

# Chapter 4

---

## *Parametric Effect Size Indexes*

Statistician: A man who believes figures don't lie, but admits that under analysis some of them won't stand up either.

—Evan Esarity

### Overview

- Contexts for effect size estimation
- Concepts about effect size
- Standardized mean differences for two-sample designs
- Measures of association for two-sample designs
- Limitations of effect size indexes
- Case-level effect size estimation
- Interval estimation for effect sizes

## Contexts for effect size estimation

---

1. When units of the outcome variable are arbitrary rather than meaningful—there is little need for standardized (metric-free) measures of effect size when the original metric is meaningful
2. When results are compared across outcomes measured on different scales—standardized effect sizes provide a common language for comparison in this case
3. A priori power analysis, which requires specification of the expected population effect size magnitude

## Contexts for effect size estimation

---

4. Meta-analysis, which typically analyzes the central tendency and variability of standardized effect sizes across a set of studies
5. Resolving interpretational problems of statistical tests, such as when a trivial effect size is statistically significant in a large sample

# Concepts about effect size

---

- *Effect size* refers to the magnitude of the impact of the independent variable (factor) on the outcome variable
- But *cause size* refers to the amount of change in the independent variable that produces a given effect on the outcome variable (Abelson, 1997)
- The idea of cause size is most relevant in experimental designs

# Concepts about effect size

---

- Families of effect size indexes for continuous outcomes (Huberty, 2002; Maxwell & Delaney, 1990; Rosenthal, 1994):

- ✓ Group (variable) level:

*d* family (*group-difference indexes*): Standardized mean differences

*r* family (*relationship indexes*): Measures of association (correlations, variance-accounted-for effect sizes)

- ✓ Case level (*group overlap indexes*):

Proportion of scores from different samples above or below certain reference points

## Concepts about effect size

---

- A *standardized mean difference* expresses a mean contrast as the proportion of a standard deviation among cases, *not* of a standard error (e.g.,  $t$ )
- A population standardized mean difference has the following general form:

$$\delta = \frac{\mu_1 - \mu_2}{\sigma^*}$$

- The term  $\sigma^*$  is the *standardizer*, a population standard deviation for a comparative study
- Because more there is more than one population standard deviation for a comparative study,  $\sigma^*$  has more than one form

## Concepts about effect size

---

- A sample standardized mean difference has the following general form:

$$d = \frac{M_1 - M_2}{\hat{\sigma}^*}$$

- The term  $\hat{\sigma}^*$  estimates a population standard deviation and has more than one form

## Concepts about effect size

---

- A *measure of association* describes the amount of the covariation between the independent and dependent variables
- It is expressed in an unsquared metric or a squared metric—the former is usually a correlation, the latter a variance-accounted-for effect size
- A squared multiple correlation ( $R^2$ ) calculated in ANOVA is called the *correlation ratio* or *estimated eta-squared*,  $\hat{\eta}^2$

## Concepts about effect size

---

- The term  $\eta^2$  estimates the following parameter for effects of fixed factors:

$$\eta_{\text{effect}}^2 = \frac{\sigma_{\text{effect}}^2}{\sigma_{\text{tot}}^2}$$

$\sigma_{\text{tot}}^2$  is the total population variance

- The term  $\hat{\eta}^2$  has the following general form:

$$\hat{\eta}_{\text{effect}}^2 = \frac{SS_{\text{effect}}}{SS_T}$$

$SS_T$  is the total sample sum of squares

## Concepts about effect size

---

- The term  $\hat{\eta}^2$  capitalizes on chance variation in a particular sample, which means that it estimates  $\eta^2$  with positive bias
- Other variance-accounted-for effect sizes have “built-in” corrections for bias:
  - ✓ Estimated omega-squared ( $\hat{\omega}^2$ ) for fixed factors
  - ✓ Intraclass correlation ( $\hat{\rho}_1$ ) for random factors

## Concepts about effect size

---

- This is not to say that inferential measures of association, such as  $\hat{\omega}^2$  or  $\hat{\rho}_1$ , are bias-free
- There are also many other bias-adjusted measures of association (e.g., Olejnik & Algina, 2000; Snyder & Lawson, 1993)
- But it is not always clear which bias-adjusted statistic is best for a particular study

## Standardized mean differences for independent contrasts

---

- Perhaps the most generally useful is Hedges's  $g$ :

$$g = \frac{M_1 - M_2}{s_P}$$

- The standardizer of  $g$  is the square root of the pooled within-groups variance (which assumes homogeneity of variance):

$$s_P^2 = \frac{SS_W}{df_W} = \frac{(n_1 - 1) s_1^2 + (n_2 - 1) s_2^2}{n_1 + n_2 - 2}$$

## Standardized mean differences for independent contrasts

---

- Hedges's  $g$  can be calculated from the independent-samples  $t$  for the contrast:

$$g = t \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- Hedges's  $g$  can also be converted from the point-biserial correlation between group membership and the outcome variable:

$$g = r_{pb}^2 \sqrt{\left(\frac{df_w}{1 - r_{pb}^2}\right) \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

## Standardized mean differences for independent contrasts

---

- The usual standardizer for Glass's  $\Delta$  is the standard deviation of the control group ( $s_1$  below):

$$\Delta = \frac{M_1 - M_2}{s_1}$$

- Glass's  $\Delta$  does not assume homogeneity of variance
- Glass's  $\Delta$  reflects the effect of a treatment on central tendency only, not also on variability
- When the group variances are similar,  $g$  is preferred over  $\Delta$

# Standardized mean differences for dependent contrasts

---

- A  $d$  statistic for a dependent mean contrast is called a *standardized mean change*
- There are two general standardizers:

A standard deviation in the metric of the

1. original scores
2. difference scores ( $D$ )

## Standardized mean differences for dependent contrasts

---

- For the same contrast, a standardized mean change in the metric of the difference scores can be much larger than in the metric of the original scores
- The choice of standardizers depends on how change should be described
- The most general choice is a standardizer in the metric of the original scores, which is directly comparable with a standardized mean difference from an independent-samples design

## Standardized mean differences for dependent contrasts

---

- A standardized mean change in the metric of the original scores can be calculated as Hedges's  $g$  or Glass's  $\Delta$
- An exception is when  $g$  is to be calculated from the dependent-samples  $t$  (Dunlap, Cortina, Vaslow, & Burke, 1996):

$$g = t \sqrt{\frac{2 s_D^2}{n (s_1^2 + s_2^2)}}$$

$$s_D^2 = s_1^2 + s_2^2 - 2 \text{cov}_{12}$$

$s_D^2$  is the variance of the difference ( $D$ ) scores

## Measures of association for independent contrasts

---

- The point-biserial correlation,  $r_{pb}$ , is the Pearson correlation between membership in one of two groups and a continuous outcome variable
- It is a special case of  $\hat{\eta} = \sqrt{SS_A/SS_T}$ , where  $SS_A$  is the sum of squares for the dichotomous factor  $A$
- However,  $r_{pb}$  is a signed correlation, but  $\hat{\eta}$  is not

## Measures of association for independent contrasts

---

- A general formula for  $r_{pb}$  is:

$$r_{pb} = \left( \frac{M_1 - M_2}{\sqrt{SS_T/N}} \right) \sqrt{pq}$$

$p$  and  $q$  are the proportions of cases in each group

- The correlation  $r_{pb}$  can also be calculated from the independent-samples  $t$  for the contrast:

$$r_{pb} = \frac{t}{\sqrt{t^2 + df_W}}$$

## Measures of association for dependent contrasts

---

- The correlation  $r_{pb}$  is for two-group designs where the groups are independent (not matched)
- However, the correlation ratio  $\hat{\eta}^2$  can be calculated for correlated designs with a dichotomous factor  $A$

## Measures of association for dependent contrasts

---

- There are two general forms of  $\hat{\eta}^2$  in such designs:

1. The proportion of total variance explained by A:

$$\hat{\eta}^2 = \frac{SS_A}{SS_T} = \frac{SS_A}{SS_A + SS_W}$$

2. The proportion of variance explained by A controlling for the subjects effect (nonadditive model assumed):

$$\text{partial } \hat{\eta}^2 = \frac{SS_A}{SS_T + SS_S} = \frac{SS_A}{SS_A + SS_{A \times S}}$$

- For the same contrast, partial  $\hat{\eta}^2$  can be much larger than  $\hat{\eta}^2$ , but the latter may be more directly comparable with  $\hat{\eta}^2$  from an independent-samples design

# Limitations of effect size indexes

---

- Standardized mean differences:
  - ✓ Heterogeneity of within-conditions variances across studies can limit their usefulness—the unstandardized contrast may be better in this case
  - ✓ Bias-adjusted forms of measures of association are more readily available

# Limitations of effect size indexes

---

- Measures of association:

- ✓ Correlations can be affected by sample variances and whether the samples are independent or not, the design is balanced or not, or the factors are fixed or not
- ✓ Also affected by artifacts such as missing observations, range restriction, categorization of continuous variables, and measurement error—see Hunter and Schmidt (1994) for various corrections
- ✓ Variance-accounted-for indexes can make some effects look smaller than they really are in terms of their substantive significance

# Limitations of effect size indexes

---

- General (i.e., how to fool yourself with effect size estimation):
  1. Measure effect size at the group level only
  2. Apply generic definitions of effect size magnitude without first looking to the literature in your area
  3. Believe that an effect size judged as “large” according to generic definitions must be an important result and that a “small” effect is unimportant—see Prentice and Miller (1992)
  4. Ignore the question of how substantive (theoretical, clinical, or practical) significance should be gauged in your research area
  5. Estimate effect size only for statistically significant results

# Limitations of effect size indexes

---

- General (i.e., how to fool yourself with effect size estimation):
  6. Believe that finding large effects somehow lessens the need for replication
  7. Forget that effect sizes are subject to sampling error, too
  8. Forget that effect sizes for fixed factors are specific to the particular levels selected for study
  9. Forget that standardized effect sizes encapsulate other quantities, such as the unstandardized effect size, error variance, and experimental design
  10. As a journal editor or reviewer, substitute effect size magnitude for statistical significance as a criterion for whether a work is to be published

## Case-level effect size estimation

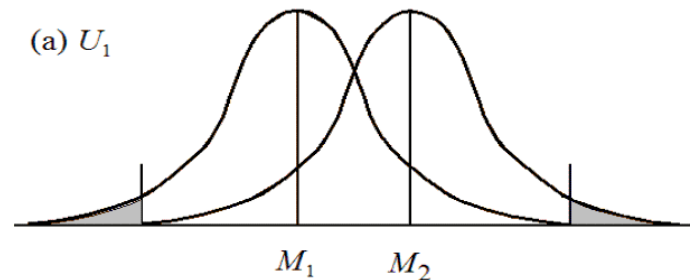
---

- Indexes such as Hedges's  $g$  and  $\hat{\eta}^2$  estimate effect size at the group or variable level only
- However, it is often of interest to estimate differences at the case level
- Case-level indexes of group distinctiveness are proportions of scores from one group versus another that fall above or below a reference point
- Reference points can be relative (e.g., a certain number of standard deviations above or below the mean in the combined frequency distribution) or more absolute (e.g., the cutting score on an admissions test)

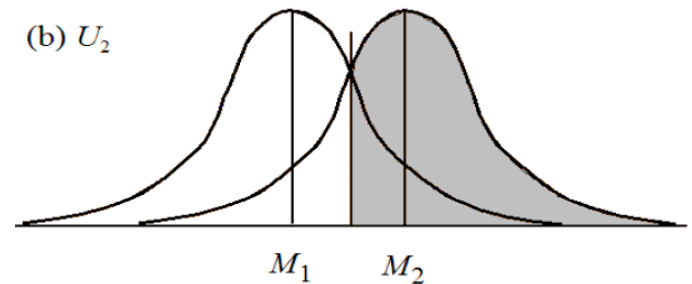
# Case-level effect size estimation

- Cohen's (1988) measures of distribution overlap:

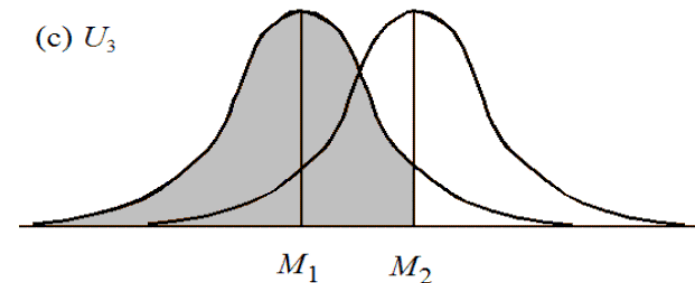
$U_1$ : Proportion of nonoverlap



$U_2$ : Proportion of scores in lower group exceeded by same proportion in upper group



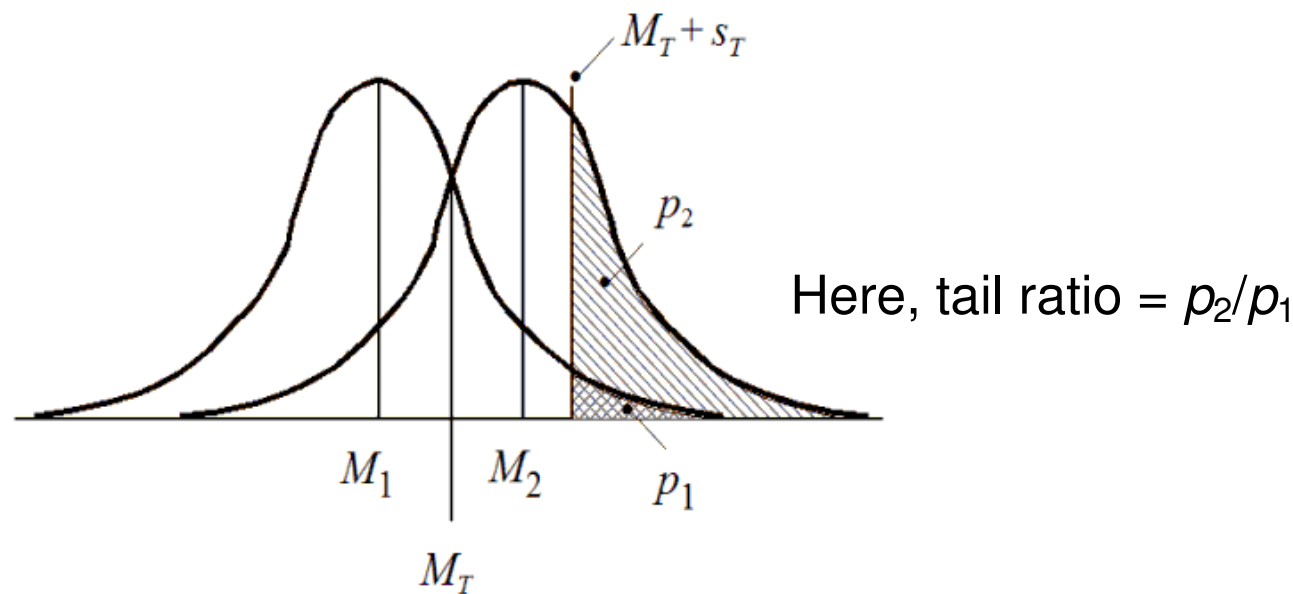
$U_3$ : Proportion of scores in lower group exceeded by typical score in upper group—probably the most generally useful



## Case-level effect size estimation

---

- Tail ratios (Feingold, 1995): Relative proportion of scores from two different groups that fall in the upper extreme (i.e., either the left or right tail) of the combined frequency distribution
- “Extreme” is usually defined relatively in terms of the number of standard deviations away from the grand mean
- Tail ratio  $> 1.0$  indicates one group has relatively more extreme scores



## Case-level effect size estimation

---

- Common language effect size (CL; McGraw & Wong, 1992) is the *predicted* probability that a random score from the upper group exceeds a random score from the lower group
- CL is calculated for the case  $M_2 > M_1$  as the proportion of cases in a normal curve to the right of:

$$z_{CL} = \frac{-(M_2 - M_1)}{\sqrt{s_1^2 + s_2^2}}$$

$\sqrt{s_1^2 + s_2^2}$  estimates the standard deviation in the *theoretical* distribution of differences between pairs of independent scores

- This computational method assumes normality and homogeneity of variance, but seems reasonably robust against violations of these assumptions

## Case-level effect size estimation

---

- Hess, Olejnik, & Huberty (2001) describe the use of *classification analysis* to estimate the degree of group overlap
- Classification is an optional part of both discriminant function analysis (DFA) and logistic regression (LR)
- Both techniques can analyze a group contrast on one or more continuous outcomes (i.e., the analysis can be univariate)
- DFA is better known but has more assumptions, such as homogeneity of variances and covariances across the groups

## Case-level effect size estimation

---

- This method estimates the reduction in classification error rates compared with chance classification with the following index:

$$I = \frac{H_o - H_e}{1 - H_e}$$

$H_o$  is the observed hit rate

$H_e$  is the hit rate expected by chance

## Interval estimation for effect sizes

---

- Some effect size statistics, such as Hedges's  $g$  and  $\hat{\eta}^2$ , have complex sampling distributions
- Traditional methods of interval estimation rely on asymptotic standard errors (i.e., approximate standard errors assuming large sample sizes)
- Examples: See formulas in Table 4.5 for  $d$  statistics

## Interval estimation for effect sizes

---

- However, formulas for approximate standard errors that can be calculated by hand are not always available (e.g., for  $\hat{\eta}^2$ )
- An alternative is to calculate noncentral confidence intervals for effect sizes, which are more exact but require computer programs
- These programs estimate values of the appropriate noncentrality parameter for the appropriate distribution, such as  $t$ , that correspond to lower and upper bounds of the confidence interval

## Interval estimation for effect sizes

---

- Examples of computer programs or macros for calculating noncentral confidences for effect sizes:

- ✓ ESCI ( $d$  only; see Cumming & Finch, 2001):

<http://www.latrobe.edu.au/psy/esci/index.html>

- ✓ Power Analysis module in STATISTICA ( $g$ ,  $\hat{\eta}^2$ ; see Steiger & Fouladi, 1997):

<http://www.statsoftinc.com/>

## Interval estimation for effect sizes

---

- Examples of computer programs or macros for calculating noncentral confidences for effect sizes:
  - ✓ SAS/STAT macros in text (see Tables 4.6, 4.7, 6.6)
  - ✓ SPSS macros by Smithson (2001;  $\hat{\eta}^2$  and partial  $\hat{\eta}^2$  only):

<http://www.anu.edu.au/psychology/people/smithson/details/CIstuff/CI.html>

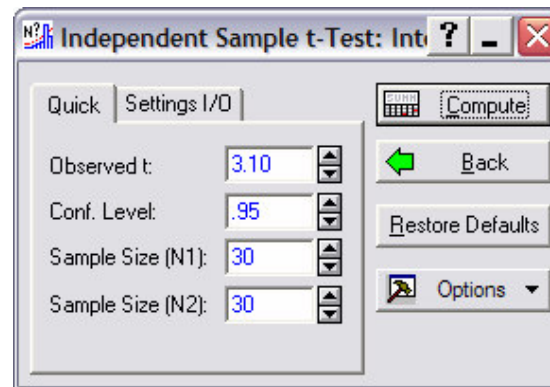
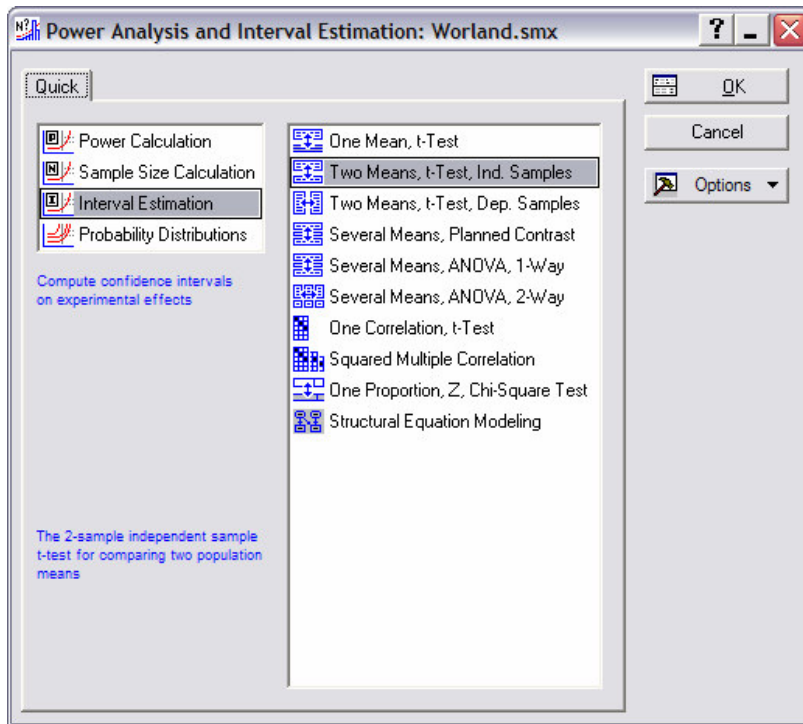
- ✓ See also Thompson (2002) and Fidler and Thompson (2001)

# Interval estimation for effect sizes

- Example with Power Analysis module in STATISTICA:

$$n_1 = n_2 = 30, g = .80, t(58) = 3.10$$

- Screenshots for automatic calculation of the 95% noncentral confidence interval .2707 to 1.3237 in  $d$  units in three steps:



Interval Estimation	
Two Means, t-Test	
	Value
Observed t-Statistic	3.1000
Sample Size N1	30.0000
Sample Size N2	30.0000
Confidence Level	0.9500
Confidence Limits:	
Delta:	
Lower Limit	1.0484
Upper Limit	5.1268
Standardized Effect:	
Lower Limit	0.2707
Upper Limit	1.3237

# References

---

- Abelson, R. P. (1997). A retrospective on the significance test ban of 1999 (If there were no significance tests, they would be invented). In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 117-141). Mahwah, NJ: Erlbaum.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.
- Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61*, 532-574.
- Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods, 1*, 170-177.
- Feingold, A. (1995). The additive effects of differences in central tendency and variability are important in comparisons between groups. *American Psychologist, 50*, 5-13.
- Fidler, F., & Thompson, B. (2001). Computing correct confidence intervals for ANOVA fixed- and random-effects effect sizes. *Educational and Psychological Measurement, 61*, 575-604.
- Hess, B., Olejnik, S., & Huberty, C. J. (2001). The efficacy of two-improvement over chance effect sizes for two-group univariate comparisons under variance heterogeneity and nonnormality. *Educational and Psychological Measurement, 61*, 909-936.
- Huberty, C. J. (2002). A history of effect size indices. *Educational and Psychological Measurement, 62*, 227-240.
- Hunter, J. E., & Schmidt, F. L. (1994). Correcting for sources of artificial variation across studies. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 323-336). New York: Russell Sage Foundation.
- Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data*. Belmont, CA: Wadsworth.
- McGraw, K. O., & Wong, S. P. (1992). A common language effect-size statistic. *Psychological Bulletin, 111*, 361-365.

- Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology, 25*, 241-286.
- Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin, 112*, 160-164.
- Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231-244). New York: Russell Sage Foundation.
- Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement, 61*, 605-632.
- Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. *Journal of Experimental Education, 61*, 334-349.
- Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 221-257). Mahwah, NJ: Erlbaum.
- Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31*, 25-32.