of ineligible (excluded) studies and the reasons for exclusion for each study
  - Table of excluded studies and reasons for exclusion is required in Cochrane/Campbell reviews and PRISMA
  - Should be made available to interested parties upon request if unable to publish in journal due to page/word limits.
- A set of studies eligible for data extraction and coding
### Sample Exclusion Tables

<table>
<thead>
<tr>
<th>Inclusion Criteria</th>
<th>Number Excluded</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intervention</td>
<td>43</td>
</tr>
<tr>
<td>Population</td>
<td>15</td>
</tr>
<tr>
<td>Study Design</td>
<td>35</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Study</th>
<th>Reason for Exclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aviv (2006)</td>
<td>Study design- No comparison group was used</td>
</tr>
<tr>
<td>Baden (1990)</td>
<td>Outcome- anxiety was not measured as an outcome</td>
</tr>
<tr>
<td>Beidas et al. (2010)</td>
<td>Study design- No comparison group was used</td>
</tr>
<tr>
<td>Clayton (2012)</td>
<td>Participant and outcome- Participants did not meet criteria for school refusal; anxiety was not measured as an outcome</td>
</tr>
<tr>
<td>Gutierrez-Maldonado (2009)</td>
<td>Outcome- Attendance was not measured as an outcome</td>
</tr>
<tr>
<td>Richardson (2013)</td>
<td>Study design- Not an intervention study</td>
</tr>
</tbody>
</table>

### Development of Coding Protocol

- **Types of information to code (overview, details to follow)**
  - Setting, study context, authors, publication date and type, etc.
  - Design, methods, and “study quality”
  - Program/intervention characteristics
  - Participants/clients/sample characteristics
  - Outcomes and measures
  - Findings, effect sizes

- **Hierarchical structure is needed to capture:**
  - Studies with multiple reports, multiple outcomes in reports, repeated measures on one or more outcomes
Sources for Coding

- **Other systematic reviews**
  - Campbell reviews often include coding manuals as appendices in protocols and reviews.
  - Useful for generic coding items (e.g., research design, risk of bias)

- **The literature**
  - Start with the assumption that you will find variability in results across studies in your review
    - What does the literature tell you about the plausible sources of that variability?
    - The literature reviewed at the beginning of a primary study often provides clues
    - Theory of change for the intervention can also be helpful

Types of Information to Code:
Basic information

- **Study setting, context, authors, publications date(s) and type, etc.**
  - Multiple publications; “study” vs “report”
  - Geographical/national setting; language
  - Publication type and publication bias issue
  - Publication date vs. study date
  - Research, demonstration, practice studies
Types of Information to Code: Methods 1

- Basic research design
  - Nature of assignment to conditions (if applicable)
  - Attrition, crossovers, dropouts, other changes to assignment
  - Nature of control condition (if applicable)
  - Multiple intervention and/or control groups

- Design quality dimensions
  - Initial and final comparability of groups
  - Treatment-control contrast
    - treatment contamination
    - blinding

Types of Information to Code: Methods 2

- Other methodological aspects
  - Issues depend on specific research area
  - Procedural, e.g.,
    - Monitoring of implementation, fidelity
    - Credentials, training of data collectors
  - Statistical, e.g.,
    - Statistical controls for group differences
    - Handling of missing data
Types of Information to Code: Methods 3: Study “quality” or risk of bias

- More than 200 scales and checklists available, few if any appropriate for systematic reviews (Deeks et al., 2003)
- Overall study quality scores have questionable reliability/validity (Jüni et al., 2001)
  - Conflate different methodological issues and study design/implementation features, which may have different impacts on reliability/validity
  - Preferable to examine potential influence of key components of methodological quality individually
- Weighting results by study quality scores is not advised!

Cochrane Risk of Bias Framework

- Focus on identifying potential sources of bias in studies of intervention effects:
  - **Selection bias** - Systematic differences between groups at baseline
  - **Performance bias** - Something other than the intervention affects groups differently
  - **Attrition bias** - Participant loss affects initial group comparability
  - **Detection bias** - Method of outcome assessment affects group comparisons (e.g., assessments aren’t blind)
  - **Reporting bias** - Selective reporting of outcomes
- Coded as low risk, high risk, or unclear risk
GRADE System

- Quality of evidence across trials
- Outcome-specific
- Considers:
  - Amount of data available (sparse data)
  - Consistency/inconsistency of results across trials
  - Study designs and risk of bias
  - Reporting bias
  - Possible influence of confounding variables
- Software available at: www.ims.cochrane.org/revman/gradepro
- Also see: www.gradeworkinggroup.org

What about Methodological Quality?

Our recommendation:

- Select a method of assessing and coding methodological quality
- Code all studies with the method you choose
- Examine the influence of methodological quality on effect sizes in the analysis stage, if possible
Types of Information to Code: Intervention(s)

- Program(s)/Intervention(s)
  - General program type (mutually exclusive or overlapping?)
  - Specific program elements (present/absent)
  - Any treatment(s) received by comparison group(s)
  - Treatment implementation issues
    - Integrity
    - Amount, “dose”
  - Goal is to differentiate across studies
  - Examples

Types of Information to Code: Participants

- Participants/clients/sample
  - Data usually reported at an aggregate level
  - Mean age, age range
  - Gender mix
  - Racial/ethnic mix
  - Risk, severity
  - Restrictiveness; special groups (e.g., clinical)
  - Examples
Types of Information to Code: Outcomes

- Outcome measures
  - Construct measured
  - Measure or operationalization used
  - Source of information
  - Composite or single indicator (item)
  - Scale: dichotomous, count, discrete ordinal, continuous
  - Reliability and validity
  - Time of measurement (e.g., relative to treatment)
  - Examples

Types of Information to Code: Findings

- Findings
  - Compute effect sizes when possible
  - May need to aggregate data or reconfigure findings
    - Add back the “dropouts”
    - Compute weighted means of subgroups (e.g., boys and girls)
  - Code data on which computations based (common situations)
  - We will look at this part of the coding in the next section
Development of Coding Protocol

- Iterative nature of development
- Structuring data
  - Data hierarchical (findings within studies)
  - Coding protocol needs to allow for this complexity
  - Analysis of effect sizes needs to respect this structure
  - Flat-file (example)
  - Relational hierarchical file (example)
- Pilot test the coding protocol

Double Data Extraction

- Important for at least two coders to independently code studies
  - Involves subjective judgments and difficulty interpretations of complex texts
  - Training coders is crucial
  - All or proportion of studies to be double coded?
- Assess reliability
  - Percentage of initial agreement
  - Cohen's kappa is better
- Agreement on key decisions
  - Study inclusion/exclusion, key characteristics, risk of bias, coding of results
- Pilot-test and refine codes!
Example of a Flat File

<table>
<thead>
<tr>
<th>ID</th>
<th>Paradigm</th>
<th>ES1</th>
<th>DV1</th>
<th>ES2</th>
<th>DV2</th>
<th>ES3</th>
<th>DV3</th>
<th>ES4</th>
<th>DV4</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>2</td>
<td>0.77</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>2</td>
<td>0.77</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>1</td>
<td>-0.1</td>
<td>5</td>
<td>-0.05</td>
<td>5</td>
<td>-0.2</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>2</td>
<td>0.94</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>1</td>
<td>0.96</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>82</td>
<td>1</td>
<td>0.29</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>185</td>
<td>1</td>
<td>0.65</td>
<td>5</td>
<td>0.58</td>
<td>5</td>
<td>0.48</td>
<td>5</td>
<td>0.068</td>
<td>5</td>
</tr>
<tr>
<td>186</td>
<td>1</td>
<td></td>
<td></td>
<td>0.83</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>204</td>
<td>2</td>
<td></td>
<td></td>
<td>0.88</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>229</td>
<td>2</td>
<td>0.97</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>246</td>
<td>2</td>
<td></td>
<td></td>
<td>0.91</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>274</td>
<td>2</td>
<td>0.86</td>
<td>3</td>
<td>-0.31</td>
<td>3</td>
<td>0.79</td>
<td>3</td>
<td>1.17</td>
<td>3</td>
</tr>
<tr>
<td>295</td>
<td>2</td>
<td>7.03</td>
<td>3</td>
<td>6.46</td>
<td>3</td>
<td>3</td>
<td>0.57</td>
<td></td>
<td></td>
</tr>
<tr>
<td>626</td>
<td>1</td>
<td>0.87</td>
<td>3</td>
<td>-0.04</td>
<td>3</td>
<td>0.1</td>
<td>3</td>
<td>0.9</td>
<td>3</td>
</tr>
<tr>
<td>1366</td>
<td>2</td>
<td></td>
<td></td>
<td>0.5</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Multiple ESSs handled by having multiple variables, one for each potential ESS.

Note that there is only one record (row) per study.

Example of a Hierarchical Structure

Study Level Data File

<table>
<thead>
<tr>
<th>ID</th>
<th>PubYear</th>
<th>MeanAge</th>
<th>TxStyle</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>92</td>
<td>15.5</td>
<td>2</td>
</tr>
<tr>
<td>7049</td>
<td>82</td>
<td>14.5</td>
<td>1</td>
</tr>
</tbody>
</table>

Effect Size Level Data File

<table>
<thead>
<tr>
<th>ID</th>
<th>ESSNum</th>
<th>Outcome</th>
<th>Type</th>
<th>TxN</th>
<th>CgN</th>
<th>ES</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>1</td>
<td>1</td>
<td>24</td>
<td>24</td>
<td>-0.39</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>2</td>
<td>1</td>
<td>24</td>
<td>24</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>3</td>
<td>1</td>
<td>24</td>
<td>24</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>4</td>
<td>1</td>
<td>24</td>
<td>24</td>
<td>-1.05</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>5</td>
<td>1</td>
<td>24</td>
<td>24</td>
<td>-0.44</td>
<td></td>
</tr>
<tr>
<td>7049</td>
<td>1</td>
<td>2</td>
<td>30</td>
<td>30</td>
<td>0.34</td>
<td></td>
</tr>
<tr>
<td>7049</td>
<td>2</td>
<td>4</td>
<td>30</td>
<td>30</td>
<td>0.76</td>
<td></td>
</tr>
<tr>
<td>7049</td>
<td>3</td>
<td>1</td>
<td>30</td>
<td>30</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Note that a single record in the file above is “related” to five records in the file to the right.
Common Mistakes and Challenges

- Too many coding items
- Too many subjective coding items (makes inter-rater reliability problematic)
- Coding two reports from the same study as two different studies
- Coder drift
- Coder fatigue
- Reporting in primary studies
  May need to request more information from study authors

D. Data Analysis, Synthesis, and Interpretation

Adapted from S. Wilson; Littell & Pigott; Wilson & Lipsey
Synthesizing Results of Primary Studies

- Purpose of a research synthesis is to
  - Combine results across studies
  - Test differences in results between studies
  - Identify gaps/questions for further research

- Methods for combining primary study results
  - Different methods can lead to different results
  - Qualitative methods (e.g., narrative)
  - Pseudo-quantitative methods (e.g., vote counting)
  - Meta-analysis

- Our focus today will be on meta-analysis

What is Meta-Analysis?

A set of statistical methods for summarizing and integrating results across studies

- Unit of analysis is the individual research study (or non-overlapping sample)
- Many meta-analytic techniques are available for assessing
  - Averages across studies (pooled effects)
  - Heterogeneity
  - Moderators
  - Sensitivity
  - Potential biases
Why Meta-Analysis?

- Meta-analysis is more accurate and reliable than
  - Narrative synthesis
  - Vote counting

Basic steps in meta-analysis

For each conceptually distinct outcome
1. Calculate one effect size for each study
2. Pool effect sizes across studies
   - Calculate a weighted average effect
   - Use inverse variance weights
     - Studies with more precise estimates (smaller SE) are given more weight than others
3. Assess heterogeneity (are results consistent across studies?)
4. Examine potential sources of variability across studies
   - Moderator analysis
A Brief Introduction to Effect Sizes

The results of each study are expressed using a quantitative index of effect size (ES).

ESs are measures of the strength (magnitude) and direction of a relationship of interest.

ESs have the advantage of being comparable (i.e., they estimate the same thing) across all of the studies and therefore can be summarized across studies in the meta-analysis.

Effect Size Basics

- Effect sizes can be expressed in many different metrics
  - d, r, odds ratio, risk ratio, etc.
    - So be sure to be specific about the metric you use!
  - The metric you use depends on several factors, including the number of groups and the types of outcome variables (dichotomous, continuous)
- Effect sizes can be unstandardized or standardized
  - Unstandardized = expressed in measurement units
  - Standardized = expressed in standardized measurement units
Unstandardized Effect Sizes

- Examples
  - 5 point gain in IQ scores
  - 22% reduction in repeat offending
  - €600 savings per person

- Unstandardized effect sizes are helpful in communicating intervention impacts
  - But in many systematic reviews are not usable since not all studies will operationalize the dependent variable in the same way

Standardized Effect Sizes

- Some standardized effect sizes are relatively easy to interpret
  - Correlation coefficient
  - Risk ratio

- Others are not
  - Standardized mean difference ($d$)
  - Odds ratio, logged odds ratio
Types of Effect Size

Most reviews use effect sizes from one of three families of effect sizes related to means, correlations, and proportions:

- the $d$ family, including the standardized mean difference,
- the $r$ family, including the correlation coefficient, and
- the odds ratio (OR) family, including proportions and other measures for categorical data.

Computing Effect Sizes

We will focus on three of the most common ES metrics:

- Standardized mean difference ($d$)
  - Difference between two means (continuous outcomes)
- Correlation Coefficient ($r$)
  - Association between two variables (continuous outcomes)
- Odds Ratios (OR)
  - Chance that something will happen compared to the chance that it will not happen (dichotomous outcomes)
Sample Sizes

Regardless of which metric you use, you will need to extract valid ns for each measure

- Valid ns often vary within studies. Instead of relying on initial/overall sample sizes, record
  - *Valid n* for each test or outcome or
  - *df* associated with each test
- Valid ns are needed to account for missing data and attrition, adjust for small sample bias, and compute variances

Standardized Mean Difference (SMD)

- Used when we are interested in comparing two groups on continuous measures (interval/ratio level data)
- Groups could be two experimental groups, or in an observational study, two groups of interest such as boys versus girls
- SMD is usually calculated from means, standard deviations, and valid ns -- the most direct method
- Without information on individual group sample sizes (n₁ and n₂), assume equal group n’s
Standardized Mean Difference

\[
ES_{sm} = \frac{\bar{X}_{G1} - \bar{X}_{G2}}{s_p}
\]

\[
s_p = \sqrt{(n_{G1} - 1)s_{G1}^2 + (n_{G2} - 1)s_{G2}^2}
\]

\[
SE_{sm} = \sqrt{\frac{n_{G1} + n_{G2}}{n_{G1}n_{G2}} + \frac{(ES_{sm})^2}{2(n_{G1} + n_{G2})}}
\]

Group means: \(\bar{X}_{G1}, \bar{X}_{G2}\)
Group sample sizes: \(n_{G1}, n_{G2}\)
Total sample size: \(N = n_{G1} + n_{G2}\)
Group standard deviations: \(s_{G1}, s_{G2}\)

Computing Standardized Mean Difference ES

Can be computed from various sources and types of data

1. Direct computational – uses means, SDs, \(ns\)
2. Algebraically equivalent methods use data from
   - ANOVA tables (one-way F)
   - \(t\) tests
   - tables of counts
3. Estimates from p-values, other sources (see Lipsey & Wilson, 2001)

➢ Direct and algebraically equivalent methods are preferred over estimates.
Product-Moment Correlation Coefficient ($r$)

- Used when we are interested in the association between two continuous variables
- Use $r$ as reported in (or calculated from) studies, transform to $Z$ for meta-analysis, then back to $r$

$$ES_r = r$$

$$ES_{Z_r} = 0.5 \log_e \frac{1 + ES_r}{1 - ES_r}$$

$$SE_{Z_r} = \frac{1}{\sqrt{n - 3}}$$

Odds Ratio

- Used when research results are expressed as relationship between two dichotomous variables
- Based on frequencies of a $2 \times 2$ contingency table

<table>
<thead>
<tr>
<th>Frequencies</th>
<th>Success</th>
<th>Failure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>Comparison</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

$$ES_{OR} = \frac{ad}{bc}$$
Interpretation of ES_{OR}

- $ES_{OR} = 1$, Treatment & Control equally effective
- $ES_{OR} > 1$, Higher odds of Treatment being effective compared to Control
- $0 < ES_{OR} < 1$, Lower odds of Treatment being effective compared to Control

Log Odds Ratio

- Some meta-analysts use the natural logarithm of the OR, or log odds ratio, because it has better statistical properties
- Distributed around 0 (instead of 1)
- Has an approximately normal distribution
- Positive values represent an increase in the odds and negative values represent a decrease in the odds

$$ES_{LOR} = \log_e \left( \frac{ad}{bc} \right) \quad SE_{LOR} = \sqrt{\left( \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} \right)}$$
ES Adjustments for Bias and Artifacts

- Small sample bias – Hedges’ g correction for SMD
  - Routinely used in meta-analysis
- Corrections for clustering
  - Clustered data are often analyzed/reported incorrectly (unit of analysis problem; see Higgins & Green, 2011)
- Correcting for methodological artifacts (Hunter & Schmidt, 2004)
  - Unreliability of variables involved in the effect size
  - Restriction in the range of variables
  - Dichotomization of continuous variables
  - Imperfect validity
  - Hunter-Schmidt corrections are controversial

Combining Effect Sizes Across Studies: Meta-Analysis

- Calculate weighted average ES across studies, using inverse variance weights
  - Studies with more precise estimates (larger N, smaller sd) contribute more to overall average than those with less precise estimates
- Choice between fixed and random effects models (and mixed models). Based on:
  - 1. a priori expectations, 2. statistical tests (heterogeneity)
  - Do we expect studies to estimate a single population parameter?
    - If Yes, use fixed effect model and test homogeneity assumption
    - Usually the answer is NO, and we use random effects or mixed models
Create a Set of Independent Effect Sizes

- Calculation of average effects (across studies) assumes that all of the study-level ESs are independent
  - Violations of this assumption (dependencies) should always be avoided
- But….we may have multiple effect sizes per study
  - Multiple groups (treatments, comparisons, subsamples)
  - Measures of different constructs (anxiety, attendance)
  - Multiple measures of the same construct (anxiety scales)
  - Multiple reports on the same measure (collateral sources)
  - Multiple follow-up measures

What to do with Multiple ES per Study?

1. Select one study-level ES for each conceptually distinct outcome (using plans developed a priori)
   - If there are multiple measures of the same construct
     - Select the most common or most direct measure OR
     - Compute a study-level mean of the measures (use pooled SDs)
   - If there are multiple treatment/comparison groups
     - Use the group that is most relevant OR
     - Create separate mean ES estimates from independent subsamples OR
     - Compute a mean ES across similar groups
   - If there are multiple endpoints (follow-ups)
     - Choose one endpoint that is most common across studies
2. Use robust standard errors (Tanner-Smith & Tipton, 2013)
Computing the Weighted Mean ES

- After we have our set of independent effect sizes, $ES_i$
  - Desired adjustments made (small sample bias, cluster)
  - $ES$ are statistically independent
- And calculate the inverse variance weight for each study
  $$w_i = \frac{1}{(SE_{ES})^2}$$
- We can calculate the Mean $ES$, standard error of the mean effect size and 95% confidence interval

Weighted Mean Effect Size

$$\bar{ES} = \frac{\sum_{i=1}^{k} w_i ES_i}{\sum_{i=1}^{k} w_i}$$

$$SE_{\bar{ES}} = \frac{1}{\sqrt{\sum_{i=1}^{k} w_i}}$$

$$ES_L = \bar{ES} - z_{(1-\alpha)} (SE_{\bar{ES}})$$

$$ES_U = \bar{ES} + z_{(1-\alpha)} (SE_{\bar{ES}})$$

$K = \text{number of effect sizes}$

$\alpha = \text{Significance level}$

$z = \text{Critical value from standard normal distribution}$
Assessing Heterogeneity of Effect Size Distribution

- Do the study-level effect sizes all estimate the same population parameter?
  - In a homogeneous distribution, the dispersion around the mean ES is no greater than what would be expected from sampling error alone
- Homogeneity tests performed under the fixed effect model
- Several tests
  - $Q$ statistic ($\chi^2$ distribution with N-1 df)
  - $I^2$ (represents the percentage of variation in ESs due to heterogeneity)

Interpreting the Q test

- Compare $Q$ to the $(1 - \alpha)$ critical value of the chi-square distribution with $k-1$ degrees of freedom
- Significant $Q = p < 0.05 = \text{heterogeneity}$
- Non-significant $Q = p > 0.05 = \text{homogeneity}$
\[ \chi^2(3) = 3.52, \ p = 0.32, \ n.s. \]

Graphing Effect Sizes: Forest Plots

- Easy to produce from raw data
- Cochrane’s RevMan software

p. 59 of MST pdf
Interpreting Mean ES

- Not straightforward for readers who are not used to thinking about distributions of data in SD units

- Cohen’s (1988) rules of thumb
  - SMD: Small < .20, Medium = .50, Large > .80
  - $r$: Small < .10, Medium = .25, Large > .4
  - Trying to move away from this; not very useful, as some “small” effects are practically meaningful while “large” effects are not always so

Translating the Mean ES

- Translation of effect sizes into other metrics that are more intuitive to readers

- C2 Indicated Truancy Intervention Review (Maynard et al., 2012) assessed effects of interventions on attendance
  - Mean ES = .46 (positive, significant and moderate), but what does it really mean?
  - A mean ES of .46 (.46 standard deviation units) is equivalent to 4.69 additional days of attending school.

Moderator Analysis

- When we find that the ES from a group of studies are heterogeneous, we can explore whether moderator variables explain this variation
- When we have continuous moderators, we use regression models (meta-regression)
- When we have categorical moderators, we use an ANOVA analogue
- Caution needed:
  - Moderator analysis is correlational
  - Moderators may be confounded within studies

Example from Sirin (2005)


There were 102 unique correlations that provided information about one or more components of SES. Table 3 presents the results of the methodological moderator analyses. The average ES for this distribution \((k = 102)\) was .31. This ES is significantly different from zero \((z = 144.12, p < .001)\). The test for homogeneity was significant, indicating that the correlations in this set were not estimating the same underlying population value, and therefore it is appropriate to look for possible moderators, \(Q(1, 102) = 2,068.36, p < .001\).
TABLE 3
Methodological characteristics: moderators of the relationship between SES and academic achievement

<table>
<thead>
<tr>
<th>Moderator</th>
<th>Categories</th>
<th>k</th>
<th>∑ between ES</th>
<th>Mean ES</th>
<th>±95% CI</th>
<th>±95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type of SES components</td>
<td>Education</td>
<td>30</td>
<td>.30</td>
<td>.39</td>
<td>.39</td>
<td>.31</td>
</tr>
<tr>
<td></td>
<td>Occupation</td>
<td>15</td>
<td>.28</td>
<td>.26</td>
<td>.26</td>
<td>.29</td>
</tr>
<tr>
<td></td>
<td>Income</td>
<td>14</td>
<td>.29</td>
<td>.27</td>
<td>.27</td>
<td>.31</td>
</tr>
<tr>
<td></td>
<td>Free or reduced-price lunch</td>
<td>10</td>
<td>.33</td>
<td>.32</td>
<td>.32</td>
<td>.34</td>
</tr>
<tr>
<td></td>
<td>Neighborhood</td>
<td>6</td>
<td>.25</td>
<td>.22</td>
<td>.22</td>
<td>.28</td>
</tr>
<tr>
<td></td>
<td>Home</td>
<td>4</td>
<td>.51</td>
<td>.49</td>
<td>.49</td>
<td>.53</td>
</tr>
<tr>
<td>SES range restriction</td>
<td>No restriction</td>
<td>78</td>
<td>.35</td>
<td>.35</td>
<td>.35</td>
<td>.36</td>
</tr>
<tr>
<td></td>
<td>3 to 7 SES groups</td>
<td>15</td>
<td>.28</td>
<td>.28</td>
<td>.28</td>
<td>.29</td>
</tr>
<tr>
<td></td>
<td>2 SES groups only</td>
<td>9</td>
<td>.24</td>
<td>.22</td>
<td>.22</td>
<td>.27</td>
</tr>
<tr>
<td>SES data source</td>
<td>Parents</td>
<td>62</td>
<td>.29</td>
<td>.29</td>
<td>.29</td>
<td>.30</td>
</tr>
<tr>
<td></td>
<td>Students</td>
<td>51</td>
<td>.36</td>
<td>.37</td>
<td>.37</td>
<td>.39</td>
</tr>
<tr>
<td>Achievement measures</td>
<td>General achievement</td>
<td>107</td>
<td>.24</td>
<td>.21</td>
<td>.21</td>
<td>.26</td>
</tr>
<tr>
<td></td>
<td>Verbal</td>
<td>45</td>
<td>.22</td>
<td>.22</td>
<td>.22</td>
<td>.23</td>
</tr>
<tr>
<td></td>
<td>Math</td>
<td>58</td>
<td>.32</td>
<td>.32</td>
<td>.32</td>
<td>.33</td>
</tr>
<tr>
<td></td>
<td>Science</td>
<td>57</td>
<td>.35</td>
<td>.34</td>
<td>.34</td>
<td>.36</td>
</tr>
</tbody>
</table>

Note. k = number of effect sizes; ES = effect size; CI = confidence interval for the average value of ES.
* p < .005.
Meta-Regression

- Use when we have multiple predictors and want to test all simultaneously
- Can combine continuous and categorical predictors in the form of dummy codes
- Can use standard statistical software BUT WITH CAVEATS

Computing Meta-Regression with Statistical Packages

- We must use weighted least squares to compute a meta-regression
- Weights are the inverse of the variance of the effect sizes – the same weight used to compute the mean effect size
- Standard statistical packages such as SPSS do not provide the correct standard errors for the regression coefficients – they must be adjusted
SPSS meta-regression example

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Regression</td>
<td>595.453</td>
<td>2</td>
<td>297.727</td>
<td>6.507</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>595.453</td>
<td>23</td>
<td>46.169</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1490.906</td>
<td>25</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), percent minority, grade level
b. Dependent Variable: Fishers z-transformation
c. Weighted Least Squares Regression - Weighted by inverse of variance

Coefficients a,b

<table>
<thead>
<tr>
<th>Model (Constant)</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
</tr>
<tr>
<td>1 (Constant)</td>
<td>1.39</td>
<td>1.45</td>
</tr>
<tr>
<td>grade level</td>
<td>0.164</td>
<td>0.037</td>
</tr>
<tr>
<td>percent minority</td>
<td>-0.001</td>
<td>0.037</td>
</tr>
</tbody>
</table>

a. Dependent Variable: Fishers z-transformation
b. Weighted Least Squares Regression - Weighted by inverse of variance

INCORRECT

Table 2. Relationships between Study Characteristics and Aggressive Behavior Effect Sizes with Method Variables Controlled (n=27)

<table>
<thead>
<tr>
<th>Study Characteristic</th>
<th>β (with method controls)</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Study Characteristics</td>
<td></td>
</tr>
<tr>
<td>Year of publication</td>
<td>-.002</td>
</tr>
<tr>
<td>Published (1) vs. unpublished (0)</td>
<td>-.07</td>
</tr>
<tr>
<td>USA (1) vs. Canada, Europe (0)</td>
<td>-.26</td>
</tr>
<tr>
<td>Student Characteristics</td>
<td></td>
</tr>
<tr>
<td>Gender mix (% male)</td>
<td>-.02</td>
</tr>
<tr>
<td>Age</td>
<td>-.10</td>
</tr>
<tr>
<td>Low SES (1) vs. Middle to High SES (0)</td>
<td>-.27</td>
</tr>
<tr>
<td>General (1) vs. Special population in regular class (0) vs. Special population in special school/class (1)</td>
<td>.05</td>
</tr>
<tr>
<td>Evaluator/Researcher Role in Study</td>
<td></td>
</tr>
<tr>
<td>Routine practice program (1 = research, 2 = demo, 3 = recruiter)</td>
<td>-.10</td>
</tr>
<tr>
<td>Program developer in researcher (1) vs. not researcher (0)</td>
<td>.10</td>
</tr>
<tr>
<td>Researcher modified program, yes (1) vs. no (0)</td>
<td>-.03</td>
</tr>
<tr>
<td>Delivery Personnel</td>
<td></td>
</tr>
<tr>
<td>Teacher provider (1) vs. other (0)</td>
<td>-.08</td>
</tr>
<tr>
<td>Researcher provider (1) vs. other (0)</td>
<td>-.12</td>
</tr>
<tr>
<td>Service professional provider (1) vs. other (0)</td>
<td>.21</td>
</tr>
<tr>
<td>Amount &amp; Quality of Treatment</td>
<td></td>
</tr>
<tr>
<td>Duration of treatment (in weeks)</td>
<td>-.03</td>
</tr>
<tr>
<td>Number of sessions per week (1 = less than weekly to 9 = daily)</td>
<td>.28*</td>
</tr>
<tr>
<td>Implementation problems (1 = yes, 2 = possible, 3 = no)</td>
<td>.28*</td>
</tr>
<tr>
<td>Treatment Elements</td>
<td></td>
</tr>
<tr>
<td>Anger management component (1) vs. other (0)</td>
<td>.07</td>
</tr>
<tr>
<td>Social problem-solving component (1) vs. other (0)</td>
<td>-.07</td>
</tr>
<tr>
<td>Perspective taking component (1) vs. other (0)</td>
<td>.04</td>
</tr>
<tr>
<td>Behavioral social skills component (1) vs. other (0)</td>
<td>.08</td>
</tr>
</tbody>
</table>

*Method controls: all-reported DV, gender adjustment, attrition, non-random assignment.
* p<.05
1. Funnel plots
2. Trim and fill analysis (need ~ 10+ studies)
3. Statistical tests (Egger’s test and others)

Do NOT use Failsafe N (Becker, 2005)
See Rothstein, Sutton, & Bornstein (2005)
Assessing Risk of Publication Bias - 3

Pseudo-95% confidence intervals

Assessing Risk of Publication Bias - 4

Publication bias will result in asymmetry in the funnel, caused by missing small, negative studies
Inter-ocular analysis of funnel plots is unreliable.

Trim and fill analysis estimates missing studies and recalculates pooled ES (a form of sensitivity analysis)
Assessing Risk of Publication Bias - 7

1. Funnel plots
   - Asymmetry can be caused by publication bias OR other factors (e.g., better implementation in small studies)
   - Contour-enhanced funnel plots (Moreno et al., 2009)

2. Trim and fill analysis (need ~ 10+ studies)
   - Does not perform well with heterogeneous studies

3. Statistical tests (Egger’s test and others)
   
   Do NOT use Failsafe N (Becker, 2005)
   See Rothstein, Sutton, & Bornstein (2005)

Sensitivity Analyses - 1

- Compare results obtained under two (or more) different sets of assumptions. For example,
- Assess impact of decisions made during the review
  - If an inclusion criterion (e.g., age range) was changed during the review, did that decision alter results?
  - Compare pooled ES for the studies that met original vs. expanded criteria. Are results robust or altered by the decision?
- Assess impact of outliers (unusual studies/ES) on overall results
  - Compare pooled ES with and without outlier(s)
Sensitivity Analyses - 2

- Assess possible impact of **missing data**
  - Publication bias (missing **studies**)
    - Trim and fill is a sensitivity analysis
  - Deviations from intent-to-treat analysis (missing **cases**)
    - Estimate upper and lower bounds for ES
    - Best case scenario: all missing treatment cases succeeded, all missing control cases failed
    - Worst case scenario: all missing tx cases failed, all missing controls succeeded
  - Outcome reporting bias (missing **data**)
    - Use ORBIT or Cochrane ROB ratings

Test assumptions and any deviations from the protocol

---

E. Reporting Results

- Percentage of chart which looks like Pac-man

This section adapted from S. Wilson; Litell & Pigott; Wilson & Lipsey
Reporting Results

- Reporting standards for systematic reviews and meta-analyses
  - Institute of Medicine (IOM) Standards for Reporting Systematic Reviews:
  - Methodological Expectations of Cochrane Intervention Reviews (MECIR) Standards for Reporting:
    - [http://editorial-unit.cochrane.org/mecir](http://editorial-unit.cochrane.org/mecir)
  - Campbell Collaboration (MEC2IR) standards:
    - [http://www.campbellcollaboration.org/Methods_Resources/MEC2IR.php](http://www.campbellcollaboration.org/Methods_Resources/MEC2IR.php)

Sample MECIR Reporting Standards: Methods and Results

Methods

- Cite the protocol for the review (0/1)
- Criteria for considering studies for this review (6/0)
- Search methods for identification of studies (5/1)
- Data collection and analysis (11/4)

Results

- Description of studies (13/4)
- Risk of bias in included studies (2/1)
- Effects of interventions (12/12)

(# mandatory/# highly desirable)
Tables

Required/recommended tables:
- Characteristics of included studies
  - Sample sizes, participant characteristics, study designs, intervention characteristics, information about outcomes measured (how and when)
- Characteristics of excluded studies, studies awaiting classification, ongoing studies (3 separate tables)
- Risk of bias assessment
- Summary of findings:
  - Results for each outcome, intervention and comparison conditions, number of studies and participants for each outcome, quality of the body of evidence
  - See Cochrane Handbook (Higgins & Green, 2011)
- Results of moderator analyses

Examples of Table/Figures

### Characteristics of Included Studies Table

<table>
<thead>
<tr>
<th>Author Year</th>
<th>Program Name</th>
<th>Description/Components</th>
<th>QUS</th>
<th>N</th>
<th>% Days Absent Pre Post</th>
<th>Grad Level</th>
<th>Study Result</th>
<th>ES</th>
<th>95% CI Lower Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Converse (2000)</td>
<td>School-based mentoring program</td>
<td>Mentoring by school staff/faculty once per week over 16 weeks</td>
<td>MC7</td>
<td>14</td>
<td>10</td>
<td>7</td>
<td>2</td>
<td>10</td>
<td>0.56</td>
</tr>
<tr>
<td>Dennis (2007)</td>
<td>Family intervention pilot project</td>
<td>Student enrollment in a school-based health center for comprehensive health services and social risk of teachers with at-risk students to engage in additional relationships</td>
<td>MC7</td>
<td>10</td>
<td>20</td>
<td>30</td>
<td>63</td>
<td>62</td>
<td>0.51</td>
</tr>
</tbody>
</table>

### Risk of Bias Table

<table>
<thead>
<tr>
<th>Type of Bias</th>
<th>Judgment</th>
<th>Support for Judgment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection Bias</td>
<td>High Risk</td>
<td>Neuronaed - Children were assigned to groups by principles of the schools.</td>
</tr>
<tr>
<td>Allocation concealment</td>
<td>High Risk</td>
<td></td>
</tr>
<tr>
<td>Performance Bias</td>
<td>High Risk</td>
<td>No measures taken to blind children.</td>
</tr>
<tr>
<td>Blinding of participants</td>
<td>High Risk</td>
<td>No measures taken to blind parents or teachers.</td>
</tr>
<tr>
<td>Blinding of outcomes assessment</td>
<td>High Risk</td>
<td></td>
</tr>
<tr>
<td>Detection Bias</td>
<td>High Risk</td>
<td>No measures taken to blind outcome assessors.</td>
</tr>
<tr>
<td>Incomplete outcome data</td>
<td>High Risk</td>
<td>Missing data in the CACL, due to measures not being returned total attrition=31%; differential attrition=10%</td>
</tr>
<tr>
<td>Reporting Bias</td>
<td>Low Risk</td>
<td>Study author reports data for all outcomes.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Study</th>
<th>Randomized</th>
<th>Allocation</th>
<th>Blinding of Participants</th>
<th>Blinding of Outcome Assessment</th>
<th>Detection</th>
<th>Incomplete Outcome Data</th>
<th>Reporting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Darby 2000</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Brant 2009</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Edwards 2007</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Gardner 2008</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Hutchings 2007a</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Study Search & Selection Flow Chart

- All standards require reporting information on the flow of studies from the # of references to # of studies included in the review
- A PRISMA flow chart provides a visual display of the results at each step of the selection process.

Identification → Reviewed 2,400 citations through databases, past reviews, and websites

Screening → Screened 374 full-text articles for eligibility

Excluded 2,026 studies deemed inappropriate upon review of the title and abstract

Excluded 309 studies that did not meet inclusion criteria

Excluded 41 studies that were deemed ineligible

Eligibility → Coded 65 studies

Included → 24 studies included

Forest Plot

- The “workhorse” graph in meta-analysis
- Display effect size estimates and confidence intervals for each study included in the meta-analysis
- Effect size estimates typically shown with blocks proportionate to the weight assigned to a given study
Forest Plots

- Studies should be ordered in a meaningful way
  - ES magnitude
  - Study weight (precision)
  - Chronological order
  - Other meaningful characteristics
- Include reference lines at the null value
- State direction of results

Funnel Plots

- Exploratory tool used to visually assess the possibility of publication/small study bias in meta-analysis
- Scatter plot of effect size (x-axis) against some measure of study size (y-axis)
  - x-axis: use log scale for ratio effect size measures, e.g., ln(OR), ln(RR)
  - y-axis: the standard error of the effect size is generally recommended (see Sterne et al., 2005 for a review of additional y-axis options)
- Not recommended in small meta-analysis (k<10)
Funnel Plots

- If publication bias is present, you would expect null or ‘negative’ findings from small n studies to be suppressed (i.e., missing from the plot)
- Asymmetry in the funnel plot for small n studies may provide evidence of possible publication bias
- Symmetry in the funnel plot provides some evidence against the possibility of publication bias

IV. SUMMARY AND CONCLUSIONS
Different review methods produce different results
- Traditional narrative methods are “haphazard” (Petticrew & Roberts, 2006) and can lead to the wrong conclusions
- Scientific methods are needed to minimize bias and error

We can use scientific principles and methods to synthesize evidence...

Reviews should take reporting and publication biases into account
- Include extensive search for grey literature
- Assess risk of bias (in original studies and in review)

Reviews should use adequate synthesis methods
- Narrative reviews are unreliable
- Vote counting is inadequate and potentially misleading
- Meta-analysis is the best available method for quantitative data
  - And it’s not hard to do

Haphazard reviews may be hazardous to public health
- Over-estimate positive effects of interventions
- Under-estimate or ignore potential harmful effects
- Minimize or ignore viable alternatives
- Promote ineffective or potentially harmful interventions
Strengths and Limitations of SRs

- **Strengths**
  - Reduce bias in reviews
  - Provide valid summaries of empirical evidence
  - Identify potential moderators, patterns not detected in primary studies

- **Limitations**
  - Limited by number and quality of available studies
  - Labor intensive, time consuming, costly
  - Not immune to bias

Valid Evidence Isn’t Always Available

- **Can’t make silk purses with sows’ ears**
  - Garbage in, garbage out
  - High quality of materials (studies) needed to produce strong, reliable results
  - “Best” materials depend on what we are building (aims)

- **Important to invest in**
  - High-quality primary research
  - Infrastructure that will make research results readily available for future reviews
  - Reliable and accurate reviews

- **Role of “empty” reviews**
Using Evidence in Context

Policy and practice are informed by
- Many types of evidence (qualitative, quantitative, anecdotal) on
- Many topics

Evidence isn’t enough
- Need to consider values, preferences, resources, ethics, legal constraints, etc.

Adapted from: Gibbs (2003), Davies (2004)

Effectiveness in Context

The MST example: inconsistent evidence of effects (Littell et al., 2005)
- Program was discontinued in
  - Ontario (concerns about program costs, no evidence of effects in local trials)
  - Hawaii (concerns about cultural sensitivity)
- Continued in
  - Sweden (despite local evidence that it was no more effective than treatment-as-usual), because administrators and staff like the structure it provides
  - Denmark (for lack of better alternatives)
  - USA and Norway (promoted as more effective than alternatives)
New Developments

- Network meta-analysis for comparative effectiveness questions (Salanti, 2012)
  - Uses direct and indirect comparisons to rank effectiveness of treatments, even if they haven’t been compared in head-to-head trials
- Semi-automated screening of abstracts (Wallace et al.; Shemilt et al.)
- Multivariate (RSE) models (Hedges, Tipton, Tanner-Smith)
- Synthesizing regression coefficients (Aloe)
- Synthesizing single subject designs (Shadish)

Thank You

Questions/Comments?

Contact Information
Julia Littell: jlittell@brynmawr.edu
V. ADDITIONAL RESOURCES

Evidence-based Standards for Reviews

- Campbell Collaboration (C2) policies and guidelines for the conduct and reporting of systematic reviews
  - http://www.campbellcollaboration.org/resources/research/the_production.php
Evidence-based Standards for Reviews

- Cochrane MECIR standards (Chandler et al., 2013)
  - http://www.editorial-unit.cochrane.org/mecir
- Cochrane Handbook (Higgins & Green, 2011)
  - http://handbook.cochrane.org/
- Institute of Medicine (IOM, 2011)
- PRISMA (Moher et al., 2009)
  - http://www.prisma-statement.org/

Websites

- Cochrane Collaboration: www.cochranecollaboration.org
- Campbell Collaboration: www.campbellcollaboration.org
- David B. Wilson’s effect size calculator: http://www.campbellcollaboration.org/resources/effect_size_input.php
Books

- Borenstein et al. (2009). Introduction to meta-analysis.

Software

- Comprehensive Meta-Analysis (CMA)
- OpenMeta
- R
- RevMan
- SAS
- SPSS
- Stata
- David Wilson’s macros for several statistical packages (including Excel) - access at http://mason.gmu.edu/~dwilsonb/ma.html
Training/Workshops

- Campbell and Cochrane Collaborations
  - Online training resources at
    - [http://www.campbellcollaboration.org/resources/training.php](http://www.campbellcollaboration.org/resources/training.php)
  - Workshops/presentations offered regularly at annual colloquia, related meetings, and by request
- Borenstein/CMA offers workshops on meta-analysis
- Your university?