Introduction to Systematic Reviewing and Meta-Analysis Methods

Workshop Manual for Participants

11-12 June 2014
Drs. Joshua R. Polanin and Emily Tanner-Smith
The Campbell Collaboration
Peabody Research Institute, Vanderbilt University
joshua.r.polanin@vanderbilt.edu
e.tanner-smith@vanderbilt.edu

Hosted by Department of Education and Professional Studies (EPS), University of Limerick
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recommended Reading</td>
<td>29</td>
</tr>
<tr>
<td>Appendix 1: PRISMA Reporting Checklist</td>
<td>30</td>
</tr>
<tr>
<td>Appendix 2: FINER Criteria for Evaluating Research Questions</td>
<td>33</td>
</tr>
<tr>
<td>Appendix 3: Example Search Strategy</td>
<td>34</td>
</tr>
<tr>
<td>Appendix 4: Example Coding Manual</td>
<td>35</td>
</tr>
<tr>
<td>Appendix 5: Common Effect Size Metrics and their Standard Errors</td>
<td>54</td>
</tr>
<tr>
<td>Appendix 6: Common Meta-Analysis Equations</td>
<td>57</td>
</tr>
</tbody>
</table>
Workshop Agenda

The workshop is designed for academic and policy-oriented researchers who are interested in conducting a Campbell Collaboration systematic review, as well as policymakers and practitioners interested in learning about the purpose and rationale of systematic reviews and the role of systematic reviews in evidence-based practice.

The workshop will focus on systematic review methods for studying the effectiveness of social interventions. The workshop includes lectures and small group work, and covers the entire systematic review process, including meta-analysis methods for the quantitative synthesis of research findings identified in a systematic review.

Upon completion of this workshop, participants should be able to:

- Describe different types of research questions that might be addressed with systematic review methods
- Identify and structure a systematic review topic
- Describe the steps for conducting a comprehensive literature search and gathering relevant studies
- Develop a coding protocol for extracting relevant information from included studies
- Interpret the results of a systematic review
- Report the results of a systematic review and/or meta-analysis using standard reporting guidelines
- Have a working knowledge of the tools and resources available for conducting a systematic review and meta-analysis

Definitions

**Systematic review**: Literature review that uses systematic and rigorous methods designed to minimize bias and maximize transparency in the literature review process.

**Meta-analysis**: the quantitative synthesis of effect sizes from two or more primary research studies.
## Day One

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.00 – 9.30</td>
<td>Welcome and introductions</td>
</tr>
<tr>
<td>9.30 – 10.45</td>
<td>Introduction to systematic reviews and topic formulation</td>
</tr>
<tr>
<td>10.45 – 11.00</td>
<td>Break</td>
</tr>
<tr>
<td>11.00 – 12.30</td>
<td>Inclusion and exclusion criteria</td>
</tr>
<tr>
<td>12.30 – 13.30</td>
<td>Lunch</td>
</tr>
<tr>
<td>13.30 – 14.30</td>
<td>Systematic literature searching principles</td>
</tr>
<tr>
<td>14:30 – 14.45</td>
<td>Break</td>
</tr>
<tr>
<td>14.45 – 16.00</td>
<td>Conducting systematic literature searches</td>
</tr>
</tbody>
</table>

## Day Two

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.00 – 9.30</td>
<td>Day one review and questions</td>
</tr>
<tr>
<td>9.30 – 11.15</td>
<td>Systematic review data collection</td>
</tr>
<tr>
<td>11.15 – 11.30</td>
<td>Break</td>
</tr>
<tr>
<td>11.30 – 12.30</td>
<td>Narrative synthesis methods</td>
</tr>
<tr>
<td>12.30 – 13.30</td>
<td>Lunch</td>
</tr>
<tr>
<td>13.30 – 15.00</td>
<td>Introduction to quantitative synthesis (meta-analysis)</td>
</tr>
<tr>
<td>15.00 – 15.15</td>
<td>Break</td>
</tr>
<tr>
<td>15.15 – 16.00</td>
<td>Reporting guidelines and summary</td>
</tr>
</tbody>
</table>
What is a Systematic Review?
A systematic review seeks to identify, collate, and systematically summarize all empirical evidence on a specific research topic, using explicit, systematic, transparent, replicable methods that are designed to minimize bias. This is in contrast to traditional narrative reviews, which often lack explicit eligibility criteria and transparency, and tend to summarize the results of included studies in a subjective manner. Systematic reviews collect data from a set of primary research studies that are identified for inclusion using pre-selected inclusion and exclusion criteria. Results from systematic reviews are then synthesized to provide a comprehensive summary of best-evidence on a given research topic. In contrast to primary research studies that may use individuals, families, classrooms, or schools as the unit of analysis, the primary studies themselves are the unit of analysis in a systematic review.

Stages of a Systematic Review
Systematic review methods can be applied to a variety of research questions and topics, but all systematic reviews have the following stages:

- Problem Formulation
- Data Collection
- Data Evaluation
- Data Analysis and Interpretation
- Report Preparation

Context and Rationale
Systematic reviews have increased in prominence since the advent of the evidenced-based practice movement, which emphasizes the integration of best available evidence and clinical expertise to inform practice (APA, 2006; Sackett et al., 1996). For example, evidence-based practice in psychology posits that the process of clinical decision-making is best informed by three strands of data: best available research evidence, clinical expertise, and patient characteristics, culture, and preferences. Many types of research designs can contribute to an understanding of the “best research evidence,” but systematic reviews provide one compelling form of such evidence.

Part of the evidence-based practice movement, and the shift toward meta-analysis as a legitimate method for synthesizing research evidence, recognizes that consistent results across multiple primary studies provide more convincing evidence than results from any single primary study. Any given primary study will be subject to sampling error and other potential biases related to the internal and external validity of that study. As such, it is generally unwise to base policy or practice decisions on findings from single primary studies.
Some systematic reviews will also involve meta-analysis, a term which encompasses a range of techniques for quantitatively synthesizing findings across studies included in a review. Because different primary studies often use different measures to represent the same underlying constructs, meta-analysis involves standardizing findings across studies in order to make results comparable. This is achieved by calculating effect sizes from the data reported in each primary study. An effect size is a standardized, quantitative index representing the magnitude and direction of a relationship. By representing the findings of each study included in a meta-analysis in the same form, the effect size permits a quantitative synthesis of those findings across studies.

While many systematic reviews use meta-analysis as the method of statistically synthesizing the included studies’ findings, not all systematic reviews will include a meta-analysis. For instance, some systematic reviews may qualitatively synthesize research findings in addition to or instead of conducting a meta-analysis. There are numerous qualitative approaches to research synthesis, including narrative synthesis, meta-ethnography, and realist synthesis (Denyer & Tranfield, 2006). Conversely, meta-analyses are also not always based on systematic reviews, and may instead be based on un-systematic or non-exhaustive searches of the literature. Most meta-analyses, however, endeavor to be exhaustive in their search so that results may generalize to a broader population of studies.

**Campbell Collaboration Systematic Reviews**

The Campbell Collaboration is an international network that supports the preparation and dissemination of high quality systematic reviews of research evidence on the effectiveness of social programs, policies, and practices. The mission of the Campbell Collaboration is to promote positive social change by contributing to better-informed decisions and greater effectiveness for public and private services around the world. In
part, the Campbell Collaboration fulfills this mission through its *Campbell Systematic Reviews* monograph series, which publishes peer-reviewed systematic reviews (and their corresponding protocols) on a diverse variety of topics related to social interventions.

**What is a Campbell Systematic Review?**
Campbell systematic reviews address the effectiveness of programs, policies, and practices (and related topics) in the areas of crime and justice, education, international development, and social welfare. Campbell systematic reviews are developed through a process that helps ensure that they are accurate, methodologically sound, comprehensive, and unbiased. Every Campbell review is required to have clear criteria for eligible research, an explicit and comprehensive search strategy, systematic and replicable coding and analysis of the key features and findings of the studies reviewed, and an integrative summary of those findings.

After peer and editorial review, approved systematic reviews are published in the *Campbell Systematic Reviews* monograph series and are freely accessible worldwide on the Campbell Collaboration website (http://www.campbellcollaboration.org/). The *Campbell Systematic Reviews* publication is supported by an editorial team that provides constructive assistance for the development of the systematic review as well as quality assurance for the completed review.

**Appropriate Topics for Campbell Reviews**
Campbell systematic reviews are intended to inform policymakers, practitioners, researchers, and other interested parties about the extent, quality, and findings of the available research evidence on the effectiveness of social programs, policies, or practices. Suitable topics, therefore, involve the synthesis of research that investigates the effects of deliberate, organized social interventions intended to bring about change on some set of targeted outcomes that represents improvement in the conditions the intervention is designed to address for a population experiencing those conditions. At the discretion of the sponsoring coordinating group, reviews may also be accepted on topics that are closely related to interventions, e.g., the predictive validity of diagnostic or risk instruments for identifying individuals appropriate for intervention, factors related to successful implementation of an intervention, and the like.

The policy and practice areas in which the Campbell Collaboration is currently able to support systematic reviews include crime and justice, education, international development, and social welfare; there are Campbell Coordinating Groups organized to support reviews in each of these areas. Though the range of each of these groups is rather broad, there are some social policy domains that are outside their ambit. Most notable among those are medical and primary health care interventions, which are the
purview of the Cochrane Collaboration (http://www.cochrane.org/), the Campbell Collaboration’s larger sister organization. On overlapping topics, joint reviews with the Cochrane Collaboration can be arranged. If there are questions about the suitability of a topic for a Campbell systematic review, they are best resolved by contacting a representative of the most relevant Campbell Collaboration coordinating group to discuss the matter. Contact information can be found on the Campbell Collaboration website at http://www.campbellcollaboration.org/coordinating_groups/index.php.

Within the Campbell policy domains, the scope of the intervention(s) addressed in a systematic review may range from narrow to broad depending on the purpose of the review and the availability of research. Generally speaking, Campbell reviews may define the focal intervention at any of three levels of breadth:

- **Specific name brand programs.** The intervention of interest in this instance is one that follows a defined protocol or manual that specifies what it is and how it is to be delivered as well as distinguishing it from similar interventions that follow a different protocol. Such interventions almost always carry a specific name that refers only to that protocol program and no other. Examples of programs of this sort that appear in Campbell reviews include Brief Strategic Family Therapy (BSFT), Multisystemic Therapy (MST), Mindfulness-Based Stress Reduction (MBSR), Motivational Interviewing, and farmer field schools.

- **Generic types of programs or practices.** A common focus for Campbell reviews is a particular type of program or practice that is not limited to a brand name version but, rather, encompasses research on all programs or practices of that type. Within a generic category of this sort, the interventions will share key defining features, but their particular form may vary in any application. Examples of interventions with this scope that have been the focus of Campbell reviews include stress management interventions, cognitive behavioral therapy, DNA testing in police investigations, volunteer tutoring programs, hot spots policing, work programs for welfare recipients, and micro-credit.

- **A range of programs for a problem or population.** The reviews that typically have the broadest scope cover a range of different interventions that are included because they all address a particular problem or outcome or, perhaps, the needs or conditions of a particular population. These reviews are often comparative, that is, they compare different interventions with regard to their relative effectiveness and, perhaps, cost effectiveness. Examples of reviews of this scope include programs for reducing school dropout, interventions to reduce cyber abuse, interventions to reduce drug use among prison inmates, programs to reduce pregnancy among adolescents, and cash transfers to influence educational outcomes.
Review Team for Conducting a Campbell Review

Campbell systematic reviews should not be conducted by a single researcher. A team of individuals is required to provide the relevant expertise and perform the necessary functions. An appropriate team should represent content knowledge in the substantive area of the review, familiarity with research methods for investigating intervention effects, proficiency in information retrieval and systematic literature search techniques, knowledge of systematic review methods, and statistical expertise in meta-analysis.

Though some individuals may have competencies in more than one of these areas, it would be rare for a single individual to have sufficient background in all of them. For the more specialized functions, such as information retrieval and meta-analysis expertise, the Campbell coordinating group in the relevant topic area may be able to provide assistance or consultation if the team lacks members with that expertise.

In addition to the range of expertise required to conduct a systematic review, multiple reviewers are needed to provide essential reliability checks on important judgments that must be made during the review process such as identification of studies meeting the inclusion criteria, extraction of data from those studies, and data entry.

Proposals for Campbell reviews may be submitted by a review team or invited by one of the Campbell coordinating groups. In either case, once a review team has been organized and has a topic in mind, the Campbell Collaboration has a standard procedure for approving the topic and working with the research team to complete the steps that lead to a finished, published systematic review.

The Key Components of a Campbell Review

To publish a review in the *Campbell Systematic Reviews* monograph series, there are three documents that must be submitted in succession by the review team: (a) a Title Registration Form, (b) a protocol for the proposed review, and (c) the completed review. Completing a Campbell systematic review is a rather rigorous process designed to ensure the highest possible quality in the Campbell monograph series. Nothing less is appropriate for a publication intended to provide sound summaries of the relevant evidence to policymakers and practitioners who want to know if the respective interventions are effective. At the same time, the Campbell Collaboration intends for this to be a supportive process that assumes at the outset that every review with an approved title registration will be completed and published in acceptable form. In this spirit, editors and other members of the respective coordinating groups will make every effort to help review teams develop acceptable protocols and complete publishable reviews.

In addition, there is another procedure for publishing a Campbell systematic review. Review teams that have completed a systematic review outside of the Campbell
editorial process may submit it to the editor of the appropriate coordinating group. To be considered for publication in *Campbell Systematic Reviews* in this fashion, the review must conform to the Campbell standards for content and organization. The coordinating group editor, in consultation with the *Campbell Systematic Reviews* editors-in-chief, will determine whether the submitted review is appropriate for possible publication. If so, the editor will arrange for content and method peer reviews. The editor will then provide feedback to the author about what revisions, if any, are required for the review to be published, or will reject it if it is not judged to be publishable even with revision. Upon submission of a draft that acceptable to the editor and approved by the co-chairs of the relevant coordinating group and the editors-in-chief, the review will be published in *Campbell Systematic Reviews* in the same fashion as a review developed through the usual Campbell editorial process. Any systematic review accepted under this alternative procedure must not be a duplicate of a version published elsewhere. Though this procedure is available to interested review teams, the regular three-step process summarized above is more likely to lead to a favorable outcome for any team that is not already very familiar with the Campbell standards and procedures.
Topic Formulation for Systematic Reviews

Systematic reviews can differ widely in scope and purpose, but all generally involve the same basic steps, the first of which is developing a research question. Three broad types of research questions are particularly appropriate for systematic reviews in the social and health sciences: questions about the etiology or epidemiology of particular conditions; questions about the efficacy or effectiveness of interventions or treatments addressing those conditions; and questions about group differences, either between naturally occurring groups (i.e., males and females) or between groups defined by researchers (e.g., between different diagnostic groups). Other types of systematic review questions might focus on prevalence of conditions, changes in conditions over time, diagnostic test accuracy, or questions related to economic evidence.

Defining a Research Question

A well-formulated problem will define the variables of interest so that relevant and irrelevant studies can be distinguished from one another.

Two primary questions should guide the development of a research question: (1) What is the specific hypothesis of interest in this systematic review? and (2) What evidence is relevant to this hypothesis?

In defining the research question for a systematic review, the researcher determines the conceptual definitions relevant to the research as well as the operational definitions that are relevant to the research, and then sets the review parameters in terms of PICOS (Populations, Interventions, Comparisons, Outcomes, and Study Designs).

Scope

Systematic reviews can vary widely in scope and are useful for answering specific, narrow questions about the effects of specific interventions or broad, global questions about the common elements of effective programs. Systematic reviews are not limited to questions about the effects of interventions, nor are they limited to a particular research method. The choice of the types of research designs to include in a systematic review depends on the research question.

Types of Research Questions

- Questions about intervention effects
- Questions about associations; diagnostic/prognostic questions
- Questions about group differences
Elements of a Good Research Question: SAMPLE

- Is it specific?
- Is it answerable?
- Are there measurable constructs?
- Is it practical? Is it relevant for policy/practice?
- Is it logical? Is it based on theory/logic model?
- Is it empirical? Can answers be attained using observable evidence?

Defining Eligibility Criteria

At the topic formulation stage, the researcher will need to define eligibility criteria, which follow directly from the research questions. Eligibility criteria should explicitly define the types of studies that are eligible for inclusion in the review, and give the reader a clear idea of the nature of the literature being reviewed. The specifics of the eligibility criteria will vary depending on the research question, but they will generally include four primary components and several secondary components:

1. The general topic – distinguishing features of the research to be included.
2. The population – types of research participants and populations of interest.
3. Research methods – types of research designs that will be included.
4. Other pertinent variables – related to e.g., interventions, settings, outcomes, data types.
5. Secondary characteristics – e.g., cultural and linguistic characteristics, literature time frame, or publication characteristics.

The PICOS Framework

One useful framework for developing research questions and eligibility criteria for a systematic review includes five primary components and is described with the acronym PICOS (Populations, Interventions, Comparisons, Outcomes, Study Designs; Higgins & Green, 2011). Under the PICOS framework, the eligibility criteria should specify the types of research participants in the primary studies of interest (Population); the critical features of the intervention under study, as well as its dose, format, frequency, duration, timing, etc. (Intervention); the types of comparison conditions that are eligible, whether no treatment, treatment as usual, placebo, or some other type of intervention

Definitions

Eligibility criteria: clearly defined criteria that provide operational and conceptual definitions of the types of studies that are eligible or ineligible for inclusion in a systematic review.
(Comparisons); the outcome constructs of interest, including the timing of measurement, operationalization, and source (Outcomes); and, finally, the types of study designs eligible for inclusion such as randomized, non-randomized, pretest-posttest only, etc. (Study Design). Although the PICOS framework applies best to systematic reviews of intervention effects, it can be modified for other types of systematic review questions.

<table>
<thead>
<tr>
<th><strong>Population</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>What are the most important characteristics of the population (demographics, pre-existing conditions, risk)?</td>
</tr>
<tr>
<td>What is the setting (hospital, community)?</td>
</tr>
<tr>
<td>Should certain types of participants or special populations be excluded?</td>
</tr>
<tr>
<td>How will studies involving subsets of participants be handled?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Intervention &amp; Comparisons</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>What are the qualifying intervention and comparison groups of interest?</td>
</tr>
<tr>
<td>Does the intervention vary in terms of theory, dose, delivery mode, delivery personnel, frequency/duration of contact, timing?</td>
</tr>
<tr>
<td>How will studies including only part of an intervention be handled?</td>
</tr>
<tr>
<td>How will studies including the intervention and another modality be handled?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Outcomes</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>What are the primary and secondary outcomes?</td>
</tr>
<tr>
<td>Do outcomes cover potential and actual adverse effects?</td>
</tr>
<tr>
<td>Are the outcomes relevant to all potential decision makers?</td>
</tr>
<tr>
<td>Type and timing of outcome measures?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Study Designs</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>For intervention studies, what types of quasi-experimental designs will be included (if any)?</td>
</tr>
<tr>
<td>For correlational studies, will cross-sectional and/or longitudinal studies be included?</td>
</tr>
<tr>
<td>Are other quality assessments used to limit eligible research designs?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Other Characteristics</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Will there be geographic and/or language restrictions, and how will that influence desired inference population?</td>
</tr>
<tr>
<td>Specific time frame for eligible studies, and rationale for that time frame?</td>
</tr>
</tbody>
</table>
Example Eligibility Criteria


Interventions. Interventions were required to meet three criteria that define social information processing programs. Although other treatment components were allowed to be present (e.g., behavioral social skills training, parenting skills training), the social information processing component had to be the clear focus of the program. The definitional criteria were as follows:

1. Training was provided on one or more of the social information processing steps: (1) encoding situational and internal cues, (2) interpretation of cues, (3) selecting or clarifying a goal, (4) generating or accessing possible responses, (5) choosing a response, (6) and behavioral enactment.
2. Cognitive skills or thinking processes were emphasized rather than specific behavioral skills.
3. Structured tasks and activities were used to teach cognitive skills and their application to actual social situations.

Settings and Subjects. The interventions were delivered to school-aged children (K-12 or equivalent ages in international settings) in regular school settings during school hours. Eligible programs had to be delivered universally to all children in a school, classroom, or group of schools or classrooms. Special education classrooms and alternative schools were considered eligible school settings, although classrooms in residential facilities (e.g., psychiatric hospitals) were not. After-school programs were not eligible. Any qualifying school in any region or country was eligible.

Outcomes. The study reported intervention effects for at least one outcome variable, measured on children, representing aggressive behavior, broadly defined to include violence, aggression, fighting, person crimes, disruptive behavior problems, acting out, conduct disorder, externalizing problems, and so forth.

Study Design. Only studies using a control group design were eligible. The intervention and control groups could be randomly or nonrandomly assigned but, if nonrandom, needed to be matched or provide evidence of initial equivalence on key demographic variables and/or pretests. Control groups could represent placebo, wait-list, no treatment, or “treatment as usual” conditions. Studies without control or comparison groups were not eligible. This included one-group pretest-posttest studies and studies in which a treatment condition was compared to another treatment condition.
Conducting Systematic Literature Searches

Using the eligibility criteria as a guide, meta-analysis next involves conducting and documenting a systematic search for all studies that meet the eligibility criteria. A well-implemented literature search will use multiple methods and sources, and make a concerted effort to include unpublished or difficult-to-locate studies ("grey literature"). Internet searches, hand searches of key journals, contact with experts in the field, and reference harvesting from previous meta-analyses, systematic reviews, and narrative literature reviews are commonly used to identify grey literature that may not be indexed in standard electronic bibliographic databases.

Conducting a systematic literature search involves seven general steps:

1. **Structure the question**: Categorize the elements of the literature search, which will be guided by the eligibility criteria (e.g., PICOS criteria).

2. **Choose databases/sources**: Choices will depend on the review time, objectives, timeline, and resources.

3. **Create search strategies for each source**: Identify search terms for the different categories of the research question (e.g., PICOS). Create search strategies for each source and add relevant search filters if desired/possible. Carry out the searches.

4. **Review initial results, revise search, iterate**: If too many false positives, or omissions of true positives, revise search strategies. New literatures sources may also be considered.

5. **Process references**: Important references from all sources into a common bibliographic management software, remove duplicates, etc.

6. **Log and report the search**: Log all decisions made in the search process, document all search terms, search dates, and any search filters used.

7. **Update search**: Update search if the review needs updating or the project data collection is lengthy.
Developing Search Strategies

A and B

A OR B

A NOT B

<table>
<thead>
<tr>
<th>Population</th>
<th>Intervention</th>
<th>Comparison</th>
<th>Outcome</th>
</tr>
</thead>
</table>

OR

AND
### Main Bibliographic Database in the Social, Behavioral, and Health Sciences

<table>
<thead>
<tr>
<th>Database</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASSIA</td>
<td>Applied Social Sciences Index and Abstracts</td>
</tr>
<tr>
<td>CENTRAL</td>
<td>Cochrane Central Register of Controlled Trials</td>
</tr>
<tr>
<td>DAI</td>
<td>Dissertation Abstracts International</td>
</tr>
<tr>
<td>ERIC</td>
<td>Education Resources Information Center</td>
</tr>
<tr>
<td>IBSS</td>
<td>International Bibliography of the Social Sciences</td>
</tr>
<tr>
<td>MEDLINE</td>
<td>MEDLINE/PubMed</td>
</tr>
<tr>
<td>PsycINFO</td>
<td>PsycINFO</td>
</tr>
<tr>
<td>Sociological Abstracts</td>
<td>Sociological Abstracts</td>
</tr>
</tbody>
</table>

### Examples of Grey Literature Databases

<table>
<thead>
<tr>
<th>Database</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DissOnline</td>
<td>DissOnline (German dissertations)</td>
</tr>
<tr>
<td>Theses Canada</td>
<td>Index to Theses (British and Irish dissertations)</td>
</tr>
<tr>
<td>Theses Canada</td>
<td>Theses Canada (Canadian theses)</td>
</tr>
<tr>
<td>PAIS International</td>
<td>PAIS International</td>
</tr>
<tr>
<td>PsycEXTRA</td>
<td>PsycEXTRA</td>
</tr>
<tr>
<td>Social Care Online</td>
<td>Social Care Online</td>
</tr>
<tr>
<td>AMED</td>
<td>Allied and Complementary Medicine (AMED)</td>
</tr>
<tr>
<td>BiblioMap</td>
<td>BiblioMap</td>
</tr>
<tr>
<td>BNI</td>
<td>British Nursing Index (BNI)</td>
</tr>
<tr>
<td>CINAHL</td>
<td>Cumulative Index to Nursing and Allied Health (CINAHL)</td>
</tr>
<tr>
<td>EMBASE</td>
<td>EMBASE</td>
</tr>
<tr>
<td>WHO Clinical Trials</td>
<td>WHO Clinical Trials</td>
</tr>
</tbody>
</table>
Systematic Review Data Collection

Purpose
- Provide an accounting of the research included in the systematic review.
  - Helps identify what’s missing in the literature.
- Identify the characteristics of the interventions, subjects, and methods in the research.
- Assuming different primary studies produce different results, study coding allows a researcher to identify variables that might explain those differences.

Elements of a Coding Manual
Once the set of eligible studies is identified and obtained, a systematic review uses objective and systematic coding procedures to extract information from the eligible studies. Systematic review data collection methods should be transparent and replicable and should include both study descriptors and study findings. The types of information extracted from studies in a systematic review vary, but generally includes the following information:

1. Study identification – general study identifiers and characteristics, e.g., title, date, author, source, study setting, region.
2. Study methodology – study methods, such as research design, quality indicators, risk of bias indicators.
3. Research participants – demographic, health, and background characteristics of study participants.
4. Intervention characteristics – if appropriate, characteristics of interventions or treatments received.
5. Effect sizes and dependent variables – if appropriate, statistical information used to estimate effect sizes for specified dependent variables of interest.

Sources for Coding Items
- Other systematic reviews
  - Campbell Systematic 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 there will be variability in treatment effects across the studies in a systematic review.
  - What does the literature suggest 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 useful.
Study Quality and Risk of Bias

- Variety of options for coding study methods
  - Cochrane risk of bias framework
  - GRADE system
  - Method quality checklists
  - Direct coding of methodological characteristics
Introduction to Meta-Analysis Methods

Many systematic reviews will also include a meta-analysis, which involves the quantitative synthesis of findings across multiple primary studies. Effect sizes are used to quantify results from primary studies in a standardized metric that permits synthesis of findings across studies that might use different measurements or operational definitions for the same underlying construct of interest.

Common Effect Size Metrics
- Measure of central tendency (mean, median)
- Proportion
- Unstandardized mean difference
- Standardized mean difference
- Odds ratio
- Relative risk ratio
- Risk difference
- Correlation coefficient
- Hazard ratio

Synthesizing Effect Sizes
Most meta-analyses will present an average effect size value synthesized from the individual effect sizes extracted from the primary studies included in the review. When calculating the average effect size, each effect size is typically weighted by the inverse of its sampling variance, so that effect sizes measured with greater precision are given greater weight because they provide better estimates of the underlying population parameter(s) of interest.

Meta-analysts will typically need to decide whether to use fixed effect (FE) or random effects (RE) inverse variance weights, the choice of which should be driven by the desired universe of generalization. Fixed effect weights are used for conditional inference models that condition (hold fixed) characteristics of studies related to the effect size parameter of interest. Thus FE weights are most appropriate to use when the meta-analyst only wishes to infer to parameters in the observed set of studies or studies identical to those in the meta-analysis; the only source of sampling error in fixed effect models is due to the sampling of people within studies.
Random effects weights are used for unconditional inference models that do not condition on characteristics of the studies. Thus RE weights are most appropriate to use when the meta-analyst wishes to infer to parameters in a hypothetical population of studies from which the observed studies are drawn; random effects models incorporate sampling error as well as variability in effect size parameters across studies. In many cases, random effects weights are most appropriate in meta-analyses in the social and behavioral sciences given the presumed diversity in study populations, interventions, outcomes, or settings.

<table>
<thead>
<tr>
<th></th>
<th>Fixed Effect Model</th>
<th>Random Effects Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Universe of generalization:</td>
<td>Parameters in the observed studies, or studies identical to those observed</td>
<td>Parameters in a hypothetical population of studies from which observed studies are drawn</td>
</tr>
<tr>
<td>Source(s) of error:</td>
<td>Sampling error due to sampling of participants within studies</td>
<td>Sampling error due to sampling of participants within studies, and variability in effect size parameters across studies</td>
</tr>
<tr>
<td>Parameter(s) estimated:</td>
<td>One true common effect size across all studies</td>
<td>Mean of the distribution of multiple true effect sizes in the population</td>
</tr>
<tr>
<td>Variation in observed effects is due to:</td>
<td>Sampling error only</td>
<td>Sampling error and true variation in effects across studies</td>
</tr>
</tbody>
</table>

Study weights are more balanced in RE models, such that large studies are assigned less relative weight and small studies are assigned more relative weight (compared to FE models). RE models will yield wider confidence intervals and larger standard errors.

Meta-analysts should not base the decision of whether to use FE or RE weights on the results of a statistical test for homo/heterogeneity (nor the p-value from any such test). Rather, the decision of which model to use should be made *a priori*, given knowledge of the literature to be synthesized as well as the meta-analysts’ desired universe of generalization.
**Quantifying Heterogeneity**

In many cases, the meta-analyst may not only be interested in the average effect, but also in the variability of those effects across different types of studies, participant samples, etc. Therefore, most meta-analyses also present statistics that summarize the amount of variability between studies ($\tau^2$), test whether any observed heterogeneity may be due to chance ($Q$ or $\chi^2$), and summarize the proportion of observed heterogeneity that can be considered true heterogeneity rather than statistical noise ($I^2$).

In meta-analysis, moderator analysis refers to statistical analyses that examine whether the coded study characteristics for each study are associated with the effect sizes from those studies; that is, whether coded variables can *explain* some or all of the observed heterogeneity in the effect sizes. Moderator analysis is conducted using analogs to ANOVA and linear regression that are modified for use with meta-analytic data. The choice between the ANOVA or regression frameworks depends on the measurement level of the covariate(s) of interest. Typically, “subgroup” analysis refers to a moderator analysis of categorical covariates in the ANOVA analog framework.

There are several exploratory statistical procedures that meta-analysts can use to explore the possibility of publication bias. The most commonly reported procedures include visual inspections of funnel plots, or regression based tests for funnel plot asymmetry. Other methods that have been used to assess publication bias in the past, such as the rank correlation test and variations on the failsafe N, are no longer recommended for use given their known limitations.
Interpreting and Reporting Your Results

An important part of any meta-analysis is the substantive interpretation of results in a way that is accurate, but still meaningful for readers. The best methods for interpreting meta-analysis findings will vary on a case-by-case basis, but might involve translating mean effect sizes into absolute or relative numbers (e.g., number needed to treat), translating mean effect sizes using a common scale(s) or measure that will be widely recognized by readers, or translating results into success/failure rates.

Translating Effect Sizes into Meaningful Metrics

- Using common/modal scales that will be familiar to researchers
- Percentages, relative risk reductions, absolute risk reductions/risk differences
- Number needed to treat
- Area under the curve
- Binomial effect size display
- Empirical, clinical, or theoretical benchmarks

Reporting Guidelines and Standards

- GRADE (Grades of Recommendation, Assessment, Development, and Evaluation) approach to assess quality of evidence
- EPICOT (Evidence, Population, Intervention, Comparison, Outcome, Time Stamp) format for making research recommendations
- PRISMA (Preferred Reporting Items for Systematic Review and Meta-Analyses) and MOOSE (Meta-Analysis of Observational Studies in Epidemiology) reporting guidelines
Campbell Coordinating Group Contacts (June 2014)

Crime & Justice Coordinating Group
Martin Killias, Co-Chair, University of Zurich, martin.killias@krc.ch
David Wilson, Co-Chair and Editor, George Mason University, dwilsonb@gmu.edu
Charlotte Gill, Managing Editor, George Mason University, cgill9@gmu.edu

Education Coordinating Group
Paul Connolly, Co-Chair, Queen’s University Belfast, paul.connolly@qub.ac.uk
Gary Ritter, Co-Chair, University of Arkansas, garyr@uark.edu
Sandra Wilson, Editor, Vanderbilt University, sandra.j.wilson@vanderbilt.edu
Carlton Fong, Managing Editor, University of Texas at Austin, managing.editor.ecg@c2admin.org

International Development Coordinating Group
Peter Tugwell, Co-Chair, University of Ottawa, tugwell.bb@uottawa.ca
Howard White, Co-Chair, International Initiative for Impact Initiative (3ie), hwhite@3ieimpact.org
Birte Snilstveit, Co-Editor, International Initiative for Impact Initiative (3ie), bsnilstveit@3ieimpact.org
Hugh Waddington, Co-Editor, International Initiative for Impact Initiative (3ie), hwaddington@3ieimpact.org
Emma Gallagher, Managing Editor, International Initiative for Impact Initiative (3ie), managing.editor.IDCG@c2admin.org

Social Welfare Coordinating Group
Jane Barlow, Co-Chair, Warwick Medical School, jane.barlow@warwick.ac.uk
Brandy Maynard, Co-Chair, Saint Louis University, bmaynar1@slu.edu
Nick Huband, University of Nottingham, editor.swcg@c2admin.org
Jane Dennis, Managing Editor, managing.editor.swcg@c2admin.org

Methods Coordinating Group
Ian Shemilt, Co-Chair, University of Cambridge, ian.shemilt@medschl.cam.ac.uk
Ariel Aloe, Co-Chair, University of Northern Iowa, ariel.aloe@uni.edu
Therese Pigott, Co-Editor, Loyola University Chicago, tpigott@luc.edu
Emily Tanner-Smith, Co-Editor, Vanderbilt University, e.tanner-smith@vanderbilt.edu
Joshua Polanin, Managing Editor, Vanderbilt University, managing.editor.methods@c2admin.org
Users Coordinating Group

John Westbrook, Co-Chair, SEDL, john.westbrook@sedl.org
Robyn Mildon, Co-Chair, Parenting Research Centre, rmildon@parentingrc.org.au
Useful Resources and Links

Systematic Review Organizations
The Campbell Collaboration: http://www.campbellcollaboration.org/
Centre for Reviews and Dissemination:
http://www.crd.york.ac.uk/crdweb/ HomePage.asp
The Cochrane Collaboration: http://www.cochrane.org/
International Initiative for Impact Evaluation:
http://www.3ieimpact.org/en/evidence/systematic-reviews/
The Joanna Briggs Institute: http://joannabriggs.org/

Online Readings
Cochrane Handbook for Systematic Reviews of Interventions:
http://handbook.cochrane.org/
Institute of Medical Standards for Systematic Reviews:
http://books.nap.edu/catalog.php?record_id=13059
Centre for Reviews and Dissemination Guidance for Systematic Reviews:
www.york.ac.uk/inst/crd/pdf/Systematic_Reviews.pdf

Reporting Guidelines
Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA): http://www.prisma-statement.org/

Meta-analysis Training Resources
Campbell Collaboration Training Resources:
http://www.campbellcollaboration.org/resources/training.php
Cochrane Collaboration Training Resources: http://www.cochrane.org/training
David Wilson’s Online Effect Size Calculator:
http://www.campbellcollaboration.org/resources/effect_size_input.php

Software Resources
Abstrackr Abstract Screening: http://sunfire34.eecs.tufts.edu/account/login
EndNote: http://endnote.com/
metafor Package in R: http://www.metafor-project.org/doku.php
OpenMeta [Analyst]: http://www.cebm.brown.edu/open_meta
SPSS, SAS, Stata macros: http://mason.gmu.edu/~dwilsonb/ma.html
Stata: http://www.stata.com/support/faqs/statistics/meta-analysis/
WinBUGS: http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml
References


Recommended Reading


## Appendix 1: PRISMA Reporting Checklist

<table>
<thead>
<tr>
<th>Section/topic</th>
<th>#</th>
<th>Checklist item</th>
<th>Reported on page #</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TITLE</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Title</td>
<td>1</td>
<td>Identify the report as a systematic review, meta-analysis, or both.</td>
<td></td>
</tr>
<tr>
<td><strong>ABSTRACT</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Structured summary</td>
<td>2</td>
<td>Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.</td>
<td></td>
</tr>
<tr>
<td><strong>INTRODUCTION</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rationale</td>
<td>3</td>
<td>Describe the rationale for the review in the context of what is already known.</td>
<td></td>
</tr>
<tr>
<td>Objectives</td>
<td>4</td>
<td>Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).</td>
<td></td>
</tr>
<tr>
<td><strong>METHODS</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protocol and registration</td>
<td>5</td>
<td>Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.</td>
<td></td>
</tr>
<tr>
<td>Eligibility criteria</td>
<td>6</td>
<td>Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.</td>
<td></td>
</tr>
<tr>
<td>Information sources</td>
<td>7</td>
<td>Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.</td>
<td></td>
</tr>
<tr>
<td>Search</td>
<td>8</td>
<td>Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.</td>
<td></td>
</tr>
<tr>
<td>Study selection</td>
<td>9</td>
<td>State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).</td>
<td></td>
</tr>
<tr>
<td>Data collection process</td>
<td>10</td>
<td>Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.</td>
<td></td>
</tr>
<tr>
<td>Data items</td>
<td>11</td>
<td>List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.</td>
<td></td>
</tr>
<tr>
<td>Risk of bias in individual studies</td>
<td>12</td>
<td>Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.</td>
<td></td>
</tr>
<tr>
<td>Summary measures</td>
<td>13</td>
<td>State the principal summary measures (e.g., risk ratio, difference in means).</td>
<td></td>
</tr>
<tr>
<td>Synthesis of results</td>
<td>14</td>
<td>Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., $I^2$) for each meta-analysis.</td>
<td></td>
</tr>
<tr>
<td>Section/topic</td>
<td>#</td>
<td>Checklist item</td>
<td>Reported on page #</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>---</td>
<td>-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>Risk of bias across studies</td>
<td>15</td>
<td>Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).</td>
<td></td>
</tr>
<tr>
<td>Additional analyses</td>
<td>16</td>
<td>Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.</td>
<td></td>
</tr>
<tr>
<td><strong>RESULTS</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study selection</td>
<td>17</td>
<td>Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.</td>
<td></td>
</tr>
<tr>
<td>Study characteristics</td>
<td>18</td>
<td>For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.</td>
<td></td>
</tr>
<tr>
<td>Risk of bias within studies</td>
<td>19</td>
<td>Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).</td>
<td></td>
</tr>
<tr>
<td>Results of individual studies</td>
<td>20</td>
<td>For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.</td>
<td></td>
</tr>
<tr>
<td>Synthesis of results</td>
<td>21</td>
<td>Present the main results of the review. If meta-analyses are done, include for each, confidence intervals and measures of consistency.</td>
<td></td>
</tr>
<tr>
<td>Risk of bias across studies</td>
<td>22</td>
<td>Present results of any assessment of risk of bias across studies (see Item 15).</td>
<td></td>
</tr>
<tr>
<td>Additional analysis</td>
<td>23</td>
<td>Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).</td>
<td></td>
</tr>
<tr>
<td><strong>DISCUSSION</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Summary of evidence</td>
<td>24</td>
<td>Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).</td>
<td></td>
</tr>
<tr>
<td>Limitations</td>
<td>25</td>
<td>Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).</td>
<td></td>
</tr>
<tr>
<td>Conclusions</td>
<td>26</td>
<td>Provide a general interpretation of the results in the context of other evidence, and implications for future research.</td>
<td></td>
</tr>
<tr>
<td><strong>FUNDING</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Funding</td>
<td>27</td>
<td>Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.</td>
<td></td>
</tr>
</tbody>
</table>

For more information, visit: [www.prisma-statement.org](http://www.prisma-statement.org).*
PRISMA 2009 Flow Diagram

Records identified through database searching (n = )

Additional records identified through other sources (n = )

Records after duplicates removed (n = )

Records screened (n = )

Records excluded (n = )

Full-text articles assessed for eligibility (n = )

Full-text articles excluded, with reasons (n = )

Studies included in qualitative synthesis (n = )

Studies included in quantitative synthesis (meta-analysis) (n = )


For more information, visit www.prisma-statement.org.
### Appendix 2: FINER Criteria for Evaluating Research Questions

<table>
<thead>
<tr>
<th>FINER Criteria</th>
<th>Adequate number of subjects</th>
<th>Adequate technical expertise</th>
<th>Affordable in time and money</th>
<th>Manageable in scope</th>
<th>To the investigator</th>
<th>Confirms/refutes previous findings</th>
<th>Extends previous finding</th>
<th>Provides new findings</th>
<th>Abides by ethical standards of scientific conduct</th>
<th>To scientific knowledge</th>
<th>To clinical and health policy</th>
<th>To future research directions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feasible</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interesting</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Novel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ethical</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relevant</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix 3: Example Search Strategy

The following search strategy was reported in a Campbell Collaboration systematic review on school dropout prevention and intervention programs for school aged children and youth.

Search terms for PsycINFO and ERIC within the CSA database, limited by the date range of 1985-2010 using CSA age limits from childhood through young adulthood:

((DE=((school dropouts) or (school attendance) or (truancy) or (school graduation) or (high school graduates))) or (KW=((school complet*) or (GED) or (general education development) or (high school diploma) or (dropout*) or (alternative high school*) or (drop* out) or (career academ*) or (school NEAR absen*) or (chronic* NEAR absen*) or (school enrollment) or (high school equivalency) or (school failure) or (high school reform) or (educational attainment) or (grade promotion) or (grade retention) or (school nonattendance) or (graduation rate) or (school refusal)))
and
(KW=((intervention) or (quantitative) or (program evaluation) or (random*) or (prevent*) or (pilot project*) or (youth program*) or (counseling) or (guidance program*) or (summative evaluation) or (RCT) or (clinical trial) or (quasi- experiment*) or (treatment outcome*) or (program effect*) or (treatment effect*) or (evaluation) or (experiment*) or (social program) or (effective*))
and not
((KW=((post-secondary) or (undergraduate) or (doctoral) or (inmate) or (schizophrenia) or (traumatic brain injury) or (autis*) or (abuse) or (antidepressant*) or (unipolar depression) or (risperidone) or (chronic illness) or (major depressive disorder) or (bulimia) or (buprenorphine) or (malaria) or (heroin) or (cancer) or (major depression) or (Massage therapy) or (fibromyalgia) or (Paroxetine) or (clomipramine) or (olanzapine) or (tuberculosis) or (spinal cord injury) or (epilep*) or (antiepileptic) or (HIV) or (psychosis) or (OCD) or (obsessive-compulsive) or (EEG) or (PTSD) or (tourette*) or (insomnia) or (obes*) or (anorexia) or (methadone) or (borderline personality disorder) or (mental retardation)) or (DE=((higher education) or (college students) or (treatment dropouts) or (employee absenteeism))))

10.4073/csr.2011.8

For more information, visit http://campbellcollaboration.org/.
Appendix 4: Example Coding Manual

The following coding manual was reported in a protocol for a Campbell Collaboration systematic review on school dropout prevention and intervention programs for school aged children and youth.

STUDY LEVEL VARIABLES

Step 1. Study Identifiers, Study Context, Group Identification, and Study-Level Coding

STUDY IDENTIFIERS

The "unit" you will code here consists of a study, i.e., one research investigation of a defined subject sample or subsamples compared to each other, and the treatments, measures, and statistical analyses applied to them. Sometimes there are several different reports (e.g., journal articles) about a single study. In such cases, the coding should be done from the full set of relevant reports, using whichever report is best for each item to be coded; BE SURE YOU HAVE THE FULL SET OF RELEVANT REPORTS BEFORE BEGINNING TO CODE. Sometimes a single report describes more than one study, e.g., one journal article could describe a series of similar studies done at different sites. In these cases, each study should be coded separately as if each had been described in a separate report.

Each study has its own study identification number, or StudyID (e.g., 619). Each report also has an identification number (e.g., 619.01), which you will find printed on the folder holding the report. The ReportID has two parts; the part before the decimal is the StudyID, and the part after the decimal is used to distinguish the reports within a study. (These two types of ID numbers, along with bibliographic information, are assigned and tracked using the bibliography.) When coding, use the study ID (e.g., 619) to refer to the study as a whole, and use the appropriate report ID (e.g., 619.01) when referring to an individual report.

While reading reports for coding, be alert to any references to other dropout studies that may be appropriate to include in this meta-analysis. If you find appropriate-looking references that are not currently entered into the bibliography, the references may need to be entered.

[StudyID] Study identification number of the study you are coding, e.g., 1923.
[Coder] Coder's initials (select from menu)
[CodeDate] Date you began coding this study (will be inserted automatically)

STUDY CONTEXT

[H1] Year of publication (four digits): If more than one report, choose earliest date.

[H2] Country in which study conducted.
   1. USA
   2. Great Britain
   3. Canada
   4. Scandinavia: Denmark, Finland, Norway, Sweden
   5. Australia/New Zealand
   6. Other Western European Country: __________
   7. Other: ________________

[H3] Type of publication. If you are using more than one type of publication to code your study, choose the publication that supplies the effect sizes (in cases where more than one report provides effect sizes, choose a “peer reviewed” choice over another option, or choose the report that provides the most effect sizes).


35
GROUP IDENTIFICATION AND SELECTION

At this stage, you will need to identify the aggregate treatment and/or comparison groups used in the study for which effect size statistics can be computed. To do this, you will need to distinguish aggregate groups, which you will code here, from subgroups (or breakouts), which you will code later:

(1) Aggregate treatment and/or comparison groups. The largest participant groupings on which contrasts between experimental conditions or contrasts between time points can be made. Note that the designations “comparison group” and “control group” refer to any group with which the treatment of interest is compared that is presumed to represent conditions in the absence of that treatment, whether a true random control or not. Often there is only one aggregate treatment group and one aggregate control group, but it is possible to have a design with numerous treatment variations (e.g., different levels) and control variations (e.g., placebos) all compared (e.g., in ANOVA format) to each other.

(2) Breakouts. Sometimes researchers will present data for some subset(s) of the participants from an aggregate group; e.g., for an aggregate group composed of males and females, the researchers may present some results for the males and females separately. You will code information about breakouts later.

Identifying the Aggregate Groups
Type in the name or identifier for each aggregate treatment group and each aggregate comparison group described in the study, whether you believe the group is eligible for coding or not.

Group labels used by researchers do not necessarily conform to the definitions of group types used in this project. In some cases, for example, researchers may compare one treatment with another treatment, and may call this “other” treatment a comparison or control group. For our purposes, if this “other” treatment group can realistically be expected to be effective, list it as a treatment group below; if it is a minimal or placebo treatment, not expected to produce an effect, list it as a comparison group.

Treatment Groups [H4a-d]
1 ______________________________
2 ______________________________
3 ______________________________
4 ______________________________

Comparison Groups [H5a-d]
1 ______________________________
2 ______________________________
3 ______________________________
4 ______________________________

[H4] Total number of treatment groups: ____
[H5] Total number of control groups: ____

ASSIGNMENT OF PARTICIPANTS

[H6] Unit of group assignment. The unit on which assignment to groups was based.

1. individual (i.e., some children assigned to treatment group, some to comparison group)
2. group (e.g., whole classrooms, schools, sites, facilities assigned to treatment and comparison groups)
3. program area, regions, school districts, counties, etc. (i.e., region assigned as an intact unit)
9. cannot tell

[H7] Method of group assignment. How participants/units were assigned to groups. This item focuses on the initial method of assignment to groups, regardless of subsequent degradations due to attrition, refusal, etc. prior to treatment onset. These latter problems are coded elsewhere.

**Random or quasi-random:**
1. randomly after matching, yoking, stratification, blocking, etc. The entire sample is matched or blocked first, then assigned to treatment and comparison groups within pairs or blocks. This does not refer to blocking after treatment for the data analysis.
2. randomly without matching, etc. This also includes cases when every other person goes to the control group.
3. regression discontinuity design: quantitative cutting point defines groups on some continuum (this is rare).
4. quasi-random procedure presumed to produce comparable groups (no obvious differences). This applies to groups which have individuals apparently randomly assigned by some naturally occurring process, e.g., next person to walk in the door. The key here is that the procedure used to select groups doesn’t involve individual characteristics of persons so that the groups generated should be essentially equivalent.

**Non-random, but matched or statistically controlled:** Matching refers to the process by which comparison groups are generated by identifying individuals or groups that are comparable to the treatment group using various characteristics of the treatment group. Statistical control refers to inclusion of the matching variable as a covariate in an ANCOVA or multiple regression analysis. Matching can be done individually, e.g., by selecting a control subject for each intervention subject who is the same age, gender, and so forth, or on a group basis, e.g., by selecting comparison schools that have the same demographic makeup and academic profile of treatment schools. Similarly, statistical control variables can be used at either the individual or school level.
5. matched or statistically controlled ONLY on pretest measures of some or all variables used later as outcome measures.
6. matched or statistically controlled on pretest measures AND other personal characteristics, such as demographics.
7. matched or statistically controlled ONLY on demographics: big sociological variables like age, sex, ethnicity, SES.

**Nonrandom, no matching prior to treatment but descriptive data, etc. regarding the nature of the group differences:**
8. Non-random, not matched, but pretreatment equivalence information is available.
99. cannot tell

[H8] Confidence in assignment ratings. Overall confidence of judgment on how participants were assigned
1. Very Low (Little Basis)
2. Low (Best Estimate)
3. Moderate (Weak Inference)
4. High (Strong Inference)
5. Very High (Explicitly Stated)

**Equivalence of the groups being compared**
At this point, you should go to the Effect Size Database to code group equivalence effect sizes and descriptive information about initial group differences for the study. See the Effect Size Coding Sheet section of this manual for more information on effect size calculation.


[H9] Number of variables on which treatment and comparison group differences were statistically compared prior to the intervention. A statistical comparison is one in which a statistical test was performed by the authors, whether they provide data or not (e.g., “no statistically significant differences were found”). Do not include here any comparisons on pretest variables, that is, measures of a dependent variable taken prior to treatment, e.g., prior number of absences when subsequent number of absences is used as an outcome measure.

[H10] Results of statistical comparisons.
   1. no comparisons made
   2. no statistically significant differences
   3. significant differences judged unimportant by coder. See note below regarding “importance” judgment.
   4. significant differences, judged of uncertain importance by coder
   5. significant differences, judged important by coder

[H11] Number of variables on which treatment and comparison group differences were or can be descriptively compared prior to the intervention. A descriptive comparison is any comparison across treatment and control groups that does not involve a statistical test (e.g., the actual number of males and females in each group or a statement by the author(s) about group similarity).

[H12] Results of descriptive comparisons.
   1. no comparisons made or available
   2. negligible differences, judged unimportant by coder. See note below regarding “importance” judgment.
   3. some differences, judged of uncertain importance by coder
   4. some differences, judged important by coder

Note: An “important” difference means a difference on several variables relevant to the outcome variables, or on a major variable, or large differences; major variables are those likely to be related to dropout, e.g., SES or poor academic performance.

[H13] Rating of similarity of treatment and control groups. Using all the available information, rate the overall similarity of the treatment group and the comparison group, prior to treatment, on factors likely to have to do with dropout or responsiveness to treatment (ignore differences on any irrelevant factors).

Note: Greatest equivalence from “clean randomization” with prior blocking on relevant characteristics and no subsequent attrition/degradation; least equivalence with some differential selection of one “type” of individual vs. another on some variable likely to be relevant to dropout.

Guidelines: Use ratings in the 1-3 range for good randomizations and matchings, e.g., 1=clean random, 2=nice matched. Use ratings in the 5-7 range for selection with no matching or randomization or instances where it has been seriously degraded, e.g., by attrition before posttest. Within this bracket, the question is whether the selection bias is pertinent to the outcomes being examined. Were participants selected explicitly or implicitly on a variable that might make a big difference in dropout? The middle three points are for sloppy matching designs, degradations, bad wait list designs, and the like. If the data indicate equivalence but the assignment procedure was not random give it a 4 or thereabouts since not all possible variables were measured for equivalence between groups.

1. Very similar, equivalent
2.
3.
4.
5.
6.
7. Very different, not equivalent
[H14] Overall confidence on rating of group similarity
1. Very Low (Little Basis)
2. Low (Best Estimate)
3. Moderate (Weak Inference)
4. High (Strong Inference)
5. Very High (Explicitly Stated)

[H15] Click here to record any problems you encountered while coding this section.

GROUP EQUIVALENCE EFFECT SIZE CODING

At this point, you should go to the Effect Size Database to code group equivalence effect sizes and descriptive information about initial group differences. See the Effect Size Coding section of this manual for more information on effect size calculation.

For each measure you can identify on which the treatment and control group were compared prior to treatment (other than dependent variables) or on which you can tell equivalence (e.g. if all males then code it here), determine which group is favored and if possible, calculate an effect size (ES, standardized difference between means or odds ratio). Do not include here any comparisons on pretest variables, that is, measures of a dependent variable taken prior to treatment. In such cases the pretreatment ES is coded later as pretest information, not here as group equivalence information.

The only eligible variables for group equivalence effect sizes are: (a) gender, (b) age, (c) race/ethnicity, and (d) variables relating to risk for school dropout. A pretest that is used later in the study as a posttest would not be coded here – you would code it as a pretest effect size. If the study reports group equivalence outcome data for multiple risk variables, group equivalence effect size information should be coded for up to four variables. If more than four variables are available for any of the risk factors, code the four most relevant ones. When deciding which are most relevant, use the following criteria:
1. First preference should be given to behavioral measures (e.g., prior absences, school performance).
2. Second preference should be given to measures of psychological conditions, predispositions, or attitudes (e.g., school engagement, school bonding, etc.).
3. Lowest preference should be given to broad measures of social disadvantage or family history (e.g., socioeconomic status of parents, residence in inner-city).

[StudyID] Indicate the Study ID for the study you are coding.
[ReportID] Enter the Report ID for the report in which you found the information on group equivalence. Use the complete Report ID, e.g. 1973.01.
[pagemum] Enter the page number on which you found the information on group equivalence.

[ES24] Type of effect size:
5. Group Equivalence (for baseline treatment-control comparisons on variables other than the dependent variables)

[ES19] Wave number. Pretests and group equivalence effect sizes always get a 1; each wave thereafter gets numbered consecutively, beginning with 1. Some studies involve more than one posttest measurement and we need to be able to distinguish one from another. Give the first posttest after treatment a 1, the second a 2, and so on.

[ES15] Variable on which comparison is made: ____________________________ (e.g., gender, age, etc.)

[ES17] Which group is favored? Whichever group has more of the characteristic that presumably makes them better off or more amenable to treatment (e.g., less truant, higher SES, smarter, etc.) is considered


39
favored. NOTE: You should code this item even for cases in which you are unable to calculate a numeric effect size but have information about which group is favored.

1. Treatment (fewer males, younger, fewer minorities, less risk)
2. Control (fewer males, younger, fewer minorities, less risk)
3. Neither, exactly equal
4. Cannot tell, no report

Data Fields: Fill in the data fields using the relevant statistical information provided in the report(s). You do not need to fill in all the fields; fill in only the information necessary to calculate an effect size. Thus, if the report provides sample sizes, means, standard deviations, and t-test scores, you need only enter the sample sizes, means, and standard deviations.

ONCE YOU HAVE FINISHED CODING THE GROUP EQUIVALENCE EFFECT SIZE INFORMATION, YOU SHOULD RETURN TO THE STUDY LEVEL FILE TO COMPLETE THE CODING OF FOR THAT SECTION.

TREATMENT AND CONTROL GROUPS CODING

Create one record in this database for each of the aggregate treatment and/or control groups that you selected earlier for coding. Studies with a treatment group and a control group will have two records, etc.

Group Identification and General Nature of Treatment

[StudyID] Type in the StudyID for the study you are coding if it does not appear automatically.
[GroupID] Number each group consecutively within a study, starting with 1.

[G1] Select the type of group you are coding.
1. Treatment group
2. Control group

[G2] What general type of “treatment” does this group receive?

Intervention Condition
1. Focal program or treatment. There may be several focal programs in a study, as when two different types of treatments, both of which could be expected to be effective, are compared.

Control Condition
2. “Straw man” alternate program or treatment, diluted version, less extensive program, etc., not expected to be effective but used as contrast for treatment group of primary interest. If the alternate treatment is not minimal and could realistically be expected to be effective, it is not a control condition and should be classified as a focal treatment instead.
3. Placebo (or attention) treatment. Group gets some attention or sham treatment (e.g., watching Wild Kingdom videos while treatment group gets therapy)
5. No treatment. Group gets no treatment at all. Note: The difference between “no treatment” and “treatment as usual” hinges on whether or not the treatment and control groups in this study have an institutional framework or experience in common.

[G3] Program name. Write in program or treatment label for this group (e.g., Dropout Prevention Curriculum, waiting list control, etc.). REMEMBER: YOU MUST CREATE A PROGRAM LABEL FOR CONTROL GROUPS AS WELL AS TREATMENT GROUPS.

[G4] Program description. Write in a brief description of the treatment this group receives. Please try to keep the description short by focusing on the key elements of treatment, but make sure you include ALL

treatment elements in your description. As much as possible, quote or give a close paraphrase of the relevant descriptive text in the study report. REMEMBER: YOU MUST CREATE A DESCRIPTION FOR CONTROL GROUPS AS WELL AS TREATMENT GROUPS.

TREATMENT CHARACTERISTICS

[G5] Intervention type:
1. School-based (administered under the auspices of school authorities and delivered during school hours)
2. School affiliated (delivered with the collaboration of school authorities, possibly by other agents, e.g., community service providers; may take place before or after school hours and/or off the school grounds)
3. Community-based (explicitly presented as dropout prevention/intervention programs; may or may not have a school affiliation)
4. Not applicable (control condition)


For each treatment AND control condition:
First check all program types that apply to a given intervention (e.g., a program may include GED preparation, tutoring, and contingency management).
Second, choose the one program type that can be considered the focal program characteristic. Most programs will arguably deliver multiple service types, but do your best to narrow the focal type down to one category. It may be helpful to examine the amount of each service type delivered. For instance, if a program delivered 1 hour/week of skills training to parents and 5 hours/week of vocational training to students, you would code vocational training as the focal program component. If a program contains too many service types to distinguish a focal type, choose “multi-service” package as the focal component.

ACADEMIC:
1. Curriculum
2. Academic program
3. Remedial education (e.g., reading remediation)
4. GED preparation
5. Computer-assisted learning
6. Test-taking and study skills assistance
7. Tutoring
8. Homework assistance
9. Extracurricular activities (e.g., after school club). NOTE: just because a program is delivered after school does not mean it should be coded here; this program component should include academic, social, or sport activities that are separate from regular school activities.
10. Professional development for school staff

SCHOOL STRUCTURE
11. Class or grade reorganization (schools within schools)
12. Small class sizes/small “learning communities”
13. Alternative school

FAMILY ENGAGEMENT:
14. Family outreach
15. Feedback to parents and students on performance
16. Parent or teacher consultation enhancement
17. Parenting skills program

COLLEGE FOCUSED/CONNECTING STUDENTS TO ATTAINABLE FUTURE:
18. Academic advising  
19. College-preparatory curriculum  
20. Academic summer program  
21. College campus visits  
22. College and financial aid application assistance  
23. College scholarships  

**WORK RELATED/ FINANCIAL SUPPORT:**  
24. Internships  
25. Career exploration  
26. Vocational training  
27. Job placement assistance  
28. Living allowance  
29. Bonuses and sanctions applied to welfare grant  

**LINKING TO SERVICES:**  
30. Case management  
31. Health services  
32. Transportation assistance  
33. Child care/day care  
34. Residential living services  

**SOCIAL RELATIONSHIPS:**  
35. Mentoring  
36. Peer support  
37. Social events  
38. Community service/volunteer service/tutoring (“helper-therapy”)  
39. Recreational, wilderness, etc. program  

**PERSONAL/AFFECTIVE:**  
40. Counseling  
41. Skills training (life skills, social skills/social competence)  
42. Cognitive behavioral therapy (e.g., problem solving skills)  

**BEHAVIORAL:**  
43. Attendance monitoring  
44. Contingency management, financial incentives, token economy, extrinsic reward system to promote attendance/academic achievement  

**OTHER:**  
45. Multi-service package *(NOTE: Only choose this program code if the group receives an amorphous, broadly defined program with components that cannot be clearly identified otherwise. Use this program code as focal if a group has multiple “focal” treatment components and you cannot make a distinction otherwise.)*  
46. OTHER *(Please, describe [prog50a]___________)*  
88. Control group  

**[G9] Treatment Site.** Nature of the site in which treatment generally delivered: (select one)  

**School Sites**  
1. Regular Class Time (this includes interventions delivered during regularly scheduled classes AND in the regular classroom for youths in the group)

2. Special Class (e.g., youth in treatment are in a classroom-type setting that is different from a typical classroom, although it may be the subjects’ usual classroom – includes such settings as special education classrooms, schools-within-schools, alternative schools, etc.)
3. Resource Room, School Counselor’s Office, or other similar setting that is NOT the student’s regular classroom; the idea here is that students are removed from class for treatment
4. Treatment delivered at school facility, but not during regular school hours (e.g., afterschool programs)

Home
9. Treatment delivered in the subject’s home

Community-based, Non-residential
10. Private office, clinic, center (e.g., YMCA, university, therapist’s office)
11. Public office, clinic, center (e.g., human services department, public health agency)
12. Work site (e.g., community service, trash collection on roadside, etc.)
13. Park, playground, wilderness area, etc.

Institutional, Residential
14. Private institution, residential
15. Public institution, residential (e.g., camp, reformatory)

Mixed or Multiple Sites
16. School and home
17. Other mixed, some combination of above sites (NOTE: if all sites are school based, use 16 above)
88 N/A: control group
99 Cannot tell

[G10] Role of the evaluator(s)/author(s)/research team or staff in the program. This item focuses on the role of the research team working on the evaluation, regardless of whether they are all listed as authors.
1. evaluator delivered therapy/treatment
2. evaluator involved in planning or controlling treatment or is designer of program
3. evaluator influential in service setting but no direct role in delivering, controlling, or supervision
4. evaluator independent of service setting and treatment; research role only
9. cannot tell

[G11] Role of program developer in the research project. This items focuses on the individual (or group of individuals) who created or developed the program and their role in the delivery of the program under study. Is the program developer the researcher conducting the study, or is the program developer not participating in the research project?
1. Program developer is author/evaluator/delivery agent
2. Delivery agent/author/evaluator modified existing program, but original program developer is not involved (note: this response suggests that the author/evaluator/delivery agent takes on a sort of quasi-developer status by modifying a program)
3. Program developer is not affiliated with research study and program is delivered as originally intended by developer

[G12] Routine practice or program vs. research project. Indicate the appropriate level for the treatment you are coding: at one end of the continuum are research projects (option 1), in which a researcher decides to implement and evaluate a particular program for research purposes; in many cases, the program may require the cooperation of a service agency (school, clinic, etc.), but the intervention is delivered primarily so the researcher can conduct research. At the other end of the continuum are evaluations of “real-world” or routine programs (option 3): a service agency implements a program on its own, and also decides to conduct an evaluation of the program; the evaluation may or may not be conducted by outside researchers. In the middle of the continuum are demonstration projects (option 2),
which are conducted primarily for research purposes, but generally have more elements of “real world” practice than typical research projects as defined under option 1. Demonstration projects generally involve a program that has been studied in prior research but is being tested for effectiveness in different settings than the original research, or on a larger scale than the original research.

If a researcher is a school principal or some other school staff person and is conducting the evaluation as part of his/her dissertation, the decision depends on the extent of the program. If the program is small-scale and implemented in, say, a classroom or two, and supervised by the researcher/principal, code it as a research project. If the program is a broader school-wide program that the researcher/principal happens to be evaluating, code it as either a demonstration or routine program, depending on whether the program is a special program being tested (demonstration) or something that the school does on a routine basis (routine practice).

1. research project: The intervention would not have been implemented without the interest or initiative of the researcher(s). The intervention is delivered by the research staff or by service providers (regular agency personnel, teachers, etc.) trained by the researchers.

2. demonstration project: A research project that involves a new or special program being tested, rather than a routine program. Although generally implemented by researchers for research purposes, a demonstration project has more elements of actual practice than a research project. Demonstration projects usually involve programs that have been studied previously, either in small-scale pilot projects or tightly controlled efficacy trials; demonstration projects would serve as a larger scale or quasi-real-world test of a promising program.

3. evaluation of a “real-world” or routine program: A service agency implemented the program using routine personnel and the typical clients for that program; there may be outside researchers who conduct the evaluation, but the program they are evaluating was already in place before the research began and is presumed to continue after the research has ended.

**[G13]** Treatment provider’s discipline. Indicate the discipline or type of treatment provider for the treatment. This item focuses on the individual(s) who have direct contact with the subjects in treatment, not necessarily the persons conducting the data analysis or evaluation.

*1. Teacher*
*2. School guidance counselor*
*3. School psychologist*
*4. School personnel, other than school counselor or teacher (e.g., principal, school nurse)*
*5. Counselor*
*6. Social worker*
*7. Researcher or researcher’s staff, graduate students*
*8. Other*
*88. N/A: no treatment received*
*99. Cannot tell*

**[G14]** Did treatment personnel receive special training in this specific program, intervention, or therapy? If the treatment is delivered by the researcher, use “yes” below, unless the report indicates otherwise.

*1. yes*
*2. no*
*9. cannot tell*

**[G15]** If yes, write in amount of training of personnel for providing this treatment: _______________

**[G16]** Treatment Format:  
*For each treatment AND control condition:*
First check all formats that apply to a given intervention (e.g., a program may include group and individual components, or have a family component).

Second, choose the **one** format type that can be considered the focal format. This selection should match the format of the focal program type you selected above under G6. If you selected multi-service package above, select the format for the most frequent or most focal piece of the package; if this is impossible, select multiple format program.

1. Subject alone (self-administered treatment)
2. Subject & provider, one-on-one
3. Subject group and provider, not classroom
4. Subject group and provider, classroom
5. Parents only and provider, child not present
6. Group of parents and provider, children not present
7. Child & parents with provider
8. Group of families with provider
9. Child & parents, no provider (self-administered treatment)
10. Teachers, treatment professional, no children
11. Parents alone (self-directed)
12. Multiple format program; no focal format
88. N/A: control group

**Focal Treatment Implementation/Length/Integrity**

[G20] **Duration of treatment.** Approximate (or exact) number of weeks that subjects received treatment, from first treatment event to last excluding follow-ups designated as such. Divide days by 7; multiply months by 4.3. Code 888 if a control group that receives nothing Code 999 if cannot tell. Estimate for this item if necessary, and if you can come up with a reasonable order of magnitude number.

[G22] Approximate (or exact) frequency of contact between subjects and provider or treatment activity. This refers only to the element of treatment that is different from what the control group receives.

1. less than weekly
2. Once a week
3. 2 times a week
4. 3-4 times a week
5. daily contact (not 24 hours of contact per day but some treatment during each day, perhaps excluding weekends)
6. continuous (e.g. residential living)
9. cannot tell
88. N/A: control group

[G24] ___________ Approximate (or exact) mean hours actual contact time between subject and provider or treatment activity per week if reported or calculable. Assume that high school classes, counseling, or therapy sessions are an hour unless otherwise specified. **Round to one decimal place.** Code 8888 for institutional, residential, or around the clock program; code 9999 if not available.

[G26] ___________ Approximate (or exact) mean number of hours total contact between subject and provider or treatment activity over full duration of treatment per subject if reported or calculable. **Round to whole number.** Code 8888 for institutional, residential, or around the clock program; code 9999 if not available.

[G29] **Monitored treatment implementation.** Was the implementation of the program monitored by the author/researcher or program personnel to assess whether it was delivered as intended?

1. Yes. Do not infer that monitoring happened. Select “yes” ONLY if the report specifically indicates that implementation was monitored.
0. No
9. Cannot Tell

[G30] Based on evidence or author acknowledgment, was there any uncontrolled variation or degradation in implementation or delivery of treatment, e.g., high dropouts, erratic attendance, treatment not delivered as intended, wide differences between settings or individual providers, etc.? Assume that there is no problem if one is not specified.

This question has to do with variation in treatment delivery, not research contact. That is, there is no “dropout” if all subjects complete treatment, even if some fail to complete the outcome measures.

1. yes (describe below)
2. possible (describe below)
3. no, apparently implemented as intended

[G31] Implementation Problems. Click to describe implementation problems, if any.

**Subject Characteristics**

[G40] Gender composition of group.

1. no males (<5%)
2. some males (<50%)
3. 50% to 60% male
4. mostly males (>60%)
5. all males (>95%)
6. cannot tell

[G42] Enter percent male: _________

**ETHNICITY CODING:**

[G43a] Percent white
[G43b] Percent black
[G43c] Percent Hispanic
[G43d] Percent other minority
[G43e] Percent non-white (ONLY use this category if specific minority groups are not mentioned; if you use this category, there should only be numbers in the white and non-white categories)

Rankings: 1=clear majority; 2=present but proportion unknown; 3=clear minority; 0=not present.

[G44a] White rank
[G44b] Black rank
[G44c] Hispanic rank
[G44d] Other minority rank
[G44e] Non-white rank (ONLY use this category if specific minority groups not mentioned; if you use this category, there should only be numbers in the white and non-white categories)

[G45] Describe others and/or non-whites:_____________________________________.

[G46] Enter the average age of the sample using number of years.

[G46a] and [G46b] High and low age using years.

[G47] Enter the average grade level of the sample. (dropdown menu)

[G47a] and [G47b] High and low grades (dropdown menu)

[G48] Predominant level of “risk” of youths in the sample: __________________________. Think of the reason that the subjects in this group ended up in this group; did the researchers select potential dropouts for treatment; if yes, how were the potential dropouts identified?

[G49] Socioeconomic status: Type in a brief description of the socioeconomic composition of the sample. This might include information on the percentage of children eligible for free lunches, the income level of the children’s parents, or a description of poverty in the community. Quote or closely paraphrase the relevant descriptive information in the report.

[G50] Please describe any problems you encountered while coding this record.

**DEPENDENT VARIABLES CODING**

Select the general construct group for the dependent variable you are coding, then select the specific construct category that best matches the dependent variable.

**[DV1] Construct Group**

- 100. Dropout
- 101. Attendance, truancy
- 102. Academic performance
- 103. School conduct
- 104. School engagement

**[DV2] Specific Construct**

**Dropout**

- 200. Dropout
- 201. Graduation
- 202. GED completion
- 203. Enrolled in post-secondary education

**Attendance**

- 204. Absences/truancy
- 205. Tardies
- 206. Attendance

**Academic performance**

- 207. GPA, grades
- 208. Standardized test scores
- 209. Academic track
- 210. Grade retention
- 211. Unstandardized, generic academic achievement score

**School conduct**

- 212. Suspensions
- 213. Expulsions
- 214. Detention
- 215. Classroom behavior

**School engagement**

- 216. School self-concept
- 217. Academic expectations/goal setting
- 218. Attitude toward school/school bonding
- 219. Attitude toward teachers

**[DV3] Source of information. Who provided the information for this dependent variable?**

1. Participants, self-report
2. Parents
3. Peers
4. Teachers
5. Principal
6. Service Provider (treatment agent)
7. School Records

8. Researcher or interviewer
9. Involved other (not treatment agent, not researcher), e.g., school counselor.
10. Multiple sources, cannot tell which is dominant
99. Cannot tell

**[DV4] Type of Measure.**
1. Survey, questionnaire, or interview
2. Standardized test (e.g., standardized achievement test)
3. School records
4. Other: __________
9. Cannot Tell

**BREAKOUT/SUBGROUP CODING**

Breakouts are comparisons involving subgroups of an aggregate treatment and/or control group. For example, the males in a treatment group might be compared with the males in a comparison group, or pretest-posttest results might be presented for males and females separately. Each variable (e.g., gender, age) by which the aggregate group(s) are subdivided constitutes one breakout, and each value of that variable defines one subgroup; i.e., a males vs. females stratification is one breakout (gender) with two subgroups, one male and one female. If only the male subgroup is reported, there is still one breakout, but only one subgroup.

Note that a simple report of the number of males and females in the treatment and control groups does not constitute a breakout (though it is relevant to group equivalence issues). To be a breakout, outcome data must be reported for the treatment-control or pretest-posttest comparison for at least one subgroup of the breakout variable. Breakouts are usually presented because the authors think that subgroups (e.g., males and females) are sufficiently different to warrant separate presentation of results (because, for example, males may be more likely to dropout than females).

**NOTE:** Only certain breakout variables are eligible for coding. These include gender, age, ethnicity, and prior school completion/dropout, GED completion, or absences/truancy. If you encounter another breakout variable that may be relevant to dropout, please check with Sandra. Create a new record for each subgroup that you will be coding for this study.

**[StudyID]** Study ID for the study you are coding.

**[BreakID]** Subgroup number. Assign a number to the subgroup such that the first subgroup you code is numbered 1, the second is numbered 2, and so on. These numbers are used within a study, so when you code subgroups from another study, you would start over with 1 again.

**[Labels:B2]** Write in descriptor for the subgroup you are coding, e.g., males, 8 year olds, whites, etc.

**EFFECT SIZE CODING**

**Step 1. General Information**

**[StudyID]** Type in the appropriate StudyID if it does not appear automatically.

**[ReportID]** Report ID for this effect size. Indicate the report number (e.g., 2098.01) for the report in which you found the information for this effect size. This is important so that we can find the source information for the effect sizes later on, if necessary, and is especially important for studies with multiple reports.

**[ESID]** Effect size ID. FileMaker will automatically generate unique effect size ID numbers ACROSS studies.

There are 3 types of effect sizes that can be coded: pretest, posttest, and group equivalence (or baseline similarity) effect sizes. They are defined as follows:

- **Pretest effect size.** This effect size measures the difference between a treatment and comparison group before treatment (or at the beginning of treatment) on the same variable used as an outcome measure, e.g., school attendance measured before the treatment begins is used as a pretest for school attendance measured the same way after the treatment ends.

- **Group equivalence effect size.** Group equivalence effect sizes are used to code the equivalence of two groups prior to treatment delivery on variables that might be related to outcome. See the Group Equivalence Coding section for more information.

- **Posttest effect size.** This effect size measures the difference between two groups after treatment on some outcome variable.

This is very important!!!! These three types of effect sizes are different from the multiple breakouts and multiple dependent variables that you might have in a study. For example, you might have a study that measures the treatment and comparison groups at pretest and posttest at 6 months after treatment on 3 different dependent variables. The results might be presented for the entire sample and broken down by gender. In this case you would have 6 group comparison effect sizes for the entire sample – three for the pretest and 3 for the 6 month posttest (the three is for your three dependent variables). In addition to these 6 aggregate effect sizes, you will have 6 more for the girls (the same as for the aggregate groups but just for the subgroup of girls) and 6 for the boys (also the same as for the aggregate groups but just for the subgroup of boys).

**[ES24]** Type of effect size:
1. Pretest (for treatment-control baseline comparison on a dependent variable)
2. Posttest (for the first treatment-control outcome comparison on a dependent variable)
5. Group Equivalence (for baseline treatment-control comparisons on variables other than the dependent variables)

**[ES19]** Wave number. Pretests and group equivalence effect sizes always get a 1; each wave thereafter gets numbered consecutively, beginning with 1. Some studies involve more than one posttest measurement and we need to be able to distinguish one from another. Give the first posttest after treatment a 1, the second a 2, and so on.

**[ES47]** Timing of measurement. Approximate (or exact) number of weeks after treatment when measure was taken. Divide days by 7; multiply months by 4.3. Enter 999 if cannot tell, but try to make an estimate if possible. Enter 0 if pretest. [es47_ck]

**Step 2. Group Selection**

- **[GroupID1]** Group 1: Treatment group
- **[GroupID2]** Group 2: Control group

**[BreakID]** Select Breakout group if relevant.

**Step 3. Dependent Variable Selection**

- **[VarNo]** Select the dependent variable for this effect size.

**Step 4. Effect Size Calculation and Data Entry**

It is now time to identify the data you will use to calculate the effect size and to calculate the effect size yourself if necessary (see below). Effect sizes can be calculated ONLY from data based on the number of subjects, e.g., average number of days absent per subject and the corresponding standard deviation) or proportion of subjects who were chronic truants during a given time period. Effect sizes can NOT be calculated from data based solely on the incidence of events, e.g., total number of days absent per group. THIS IS VERY IMPORTANT—BE SURE YOU KNOW WHICH KIND OF DATA YOU HAVE.

You need to determine what effect size format you will use for each effect size calculation. There are two general formats you can use, each with its own section in FileMaker:
1. Compute ES from means, sds, variances, test statistics, etc.
2. Compute ES from frequencies, proportions, contingency tables, odds, odds ratios, etc.

Also note that within each of the above effect size formats, effect sizes can be calculated from a variety of statistical estimates; to determine which data you should use for effect size calculation, please refer to the following guidelines in order of preference:
1. Compute ES from descriptive statistics if possible (means, sds, frequencies, proportions).
2. If adequate descriptive statistics are unavailable, compute ES from significant test statistics if possible (values of t, F, Chi square, etc.).
3. If significance tests statistics are unavailable or unusable but p value and degrees of freedom (df) are available, determine the corresponding value of the test statistic (e.g., t, chi-square) and compute ES as if that value had been reported.

Note that if the authors present both covariate adjusted and unadjusted means, you should use the covariate adjusted ones. If adjusted standard deviations are presented, however, they should not be used.

[ES17] Which group is favored?
Select the group that has done "better":
1. Treatment
2. Control
3. Neither, Exactly Equal
4. Cannot tell

For treatment-control comparisons, the treatment group is favored when it does "better" than the control group. The control group is favored when it does "better" than the treatment group.
Remember that you cannot rely on simple numerical values to determine which group is better off. For example, a researcher might assess the attendance and report this variable in terms of the average number of absences in the last semester. Fewer absences are better than more, so in this case a lower number, rather than a higher one, indicates a more favorable outcome.

Sometimes it may be difficult to tell which group is better off because a study uses multi-item measures in which it is unclear whether a high score or a low score is more favorable. In these situations, a thorough reading of the text from the results and discussion sections usually can bring to light the direction of effect—e.g., the authors will often state verbally which group did better on the measure you are coding, even when it is not clear in the data table. Note that if you cannot determine which group has done better, you will not be able to calculate a numeric effect size. (You will still be able to create an effect size record—just not a numeric effect size.)

[ES23] Effect size derived from what type of statistics?
1. Means and SDs; means and variances; means and standard errors
2. N successful/ unsuccessful (frequencies)
3. Proportion successful/unsuccessful (percentage successful or not)
4. Multi-category (polychotomous) frequency or %
5. Independent t-test
6. Probability (p-value) with N or degrees of freedom
7. One-way ANOVA (2 groups, 1 degree of freedom)
8. One-way ANOVA (>2 groups, >1 degree of freedom)

9. Factorial Design (Repeated measures ANOVA, 2x2 ANOVA, MANOVA, etc.)
10. Covariance Adjusted (ANCOVA)
11. Chi-square statistic (1 degree of freedom; from 2x2 table)
12. Chi-square (from larger than 2x2 table)
13. Nonparametric statistics (Mann Whitney, etc.)
14. Correlation coefficient (zero-order)
15. Effect sizes as reported directly in the study
16. Other (please specify)

[ES50] For this effect size, did you use adjusted data (e.g., covariate adjusted means) or unadjusted data? If both unadjusted and adjusted data are presented, you should use the adjusted data for the group means or mean difference, but use unadjusted standard deviations or variances. Adjusted data are most frequently presented as part of an analysis of covariance (ANCOVA). The covariate is often either the pretest or some personal characteristic such as socioeconomic status. If you encounter data that is adjusted using something other than a covariate, please see Sandra or Mark.

1. Unadjusted data
2. Pretest adjusted data (or other baseline measure of an outcome variable construct)
3. Data adjusted on some variable other than the pretest (e.g., socioeconomic status)
4. Data adjusted on pretest plus some other variables

[ES22] Confidence in effect size calculation
1. High Estimate (e.g., have N and crude p values only, e.g., p<.10, and must reconstruct via rough t-test equivalence)
2. Moderately Estimated (e.g., have complex but relatively complete statistics, e.g., multiple regression, multifactor ANOVA, etc. as basis for estimation)
3. Some Estimation (e.g., have unconventional statistics and must convert to equivalent t-values or have conventional statistics but incomplete, such as exact p values only)
4. Slight Estimation (e.g., must use significance testing statistics rather than descriptive statistics, but have complete statistics of the conventional sort, such as a t-value or F-value)
5. No Estimate (e.g., have descriptive data: means, sds, frequencies, proportions, etc.; can calculate an ES directly.)

[ES44] Significance information for this comparison.
Did the authors provide any information about the statistical significance of the difference between the two groups you selected on the dependent variable you selected for the time point you have selected for this comparison? Sometimes authors will state that a particular comparison was not significant, but not provide any calculable effect size data. In these cases, you should select “5” for this item. The effect size field should remain blank. In other cases, authors will state that a particular comparison was significant, but not provide any calculable effect size data. In these cases, you should select “4” for this item. Again, the effect size field should remain blank. NOTE: the last three options (4, 5, and 6) are for cases for which you have direction (i.e., you know which group is favored) but no effect size information.

1. Significant result, ES data below
2. Non-significant result, ES data below
3. Significance not reported, ES data below
4. Significant result, no ES data
5. Non-significant result, no ES data
6. Significance not reported, no ES data

[ES55] Intent-to-treat analysis: Are results for this effect size based on an intent-to-treat analysis? Experimental and quasi-experimental designs may employ “intent-to-treat” (ITT) or “completer” analyses. An intent-to-treat analysis is one that (attempts to) include outcome data from all the participants initially assigned to the treatment and comparison conditions regardless of their compliance with the entry criteria, the treatment they actually received, or any subsequent withdrawal from treatment (non-completers) or deviation from the protocol. A true ITT is possible only when the authors (attempt to) use outcome data for all randomized (or otherwise assigned) subjects; if all assigned subjects are used to...
present outcome results, then code as ITT, regardless of whether authors call the analysis an ITT. If the authors attempt to collect outcome data on non-completers and even if they are not 100% successful in this attempt, still code as ITT (as the missing data for non-completers is then coded as attrition). Sometimes researchers will use a modified ITT, in which they estimate missing data on non-completers, or include all subjects with pretests but not all who were randomized. These modified ITTs would be coded as “2” below. Completer analyses (AKA ‘treatment on the treated (TOT)’ analyses) involve only the participants who completed treatment or met some other criteria indicating an acceptable level of participation.

1. Intent-to-treat analysis (all subjects who were assigned are used in posttest)
2. Modified intent-to-treat (not all assigned subjects are used in posttest, but authors have done some modifications to approximate a true ITT)
3. Completer analysis (only those subjects who completed treatment or who stayed in the study are used in posttest)

[ES15] Significance of group equivalence comparison (ONLY).
   1. No statistically significant differences
   2. Statistically significance differences
   3. Negligible descriptive differences
   4. Significant descriptive differences
   98. N/A: No comparison made

Assigned and Observed N

Assigned N, Observed N. These fields refer to the number of subjects who were originally assigned to the group(s) involved in this effect size (Assigned N) and to the number of subjects who were actually “observed” or “measured” (Observed N). If you cannot tell how many subjects were originally assigned to a group, look at the number of subjects (Observed N) at pretest; you can frequently use pretest sample sizes for assigned N. However, in cases where the authors have removed the subjects who do not have both pretest and posttest measures (such that the pretest N and the posttest N are the same), do not assume that the number of subjects at pretest is the correct number for Assigned N and, instead, leave this field blank. In cases where there is no attrition, the Assigned N is the same as the Observed N. Only use the same numbers for Assigned N and Observed N when you are SURE that there is no attrition.

[ES36] Assigned N for the treatment group (or pretest, if this is a pretest-posttest effect size).
[ES37] Assigned N for the comparison or second treatment group (or posttest, if this is a pretest-posttest effect size; if this is a pretest-posttest effect size, this value should be the same as the assigned N for the pretest).
[ES38] Total Assigned N.
[ES1] Observed N for the treatment group (or pretest, if this is a pretest-posttest effect size).
[ES2] Observed N for the comparison or second treatment group (or posttest, if this is a pretest-posttest effect size).
[ES3] Total Observed N.

[ES51] Number of units assigned for treatment group (for cluster-assigned studies): ____
[ES52] Number of units assigned for control group (for cluster-assigned studies): ____
[ES53] Intra-class correlation (ICC) for outcome measure (for cluster-assigned studies): ____

Other Effect Size Data Fields

[ES9] Mean for treatment group
[ES10] Mean for comparison group
[ES11] Difference in group means
[ES12] Standard deviation for treatment group
[ES13] Standard deviation for comparison group
[ES14] Pooled sd
[ES31] N successful for treatment group

[ES32] N successful for comparison group
[ES33] N failed for treatment group
[ES34] N failed for comparison group
[ES4] Dependent t-value
[ES5] Independent t-value
[ES6] \( \chi^2 \) (df=1)
[ES20] Effect size reported by authors
[ES60] Odds ratio reported by authors

Final Effect Size Determination
[ES21] Effect size value- standardized mean difference
[ES81] Effect size value- odds ratio

Remember that you cannot rely on simple numerical values to determine which group has done better. For treatment-control comparisons, a positive effect size should indicate that the treatment group did “better” on the outcome measure than the comparison group, while a negative effect size indicates that the comparison group did “better” than the treatment group, and a zero effect size means that the two groups are exactly equal on the measure. You must make sure that the sign of the effect size matches the way we think about direction, such that the effect size is positive when the treatment group is better and negative when the comparison group is better.

Effect sizes can range anywhere from around –3 to +3. However, you will most commonly see effect sizes in the –1 to +1 range. Odds ratios smaller than 1 indicate that the control group is better off; those greater than 1 indicate that the treatment group has the better outcome.

Note: If the authors report an effect size, include that in your coding and use it for the final effect size value if no other information is reported. However, if the authors also include enough information to calculate the effect size, always calculate your own and report it in addition to that reported in the study.

[ES39] Any problems coding this effect size?
Appendix 5: Common Effect Size Metrics and their Standard Errors

Effect Sizes for Comparing Continuous Outcomes Between Two Groups

Standardized mean difference effect size (Cohen's d)

\[ d = \frac{\overline{X}_{G1} - \overline{X}_{G2}}{\sqrt{s^2_{G1}(n_{G1} - 1) + s^2_{G2}(n_{G2} - 1)}} \]

\[ SE_d = \sqrt{n_{G1} + n_{G2} + \frac{d^2}{2(n_{G1} + n_{G2})}} \]

Small-sample corrected standardized mean difference effect size (Hedges’ g)

\[ g = \left[ 1 - \left( \frac{3}{4N - 9} \right) \right] \times d \]

\[ SE_g = \sqrt{n_{G1} + n_{G2} + \frac{g^2}{2(n_{G1} + n_{G2})}} \]

Key Terms

- \( \overline{X}_{G1} \): Mean for intervention group
- \( \overline{X}_{G2} \): Mean for comparison group
- \( s^2_{G1} \): Variance for intervention group
- \( s^2_{G2} \): Variance for comparison group
- \( n_{G1} \): Sample size for intervention group
- \( n_{G2} \): Sample size for comparison group
- \( N \): Total sample size across groups

Definitions

Effect size: standardized measure used to index the magnitude and direction of a quantity or relationship of interest. Used as the dependent (outcome) variable in a meta-analysis.

Standard error: each effect size in a meta-analysis has a corresponding standard error. The standard error is an index of the precision with which an effect size is estimated and is used to compute confidence intervals around an effect size.

Sampling variance: The square of the standard error. The inverse of the sampling variance is used to weight effect sizes in a meta-analysis, such that each effect size is weighted by an estimate of its precision.
Effect Sizes for Comparing Binary Outcomes Between Two Groups

Log odds ratio effect size (LOR)

\[ LOR = \ln \left( \frac{A \times D}{B \times C} \right) \]

\[ SE_{LOR} = \sqrt{\frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}} \]

Log relative risk/risk ratio effect size (LRR)

\[ LRR = \ln \left( \frac{A/n_{G1}}{C/n_{G2}} \right) \]

\[ SE_{LOR} = \sqrt{\frac{1}{A \times n_{G1}^{-1}} + \frac{1}{C \times n_{G2}^{-1}}} \]

Key Terms

- **A**: Number of events in intervention group
- **B**: Number of non-events in intervention group
- **C**: Number of events in comparison group
- **D**: Number of non-events in comparison group
- \( n_{G1} \): Sample size for intervention group
- \( n_{G2} \): Sample size for comparison group
Effect Sizes for Examining Correlation Between Two Continuous Measures

Correlation coefficient effect size \( (r) \)

\[
\begin{align*}
\rho &= \frac{\sigma_{xy}^2}{\sigma_x \sigma_y} \\
SE_r &= \sqrt{\frac{1 - r^2}{n - 2}}
\end{align*}
\]

Fisher's Z-transformed correlation effect size \( (z) \)

\[
\begin{align*}
z &= 0.5 \ln \left( \frac{1 + r}{1 - r} \right) \\
SE_r &= \sqrt{\frac{1}{n - 3}}
\end{align*}
\]

Key Terms

- \( \sigma_{xy}^2 \): Covariance between two variables
- \( \sigma_x \): Standard deviation of predictor variable
- \( \sigma_y \): Standard deviation of outcome variable
- \( n \): Sample size
Appendix 6: Common Meta-Analysis Equations

Estimating a Mean Effect Size Across Studies

$$\bar{ES} = \frac{\sum (w \times ES)}{\sum w}$$

$$SE_{\bar{ES}} = \sqrt{\frac{1}{\sum w}}$$

Key Terms

ES: Effect size (e.g.: d, g, LOR, LRR)

w: Inverse variance weight (fixed or random)

Fixed Effect Inverse Variance Weight

$$w_{FE} = \frac{1}{V_{ES}}$$

Random Effects Inverse Variance Weight

$$w_{RE} = \frac{1}{V_{ES} + \tau^2}$$
Estimating Heterogeneity

\[ \tau^2 = \frac{Q - df}{\sum w - \sum w^2} \]

\[ Q = \sum (w \times ES^2) - \frac{\left( \sum (w \times ES) \right)^2}{\sum w} \]

\[ df = k - 1 \]

\[ I^2 = \frac{(Q - df)}{Q} \times 100\% \]

**Key Terms**

- ES: Effect size (e.g.: d, g, LOR, LRR)
- w: Fixed effect inverse variance weight
- k: Number of studies