Introductory:

Coding

Sandra Jo Wilson
Editor, Education Coordinating Group
Associate Director, Peabody Research Institute
Research Assistant Professor, Dept. of Special Education
Vanderbilt University, USA

Workshop Overview

- Levels of Study Coding
  - Study eligibility screening
    - Abstract level
    - Full-text level
  - Study content coding
  - Effect size coding (tomorrow at 10:30am)
- Common mistakes
Why do study coding?

- Provide an accounting of the research included in your review.
  - Also helps identify what’s missing.
- Identify the characteristics of the interventions, subjects, and methods in the research.
- Assuming different primary studies produce different results, study coding allows you to identify variables that might explain those differences.

Describing the literature (example 1)

Table 1: Characteristics of Included Studies

<table>
<thead>
<tr>
<th>Study Characteristic</th>
<th>k</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Publication Year</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1980s</td>
<td>5</td>
<td>16.7</td>
</tr>
<tr>
<td>1990s</td>
<td>10</td>
<td>33.3</td>
</tr>
<tr>
<td>2000s</td>
<td>15</td>
<td>50.0</td>
</tr>
<tr>
<td>Method of Assignment</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random (indiv.)</td>
<td>2</td>
<td>6.7</td>
</tr>
<tr>
<td>Random (cluster)</td>
<td>8</td>
<td>26.7</td>
</tr>
<tr>
<td>Non-random</td>
<td>20</td>
<td>66.7</td>
</tr>
<tr>
<td>Duration of treatment (weeks)</td>
<td>23.5</td>
<td>10.2</td>
</tr>
<tr>
<td>Age of intervention participants</td>
<td>12.3</td>
<td>8-13</td>
</tr>
</tbody>
</table>
Describing the literature (example 2)

Table 1: Characteristics of Included Studies

<table>
<thead>
<tr>
<th>Study</th>
<th>Intervention</th>
<th>Personnel</th>
<th>Design</th>
<th>Outcomes</th>
<th>% Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wilson, 2013</td>
<td>CBT</td>
<td>Teachers</td>
<td>RCT</td>
<td>Aggression</td>
<td>56%</td>
</tr>
<tr>
<td>Lipsey, 2009</td>
<td>Behavioral</td>
<td>Psychologists</td>
<td>QED</td>
<td>Suspensions Fighting</td>
<td>75%</td>
</tr>
<tr>
<td>Tanner-Smith, 1999</td>
<td>CBT</td>
<td>Psychologists</td>
<td>RCT</td>
<td>Conduct disorder</td>
<td>89%</td>
</tr>
<tr>
<td>Allen, 1978</td>
<td>Anger Mgmt.</td>
<td>Teachers</td>
<td>Cluster RCT</td>
<td>Conduct disorder</td>
<td>45%</td>
</tr>
<tr>
<td>Jones, 2000</td>
<td>Anger Mgmt.</td>
<td>Teachers</td>
<td>QED</td>
<td>Conduct disorder</td>
<td>46%</td>
</tr>
<tr>
<td>Site A</td>
<td>Anger Mgmt.</td>
<td>Teachers</td>
<td>QED</td>
<td>Conduct disorder</td>
<td>82%</td>
</tr>
<tr>
<td>Site B</td>
<td>Anger Mgmt.</td>
<td>Teachers</td>
<td>QED</td>
<td>Conduct disorder</td>
<td>100%</td>
</tr>
</tbody>
</table>

Study Characteristics as Moderators

<table>
<thead>
<tr>
<th>Meta-regression Model</th>
<th>b</th>
<th>se</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methodological Characteristics</td>
<td>Random assignment (1=yes)</td>
<td>-.03</td>
<td>.15</td>
</tr>
<tr>
<td>Matched groups design (1=yes)</td>
<td>-.01</td>
<td>.15</td>
<td>-31, .29</td>
</tr>
<tr>
<td>Participant Characteristics</td>
<td>Percent male</td>
<td>.07</td>
<td>.20</td>
</tr>
<tr>
<td>Program Characteristics</td>
<td>Implementation quality (1-3)</td>
<td>.30*</td>
<td>.06</td>
</tr>
<tr>
<td></td>
<td>Classroom program (1=yes)</td>
<td>.45*</td>
<td>.12</td>
</tr>
</tbody>
</table>

ANOVA Model

<table>
<thead>
<tr>
<th>k</th>
<th>mean</th>
<th>ES</th>
<th>se</th>
<th>CI</th>
<th>Q_{F,0}</th>
</tr>
</thead>
<tbody>
<tr>
<td>RCTs</td>
<td>20</td>
<td>.32</td>
<td>.21</td>
<td>.03-.67</td>
<td>1.6*</td>
</tr>
<tr>
<td>cluster</td>
<td>13</td>
<td>.42</td>
<td>.10</td>
<td>.20-.30</td>
<td>2.5*</td>
</tr>
<tr>
<td>QEDs</td>
<td>10</td>
<td>.75</td>
<td>.32</td>
<td>.40-.85</td>
<td>2.4*</td>
</tr>
</tbody>
</table>
Eligibility Criteria

- The PICOS framework:
  - Population/Participants (problems/conditions)
  - Interventions
  - Comparison group (e.g., absolute vs. relative effects, counterfactual conditions)
  - Outcomes (primary and secondary outcomes, acceptable outcome measures)
  - Study Design
  - Geographic area, time, language, other criteria

Abstract Level Relevance Screening

- Does the document look like it might be relevant?
  - Based on reading titles and abstracts
    - If yes, retrieve full text of article
    - Exclude obviously irrelevant studies, but don’t assume the title and abstract are going to give reliable information on study design, outcomes, subject characteristics, or even interventions
  - When in doubt, double code
    - At least two trained raters working independently
Abstract Screening Example

**Full-Text Eligibility Screening**

- Develop an eligibility screening form with criteria.
- Before screening, link together all reports from the same study.
- Complete form for all studies retrieved as potentially eligible.
- Modify criteria after examining sample of studies (controversial).
- Double-code eligibility.
- Maintain database on results for each study screened.
Eligibility Database Example

<table>
<thead>
<tr>
<th>StudyID</th>
<th>Surname</th>
<th>Date Eligibility Determined</th>
<th>Eligibility</th>
<th>Status (from SIs)</th>
<th>Status</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>2319</td>
<td>Surname</td>
<td>11/10/2010</td>
<td>No</td>
<td>Not English</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**References**


**Review Articles**

- No
- Yes

**Eligibility Criteria**

**Outcome**

- Developing Country Study:

**Screening**

**Eligibility Criteria**

- Interventions
  - No
  - Yes
  - Developed, school-affiliated, or community-based intervention?

**Subjects**

- 1a.
  - Study description only (not N/S)
  - Yes
  - No

**Research Design**

- Yes
- No
  - Study has a control group?
  - Yes
  - No

**Outcome Variables**

- Yes
- No
  - Includes at least one outcome variable? (CLICK for details)

**Date of Publication**

- Yes
- No
  - Published prior to 1985, or research conducted prior to 1985?

**Effect Sizes**

- Yes
- No
  - Includes sufficient data to calculate effect sizes?

Eligibility Screening Results

- Once screening of all relevant full-text reports is complete, you will have:
  - An accounting of the ineligible studies and the reasons for their ineligibility.
    - Campbell and Cochrane reviews often include a table of ineligible studies as an appendix.
  - A set of studies eligible for coding.
### Sample Exclusion Table

<table>
<thead>
<tr>
<th>Inclusion Criteria</th>
<th>Number Excluded</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intervention A school-based or affiliated psychological, educational, or behavioral prevention/intervention program that involves actions performed with the expectation that they will have beneficial effects on student recipients.</td>
<td>317</td>
</tr>
<tr>
<td>Population Intervention directed toward school-aged youth (ages 4-18) or recent dropouts between the ages of 18-22 for programs explicitly oriented toward secondary school completion or the equivalent.</td>
<td>2</td>
</tr>
<tr>
<td>Population General population samples of school-age children. Samples consisting exclusively of specialized populations, such as students with mental disabilities or other special needs, are not eligible.</td>
<td>9</td>
</tr>
</tbody>
</table>

### Sample Exclusion Table

<table>
<thead>
<tr>
<th>Citation</th>
<th>Reason Excluded</th>
</tr>
</thead>
</table>
Study Content and Effect Size Coding

• Develop a coding manual that includes:
  – Setting, study context, authors, publication date & type
  – Methods and method quality
  – Program/intervention
  – Participants/clients/sample
  – Outcomes
  – Findings, effect sizes
• Coding manual should have a “paper” version and an electronic version.

Sources for Coding Items

• Other systematic reviews
  – Campbell reviews often include coding manuals as appendices in protocols and reviews.
  – Useful for generic coding items (e.g., research design, risk of bias)
• The literature
  – Start with the assumption that you will find variability in treatment effects across the studies in your review.
  – What does the literature tell you about the plausible sources of that variability?
    • The literature reviewed at the beginning of a primary study often provides clues.
    • Theory of change for the intervention can also be useful.
Study Context Coding

- Setting, study context, authors, publication date and type, etc.
  - Multiple publications; “study” vs. “report”
  - Geographical/national setting; language
  - Publication type and publication bias
  - Publication date vs. study date
  - Research, demonstration, practice studies

Publication Type

[SH20] Type of publication. If you are using more than one type of publication to code your study, choose the publication that supplies the effect size data.

1. Book
2. Journal article
4. Thesis or dissertation
5. Technical report
6. Conference paper
7. Other
8. Cannot tell
Study Method Coding

- Session on Risk of Bias (Thursday at 10:45)
- Variety of options for coding study methods
  - Cochrane risk of bias framework
  - GRADE system
  - Method quality checklists
  - Direct coding of methodological characteristics

Cochrane Risk of Bias Framework

- Focus on identifying potential sources of bias:
  - Selection bias - Systematic differences between groups at baseline (allocation concealment and sequence generation)
  - Performance bias - Something other than the intervention affects groups differently, e.g., contamination (blinding of participants)
  - Attrition bias - Participant loss affects initial group comparability
  - Detection bias - Method of outcome assessment affects group comparisons (blinding of data collectors)
  - Reporting bias - Selective reporting of outcomes
- Coded as low risk, high risk, unclear risk
Modifications of Cochrane Framework

• Quasi-experimental studies may need additional coding with regard to selection bias, e.g., potential third variables that might account for results when groups were not randomized.

• Some risk of bias items don’t make sense in social science research
  – Blinding of study participants to what condition they’re in
    • Code performance bias directly, e.g., was contamination evident? Were there refusals?
  – Blinding of data collectors

Method Quality Checklists

• Method quality ratings (or not)
• More than 200 scales and checklists available, few are appropriate for systematic reviews
• Overall study quality scores have questionable reliability and validity (Jüni et al., 2001)
  – Conflate different methodological issues with other study features, which may have different impacts on reliability/validity
  – Preferable to examine potential influence of key components of methodological quality individually
• Weighting results by study quality scores is not advised!
Direct Method Coding

• Methods: Basic research design
  – Nature of assignment to conditions
  – Attrition, crossovers, dropouts, other changes to assignment
  – Nature of control condition
  – Multiple intervention and/or control groups

• Design quality dimensions
  – Initial & final comparability of groups (pretest effect sizes)
  – Treatment-control contrast
    • Treatment contamination
    • Blinding

Direct Method Coding

• Methods: Other aspects
  – Issues depend on specific research area
  – Procedural, e.g.,
    • monitoring of implementation, fidelity
    • credentials, training of data collectors
  – Statistical, e.g.,
    • statistical controls for group differences
    • handling of missing data
Sample Coding Item

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 (i.e., whole classrooms, schools, therapy groups, sites, residential facilities assigned to treatment and comparison groups)
3. program area, regions, school districts, counties, etc. (i.e., region assigned as an intact unit)
4. cannot tell

Sample Coding Item

[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, which are coded elsewhere.

Random or near-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. wait list control or other 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., first 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:
1. matched ONLY on pretest measures of some or all variables used later as outcome measures.
2. matched on demographic characteristics.

etc.
GRADE System for Method Quality

- Quality of evidence across studies included in a systematic review
- Outcome-specific
- Considers: sparse data, consistency/inconsistency of results across trials, study designs, reporting bias, possible influence of confounding variables
- Software available at: [www.ims.cochrane.org/revman/gradepro](http://www.ims.cochrane.org/revman/gradepro)
- Also see: [www.gradeworkinggroup.org](http://www.gradeworkinggroup.org)

What to do with method quality ratings?

- Select a method of assessing and coding methodological quality or design your own items.
- Code all the studies with whatever method you select.
- Examine the influence of the methodological characteristics on effect sizes at the analysis stage.
  - Use the method variables as moderators.
- Summarize the quality of evidence (e.g., GRADE).
- The important thing is to USE the information about quality to inform your conclusions!
Intervention Coding

- General program type (mutually exclusive or overlapping?)
- Specific program elements (present/absent)
- Any treatment received by the comparison group
- Treatment implementation issues
  - Integrity, implementation fidelity
  - Amount, length, frequency, “dose”
- Goal is to differentiate across studies
- What does the literature tell you about what’s likely to be important?

Intervention Coding Example

<table>
<thead>
<tr>
<th>Group Identification</th>
<th>TX Characteristics</th>
<th>TX Dimensions</th>
<th>Dosage</th>
<th>Subject Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study ID: 2011</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group ID: 39</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Code: 452</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **ACADEMIC:**
  - Curriculum
  - EL/ESL program
  - Remedial education
- **GED preparation**
- **Computer-assisted learning**
- **Test-taking, study skills**
- **Tutoring**
- **Homework assistance**
- **Extracurricular**
- **PD for school staff**
- **Individualized teaching**

- **SCHOOL STRUCTURE:**
  - Class size
  - Alternative school

- **Focal modality:**
  - 11 Class or grade reorganization schools within 11

- **GROUP NAME:**
  - Cohort 3 Computech Peninsular Academy (entering fall 1987)

Coding problems? 1. Dosage is based on the number of instructional
Participant Coding

- Participants/clients/sample
  - Data are at aggregate level (you’re coding the characteristics of the entire sample, not a single individual)
  - Mean age, age range
  - Gender mix
  - Racial/ethnic mix
  - Risk, severity
  - Restrictiveness; special groups (e.g., clinical)

Study Outcome Coding

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

- Hierarchical structure of meta-analytic data
  - Multiple outcomes within studies
  - Multiple measurement points within outcomes
  - Multiple subsamples within studies
  - Multiple effect sizes within studies

- You need to design coding scheme and eventual analytic datasets in a way that handles the hierarchical nature of your data.

Coding Studies: Relational Files

Multiple rows per study

<table>
<thead>
<tr>
<th>Study Level File</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Study ID</td>
<td>PubYear</td>
<td>MeanAge</td>
<td></td>
</tr>
<tr>
<td>175</td>
<td>2001</td>
<td>13.4</td>
<td></td>
</tr>
<tr>
<td>188</td>
<td>2006</td>
<td>15.9</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Effect Size Level File</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Study ID</td>
<td>DV Num</td>
<td>ES Num</td>
<td>Time</td>
<td>ES</td>
</tr>
<tr>
<td>175</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1.61</td>
</tr>
<tr>
<td>175</td>
<td>2</td>
<td>2</td>
<td>6</td>
<td>-0.99</td>
</tr>
<tr>
<td>175</td>
<td>5</td>
<td>3</td>
<td>9</td>
<td>0.96</td>
</tr>
<tr>
<td>175</td>
<td>5</td>
<td>4</td>
<td>12</td>
<td>0.12</td>
</tr>
<tr>
<td>175</td>
<td>11</td>
<td>5</td>
<td>18</td>
<td>0.54</td>
</tr>
<tr>
<td>188</td>
<td>5</td>
<td>6</td>
<td>1</td>
<td>-0.06</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Outcome Level File</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Study ID</td>
<td>DV Num</td>
<td>Construct</td>
<td></td>
</tr>
<tr>
<td>175</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>175</td>
<td>5</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>175</td>
<td>11</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>188</td>
<td>5</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>188</td>
<td>7</td>
<td>11</td>
<td></td>
</tr>
</tbody>
</table>
Procedural Details

• Double coding:
  • Cohen’s kappa
  • Agreement on key decisions
    – Study inclusion/exclusion, key characteristics, risk of bias, coding of results
  • Pilot-test and refine coding manual

Creating your coding manual

• Write up a detailed document that includes all the coding items.
  – Organize the manual hierarchically
    • One section for study-level, one for group-level, one for dependent variables, one for effect sizes
  – Include coding manual in Campbell Collaboration protocol
• Translate the coding manual into “forms” to use for actual coding
  – Paper forms, spreadsheets
  – Databases
### Sample Coding Database

<table>
<thead>
<tr>
<th>Treatment Info</th>
<th>Focal Treatment</th>
<th>Tx Characteristics</th>
<th>Tx Characteristics Recodes</th>
<th>Implementation</th>
<th>Subjects</th>
<th>Subjects (cont.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study ID</td>
<td>3792</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group ID</td>
<td>1</td>
<td>Number groups consecutively within a study.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Duration of treatment in WEEKS**
  - Determined by (select one)
  - 12.9

- **Frequency of treatment event/contact**
  - Determined by (select one)
  - 2
  - Format
  - 2 Format

- **Frequency of treatment RECODE**
  - 4
  - Continuous
  - 1 Continuous

- **Mean HOURS of contact per WEEK**
  - Determined by (select one)
  - 1
  - 3-4x per week
  - 1 3-4x per week

- **Minutes per treatment session**
  - Approx. mean HOURS of TOTAL contact
  - 12.9
  - 9 Cannot Tell
  - 2 Format

- **Confidence in estimates of duration, frequency,**
  - Monitored treatment implementation?
  - 5 Very High (explicitly stated)

- **Evidence of implementation problems?**
  - 3 No

- **Describe implementation problems:**
  - 3 No

---

### Sample Coding Database

<table>
<thead>
<tr>
<th>General</th>
<th>Group Selection</th>
<th>DV Selection</th>
<th>Breakdown</th>
<th>ES Calc</th>
<th>group Equivalence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confirm DV and Groups Information</td>
<td>Study ID</td>
<td>ESID</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dependent Variable</td>
<td>Behavioral Conduct Subscale (Self-Perception Profile for Children) measures</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type of Effect Size</td>
<td>2 Pre-Test Followup Number</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Conflict Resolution Skills Training</td>
<td>Traditional School Services (Control)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Groups</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Breakdown</td>
<td>Assigned</td>
<td></td>
<td>Assigned</td>
<td>Reviewed</td>
<td>Mean</td>
</tr>
<tr>
<td>Tx Group or Pretreatment</td>
<td>49</td>
<td>49</td>
<td>2.7312</td>
<td>2.7410</td>
<td>7.7064</td>
</tr>
<tr>
<td>Ct Group or Posttreatment</td>
<td>49</td>
<td>49</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total/Difference</td>
<td>98</td>
<td>98</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Which Group is Favored? | 2 Control (or Pretreat) | F-value (df=1): |
| What kind of data used to calculate ES? | 4 Means and SDs or variances or |
| Did you use adjusted data? | 2 Adjusted data |
| Confidence in ES calculation | 5 No Estimate |
| Significance for this comparison | 2 Non-significant result, ES data |
| Variance control techniques | 4 Variance control techniques used |
| Pre-Posttest Correlation | P-value (df=1): |

| Effect Size (from Toolkit) | 0.0133 |
| Effect Size (from FileMaker) | 0.0133 |
Common Mistakes

- Too many coding items (my biggest problem!)
- Too many subjective coding items
- Coding two reports from the same study as two different studies
- Coder drift
- Failure to ask questions

Final Remarks

- Remember the purposes of coding:
  - To describe the methods, interventions, participants, outcomes, and findings common in the literature you’re reviewing.
  - To identify gaps in the literature.
    - What’s NOT covered
    - What’s NOT reported
  - To explain observed variation in treatment effects.
Thank you!
Any Questions?

Sandra Jo Wilson
email: sandra.j.wilson@vanderbilt.edu