Design and Analysis of Replication Studies

Douglas G. Bonett

Center for Statistical Analysis in the Social Sciences

April 29, 2013

Overview

- Advantages of replication research
- Basic replication designs
- Statistical methods for replication studies
- Sample size planning for replication studies

Advantages of Replication Research

- Replication research can expose misleading results in prior studies
- Results from prior studies can be combined with results from a new study to obtain more a precise interval estimate along with greater generalizability of results
- Results from prior studies can be compared with results from a new study to assess possible interaction effects
- Ignoring data from prior studies is wasteful

Incomplete Two-sample Multi-study Design

Study	Treatment 1	Treatment 2
1	θ_1	---
2*	θ_2	θ_3

Treatment 1 replication effect: $\theta_1 - \theta_2$

Multi-study Treatment 1: $(\theta_1 + \theta_2)/2$

Multi-study Treatment effect: $(\theta_1 + \theta_2)/2 - \theta_3$

θ_k is a population parameter such as a population mean, population median, or population proportion

Two-sample Multi-study Design

Study	Treatment 1	Treatment 2
1	θ_1	θ_2
2*	θ_3	θ_4

Treatment Effect replication: $(\theta_1 - \theta_2) - (\theta_3 - \theta_4)$

Treatment 1 replication: $\theta_1 - \theta_3$

Treatment 2 replication: $\theta_2 - \theta_4$

Multi-study Treatment effects

assuming Treatment Effect replication: $(\theta_1 + \theta_3)/2 - (\theta_2 + \theta_4)/2$

assuming only Treatment 1 replication: $(\theta_1 + \theta_3)/2 - \theta_4$

assuming only Treatment 2 replication: $\theta_3 - (\theta_2 + \theta_4)/2$

Extensions of the Two-sample Multi-study Design

Study	Treatment 1	Treatment 2
1	θ_1	---
2	---	θ_2
3	θ_3	θ_4
4*	θ_5	θ_6

Multi-study designs also can be extended to have more than two treatments per study.

Factorial Multi-study Design (*Example 1*)

Study	a1	a2	b1	b2	a1b1	a2b1	a1b2	a2b2
1	θ_1	θ_2	---	---	---	---	---	---
2	---	---	θ_3	θ_4	---	---	---	---
3*	---	---	---	---	θ_9	θ_{10}	θ_{11}	θ_{12}

Factor A replication: $(\theta_1 - \theta_2) - [(\theta_9 + \theta_{11})/2 - (\theta_{10} + \theta_{12})/2]$

Factor B replication: $(\theta_3 - \theta_4) - [(\theta_9 + \theta_{10})/2 - (\theta_{11} + \theta_{12})/2]$

A main effect: $\{(\theta_1 - \theta_2) + [(\theta_9 + \theta_{11})/2 - (\theta_{10} + \theta_{12})/2]\}/2$

B main effect: $\{(\theta_3 - \theta_4) - [(\theta_9 + \theta_{10})/2 - (\theta_{11} + \theta_{12})/2]\}/2$

AB interaction: $(\theta_9 - \theta_{10}) - (\theta_{11} - \theta_{12})$

Factorial Multi-study Design *(Example 2)*

Study	a_1b_1	a_2b_1	a_1b_2	a_2b_2
1	θ_1	θ_2	---	---
2	---	---	θ_3	θ_4
3*	θ_5	θ_6	θ_7	θ_8

A at b_1 replication: $(\theta_1 - \theta_2) - (\theta_5 - \theta_6)$

A at b_2 replication: $(\theta_3 - \theta_4) - (\theta_7 - \theta_8)$

A main effect: $(\theta_1 + \theta_5 + \theta_3 + \theta_7)/4 - (\theta_2 + \theta_6 + \theta_4 + \theta_8)/4$

B main effect: $(\theta_1 + \theta_5 + \theta_2 + \theta_6)/4 - (\theta_3 + \theta_7 + \theta_4 + \theta_8)/4$

AB interaction: $[(\theta_1 + \theta_5)/2 - (\theta_2 + \theta_6)/2] - [(\theta_3 + \theta_7)/2 - (\theta_4 + \theta_8)/2]$

Multi-study Designs for Effect Size Parameter

Study	Effect Size
1	θ_1
2	θ_2
⋮	⋮
s - 1	θ_{s-1}
s*	θ_s

Replication effect: $[\sum_j^{s-1} \theta_j/(s-1)] - \theta_s$

Multi-study effect: $\sum_j^s \theta_j/s$

Examples of effect size measures: mean difference, standardized mean difference, correlation, slope, risk difference, risk ratio, and odds ratio

Multiple Replication Designs

The *Association for Psychological Science* is encouraging the replication of specific important psychological studies by *multiple* investigators. A simple example of this type of multiple replication design is illustrated below where the original two-treatment study (Study 1) is replicated by two other investigators (Study 2 and Study 3).

Study	Treatment 1	Treatment 2
1	θ_1	θ_2
2*	θ_3	θ_4
3*	θ_5	θ_6

Multiple Replication Designs (*continued*)

Replication effects:

$$(\theta_1 - \theta_2) - (\theta_3 - \theta_4)$$
$$(\theta_1 - \theta_2) - (\theta_5 - \theta_6)$$
$$(\theta_3 - \theta_4) - (\theta_5 - \theta_6)$$

Multi-study treatment effect:

$$(\theta_1 + \theta_3 + \theta_5)/3 - (\theta_2 + \theta_4 + \theta_6)/3$$

Statistical Evidence of Replication

The current approach (in the social sciences) simply uses significance testing results and declares a study to have been replicated if the *p*-value based test result – "significant" or "nonsignificant" – is the same in the original study and the replication study.

This approach is flawed because taking a large sample in the replication study will tend to replicate a "significant" result in the original study, and taking a small sample in the replication study will tend to replicate a "non-significant" result in the original study.

Statistical Evidence of Replication (*continued*)

Instead of assessing replication evidence dichotomously, replication evidence should be assessed quantitatively.

For example, if θ_1 the population parameter value for the original study (study 1) and θ_2 is the population parameter for the replication study (study 2), then the magnatude of $\theta_1 - \theta_2$ describes the *degree* to which study 2 replicates study 1.

θ_1 and θ_2 will each be estimated from random samples and an interval estimate of $\theta_1 - \theta_2$ provides statistical evidence of replication.

Interval Estimates of Linear Contrasts

All of the replication effects described in the illustrative multi-study designs can be expressed in terms of a linear contrast of population parameters (treatment-specific parameters or effect sizes) and we can compute an interval estimate of the linear contrast using sample estimates of the parameters and their standard errors.

Means (*general independent-group designs*)

$100(1-\alpha)\%$ interval estimate:

$$\sum_{j=1}^{k} c_j \hat{\theta}_j \pm t_{\alpha/2;df} \sqrt{\sum_{j=1}^{k} c_j^2 \hat{\sigma}_j^2 / n_j}$$

where $df = \left[\sum_{j=1}^{k} \frac{c_j^2 \hat{\sigma}_j^2}{n_j}\right]^2 / [\sum_{j=1}^{k} \frac{c_j^4 \hat{\sigma}_j^4}{n_j^2 (n_j - 1)}]$ and k = total number of estimates (sample means).

Proportions (*general independent-group designs*)

$100(1-\alpha)\%$ interval estimate:

$$\sum_{j=1}^{k} c_j \hat{\theta}_j \pm z_{\alpha/2} \sqrt{\sum_{j=1}^{k} c_j^2 \frac{\hat{\theta}_j(1-\hat{\theta}_j)}{n_j + 4/m}}$$

where $\hat{\theta}_j = (f_j + 2/m)/(n_j + 4/m)$, and m is the number of nonzero c_j values.

(Price & Bonett, 2004)

Medians (*general independent-group designs*)

$100(1-\alpha)\%$ interval estimate:

$$\sum_{j=1}^{k} c_j \hat{\theta}_j \pm z_{\alpha/2}\sqrt{\sum_{j=1}^{k} c_j^2 \, var(\hat{\theta}_j)}$$

where $var(\hat{\theta}_j) = [y_{(n_j - a + 1)} - y_{(a)}]^2/16$ and

$a = (n_j + 1)/2 - \sqrt{n}$ (a is rounded down to nearest integer).

(Bonett & Price, 2002)

Effect Size Designs

General $100(1-\alpha)\%$ interval estimate:

$$\sum_{j=1}^{s} c_j \hat{\theta}_j \pm z_{\alpha/2} \sqrt{\sum_{j=1}^{s} c_j^2 var(\hat{\theta}_j)}$$

where $\hat{\theta}_j$ is some measure of effect size and s is the total number of studies.

Effect Size: Linear contrast of means in repeated measures designs

Let $\hat{\theta}_j = \boldsymbol{h}'\hat{\mu}_j$ and $var(\hat{\theta}_j) = \boldsymbol{h}'cov(\boldsymbol{y}_j)\boldsymbol{h}/n_j$ where $cov(\boldsymbol{y}_j)$ is a q x q covariance matrix of the q measurements from each sampling unit in study j and $\boldsymbol{h}' = [h_1\ h_2\ \ldots\ h_q]$.

$z_{\alpha/2}$ can be replaced with $t_{\alpha/2;df}$ where

$$df = [\sum_{j=1}^{s} c_j^2 var(\hat{\theta}_j)]^2 / [\sum_{j=1}^{s} \frac{[c_j^2 var(\hat{\theta}_j)]^2}{(n_j - 1)}]$$

Note that $var(\hat{\theta}_j)$ can be computed from prior studies that report $t = \boldsymbol{h}'\hat{\boldsymbol{\mu}}_j / \sqrt{var(\boldsymbol{h}'\hat{\boldsymbol{\mu}}_j)}$ and the sample means.

Other Effect Sizes

Risk difference (paired samples):

$$\hat{\theta}_j = (f_{12j} - f_{21j})/(n_j + 2)$$

$$var(\hat{\theta}_j) = [f_{12j} + f_{21j} + 2 - (f_{12j} - f_{21j})^2]/(n_j + 2)$$

(Bonett & Price, 2012, 2013a)

Odds ratio:

$$\hat{\theta}_j = log[(f_{11} + \tfrac{1}{2})(f_{22} + \tfrac{1}{2})/\{\left(f_{12} + \tfrac{1}{2}\right)\left(f_{21} + \tfrac{1}{2}\right)\}]$$

$$var(\hat{\theta}_j) = 1/(f_{11j} + 0.5) + 1/(f_{22j} + 0.5) + 1/(f_{12j} + 0.5) + 1/(f_{21j} + 0.5)$$

(Bonett & Price, 2013b)

Other Effect Sizes (*continued*)

Risk ratio (independent samples):

$$\hat{\theta}_j = log\{(f_{11} + \frac{1}{4})/(n_1 + \frac{7}{4})\}/\{(f_{12} + \frac{1}{4})/(n_2 + \frac{7}{4})\}$$

$$var(\hat{\theta}_j) = \frac{1}{\left\{f_{11} + \frac{1}{4} + \frac{\left(f_{1\,1} + \frac{1}{4}\right)^2}{n_1 - f_{11} + \frac{3}{2}}\right\}} + \frac{1}{\left\{f_{12} + \frac{1}{4} + \frac{\left(f_{12} + \frac{1}{4}\right)^2}{n_2 - f_{12} + \frac{3}{2}}\right\}}$$

(Price & Bonett, 2008; Bonett & Price, 2013b)

Other Effect Sizes (*continued*)

Standardized mean difference:

$$\hat{\theta}_j = (\hat{\mu}_1 - \hat{\mu}_2) / \sqrt{(\hat{\sigma}_1^2 + \hat{\sigma}_2^2)/2}$$

$$var(\hat{\theta}_j) = \frac{\hat{\theta}_j{}^2\left(\frac{1}{n_1-1} + \frac{1}{n_2-1}\right)}{8} + \frac{1}{n_1} + \frac{1}{n_2}$$ independent samples

$$var(\hat{\theta}_j) = \frac{\hat{\theta}_j{}^2(1+\hat{\rho}_{12}^2)}{4(n-1)} + \frac{2(1-\hat{\rho}_{12})}{n}$$ paired samples

(Bonett, 2009)

Effect size: Pearson correlation

Interval estimates for two special types of linear contrasts are currently available:

Case 1 $(\theta_1 + \theta_2 + \ldots + \theta_s)/s$

Case 2 $(\theta_1 + \theta_2 + \ldots + \theta_{s-1})/(s-1) - \theta_s$

Case 1

$100(1-\alpha)\%$ interval estimate for $(\sum_{j=1}^{s}\theta_j)/s$:

$$tanh[tanh^{-1}(\bar{\theta}) \pm z_{\alpha/2}\sqrt{var\{tanh^{-1}(\bar{\theta})\}}]$$

where $var\{tanh^{-1}(\bar{\theta})\} = var(\bar{\theta})\ /(1-\ \bar{\theta}^2)^2$

$$var(\bar{\theta}) = s^{-2}\sum_{j=1}^{s} var(\hat{\theta}_j)$$

$$var(\hat{\theta}_j) = \left(1-\ \hat{\theta}_j^2\right)^2/(n_j-3)$$

$$\bar{\theta} = (\sum_{j=1}^{s}\hat{\theta}_j)/s$$

(Bonett, 2008)

Case 2

Let L_1 and U_1 be the $100(1-\alpha)\%$ interval estimate for $(\sum_{j=1}^{s-1}\theta_j)/(s-1)$ and let L_2 and U_2 be the $100(1-\alpha)\%$ interval estimate for θ_s. A $100(1-\alpha)\%$ interval estimate for $(\sum_{j=1}^{s-1}\theta_j)/(s-1)-\theta_s$ is:

$$L = \hat{\theta}_1 - \hat{\theta}_2 - \sqrt{(\hat{\theta}_1 - L_1)^2 + (\hat{\theta}_2 - U_2)^2}$$

$$U = \hat{\theta}_1 - \hat{\theta}_2 + \sqrt{(\hat{\theta}_1 - U_1)^2 + (\hat{\theta}_2 - L_2)^2}$$

(Zou, 2007; Bonett, 2008)

Sample Size Planning - Proportions

The required sample size per group in a new study to estimate a linear contrast of population proportions with desired precision and confidence using information from prior studies is approximately:

$$n = [\sum_{j=k-q+1}^{k} c_j^2 \tilde{\theta}_j(1-\tilde{\theta}_j)]/[\left(\frac{w}{2z_{\alpha/2}}\right)^2 - \sum_{j=1}^{k-q} c_j^2 SE(\hat{\theta}_j)^2]$$

where $\tilde{\theta}_j$ is a planning value of the proportion in group j of the new study, w is the desired confidence interval width, $SE(\hat{\theta}_j)$ is the standard error of the j^{th} sample proportion in prior studies

Sample Size Planning - Means

The required sample size per group in a new study to estimate a linear contrast of population means with desired precision and confidence using information from prior studies is approximately:

$$n = [\sum_{j=k-q+1}^{k} c_j^2 \tilde{\sigma}_j^2]/[\left(\frac{w}{2z_{\alpha/2}}\right)^2 - \sum_{j=1}^{k-q} c_j^2 \, SE\left(\hat{\theta}_j\right)^2]$$

where $\tilde{\sigma}_j^2$ is a planning value of the variance in group j of the new study, w is the desired confidence interval width, $SE\left(\hat{\theta}_j\right)$ is the standard error of the j^{th} sample mean in prior studies.

Sample Size Planning - Medians

Use the sample size formula for means but replace $\tilde{\sigma}_j^2$ with $f\tilde{\sigma}_j^2$ where f is given below for several different distributions.

Distribution	f	skew	kurtosis
Normal	1.57	0	3
Logistic	1.17	0	4.2
Laplace	0.50	0	6
Rayleigh	1.68	0.63	3.3
Gamma(5)	1.48	0.89	4.2
Gamma(2)	1.28	1.41	6
Exponential	1.00	2	9

Sample Size Planning – Average effect size

The required sample size in a new study to estimate an average effect size with desired precision and confidence using estimates from $s - 1$ prior studies is approximately:

$$n = v/[\left(\frac{w}{2z_{\alpha/2}}\right)^2 - (s-1)\sum_{j=1}^{s-1} SE\left(\hat{\theta}_j\right)^2]$$

where (e.g.)

$v = (1 - \tilde{\rho}^2)^2$	correlation
$v = 2\tilde{\sigma}^2$	mean difference
$v = \tilde{\pi}_1(1-\tilde{\pi}_1) + \tilde{\pi}_2(1-\tilde{\pi}_2)$	risk difference
$v = \frac{\tilde{\theta}^2}{4} + 2$	STD mean difference

Example

Study	Sample Proportions (*n*) Treatment 1	Treatment 2	
1	.825 (40)	---	Milgram
2*	.700 (40)	.633 (30)	Burger (2009)

	95% Interval Estimates
Treatment 1 replication effect:	[-0.06, 0.30]
Multi-study Treatment 1:	[0.67, 0.85]
Multi-study Treatment effect:	[-0.06, 0.32]

Example (*continued*)

Burger (2009) also included a Sex factor in his study.

a1 = Milgram condition
a2 = modeled refusal

b1 = male
b2 = female

Study	a1b1	a2b1	a1b2	a2b2		
1	θ_1	---	---	---	Milgram	n = 40
2	θ_2	θ_3	θ_4	θ_5	Burger (2009)	n = 70
3*	θ_6	θ_7	θ_8	θ_9	(planned study)	n = ?

"Societal Change": $\theta_1 - (\theta_2 + \theta_6)/2$

Main effect of A: $(\theta_1 + \theta_2 + \theta_6 + \theta_4 + \theta_8)/5 - (\theta_3 + \theta_7 + \theta_5 + \theta_9)/4$

Main effect of B: $(\theta_1 + \theta_2 + \theta_6 + \theta_3 + \theta_7)/5 - (\theta_4 + \theta_8 + \theta_5 + \theta_9)/4$

How many participants should be used in Study 3?

Example (*continued*)

Using the Milgram (Study 1) and Burger (Study 2) results as planning values, the required sample size to estimate the main effects of Factors A or B in Study 3 with 95% confidence and an interval width of 0.25 is **31** per group

The required sample size in the a_1b_1 condition of Study 3 to estimate "Societal Change" with 95% confidence and an interval width of 0.35 is **38**

References

Bonett, D.G. (2008). Confidence intervals for standardized linear contrasts of means. *Psychological Methods, 13*, 99-109.

Bonett, D.G. (2008). Meta-analytic interval estimation for bivariate correlations. *Psychological Methods, 13*, 173-189.

Bonett, D.G. (2009). Meta-analytic interval estimation for standardized and unstandardized mean differences. *Psychological Methods, 14*, 225-238.

Bonett, D.G. & Price, R.M. (2002). Statistical inference for a linear function of medians: Confidence intervals, hypothesis testing, and sample size requirements. *Psychological Methods*, *7*, 370-383.

Bonett, D.G. & Price, R.M. (2012). Adjusted Wald interval for a difference in binomial proportions based on paired samples. *Journal of Educational and Behavioral Statistics, 37*, 479-488.

Bonett, D.G. & Price, R.M. (2013a). Meta-analysis for risk differences. *British Journal of Mathematical and Statistical Psychology*, revision under review.

Bonett, D.G. & Price, R.M. (2013b). Varying coefficient meta-analysis methods for odds ratios and risk ratio, under review.

Burger, J. M. (2009). Replicating Milgram: Would people still obey today? *American Psychologist, 64*, 1-11.

Price, R.M. & Bonett, D.G. (2004). Improved confidence interval for a linear function of binomial proportions. *Computational Statistics & Data Analysis, 45*, 499-456.

Price, R.M. & Bonett, D.G. (2008). Confidence intervals for a ratio of two binomial proportions. *Statistics in Medicine, 27*, 5497-5508.

Zou, G.Y. (2007). Toward using confidence intervals to compare correlations. *Psychological Methods*, *12*, 399-413.