The Campbell Collaboration

Effect Size Calculations and Elementary Meta-Analysis

David B. Wilson, PhD
George Mason University
August 2011

The End-Game

- Forest-Plot of Odds-Ratios and 95% Confidence Intervals for the Effects of Cognitive-Behavioral Programs on Recidivism

Porporino & Robinson, 1995
Johnson & Hunter, 1995
Robinson, D., 1995
Porporino, Fabiano, & Robinson, 1991
Little, Robinson, & Burnett, 1991a
Little & Robinson, 1999
Little, Robinson, & Burnett, 1994
Burnett, 1996
Ross, Fabiano, & Ewes, 1988

Mean

Odds-Ratio Effect Size
Overview

• Logic of Meta-analysis
• Effect sizes
  – What are they
  – Common types
• Basic analysis
  – Mean effect size
  – Confidence interval
  – Homogeneity analysis
• Random-effects versus Fixed-effect model
• Moderator analysis
• A note about software

Logic of Meta-analysis

• Narrative review methods:
  – Focuses on statistical significance
  – Lacks transparency and replicability
• Weakness of statistical significance:
  – Significant effect is a strong conclusion
  – Non-significant effect is a weak conclusion
  – How do you balance a collection of significant and non-significant effects?
Logic of Meta-analysis

- Meta-analysis:
  - Focuses on direction and magnitude of effect
  - Approaches task as a research endeavor
  - Examines pattern of evidence across studies
    - Average effect
    - Consistency of effects
    - Relationship between study features and effects

Effect Size

- Encodes relationship of interest into a common index
- Must be:
  - comparable across studies
  - independent of sample size
  - have a computable standard error
- Many different effect size indexes
- Multiple methods of computing each
- Most common:
  - Correlation coefficient (r )
  - Standardized mean difference (d or g)
  - Odds-ratio and Risk-ratio
Computing Effect Sizes

- Must compute effect size from information provided
  - Conversions from other statistics
    - t-test
    - p-value
    - descriptive statistics
    - etc.
  - Manipulation of data
    - collapsing across subgroups
    - adding ‘drop-outs’ back into the treatment condition
  - Some conversions better than others (algebraic equivalents; rough approximations)
- Some studies simply do not provide necessary information

Standardized Mean Difference

- Fundamental relationship:
  - Group contrast
  - Continuous dependent variable
- Logic: scaling effects based on standard deviation
- Definitional equation:
  \[ ES_{sm} = \frac{\bar{X}_1 - \bar{X}_2}{s_{pooled}} \]
Standardized Mean Difference

- Based on a $t$-test
  \[ ES_{tm} = t \sqrt{\frac{n_1 + n_2}{n_1 n_2}} \]
- Based on a correlation
  \[ ES_{sm} = \frac{2r}{\sqrt{1-r^2}} \]
- Based on 2 by 2 table (dichotomous outcome; logit method)
  \[ ES_{sm} = \ln \left( \frac{ad}{bc} \right) \sqrt{\frac{3}{\pi}} \]

Correlation as Effect Size

- Fundamental relationship:
  - Two inherently continuous constructs
- Correlation “comes” standardized
  \[ ES_r = r \]
- Example: Relationship between GRE scores and performance in graduate school
Odds-Ratio

- Fundamental relationship:
  - Group contrast
  - Dichotomous dependent variable
- Data can be represented in a 2 by 2 contingency table

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

- Odds-ratio effect size computed as:

\[ ES_{OR} = \frac{ad}{bc} \]

Basics of Meta-Analysis

- Goal:
  - Describe the distribution, including its mean
  - Establish a confidence interval around the mean
  - Test that the mean differs from zero
  - Test whether studies tell a consistent story (are homogeneous)
  - Explore the relationship between study features and effect size
Determining the Mean Effect Size

- Problem: some effect sizes are more accurate than others
- What we need is an index of precision
- Standard error is a direct measure of precision
- Hedges and Olkin solution:
  - Weight by the inverse variance
  - Provides a statistical basis for:
    - Standard error of the mean effect size
    - Confidence intervals
    - Homogeneity testing

Some Preliminary Transformations

- Small sample size bias correction for the standardized mean differences: 
  \[ ES'_{sm} = \left(1 - \frac{3}{4N-5}\right) ES_{sm} \]
- Fisher's Z transform of correlations (ES_r):
  \[ ES_{Zr} = \frac{1}{2} \log \left(\frac{1+r}{1-r}\right) \]
- Log transform of Odds-ratios ES_{OR} (also for Risk-ratio):
  \[ ES_{log(OR)} = \log (ES_{OR}) \]
Inverse Variance Weights

- Standardized mean difference \( ES_{sm} \):
  \[
  se_{sm} = \sqrt{\frac{n_1+n_2}{n_1n_2}} + \frac{ES_{sm}^2}{2(n_1+n_2)}
  \]

- Correlation \( ES_r \), (actually, the Fisher’s \( Z_r \)):
  \[
  se_r = \frac{1}{\sqrt{n-3}}
  \]

- Odds-ratio \( ES_{OR} \) (actually, the logged odds-ratio):
  \[
  se_{OR} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}
  \]

- Inverse variance weight \( w \):
  \[
  w = \frac{1}{se^2}
  \]

Almost ready

- At this point, we have for each study:
  - An effect size
  - An inverse variance weight
- Problem: statistical models assume independence
- Only include one effect size per study (or independent sample)
- Multiple analyses for different subsets of independent effects
  - Different outcome constructs
  - Different time periods
Inverse Variance Weighted Mean Effect Size

Meta-analytic mean effect size is:

\[ \overline{ES} = \frac{\sum w_i ES_i}{\sum w_i} \]

where \( ES_i \) is the effect size for each study (\( i \)) and \( w_i \) is the inverse variance weight

Standard error of the mean effect size is:

\[ se_{\overline{ES}} = \frac{1}{\sum w_i} \]

Some Basic Inferential Statistics

• Confidence intervals can be constructed in the usual manner:

\[ \overline{ES}_{lower} = \overline{ES} - se_{\overline{ES}} 1.96 \]
\[ \overline{ES}_{upper} = \overline{ES} + se_{\overline{ES}} 1.96 \]

• And a z-test can be performed as:

\[ z = \frac{\overline{ES}}{se_{\overline{ES}}} \]
An Example: Group-Based Cognitive-Behavioral Programs for Adult Offenders

<table>
<thead>
<tr>
<th>Author</th>
<th>Sample Size</th>
<th>Odds-Ratio</th>
<th>Logged OR</th>
<th>$w$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Burnett, 1996</td>
<td>60</td>
<td>2.25</td>
<td>0.81</td>
<td>1.727</td>
</tr>
<tr>
<td>Johnson &amp; Hunter, 1995</td>
<td>98</td>
<td>1.22</td>
<td>0.20</td>
<td>4.843</td>
</tr>
<tr>
<td>Little &amp; Robinson, 1989</td>
<td>180</td>
<td>1.52</td>
<td>0.42</td>
<td>7.614</td>
</tr>
<tr>
<td>Little et al, 1991</td>
<td>152</td>
<td>1.49</td>
<td>0.40</td>
<td>8.466</td>
</tr>
<tr>
<td>Little et al, 1994</td>
<td>1381</td>
<td>1.86</td>
<td>0.62</td>
<td>45.742</td>
</tr>
<tr>
<td>Porporino et al, 1991</td>
<td>63</td>
<td>1.33</td>
<td>0.28</td>
<td>3.633</td>
</tr>
<tr>
<td>Porporino &amp; Robinson, 1995</td>
<td>757</td>
<td>1.08</td>
<td>0.08</td>
<td>19.919</td>
</tr>
<tr>
<td>Robinson, D., 1995</td>
<td>2125</td>
<td>1.25</td>
<td>0.20</td>
<td>56.895</td>
</tr>
<tr>
<td>Ross et al, 1988</td>
<td>45</td>
<td>10.29</td>
<td>2.33</td>
<td>1.958</td>
</tr>
</tbody>
</table>

Note: These studies are a subset of studies included in Wilson et al. (2005) and represent two specific treatment programs (Moral Reconciliation and Reasoning and Rehabilitation) and studies that were randomized or used high quality quasi-experimental designs.

An Example: Group-Based Cognitive-Behavioral Programs for Adult Offenders

Stata output from “meanes.ado”

```
.Call (analytic weights assumed)
No. of obs = 9
Homogeneity Analysis
  Minimum obs = .0764
  Maximum obs = 2.331
  df = 8
  Q = 14.19
  p = 0.07695

                  | Mean  | 95% CI    | 95% CI    | SE   | Z    | P
Fixed effect      | 0.37107 | 0.21164   | 0.53087   | 0.06143 | 4.58671 | 0.00001
Random effect 1    | 0.40218 | 0.14349   | 0.66086   | 0.13198 | 3.04718 | 0.00231
Random effect 2    | 0.38438 | 0.17673   | 0.59203   | 0.10565 | 3.62808 | 0.00029

1 Random effects variance component (method of moments) | 0.00042
2 Random effects variance component (full information ML) | 0.02054
```
An Example: Group-Based Cognitive-Behavioral Programs for Adult Offenders

Stata output from “meanes.ado"

```
. means lgor U=1, print(exp)
(analytic weights assumed)
Version 2005.03.23 of means.ado

No. of obs = 9
Minimum obs = 9.00
Maximum obs = 10.250
Weighted SD =

Homogeneity Analysis

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>95% CI</th>
<th>SE</th>
<th>Z</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed effect</td>
<td>1.44928</td>
<td>1.28548</td>
<td>1.70008</td>
<td>4.55671</td>
<td>0.00001</td>
</tr>
<tr>
<td>Random effect 1</td>
<td>1.48047</td>
<td>1.15430</td>
<td>1.93845</td>
<td>3.04719</td>
<td>0.00231</td>
</tr>
<tr>
<td>Random effect 2</td>
<td>1.46870</td>
<td>1.19331</td>
<td>1.80765</td>
<td>3.62808</td>
<td>0.00002</td>
</tr>
</tbody>
</table>

1 Random effects variance component (method of moments) = 0.05542
2 Random effects variance component (full information ML) = 0.02094
Results are the exponent of computed values (i.e., results are odds-ratios)
```

Homogeneity Testing

- Homogeneity analysis tests whether the assumption that all of the effect sizes are estimating the same population mean is a reasonable assumption.
- If homogeneity is rejected, the distribution of effect sizes is assumed to be heterogeneous.
  - Single mean ES not a good descriptor of the distribution
  - There are real between study differences, that is, studies estimate different population mean effect sizes.
  - Three options:
    - model between study differences
    - fit a random effects model
    - do both
Computation of the Homogeneity $Q$ Statistic

- $Q$ is simply a weighted sums-of-squares:
  $$Q = \sum w_i (ES_i - ES)^2$$

- There are easier computational formulas:
  $$Q = \sum w_i ES_i^2 - \left( \frac{\sum w_i ES_i}{\sum w_i} \right)^2$$

- It is distributed as a chi-square with $k - 1$ degrees of freedom, where $k$ is the number of effect sizes.

Alternative to $Q$

- $Q$ is statistically under-powered when the number of studies is low and when the sample size within the studies is low.

  - $I^2 = 100\% \times \frac{Q - df}{Q}$
  - Larger values of $I^2$, the more heterogeneity
    - 75%: large heterogeneity
    - 50%: moderate heterogeneity
    - 25%: low heterogeneity
Random versus Fixed Effects Models

• Fixed effects model assume:
  – there is one true population effect that all studies are estimating
  – all of the variability between effect sizes is due to sampling error
• Random effects model assume:
  – there are multiple (i.e., a distribution) of population effects that the studies are estimating
  – variability between effect sizes is due to sampling error + variability in the population of effects

Fixed versus Random: Which to Use?

• A random-effects model becomes a fixed-effect model when distributions is homogeneous
• Assumptions of fixed effects model rarely plausible
  – Consequence: standard error that is too small; confidence intervals that are too narrow
• Bottom-line: Use the random-effects model
Computing a Random Effects Model

- Fixed effects model: weights are a function of sampling error
- Random effects model: weights are a function of sampling error + study level variability
- Thus, we need a new set of weights
- First, compute $\tau^2$ (random effects variance component):
  $$\tau^2 = \frac{Q - df \, Q}{\sum w_i - \frac{df \, Q}{w_i}}$$
- Second, re-compute the inverse variance weights:
  $$w_i = \frac{1}{se_i^2 + \tau^2}$$
- Third, re-compute meta-analytic results using new weight

Moderator Analysis

- Modeling between study variability
  - Categorical models (analogous to a one-way ANOVA)
  - Regression models
- Fixed and random effects versions of each
  (latter often called mixed" models)
Analog to the ANOVA

- Useful for a single categorical independent variable
- Produce a separate mean effect size for each category
- Recall that $Q$ is a sum-of-squares
- The total sum-of-squares ($Q$) can be partitioned
  - Variability between groups ($Q_{between}$)
  - Residual variability within groups ($Q_{within}$)
- $Q_{between}$ analogous to an $F$-test between means
- $Q_{within}$ assesses whether residual distribution homogeneous
- Note: in a random effects (mixed effects) version of this, the $Q_{within}$ is not meaningful

Analog to the ANOVA Example: Experimental versus Quasi-experimental Studies in the Domestic Violence
Meta-analytic Regression

- Conceptually identical to multiple regression
  - Effect size is the dependent variable
  - Study moderator variables are the independent variables
- Can handle multiple variables simultaneously
- Don't use standard OLS regression procedures (even if weighted)
- Must use specialized software

Meta-analytic Regression: Example

```
***** Inverse Variance Weighted Regression *****
***** Random Intercept, Fixed Slopes Model *****

------- Descriptives -------
Mean Eta   R-Square   k
.1483      .2225       38.0000

------- Homogeneity Analysis -------
Q     df  p
Model 14.7731      3.0000  .0020
Residual 51.5274     34.0000 .0000
Total  66.0495     27.0000  .0021

------- Regression Coefficients -------
   S  SE  -2.5% CI  +2.5% CI  Z  P  Beta
Constant -.6752 .2392 -.1.1439 -.1.065 -.2.8233 .0048 .0000
RANDOM .0790 .0834 -.0906 .2863 .8746 .3818 .1107
TYBAR1 .3790 .1436 .0972 .6508 2.6354 .0004 .3264
TYBAR2 .1986 .0821 .0838 .3966 2.4204 .0155 .3091

------- Method of Moments Random Effects Variance Component -------
v = .04715
```
**Forest Plots**

- Visual representation of results
- Row for each study that shows
  - study label
  - sample size; may include other information
  - effect size (dot, square, diamond)
  - confidence interval (horizontal line)
- Row for the overall mean results
  - effect size (dot, square, diam-and)
  - confidence interval (horizontal line)

---

**Example Forest-Plot**

- Forest-Plot of Odds-Ratios and 95% Confidence Intervals for the Effects of Cognitive-Behavioral Programs on Recidivism

```
Porporino & Robinson, 1995
Johnson & Hunter, 1995
Robinson, D., 1995
Porporino, Fabiano, & Robinson, 1991
Little, Robinson, & Burnett, 1991a
Little & Robinson, 1999
Little, Robinson, & Burnett, 1994
Burnett, 1996
Ross, Fabiano, & Ewles, 1988
```

---

The Campbell Collaboration  
www.campbellcollaboration.org
Comments on Software

- Specialized software
  - RevMan (developed for Cochrane)
  - CMA - Comprehensive Meta-analysis
- Add-ons for Common Statistical Programs
  - Stata - lots of macros available
  - SPSS - macros available from my website
  - SAS - macros available from my website
  - R - lots of procedures available
- For computing effect sizes
  - CMA
  - RevMan
  - ES Calculator by Wilson
    - http://www.campbellcollaboration.org/resources/effect_size_input.php

Final Comments

- Methods continue to advance
- Publication selection-bias should be assessed (not addressed in this course)
- Methods for analyzing dependent effect sizes actively advancing
- Common errors
  - Incorrectly computing effect sizes
  - Not recognizing situations were effect sizes can be computed
  - Using fixed-effect models
  - Not using moderator analysis to compare mean effect sizes for study subsets
  - Focusing too much on statistical significance and not size and direction of effect