Institute of Education Sciences
National Center for Education Statistics

NATIONAL INSTITUTE OF STATISTICAL SCIENCES TASK FORCE REPORT

EFFECT SIZE

TABLE OF CONTENTS

NATIONAL INSTITUTE OF STATISTICAL SCIENCES
TASK FORCE REPORT

EFFECT SIZE, UNCERTAINTY, COMPLETENESS

EXECUTIVE SUMMARY

The Task Force was convened at ESSI in Washington, DC on December 11, 2006. Also present were NCES Chief Statistician Marilyn Seastrom and Special Assistant to the Commissioner Andrew White. Presentations to the Task Force were made by NCES staff members Chris Chapman, Bill Tirre and John Wirt.

Following the meeting the Task Force formulated an initial set of recommendations, which were refined and finalized through a series of e-mail interchanges. These appear in final form in this report.

Summary: The Task Force strongly supports reporting of effect sizes by NCES while recognizing that in some instances this may be inappropriate, ineffective or even impossible. Also, the Task Force acknowledges that effect sizes are attractive because they are dimensionless, this can also raise issues of interpretability that cannot be dismissed lightly.

Recommendation: NCES should routinely report effect sizes for differences in means and for categorical comparisons, unless there is compelling reason not to.

Recommendation: A chief exception for reporting is when the absolute difference is not statistically significant or when it is below an a priori designated detection threshold. In such cases, effect sizes should not be reported.

Recommendation: NCES tables that contain effect sizes should not also contain the actual differences in means or proportions (categorical comparisons).

Recommendation: NCES should identify the circumstances in which reporting uncertainties associated with effect sizes improves quality and usability and implement this practice.

Recommendation: NCES should evaluate the feasibility of developing and implementing a sensible, consistent mechanism for calling attention to large effect sizes.

NATIONAL INSTITUTE OF STATISTICAL SCIENCES TASK FORCE REPORT

PREFACE

The Task Force was charged by NCES to assess whether - and if so, how - the NCES should report effect sizes in its publications. Specific questions to be addressed were:

- For which results should which NCES data collection programs report effect sizes?
- What are appropriate measures of effect sizes for particular results?
- In what way(s) should effect sizes be presented (including visualizations) and interpreted in NCES publications?

The Task Force met in person at ESSI in Washington, DC on December 11, 2006. Subsequently the Task Force discussed and finalized a series of recommendations via e-mail.

NATIONAL INSTITUTE OF STATISTICAL SCIENCES
TASK FORCE REPORT

EFFECT SIZE, UNCERTAINTY, COMPLETENESS

I. INTRODUCTION

The Task Force was charged by NCES to assess whether - and if so, how - the NCES should report effect sizes in its publications. The Task Force met in person at ESSI in Washington, DC on December 11, 2006. Also present were NCES Chief Statistician Marilyn Seastrom and Special Assistant to the Commissioner Andrew White. Presentations to the Task Force were made by NCES staff members Chris Chapman, Bill Tirre and John Wirt.

Following the meeting the effect size literature was surveyed and the Task Force formulated an initial set of recommendations, which were refined and finalized through a series of e-mail interchanges. These recommendations follow in final form with explanations or elaborations to provide greater detail. The survey is summarized in Appendix C.

II. RECOMMENDATIONS

In general, the Task Force strongly supports reporting of effect sizes by NCES, but realizes that there are instances in which doing so may not be appropriate, ineffective or even impossible. At the same time, the Task Force acknowledges that the very dimensionlessness that makes effect sizes attractive for some purposes raises issues of interpretability that cannot be dismissed lightly. For instance, most people can understand and interpret a difference of $250, but not an associated effect size of 0.85.

The major recommendations of the Task Force address differences in means, differences in category proportions, uncertainties in effect sizes and calling out of "large" effect sizes.

Although this report focuses reporting of effect sizes by NCES, the Task Force hopes that NCES will take leadership in encouraging users of its databases to follows its own example.

III. DIFFERENCES IN MEANS

Recommendation 1: The Task Force recommends that NCES routinely report effect sizes for differences in means, unless there is compelling reason not to.

Elaboration: The case for reporting effect sizes is strongest when the underlying values lack physical interpretability[1]: if absolute differences have no straightforward interpretation, there is only benefit reporting effect sizes instead. In this context, Recommendation 1 can be construed as "report effect sizes for differences in means instead of differences in means."

When underlying values have strong physical interpretability, Recommendation 1 might be phrased as "report effect sizes for differences in means in addition to differences in means," although this may be too inflexible. How to implement "in addition to" may be program-, report-, or audience-specific. (Sub-Recommendation 1.2 addresses a specific point associated with reporting both.)

[1] An example is assessment scores, as compared, for instance, to dollar amounts or student enrollments.

This recommendation raises no significant presentation issues. Table 1, in appendix A, shows how effect sizes measuring state-level year-by-year differences in (say) NAEP scores can be displayed in the same format as absolute differences would be displayed.

Sub-Recommendation 1.1: In general, the Task Force favors as effect sizes measures for differences of means standardized differences of the form $(\mu_1-\mu_2)/\sigma$, where μ_1 and μ_2 are the two estimated means and σ is an estimated standard deviation.

Elaboration: The Task Force finds that existing technologies (see appendix D) for calculating effect sizes for differences in means can fulfill most if not all of NCES' needs. Differences among methods of this general form arise from different estimates σ that reflect independence assumptions about the two groups being compared. Cases where dependence is an issue include changes over time in longitudinal studies and group-to-subgroup comparisons. NAEP may carry dependences too complex to be captured exactly by existing methods, but it is likely to existing methods will provide acceptable approximations.

Because they are statistical estimates, effect size measures themselves have associated uncertainties, as discussed below.

NCES provided the Task Force with program-specific material regarding effect sizes that varied dramatically with respect to completeness and specificity. Those proposals that were complete and detailed seem reasonable.

Sub-Recommendation 1.2: The Task Force recommends that NCES *not report effect sizes for differences in means if the associated (absolute) difference is not statistically significant, or if it is below a designated detection threshold.*

Elaboration: The "(absolute) differences" in this case are unstandardized differences. Detection thresholds are anticipated to depend on:

- Domain knowledge: what is a substantive difference?
- The design of the data collection: effect sizes smaller than what a survey is designed to detect should not be reported.
- Reporting practices: a n effect size that would be reported as zero because of rounding should not be reported.

Sub-Recommendation 1.3: The Task Force recommends that NCES *tables containing effect sizes for differences of means <u>not</u> also contain the actual difference in means*.

Elaboration: The rationale is that inclusion of actual differences is self-defeating: use of effect sizes implies that actual differences are misleading. If so, they should not be presented. This recommendation leaves unaddressed how it should be implemented in the "in addition to" version of Recommendation 1: it only says how it should not be implemented.

IV. DIFFERENCES IN CATEGORY PROPORTIONS

Recommendation 2: The Task Force recommends that NCES *routinely report effect sizes for categorical comparisons,* unless there is compelling reason not to.

Elaboration: The preferred measures are standardized differences in proportions, although in some circumstances there may be preferred alternatives.

Sub-Recommendation 2.1: The Task Force recommends that NCES *not report effect sizes for differences in proportions if the associated (absolute) difference is not statistically significant, or if it is below a designated detection threshold.*

Sub-Recommendation 1.3: The Task Force recommends that NCES *tables containing effect sizes for differences of proportions not also contain the actual difference in proportions*.

V. UNCERTAINTIES IN EFFECT SIZES

Recommendation 3: The Task Force recommends that NCES *identity circumstances under which reporting uncertainties associated with effect sizes improves quality and usability, and do so in such cases.*

Elaboration: This is the most complex recommendation, in part because it places a significant burden on NCES or contractors. Circumstances under which quality and usability are improved depend on multiple factors, including diverse audiences and purposes for NCES publications, in ways that seem to preclude succinct summary.

The Task Force finds that there is currently no consensus regarding how to convey uncertainties to literate but not statistically sophisticated individuals. Potential methods, which are largely identical in terms of mathematical content, include:

- Confidence intervals,
- Effect size ± uncertainty, where uncertainty, typically, is confidence interval half-width,
- Effect size ± uncertainty/effect size (i.e., effect size ± p%),
- Graphical methods (e.g., map effect uncertainty onto color, although this may conflict with existing report and web site standards).

There are counterbalancing issues of "information overload." The inclusion of uncertainties in the form of the second measure above, as in table 2, increases the "busyness" significantly, as compared to table 1, and may decrease usability for those not interested in or not able to assimilate uncertainties. In table 3, the uncertainties are colored according to size. The point is not that either of tables 2 and 3 is good, but rather to illustrate the nature of the problem.

VI. LARGE EFFECT SIZES

Recommendation 4: The Task Force recommends that NCES *evaluate the feasibility of developing and implementing a sensible, consistent mechanism for calling attention to large effect sizes.*

- **Elaboration:** Here, "large" is an analog of "statistically significant" in other contexts. This is distinct from the "above the detection threshold" criterion discussed in Sub-Recommendation 1.3. The Task Force finds two classes of measures to be meaningful:
- "Large" represents exceeding a touchstone - a scientifically defined important difference, such as one year's progress in an assessment context.
- "Large" represents extreme relative to some population of effect sizes, for example, among the *x*% of effect sizes in a report, or in a collection of similar reports.

The Task Force feels that scientifically-based touchstones, when they can be identified, are preferable. Among the problematic aspects of "extreme relative to some population" is defining that population, that whether an effect size is large depends on how many other effect sizes are reported, and that spurious large effects are possible.

The Task Force acknowledges that Recommendation 4 poses other difficulties, for example that "large" may be interpreted as "important," or even "causal." That "Large" could have different interpretations in different contexts is also a potential problem.

The implementation component of Recommendation 4 is not the difficult one. To illustrate, Table 4 in appendix C,[2] highlights using **boldface** all effect sizes whose absolute value exceeds a touchstone of 1.0.

The Task Force urges that NCES *not employ arbitrarily defined cutoffs or value-laded characterizations to define "large" effect sizes*, as has been proposed in some papers in the literature.

Other Items: Some discussions during and following the December 11, 2006 Task Force meeting raised points that the Task Force wishes to call to NCES' attention, but did not lead to specific recommendations.

- There is little conceptual, methodological or empirical research concerning how, as discussed above, "to convey uncertainties to literate but not statistically sophisticated individuals." It may be necessary for NCES to await such research prior to acting on Recommendation 3 in some settings.
- The Task Force discussed alternatives to the term "effect size." For non-statistical audiences, the term has no particular content, and "effect" always raises the possibility of "cause." The ungainly term "Comparable Differences" captures the essence of putting differences in a format in which they can be compared in a principled way, but it is difficult to imagine its being adopted widely. "Dimensionless Difference" also both captures the essence and seems unlikely to catch on.
- The Task Force also considered broader issues regarding NCES' presentation of data summaries in tabular form, which in a strict sense go beyond its charge. In addition to uncertainties these alternate views (for instance, tables sorted by effect size, or reordering of hierarchies) and interactivity.
- The Task Force notes that effect sizes carry multiplicity issues similar to those associated with hypothesis testing and statistical significance, although there seems to be only a modest literature on the subject. To some extent, the issues are attenuated - or, depending on one's point of view, obscured - in the same way as when p-values rather than significance are reported. However, they are overt if "large" effect sizes are defined relative to populations of effect sizes.

[2] Containing the same numerical values as tables 1-3.

APPENDICES

Appendix A: Displaying Effect Sizes in Tables

Example (hypothetical entries)

State	Change 00-01	Change 01-02	Change 02-03	Change 03-04	Change 04-05
AL	1.23	1.35	-2.20	1.21	3.33
AK	-1.11	-2.22	-1.11	-2.22	-1.11
AZ	0.31	0.32	0.33	0.34	0.35
AR	-1.10	-1.00	-0.90	-0.80	-0.70
CA	-1.00	1.00	-1.00	1.00	-1.00
CO	-0.50	-1.00	-2.00	-4.00	-8.00
CT	NS	NS		1.00	1.67
DE	1.23	4.56	7.89	-4.56	-1.23
…	…	…	…	…	…

Table 1: Table containing effect sizes (fictitious values) for state-by-state annual changes in mean NAEP scores.

Inclusion of Uncertainties in Tables – Two Examples

State	Change 00-01	Change 01-02	Change 02-03	Change 03-04	Change 04-05
AL	1.23 ± .05	1.35 ± .10	-2.20 ± .15	1.21 ± .02	3.33 ± .01
AK	-1.11 ± .20	-2.22 ± .25	-1.11 ± .22	-2.22 ± .23	-1.11 ± .19
AZ	0.31 ± .11	0.32 ± .12	0.33 ± .12	0.34 ± .12	0.35 ±.01
AR	-1.10 ± .30	-1.00 ± .29	-0.90 ± .25	-0.80 ± .25	-0.70 ± .27
CA	-1.00 ± .02	1.00 ± .02	-1.00 ± .01	1.00 ± .02	-1.00 ± .02
CO	-0.50 ± .20	-1.00 ± .10	-2.00 ± .15	-4.00 ± .12	-8.00 ± .24
CT	NS	NS	NS	1.00 ± .17	1.67 ± .15
DE	1.23 ± .21	4.56 ± .22	7.89 ± .22	-4.56 ± .21	-1.23 ± .20
…	…	…	…	…	…

Table 2: Table containing effect sizes and associated uncertainties (fictitious values) for state-by-state annual changes in mean NAEP scores.
NS means "not a substantive difference."

State	Change 00-01	Change 01-02	Change 02-03	Change 03-04	Change 04-05
AL	1.23 ± .05	1.35 ± .10	-2.20 ± .15	1.21 ± .02	3.33 ± .01
AK	-1.11 ± .20	-2.22 ± .25	-1.11 ± .22	-2.22 ± .23	-1.11 ± .19
AZ	0.31 ± .11	0.32 ± .12	0.33 ± .12	0.34 ± .12	0.35 ±.01
AR	-1.10 ± .30	-1.00 ± .29	-0.90 ± .25	-0.80 ± .25	-0.70 ± .27
CA	-1.00 ± .02	1.00 ± .02	-1.00 ± .01	1.00 ± .02	-1.00 ± .02
CO	-0.50 ± .20	-1.00 ± .10	-2.00 ± .15	-4.00 ± .12	-8.00 ± .24
CT	NS	NS	NS	1.00 ± .17	1.67 ± .15
DE	1.23 ± .21	4.56 ± .22	7.89 ± .22	-4.56 ± .21	-1.23 ± .20
…	…	…	…	…	…

Table 3: Table containing effect sizes and associated uncertainties (fictitious values) for state-by-state annual changes in mean NAEP scores, with uncertainties colored according to size. Blue corresponds to uncertainties less than .10, green to uncertainties from .10 to .19 and red to uncertainties of .20 or greater.

Highlighting "Large" Effect Sizes

State	Change 00-01	Change 01-02	Change 02-03	Change 03-04	Change 04-05
AL	**1.23**	**1.35**	**-2.20**	**1.21**	**3.33**
AK	**-1.11**	**-2.22**	**-1.11**	**-2.22**	**-1.11**
AZ	0.31	0.32	0.33	0.34	0.35
AR	**-1.10**	**-1.00**	-0.90	-0.80	-0.70
CA	**-1.00**	**1.00**	**-1.00**	**1.00**	**-1.00**
CO	-0.50	**-1.00**	**-2.00**	**-4.00**	**-8.00**
CT	NS	NS	NS	**1.00**	**1.67**
DE	**1.23**	**4.56**	**7.89**	**-4.56**	**-1.23**
…	…	…	…	…	…

Table 4: Table containing effect sizes (fictitious values) for state-by-state annual changes in mean NAEP scores, with "large" effect sizes exceeding 1.0 in absolute value in boldface.

Appendix B: Literature Survey on Effect Sizes

Following Kline (2004) effect size indexes are classified into parametric indexes or nonparametric indexes, corresponding to the response/outcome variable being continuous or respectively, categorical. Parametric effect size indexes are further classified into measures of association, standardized mean differences, and case-level (as opposed to group-level) effect sizes. The nonparametric effect size indexes are further classified into effect size indexes for 2 x 2 tables, and effect size indexes for larger (than 2 x 2) tables.

Presently there is no consistent terminology for the classification of effect size indexes. Alternative names for the measures of association category from Kline (2004) include relationship indexes [Huberty (2002)] and the *r* (relationship) family [Rosenthal (1994)]. Similarly, the category of standardized mean differences has been referred to as group difference indexes [Huberty (2002)] or the *d* (difference) family [Rosenthal (1994)]. The category of case-level effect sizes has been mentioned previously by Huberty (2002) as group overlap indexes.

In addition to the previously described inconsistency regarding classification, it should also be noted that the definitions and the names of the effect size indexes do not always recognize the presence of a population-level effect size index and its sample-based estimators.

The measures of association can be further classified into unsquared measures of association (referred to as correlational indexes in Huberty (2002)) and squared measures of association (referred to as explained variance indexes in Huberty (2002)). Correlational indexes include Pearson's correlation coefficient ρ (and its estimator r) for the situation when the dependent variable is continuous, and the point-biserial correlation coefficient r_{pb} [Friedman (1968)] for the situation when the dependent variable is dichotomous (i.e. the situation of two independent populations). Explained variance indexes include the squares of previously described correlational indexes, η^2 [Pearson (1905)], ε^2 [Kelley (1935)], ω^2 [Hays (1963)], and ρ_I, the intraclass correlation coefficient. The general formula for η^2 is

$$\eta^2 = \frac{\sigma^2_{effect}}{\sigma^2_{total}}$$

the proportion of the total population variance explained by the effect of interest. In the simplest ANOVA situation, the estimator for η^2 is the correlation ratio R^2.

There are also partial η^2 and ω^2 [Keppel (1991)] and generalized η^2 and ω^2 [Olejnik and Algina (2003)], the later proposed in an effort to define effect size measures that are comparable across different designs involving blocking, covariates, and additional factors. Olejnik and Algina (2003) present generalized η^2 and ω^2 for between-subject designs, analysis of covariance, repeated measures designs, and mixed designs.

The Binomial Effect Size Display (BESD) is a popular way to recast correlational indexes into the 2 x 2 table framework [Rosenthal and Rubin (1982)]. Starting from the correlation coefficient r, the method produces a 2 x 2 table where the success rates difference (SRD) equals r. The two success rates in the conceptual 2 x 2 table are

$$0.50 \pm \frac{r}{2}$$

There are also partial η^2 and ω^2 [Keppel (1991)] and generalized η^2 and ω^2 [Olejnik and Algina (2003)], the later proposed in an effort to define effect size measures that are comparable across different designs involving blocking, covariates, and additional factors. Olejnik and Algina (2003) present generalized η^2 and ω^2 for between-subject designs, analysis of covariance, repeated measures designs, and mixed designs.

The Binomial Effect Size Display (BESD) is a popular way to recast correlational indexes into the 2 x 2 table framework [Rosenthal and Rubin (1982)]. Starting from the correlation coefficient r, the method produces a 2 x 2 table where the success rates difference (SRD) equals r. The two success rates in the conceptual 2 x 2 table are

$$0.50 \pm \frac{r}{2}$$

BESD has been extended to situations involving more than 2 groups [Rosnow, Rosenthal, and Rubin (2000)]. Hsu (2004) argues that there is overestimation of the real SRDs and recommends the stochastic difference measure δ [Cliff (1993)] as an alternative.

A simple effect size index, named requivalent, has also been proposed by Rosenthal and Rubin (2003) for situations when only the sample size and the p value (especially from a nonparametric procedure for which there is no accepted effect size index) are known. The effect size equals the sample point-biserial correlation rpb from a two-treatment group experiment with equal sample sizes per group and a normally distributed outcome, which produces the same p value when the t test is applied.

For the simplest situation of two independent groups and a univariate response/outcome variable, the general formulas for the population-level effect size and a corresponding estimator are

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

$$d = \frac{\overline{X}_1 - \overline{X}_2}{\hat{\sigma}}$$

The standard deviation σ used in the definition of the population-level effect size index can be the assumed common standard deviation of the two populations, or the standard deviation for one of the populations. Even for the same population-level effect size index different estimators may arise by using different estimators of σ. Examples include Cohen's d [Cohen (1969)], Hedges' g [Hedges (1981)], and Glass's Δ [Glass (1976)]. Cohen's d and Hedges' g are using different estimators for the (assumed) common standard deviation, Glass's Δ is defined using the standard deviation of the control group. According to Rosnow and Rosenthal (1996) Cohen's d is a descriptive measure and Hedges' g is an inferential measure.

For the situation of two dependent groups there is no general agreement regarding which standard deviation should be used for standardization, the standard deviation of the original scores, as recommended by Dunlop et al. (1996), or the standard deviation of the difference scores, as recommended by Rosenthal (1991). One argument for the use of the standard deviation of the original scores is to make the effect size index directly comparable with the effect size from the situation involving two independent populations.

The case-level effect sizes will be described in more detail since most of the literature has emphasized mostly with the first two categories, i.e. the measures of association, and the standardized mean differences. This category of effect size indexes may be especially useful for ordinal variables.

The case-level effect size indexes include the three U measures of distribution overlap proposed in Cohen (1988) (U_1 is the proportion of no overlap, U_2 is the proportion of scores in the lower group exceeded by the same proportion in the upper group, and U_3 is the proportion of scores in the lower group exceeded by the typical score (e.g. median) in the upper group), tail ratios (relative proportion of scores that fall in the upper (or lower) tail of the combined distribution) [Feingold (1995)], the common language (CL) effect size indicator [McGraw and Wong (1992)], the probability of superiority (PS) [Grissom (1994), Acion et al (2006)], the A measure of stochastic superiority [Vargha and Delaney (2000), Newcombe (2006)], the stochastic difference measure δ [Cliff (1993)], Agresti's α [Agresti (1980)], Levy's probability of misclassification P(m) [Levy (1967)], better-than-chance (improvement-over-chance) classification index I [Huberty (1994), Huberty and Lowman (2000), Hess, Olejnik, and Huberty (2001)], and the probability of correct classification Q [Wilcox and Muska (1999)].

The CL effect size indicator is defined as

$$P(X_1 > X_2)$$

under assumed normality and homoscedasticity. The calculation of the (sample) CL involves the (sample) means and the (sample) standard deviations. PS is defined similarly, although in a more general context, and its estimation is based on the Mann-Whitney U statistic. The related A measure of stochastic superiority

$$A_{12} = P(X_1 > X_2) + 0.5P(X_1 = X_2)$$

explicitly deals with the ties, and is equivalent with the area under the curve (AUC) and the mean ridit.

The stochastic difference is defined as the difference

$$\delta = P(X_1 > X_2) - P(X_2 > X_1)$$

By comparison Agresti's α is defined as the ratio

$$\frac{P(X_1 > X_2)}{P(X_2 > X_1)}$$

The better-than-chance classification index I involves the observed and the expected hit rate

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

$$A_{12} = P(X_1 > X_2) + 0.5P(X_1 = X_2)$$

and it is based on predictive discriminant analysis (PDA) or logistic regression analysis (LRA). These methods have been extended to multigroup and/or multivariate situations with heterogeneous covariance matrices. By contrast, the Q index from Wilcox and Muska (1999) is based on classification methods that involve kernel density estimators. For the comparison of two groups, independent or dependent case, four estimators of Q are considered: the so-called apparent estimator, the leave-one-out cross-validation estimator, the basic bootstrap estimator, and a bias-adjusted bootstrap estimator, that the authors recommend.

Nonparametric effect size indexes for the 2 x 2 table include the four measures of effect size for categorical data described by Fleiss (1994): the difference between two probabilities (the risk difference or SRD)

$$RD = \pi_1 - \pi_2$$

the ratio of two probabilities (the risk ratio or relative risk)

$$RR = \frac{\pi_1}{\pi_2}$$

the ϕ coefficient, i.e. the Pearson correlation coefficient for two dichotomous variables, with its estimator closely related to the classical chi-squared statistic

$$\hat{\varphi} = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

and the odds ratio (the cross-product ratio)

$$OR = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}$$

Fleiss (1994) recommends the odds ratio as the effect size index of choice for 2 x 2 tables. Rosenthal (1994) includes ϕ in the *r* family and the difference between proportions in the *d* family of effect sizes.

The reciprocal of SRD, known as the number needed to treat (NNT), is also a useful effect size index in the context of clinical trials. An extended NNT can be defined as the reciprocal of the stochastic difference measure δ, since the later can be regarded as an extended SRD, according to Kraemer and Kupfer (2005).

For the case of larger tables, the most common effect size index is Cramér's V, an extension of the estimator of the ϕ coefficient.

Several authors have described the relationships and conversion formulas between various effect size indexes. Rosenthal (1994) and Fern and Monroe (1996) describe useful transformations between standardized mean difference measures, correlational measures, and the *t* statistic. Fern and Monroe (1996) also describe transformations among measures of explained variance and the *F* statistic. Rice and Harris (2005) provide a table to help conversions between AUC, Cohen's d, and the point-biserial correlation coefficient r_{pb}, the three common effect sizes used in follow-up studies.

Olejnik and Algina (2000) present effect size indexes for more complex group comparison studies. The standardized mean difference is described for univariate between-subject designs (single-factor designs, multifactor designs, and single-factor designs with covariates), the multivariate single-factor between-subject design, within-subject designs, and split-plot designs. Proportion of variance effect size indexes are described for univariate and multivariate between-subject designs (single-factor designs, multifactor designs, and single-factor designs with covariates), within-subject designs, and split-plot designs.

Additional classification categories may be needed if there is further interest to separate the original effect size indexes from their corrected versions [Hunter and Schmidt (1990), Vacha-Haase and Thompson (2004)] or robust versions [Kraemer and Andrews (1982), Hedges and Olkin (1984), Algina, Keselman, and Penfield (2005)].

Reasons for correction/adjustment include bias reduction, accounting for measurement error, etc. Hunter and Schmidt (1990) describe adjustment methods to account for reliability issues, dichotomization of continuous variables, range restriction, construct validity problems, and unequal sample sizes. The impact of correction is likely to be small if the sample size is very large, the number of measured variables is small, and the (unknown) population-level effect size is large.

The robust versions of effect size indexes replace the mean with the median, the trimmed mean, or the Winsorized mean, and also replace the standard deviation with the range, linear combination of order statistics, median absolute deviation, or the square root of the Winsorized variance. Alternative effect size indexes that under normality reduce to the standardized mean difference, in the spirit of the nonparametric estimator of effect size from Kraemer and Andrews (1982), have been proposed by Hedges and Olkin (1984). For the simplest two independent groups design one option is $\Phi^{-1}(p)$ where p is the sample proportion of units in the control group exceeded by the median of the experimental group (same as Cohen's U_3) and Φ is the standard normal distribution function.

Algina, Keselman and Penfield (2005) propose a robust version of the standardized mean difference measure obtained by replacing the population means with 20% trimmed means and the population standard deviation with the square root of a 20% Winsorized variance

$$\delta_R = 0.642\left(\frac{\mu_{t1} - \mu_{t2}}{\sigma_W}\right)$$

The factor 0.642 ensures that $\delta_R = \delta$ for normal data with equal variances.

Although the previous discussion has focused on point estimators for population-level effect size indexes, it is very important to consider interval estimation as well. The construction of parametric confidence intervals for the standardized mean difference effect size using methods based on the noncentral *t* distribution has been described by Steiger and Fouladi (1997) and Cumming and Finch (2001). The later authors provide software for the implementation of their method, the so-called ESCI (Exploratory Software for Confidence Intervals).

New parametric confidence intervals for the standardized mean difference effect size are proposed by Wu, Jiang, and Wei (2006) under assumed normality and homoscedasticity. Their confidence intervals are based on a modified signed log-likelihood method and have better coverage properties than previously proposed methods.

Kelley (2005) investigates the effect of nonnormal distributions on parametric and bootstrap confidence intervals for the standardized mean difference effect size. The parametric confidence interval considered is analogous to the method described in Steiger and Fouladi (1997). The bootstrap methods considered included the percentile method and the bias-corrected and accelerated method. The author recommends the second bootstrap confidence interval for general use, especially when normality does not hold.

Algina, Keselman and Penfield (2005) propose two types of confidence intervals for their robust effect size measure, one based on the noncentral *t* distribution and the other on the percentile bootstrap. In their simulation study the percentile bootstrap confidence intervals enjoyed better coverage probability.

Appendix C: References

Acion, L., Peterson, J. J., Temple, S., Arndt, S. (2006). Probabilistic index: an intuitive non-parametric approach to measuring the size of effect treatments. *Statistics in Medicine*, 25:591-602.

Algina, J., Keselman, H. J., and Penfield, R. D. (2005). An alternative to Cohen's standardized mean difference effect size: a robust parameter and confidence interval in the two independent groups case. *Psychological Methods*, 10:317-328.

Agresti, A. (1980). Generalized odds ratios for ordinal data. *Biometrics*, 36:59-67.

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, 114:494-509.

Cohen, J. (1969). *Statistical Power Analysis for the Behavioral Sciences.* Academic Press, New York.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd edition). Academic Press, New York.

Cumming, G., and 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.

Dunlop, W. P., Cortina, J. M., Vaslow, J. B., and 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*, 1:170-177.

Fern, E. F., and Monroe, K. B. (1996). Effect-size estimates: issues and problems in interpretation. *Journal of Consumer Research*, 23:89-105.

Fleiss, J. L. (1994). Measures of effect size for categorical data. In H. Cooper and L. V. Hedges (Eds.), *The Handbook of Research Synthesis* (pp. 245-260), Russell Sage Foundation, New York.

Friedman, H. (1968). Magnitude of the experimental effect and a table for its rapid estimation. *Psychological Bulletin*, 70:245-251.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, 5:3-8.

Grissom, R. (1994). Probability of the superior outcome of one treatment over another. *Journal of Applied Psychology*, 79:314-316.

Hays, W. L. (1963). *Statistics for Psychologists.* Holt, Rinehart and Winston, New York.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6:107-128.

Hedges, L. V., and Olkin, I. (1984). Nonparametric estimators of effect size in meta-analysis. *Psychological Bulletin*, 96:573-580.

Hess, B., Olejnik, S., and 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.

Hsu, L. M. (2004). Biases of success rate differences shown in Binomial Effect Size Displays. *Psychological Methods*, 9:183-197.

Huberty, C. J. (1994). *Applied Discriminant Analysis*. Wiley, New York.

Huberty, C. J. (2002). A history of effect size indexes. *Educational and Psychological Measurement*, 62(2):227-240.

Huberty, C. J., and Lowman, L. L. (2000). Group overlap as basis for effect size. *Educational and Psychological Measurement*, 60:543-563.

Hunter, J. E., and Schmidt, F. L. (1990). *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings.* Sage, Newbury Park, CA.

Kelley, T. L. (1935). An unbiased correlation ratio. *Proceedings of the National Academy of Sciences*, 21:554-559.

Kelley, K. (2005). The effects of nonnormal distributions on confidence intervals around the standardized mean difference: bootstrap and parametric confidence intervals. *Educational and Psychological Measurement*, 65:51-69.

Keppel, G. (1991). *Design and Analysis: A Researcher's Handbook*. Prentice-Hall, Englewood Cliffs, NJ.

Kline, R. B. (2004). *Beyond Significance Testing. Reforming Data Analysis Methods in Behavioral Research.* American Psychological Association, Washington, DC.

Kraemer, H. C., and Andrews, G. (1982). A nonparametric technique for meta-analysis effect size calculation. *Psychological Bulletin*, 91:404-412.

Kraemer, H. C., and Kupfer, D. J. (2005). Size of treatment effects and their importance to clinical research and practice. *Biological Psychiatry*, 59:990-996.

Levy, P. (1967). Substantive significance of significant differences between groups. *Psychological Bulletin*, 67:37-40.

McGraw, K. O., and Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, 111:361-365.

Newcombe, R. G. (2006). Confidence intervals for an effect size measure based on the Mann-Whitney statistic. Part 1: General issues and tail-area-based methods. *Statistics in Medicine*, 25:543-557.

Olejnik, S., and Algina, J. (2000). Measures of effect size for comparative studies: applications, interpretations, and limitations. *Contemporary Educational Psychology*, 25:241-286.

Olejnik, S., and Algina, J. (2003). Generalized eta and omega squared statistics: measures of effect size for some common research designs. *Psychological Methods*, 8:434-447.

Pearson, K. (1905). Mathematical contributions to the theory of evolution, XIV: On the general theory of skew correlation and non-linear regression. *Drapers' Company Research Memoirs*, Biometric Series II, Dulau, London.

Rice, M. E., and Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC area, Cohen's d, and r. *Law and Human Behavior*, 29:615-620.

Rosenthal, R. (1991). *Meta-analytic Procedures for Social Research*. Sage, Newbury Park, CA.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper and L. V. Hedges (Eds.), *The Handbook of Research Synthesis* (pp. 231-244), Russell Sage Foundation, New York.

Rosenthal, R., and Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology*, 74:166-169.

Rosenthal, R., and Rubin, D. B. (2003). $r_{equivalent}$: a simple effect size indicator. *Psychological Methods*, 8:492-496.

Rosnow, R.L., Rosenthal, R., and Rubin, D. B. (2000). Contrasts and correlations in effect-size estimation. *Psychological Science*, 11:446-453.

Steiger, J. H., and Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical methods. In L. L. Harlow, S. A. Mulaik, and J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 221-257). Lawrence Erlbaum, Mahwah, NJ.

Vacha-Haase, T., and Thompson, B. (2004). How to estimate and interpret various effect sizes. *Journal of Counseling Psychology*, 51:473-481.

Vargha, A., and Delaney, H. D. (2000). A critique and improvement of the CL common language effect size statistic of McGraw and Wong. *Journal of Educational and Behavioral Statistics*, 25:101-132.

Wilcox, R. R., and Muska, J. (1999). Measuring effect size: A non-parametric analogue of ω^2. *British Journal of Mathematical and Statistical Psychology*, 52:93-110.

Wu, J., Jiang, G., and Wei, W. (2006). Confidence intervals of effect size in randomized comparative parallel-group studies. *Statistics in Medicine*, 25:639-651.

Appendix D: Task Force Members

Mark Lipsey, Vanderbilt University

Stephen Olejnik, University of Georgia

Ingram Olkin, Stanford University

Bruce Spencer, Northwestern University

Bruce Thompson, Texas A&M University

Leland Wilkinson, Systat

Panel convened by National Institute of Statistical Sciences

Alan Karr, NISS

George Luta, NISS