Effect Size Indices for Artificially Dichotomized Variables Measured with Error: An Empirical Investigation of Accuracy and Precision

Isaac Li[1], Patricia Rodríguez de Gil[1], Jeanine Romano[1], Aarti P. Bellara[1], George MacDonald[1], Harold Holmes[1], Patrice Rasmussen[1], Yi-Hsin Chen[1], Jeffrey D. Kromrey[1]

[1]Educational Measurement and Research, University of South Florida, College of Education, 4202 East Fowler Avenue, EDU 105, Tampa, FL 33620

Abstract

Monte Carlo methods were used to investigate the accuracy and precision of effect size indices in estimating what the standardized mean difference from a 2 X 2 sample table of dichotomized variables would have been had the data not been dichotomized. Normally distributed, continuous data were generated for two groups and the continuous variable was dichotomized at specified cut points. The factors manipulated in the simulation study included overall sample size ($n_1 + n_2$ = 30, 60, 120, 240), reliability levels (.5, .7, .8, .9, 1), population effect size (0, .2, .5, .8), continuous score cut point for dichotomization (.10, .25, .40, .50, .70), and population variance ratio (1:1, 1:2, 1:4). For each sample generated (100,000 replications), each of seven proposed effect size indices was calculated. Both the statistical bias and the RMSE were computed across the set of replications. Although the sample standardized mean difference became substantially biased in the presence of measurement error, the performance of the seven indices was not notably affected. Results were interpreted in terms of recommendations for estimating effect sizes with dichotomized variables.

Key Words: effect sizes, simulation, reliability, statistical bias, dichotomy

1. Background

The topic of effect sizes can be a controversial issue for many journal editors, as well as for researchers. For instance, Pedhazur and Pedhazur-Schmelking (1991) argued that Cohen's (1988) convention of small, medium, and large effect sizes distorted the distinction between the magnitude of an effect and its substantive importance, i.e., researchers relegating "small" effects as less important or considering "large" effects as important. Also, the use and interpretation of a specific effect size across studies can be problematic due to the variability of research design factors, the last edition of the Publication Manual of the American Psychological Association (6th edition; 2010) as well as the 1999 report by Wilkinson and the APA Task Force on Statistical Inference have made clear the imperative for reporting effect sizes to supplement statistical hypothesis testing and ensure "accuracy of scientific knowledge" (p. 11). Effect sizes provide useful indices of the magnitudes of treatment effects in individual studies as well as representing the primary statistics that are used in synthesizing research or meta-analysis.

Cohen (1977) defined effect size (ES) as "the degree to which the phenomenon is present in the population" (pp. 9-10), or "the degree to which the H_0 is believed to be false"

(1992; p. 156). Each statistical test has its own ES index, which is scale free and continuous; however, to convey meaning of any ES index, it is necessary to have an idea of its scale, for which he proposed the small, medium, and large ES conventions.

One ES index of mean differences is *d*. That is, for studies comparing two groups on a continuous outcome variable, the standardized mean difference (also referred to as Cohen's *d*) is the most common index of effect size. In a population, this ES is obtained as

$$\delta = \frac{\mu_E - \mu_C}{\sigma}$$

where μ_E and μ_C are the means of the experimental and control populations, respectively, and σ is the common population standard deviation.

A sample estimate of this effect size (*d*) is obtained by substituting sample statistics for each of these parameters. The sampling variance of this estimate is given by

$$S_d^2 = \frac{n_E + n_C}{n_E n_C} + \frac{d^2}{2\left(n_E + n_C\right)}.$$

In medical and social science research, continuous outcome variables are often dichotomized using a cut score to delineate the two dichotomous values. For example, measures of blood pressure or depression may be obtained on a continuous scale, but are subsequently dichotomized to classify patients as hypertensive or clinically depressed. Similarly, educational achievement measures may be obtained as a continuous variable (number of test items correct or scale score on a standardized achievement test) and subsequently dichotomized into pass/fail. With such dichotomization, the standardized mean difference can no longer be directly estimated from sample data. A variety of statistics have been proposed to provide estimates of the standardized mean difference that would have been obtained if the data had not been dichotomized. These effect sizes have been compared in Monte Carlo studies reported by Sanchez-Meca, Marín-Martínez, and Chacón-Moscoso (2003), and by Kromrey and Bell (2012).

2. Alternative Methods for Computing Dichotomized Effect Sizes

The standardized proportion difference (d_p; Johnson, 1989) is a direct analogy to the standardized mean difference

$$d_p = \frac{P_E - P_C}{S'}$$

where P_E and P_C are the sample success (or failure) proportions of the experimental and control groups. The pooled standard deviation is obtained as

$$S' = \sqrt{\frac{(n_E - 1)P_E(1 - P_E) + (n_C - 1)P_C(1 - P_C)}{n_E + n_C - 2}}.$$

The sampling variance of this statistic is obtained as

$$S^2_{d_p} = \frac{n_E + n_C}{n_E n_C} + \frac{d_p^2}{2\left(n_E + n_C\right)}.$$

The phi coefficient (Φ) is commonly used as a measure of association for 2 X 2 tables. This index can be transformed to the scale of the standardized mean difference (Rosenthal, 1994). The phi coefficient is obtained as

$$\Phi = \frac{O_{1E}O_{2C} - O_{2E}O_{1C}}{\sqrt{n_E n_C m_1 m_2}}$$

where the O_{iE} and O_{iC} are the success and failure frequencies in the experimental and control groups, respectively, and the m_i are the marginal success and failure frequencies.

The transformation of Φ to the standardized mean difference metric is given by

$$d_\Phi = \frac{\Phi}{\sqrt{1-\Phi^2}}\sqrt{\frac{\left(n_E + n_C - 2\right)\left(n_E + n_C\right)}{n_E n_C}}$$

with sampling variance

$$S^2_{d_\Phi} = \frac{n_E + n_C}{n_E n_C\left(1-\Phi^2\right)^2}.$$

An arcsine transformation of the proportions in a 2 X 2 table was suggested by Cohen (1988)

$$d_{asin} = 2\arcsin\left(\sqrt{P_E}\right) - 2\arcsin\left(\sqrt{P_C}\right)$$

with sampling variance

$$S^2_{d_{asin}} = \frac{1}{n_E} + \frac{1}{n_C}.$$

Hasselblad and Hedges (1995) suggested a transformation of the log odds ratio to the metric of the standardized mean difference

$$d_{HH} = Ln(OR)\frac{\sqrt{3}}{\pi}$$

with sampling variance

$$S^2_{d_{HH}} = \frac{3}{\pi^2}\left(\frac{1}{O_{1E}} + \frac{1}{O_{2E}} + \frac{1}{O_{1C}} + \frac{1}{O_{2C}}\right).$$

A similar transformation of the log odds ratio was proposed by Cox (1970)

$$d_{Cox} = \frac{Ln(OR)}{1.65}$$

with sampling variance

$$S^2_{d_{Cox}} = 0.367\left(\frac{1}{O_{1E}} + \frac{1}{O_{2E}} + \frac{1}{O_{1C}} + \frac{1}{O_{2C}}\right).$$

Glass, McGaw, and Smith (1981) proposed a probit transformation to obtain an effect size in the metric of the standardized mean difference

$$d_{Probit} = Z_E - Z_C$$

where Z_E and Z_C are the inverse of the standard normal distribution for P_E and P_C, respectively. That is,

$$Z_i = \Phi^{-1}(P_i).$$

The sampling variance of d_{Probit} can be estimated by

$$S^2_{d_{Probit}} = \frac{2\pi P_E(1-P_E)e^{Z_E^2}}{n_E} + \frac{2\pi P_C(1-P_C)e^{Z_C^2}}{n_C}.$$

A final index is obtained by transforming the biserial-phi correlation coefficient into the metric of the standardized effect size (Becker & Thorndike, 1988). The biserial-phi coefficient is obtained as

$$\Phi_{bis} = \frac{\sqrt{p'q'}}{y'}\Phi$$

where p' and q' are the marginal success and failure proportions from the 2 X 2 table, and y' is the ordinate of the standard normal distribution at p'.

The biserial-phi coefficient is transformed to the scale of the standardized mean difference using

$$d_{bis} = \frac{\Phi_{bis}}{\sqrt{1-\Phi^2_{bis}}}\sqrt{\frac{(n_E + n_C - 2)(n_E + n_C)}{n_E n_C}}$$

with sampling variance

$$S^2_{d_{bis}} = \frac{p'q'(1-\Phi^2)(n_E + n_C)}{y'^2 n_E n_C (1-\Phi^2_{bis})^3}.$$

These indices were compared in a simulation study by Sanchez-Meca et al. (2003). The authors varied the sample sizes for the two groups, the cut points for dichotomization, and the population effect size, but maintained homogeneous variances in the two groups. The results suggested that d_{Cox} and d_{Probit} provided nearly unbiased estimates of the

population effect size δ. This research was replicated and extended by Kromrey and Bell (2012) who investigated heterogeneous variance conditions in addition to the homogeneous variance conditions studied by Sanchez-Meca et al. (2003).

In the latter study, the Cox and Probit effect sizes showed smallest bias overall, but this advantage was limited to conditions in which population variances were equal. Under heterogeneous variances, all indices evidenced similar patterns of bias associated with changes in the cut point for dichotomization. An evaluation of the sampling errors of these indices or a consideration of the RMSE (which combines sampling error and bias into a metric of total error) suggested that the standardized proportion difference, the transformed phi coefficient, and the arcsine transformation have smaller sampling errors and smaller total errors than the Cox and Probit indices. These advantages were especially pronounced at smaller sample sizes and with populations presenting smaller effect sizes.

3. Purpose

By furthering the research by Kromrey and Bell (2012), this study was designed to examine the performance of the seven effect size indices under more complex conditions with varying degrees of reliability. This design used a Monte Carlo simulation methodto compare the accuracy and precision of these effect size indices proposed for estimating a population standardized mean difference from a 2 by 2 sample table of artificially dichotomized variables.

4. Method

This Monte Carlo study investigated (a) overall sample size ($n_1 + n_2$ = 30, 60, 120, 240), (b) ratio of sample sizes in the two groups (1:1, 1:2, 1:4), (c) population effect size (0, .2, .5, .8), (d) continuous score cut point for dichotomization (.10, .25, .40, .70), (e) population variance ratio (1:1, 1:2, 1:3) and (f) reliability (.50, .70, .80, .90, 1.0). For each condition, 100,000 replications were conducted. Using SAS/IML, normally distributed, continuous data were generated for two groups and the continuous variables were dichotomized at the specified cut points. In each sample, the standardized mean difference was computed prior to dichotomization and the seven effect size indices were computed after the data were dichotomized.

5. Results

The outcomes from the simulation were analyzed by first examining box-and-whisker plots to describe the distributions of each outcome across all simulation conditions. Subsequently, variation in each outcome was analyzed using analysis of variance to obtain the eta-squared (η^2) values associated with each simulation design factor and their interactions. Design factors and interactions with large values of η^2 were further analyzed for substantive patterns. For each analysis, Cohen's *d* effect size (the sample effect size prior to dichotomization) was included as a reference statistic.

5.1 Statistical Bias of Effect Size Indices

The overall distributions of statistical bias across all simulation conditions are presented as box and whisker plots in Figure 1. All seven indices tended to be negatively biased in their estimation of the effect size, yielding an effect size index that was smaller on

average than the population effect size. Overall, the Cox, Probit, and biserial-phi indices evidenced notably smaller bias than the other indices, although the variability in the bias of the biserial-phi index was large relative to the others. This finding is consistent with Kromrey and Bell (2012).

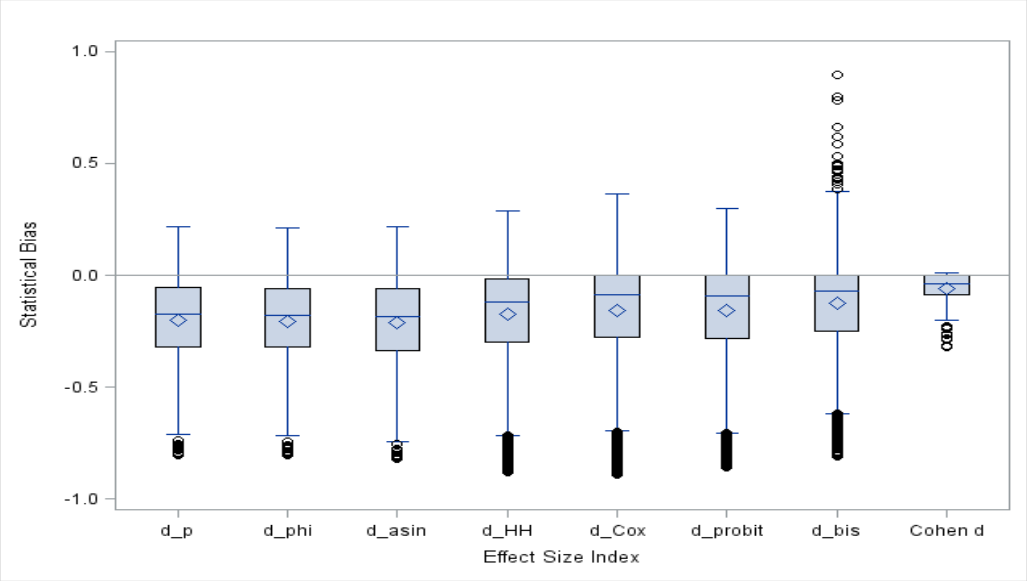

Figure 1: Distributions of statistical bias estimates

The distribution of bias associated with every index was then examined across different reliability conditions (.5, .7, .8, .9, and 1). Figure 2 shows that as expected, bias of all indices decreased as reliability improved although such decrease developed gradually. At every reliability level, the three indices with smaller bias overall (the Cox, Probit, and biserial-phi) exhibited smaller bias than the other four (first row of Figure 2) but the variability of bias remained unchanged or even increased as reliability increased. When reliability reached the optimal level of 1 with Cohen's *d* at zero, all seven indices still demonstrated bias which varied largely.

The simulation design factors associated with variation in statistical bias in the effect size indices included the population effect size (η^2 = .37), the cut point (η^2 = .35), and population variance ratio (η^2 = .09). Moreover, interactions between the design factors were also associated with statistical bias. The interaction between the population variance ratio and the cut point yielded η^2 = .14. Notably, reliability as another factor in simulation did not result in significant bias variation.

The following figures show the effect of some of these factors. Figure 3 presents the mean bias estimate by population effect size and effect size index. All of the indices became progressively more negatively biased as the population effect size increased. However, the Cox (d_{Cox}), Probit (d_{probit}), Hasselblad and Hedges (d_{HH}), and biserial-phi (d_{bis}) indices showed smaller increases in negative bias than the standardized proportion difference (d_p), the transformed phi coefficient (d_{phi}), and the arcsine transformation (d_{asin}).

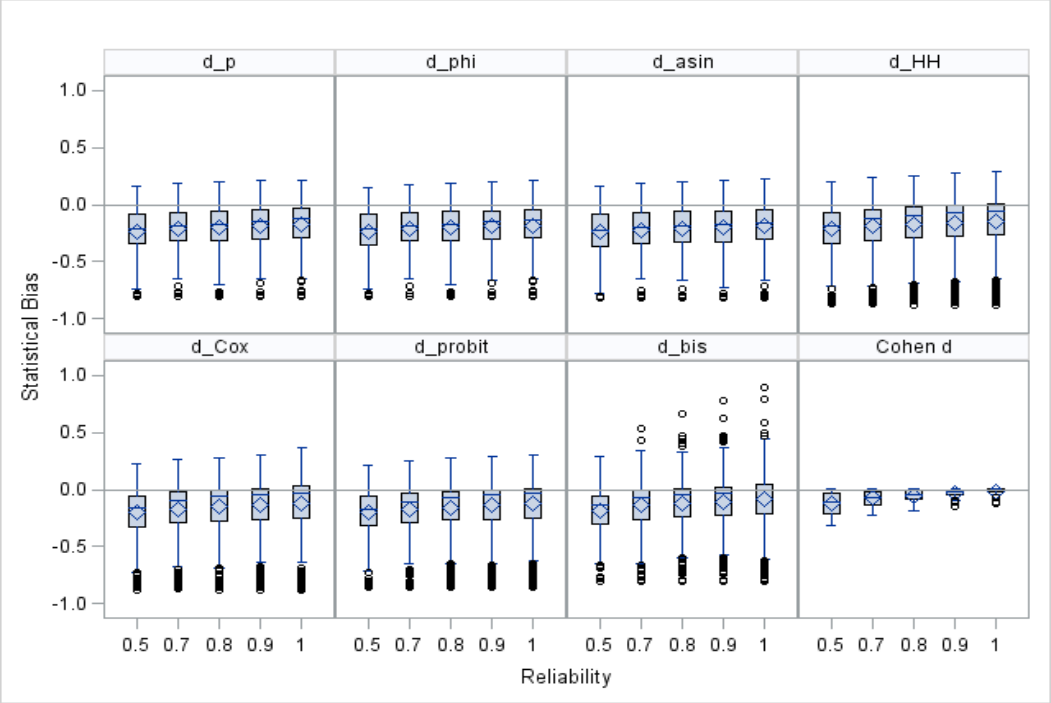

Figure 2: Distributions of statistical bias estimates across different reliability conditions

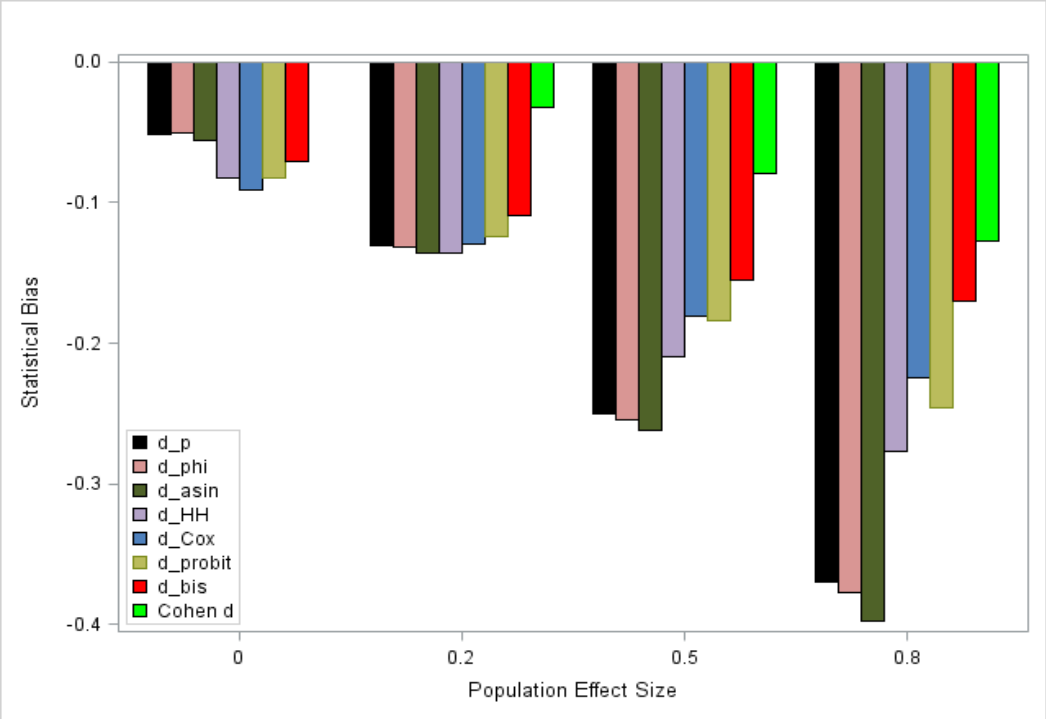

Figure 3: Distributions of statistical bias estimates across different levels of population effect size

Figure 4 shows the distributions of statistical bias estimates across different levels of population effect size under five reliability conditions. At every reliability level, the pattern of the seven indices is similar to that in Figure 3.

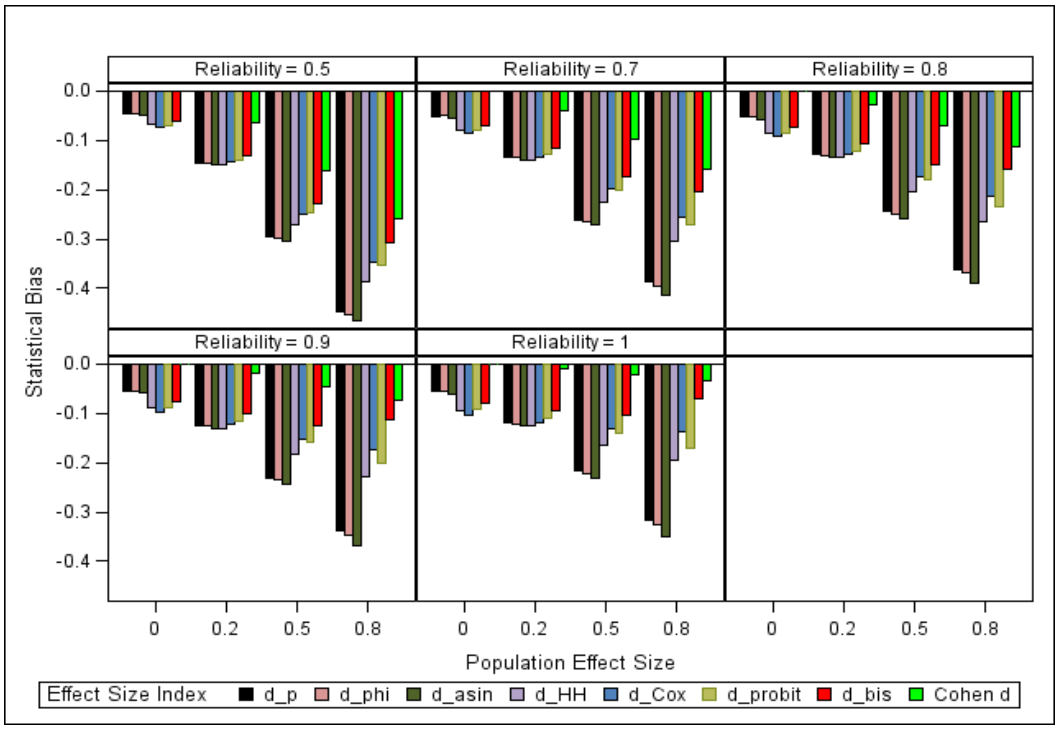

Figure 4: Distributions of statistical bias estimates across different levels of population effect size under five reliability conditions

Figure 5 presents the average bias estimates associated with the interaction between cut point and population variance ratio. Evident in this figure is that the impact of the cut point on resulting bias increased as population variances became more heterogeneous. With heterogeneous variances, all seven indices displayed substantial negative bias with a cut point of .1, but moved to small positive bias with a cut point of .7. In contrast, under homogeneous variances, the Cox, Hasselblad and Hedges, and Probit indices showed notably lower levels of bias than the other indices and their biasestimateswere less influenced by changes in the cut point for dichotomization.

5.2 RMSE of Effect Size Indices

The distributions of root mean squared error (RMSE) by each effect size index across all simulation conditions are presented in Figure 6. In contrast to the statistical bias estimates, the smallest average RMSE values were obtained for the standardized proportion difference (d_p), the transformed phi coefficient (d_{phi}), and the arcsine transformation (d_{asin}). Notable also in Figure 6 is that the biserial-phi coefficient (d_{bis}) reached extremely large values in many conditions. The analysis of variance conducted on the RMSE values indicated that variation in RMSE was associated with total sample size (η^2 = .30), effect size index (η^2 = .12), cut point (η^2 = .16), population effect size (η^2 = .08), and population variance ratio (η^2 = .17).

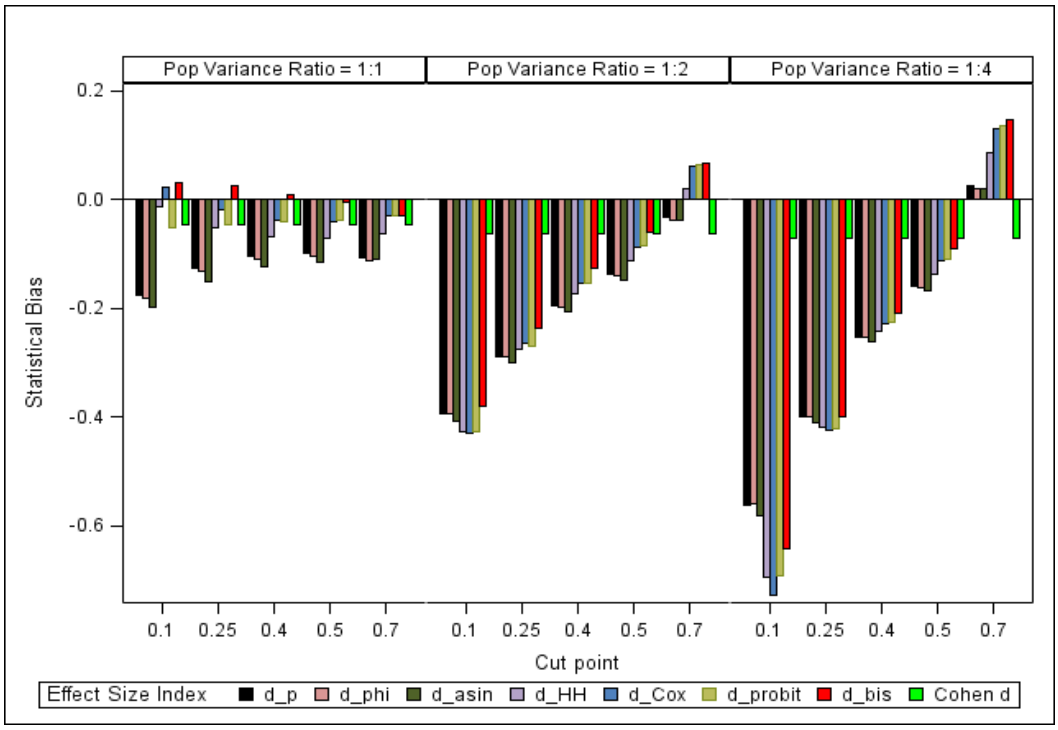

Figure 5: Distributions of statistical bias estimates across different cut points under different population variance ratios

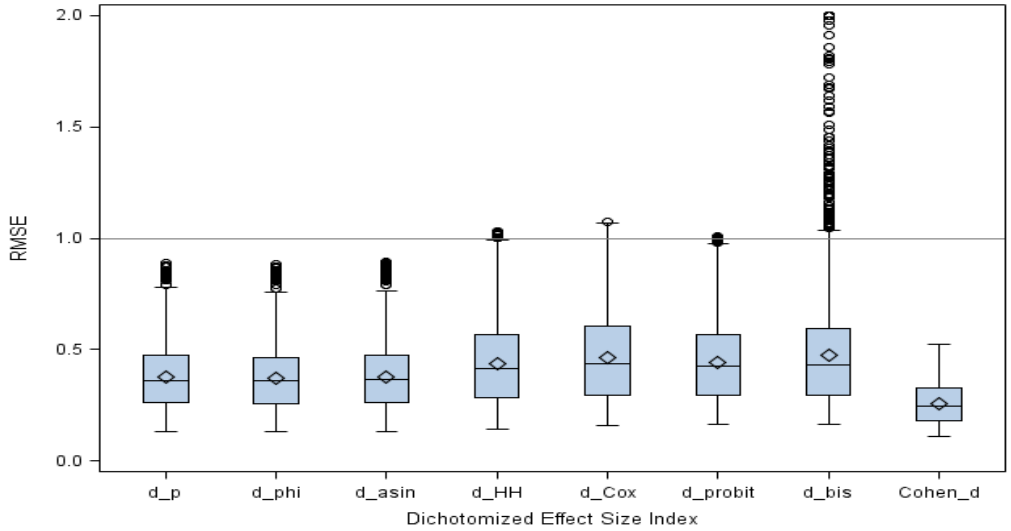

Figure 6: Distributions of RMSE estimates by effect size indices across all conditions

Figure 7 shows the variation in the average RMSE estimates for all seven effect size indices associated with every design factor. Variation due to total sample size ($.23 \leq \eta^2 \leq .42$) and cut point ($.12 \leq \eta^2 \leq .35$) for some effect size indices (d_p, d_{phi}, and d_{bis}) was large enough to merit further investigation.

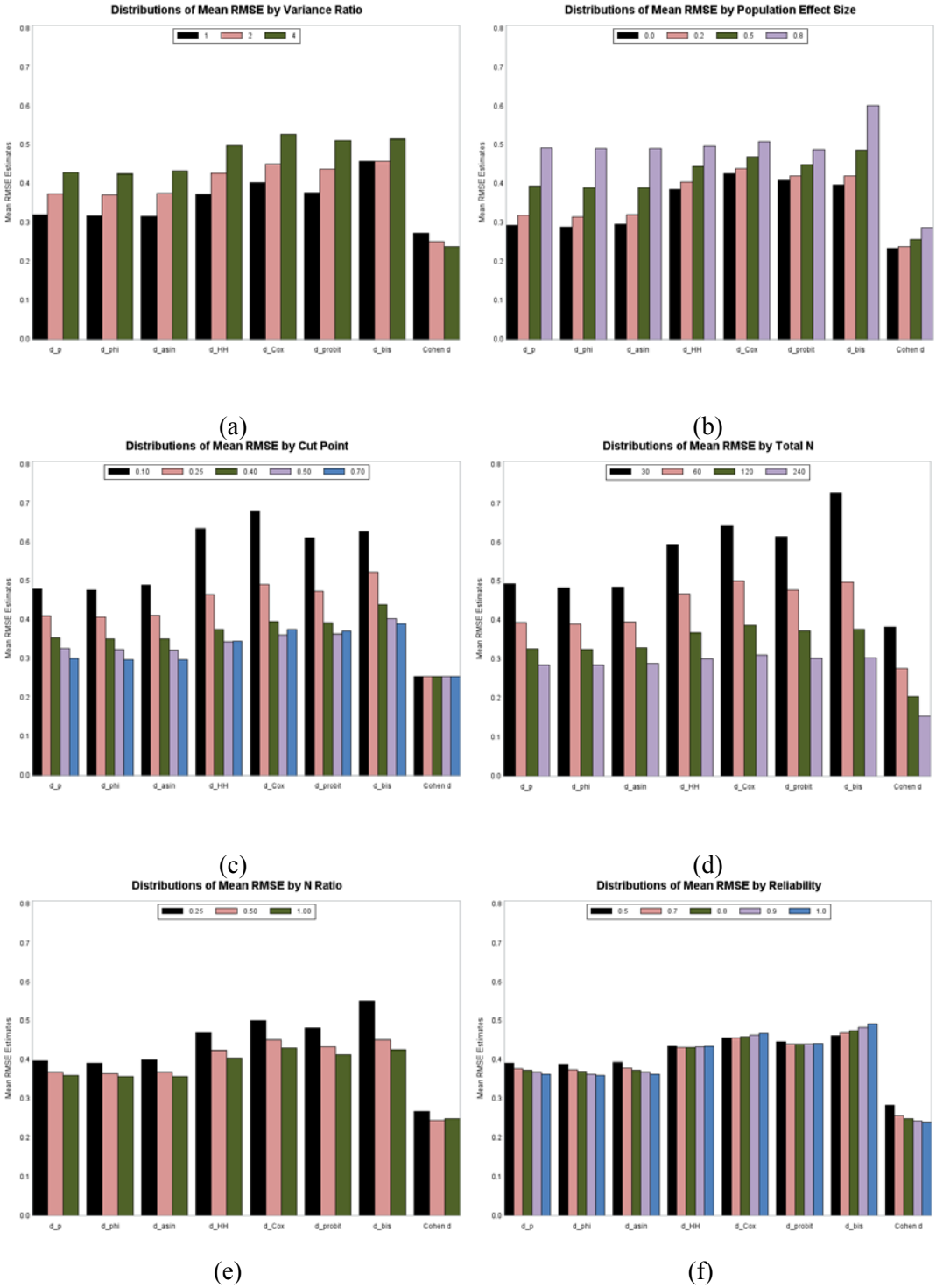

Figure 7: Distributions of mean RMSE estimates across effect size indices by each factor separately

In (*b*), variation in mean RMSE estimate increased as the population effect size increased; on the other hand, variation due to cut point (*c*) decreased with higher values of the factor. For total sample size (*d*), the variation in average RMSE decreased with larger total sample size. While high, average mean RMSE for reliability did not show sharp variations across effect sizes and levels of the factor (*f*).

5.3 Empirical Variances of Effect Size Indices

Figure 8 presents the distributions of empirical effect size variance estimates. Compared to the mean variance of Cohen's *d*, the mean variances of the standardized proportion difference, the transformed phi coefficient, and the arcsine transformation were slightly larger. However, the remaining effect size indices had average variances approximately twice or more as large. Specifically, the biserial-phi index had an average variance three times greater than that of Cohen's *d* with exceptionally large sampling variance in many conditions. The analysis of variance indicated that variation in the empirical variance values was associated with sample size (η^2 = .51), effect size index (η^2 = .15) and the interaction between sample size and effect size index (η^2 = .08).

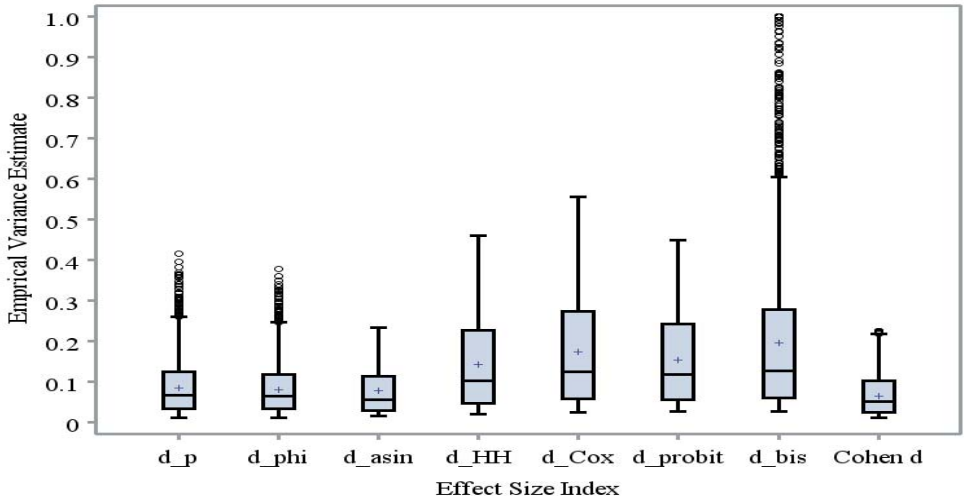

Figure 8: Distributions of empirical effect size variance estimates

Figure 9 presents the average empirical variances of the effect size indices by sample size. As expected, the variances of all the indices deceased with larger sample sizes but the rate of decrease varied among the sample sizes. The average variances of the standardized proportion difference, transformed phi coefficient, and the arcsine transformation were slightly larger than those of Cohen's *d* at all sample sizes. In contrast, the remaining indices had remarkably larger variances at smaller sample sizes, but these variances decreased rapidly with increased sample sizes and only trivial differences in variances were evident at the largest sample size examined (*N* =240).

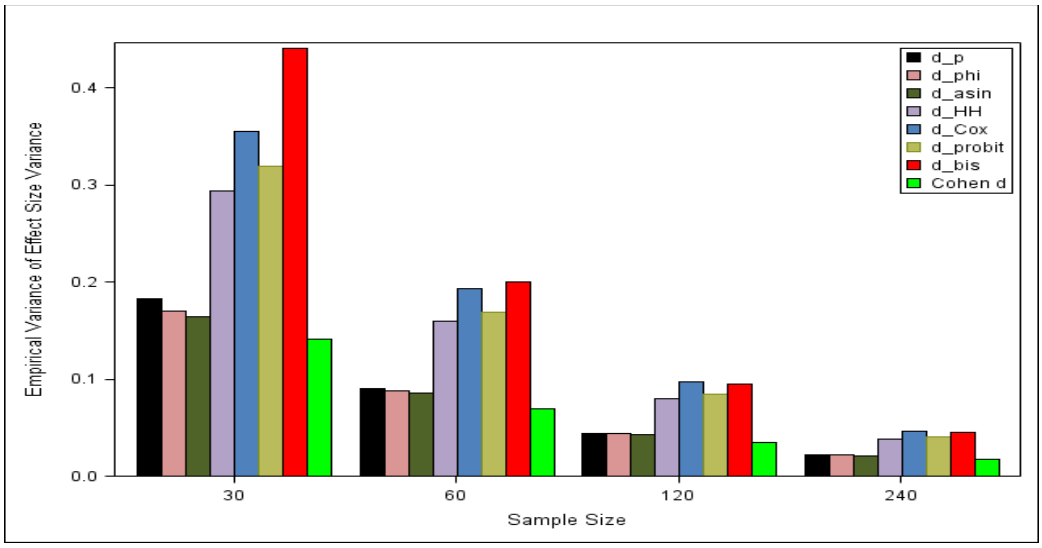

Figure 9: Average empirical variances of effect size indices by sample size

5.4 Estimated Variances of Effect Size Indices

The distributions of average estimated variances of the effect size indices are presented in Figure 10, which exhibited a similar pattern to those of empirical variance estimates (Figure 8). The average estimated variances of the standardized proportion difference, the transformed phi coefficient, and the arcsine transformation were close to those of the Cohen's *d* effect size. The average estimated variances of Hasselblad and Hedges, Cox and the Probit effect size indices were about twice as large. The average estimated variances of the biserial-phi index was more than three times as large as what was observed for Cohen's *d* effect size. The analysis of variance indicated that variation in the average estimated variances values was mainly associated with sample size (η^2 = .41) and effect size index (η^2 = .16). Significant first order interaction effects were found with sample size and effect size indices (η^2 = .15)

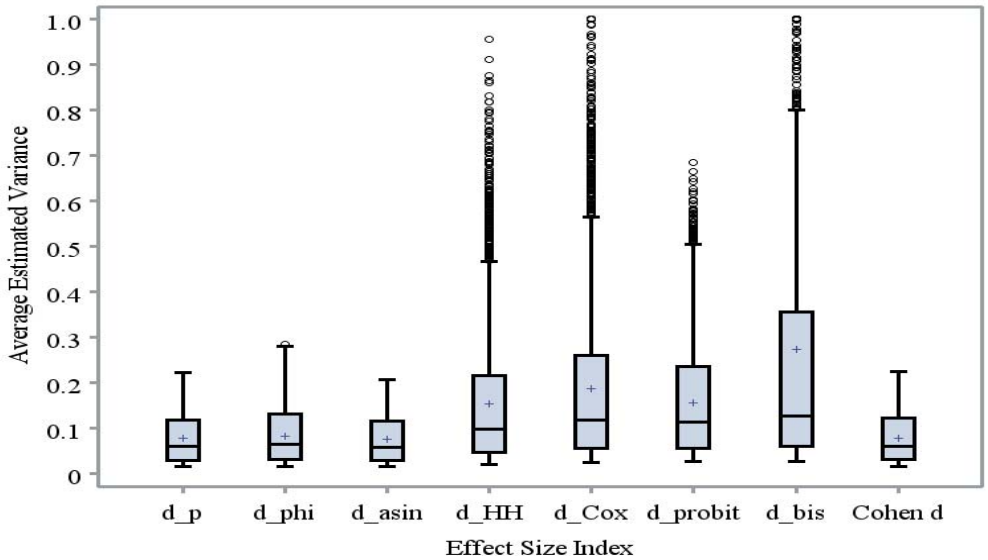

Figure 10: Distributions of average estimated effect size variances

Figure 11 displays average effect size variances by sample size. It is evident that differences between these variances and the variances themselves were at the largest with smallest sample size. They decreased rapidly as sample size increased. With the largest sample size (N = 240), variance estimates of all effect size indices essentially converged around .03. The seven indices did not behave in the same way as sample size changed. For N = 30, the estimated average variance associated with the biserial-phi index (.70) was approximately twice as large as the Probit and Hasselblad and Hedges, Cox and more than four times as large as that of the standardized proportion difference, the transformed phi coefficient, the arcsine transformation, and Cohen's *d*.

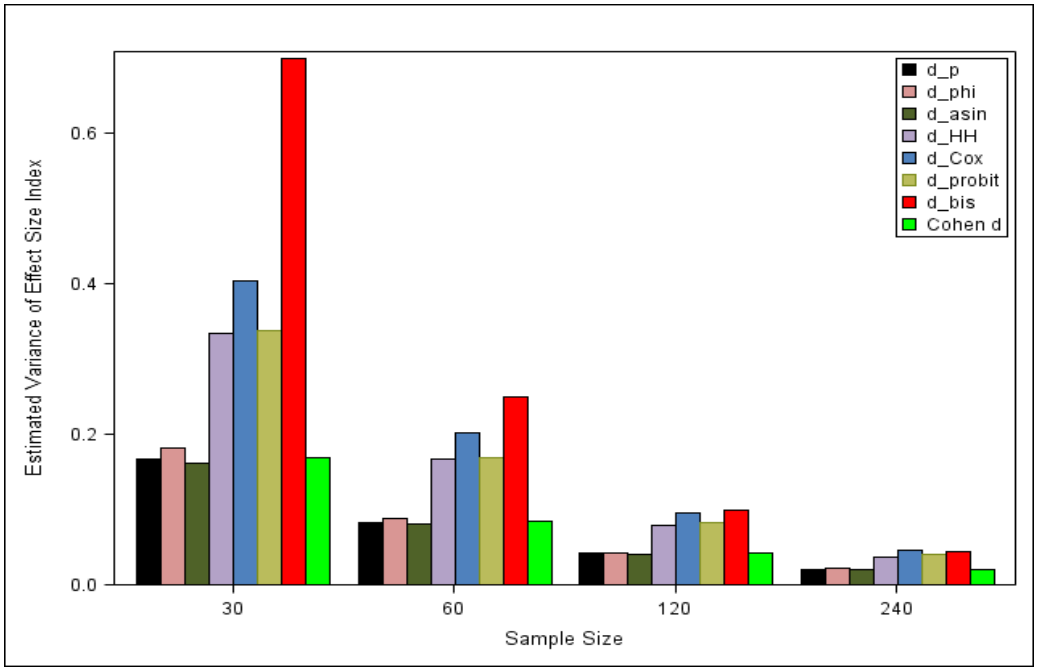

Figure 11: Average estimated effect size variance by sample size and effect size index

5.5 Estimated Variances vs. Empirical Variances of Effect Size Indices

To evaluate the accuracy of the estimated variances of the effect size indices, the ratios of the average estimated variances to the empirical variance estimates were computed for each condition in the simulation. The distributions of these ratios are presented in Figure 12. A ratio greater than one would indicate an underestimation of population variance and a ratio less than one would indicate an overestimate of variance. It can be seen that the Cohen's *d*, the standardized phi, and the biserial-phi indices overestimated the variance whereas the others underestimated it and, in general, had less variability.

Variation in the variance ratios was associated with factors of the effect size indices (η^2 = .20) and the cut point (η^2 = .07). Variation was also found in connection with interaction between the effect size index and total sample size (η^2 = .07), between the effect size index and the population ratio, and between the cut point and the total sample size (η^2 = .06). Similar to previous analyses, reliability did not lead to substantial variation in the variance ratios.

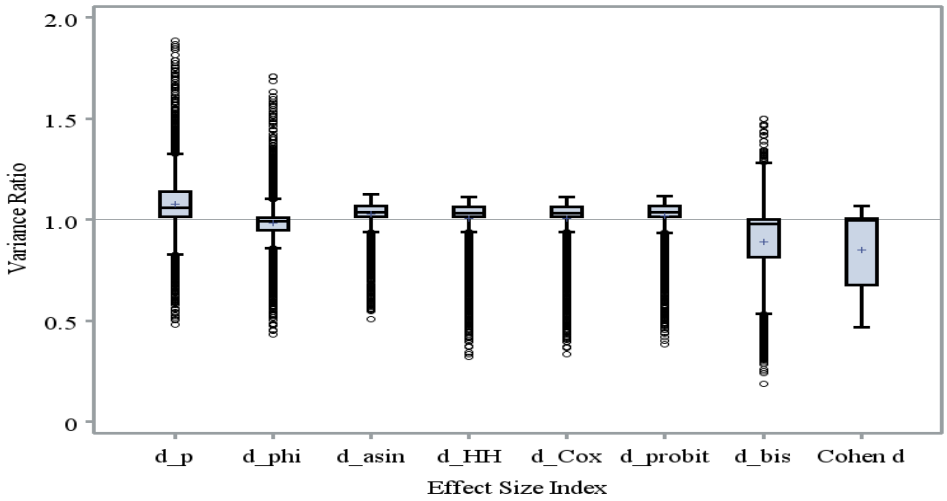

Figure 12: Distributions of the ratios of empirical estimates of effect size variance to average estimated effect size variance

6. Conclusions

The seven effect size indices discussed in this paper have been proposed in the literature as methods to provide the estimated standardized mean difference after continuous data have been dichotomized. As noted in the simulation study by Kromrey and Bell (2012), the statistical bias and sampling error for all seven indices compared to those of the effect size obtained prior to dichotomization (Cohen's *d*) clearly demonstrated tenuous estimations of the effect size parameter after continuous variables have been dichotomized. This was especially problematic with smaller sample sizes and cut points in which the percentages were farther away from the median (e.g. .10 and .70 as opposed to .25 and .40). Most notably were the results using the biserial-phi coefficient where the variance and RSME reached extremely large values in many conditions. This is most likely due to the fact that the denominator of the biserial-phi coefficient (y') is the ordinate of the standard normal distribution at p'. Under the normal distribution, the values for y' would be smaller in the tails of the distribution and would be notably larger between the inflexion points of the distribution, (i.e. the 16th and 84th percentile), which concomitantly could cause wide fluctuations for the coefficient. Working "backwards" to "undo" dichotomized continuous variables is a difficult task. A better choice is for researchers to report the Cohen's *d* value prior to the dichotomization. However, if the original data are not available, these effect size indices provide a useful approximation.

References

Becker, G. & Thorndike, R. L. (1988). The biserial-phi correlation coefficient. *Journal of Psychology*, *122*, 523-526.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cox, D. R. (1970). *Analysis of binary data*. New York: Chapman & Hall.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hill, CA: Sage.

Hasselblad, V. & Hedges, L. V. (1995). Meta-analysis of screening and diagnostic tests. *Psychological Bulletin*, *117*, 167-178.

Johnson, B. T. (1989). *DSTAT: Software for the meta-analysis review of research* [computer program and manual]. Hillsdale, NJ: Erlbaum.

Kromrey, J. D. & Bell, B. A. (2012). *Effect Size Indices for Dichotomized Outcomes under Variance Heterogeneity: An Empirical Investigation of Accuracy and Precision*. Paper presented the Joint Statistics Meeting, 2012.

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

Sanchez-Meca, J., Marin-Martinez, F., & Chacon-Moscoso, S. (2003). Effect size indices for dichotomized outcomes in meta-analysis. *Psychological Methods*, *8*, 448-467.

SAS Institute Inc. (2008). *SAS, release 9.2* [computer program]. Carey, NC: SAS Institute Inc.