Converting from d to r to z when the Design Uses Extreme Groups, Dichotomization, or Experimental Control

James E. Pustejovsky

Northwestern University

Author Note

Jamers E. Pustejovsky, Department of Statistics, Northwestern University.

This research was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B080027 to Northwestern University. The opinions expressed are my own and do not represent views of the Institute or the U.S. Department of Education. I am grateful to Chin Ming Hui and Priya Kamat for providing the example that stimulated this inquiry and to Martyna Citkowicz for critiquing the draft manuscript.

Address correspondence to James E. Pustejovsky, Department of Statistics, Northwestern University, 2006 Sheridan Road, Evanston, IL 60208. Email: pusto@u.northwestern.edu.

Abstract

Meta-analyses of the relationship between two continuous variables sometimes involves conversions between different effect sizes, but methodological literature offers conflicting guidance about how to make such conversions. This article provides methods for converting from a standardized mean difference to a correlation coefficient (and from there to Fisher's z) under three types of study designs: extreme groups, dichotomization of a continuous variable, and controlled experiments. Also provided are formulas and recommendations regarding how the sampling variance of effect size statistics should be estimated in each of these cases. The conversion formula for extreme groups designs, originally due to Feldt, can be viewed as a generalization of Hunter and Schmidt's method for dichotomization designs. A simulation study examines the finite-sample properties of the proposed methods. The conclusion highlights areas where current guidance in the literature should be amended or clarified.

Keywords: effect size; extreme groups; dichotomization; correlation

Converting from *d* to *r* to *z* when the Design Uses Extreme Groups, Dichotomization, or Experimental Control

Sophisticated consumers of research, attuned to questions of effect magnitude, sometimes need to convert between different effect size metrics, such as between standardized mean difference (*d*-type) and correlation (*r*-type) metrics. Converting between metrics is necessary because study results may be reported using statistics other than the effect size that is most useful or intuitive for interpreting the findings. For instance, a study might report a *d* statistic based on dichotomizing a continuous variable, when one's interest is in the correlation between the continuous variable and the outcome. Furthermore, effect size conversions are a crucial step in the process of quantitative research synthesis, in which the goal is to combine results (in the form of effect sizes) from multiple studies. Meta-analytic models that average or compare results from different studies become uninterpretable if the effect sizes to be modeled are not first put on a common scale. For instance, a meta-analysis might be based predominantly on correlational effect sizes but also include effect sizes from controlled experiments, which are typically reported as *d*-statistics (e.g., Hagger & Chatzisarantis, 2009; Macbeth & Gumley, 2012). In such cases, the meta-analyst will need to convert from *d* to *r*.

Despite their importance, the methods and formulas for effect size conversions can be a source of confusion for the working researcher or meta-analyst. There are several reasons for this. The first is a ambiguity regarding the estimand to which an effect size statistic is intended to correspond. For example, Borenstein (2009) and Borenstein, Hedges, Higgins, and Rothstein (2009) propose to convert standardized mean differences between two groups (*d*-type effect sizes) into point-biserial correlation coefficients, which measure the linear association between the outcome variable and a binary indicator for group membership. In contrast to Borenstein's approach, Hunter and Schmidt (1990, 2004) take the target parameter to be the correlation between the outcome and an underlying continuous variable on which group membership is based, for which the the point-biserial correlation

coefficient is a biased estimator. They therefore provide and recommend methods that correct this "study artifact" (Hunter & Schmidt, 1990, 2004). The difference between these two approaches stems from different assumptions regarding which target parameter is appropriate as an effect size measure.

A second source of confusion regarding effect size conversions stems from disagreement regarding whether and when certain effect size conversions are appropriate. Many methodologists recommend meta-analyzing correlation coefficients after applying Fisher's z-transformation (e.g., Borenstein, 2009; Hedges & Olkin, 1985) because the transformed estimates have improved statistical properties. In contrast, Hunter and Schmidt (2004) recommend against z-transformation, suggesting instead that correlation coefficients be meta-analyzed in their natural metric. Working meta-analysts are thus left without definitive guidance, in spite of the fact that controversy regarding z-transformation seems to stem largely from differences in maintained modeling assumptions (Hafdahl & Williams, 2009).

Finally, and most fundamentally, confusion regarding effect size conversions arises because conversion formulas are framed as applying across entire categories of effect size measure and as being reversible (cf. Borenstein et al., 2009). When framed this way, one may have the impression that converting from an effect size of type A to one of type B is simply a matter of inverting the conversion formula from B to A, as if it were an algebraic identity. In fact, conversion formulas are more narrowly applicable and depend on the study design used to generate a given effect size statistic. Consider the standardized mean difference d, which is calculated by taking the difference in mean outcomes between two groups, then dividing by the standard deviation of the outcome. The statistical properties of the standardized mean difference depend strongly on the study design; the d statistic calculated from a repeated measures design has a different sampling variance than that calculated from a simple between-groups design (Morris & DeShon, 2002), and both differ from the various d statistics that can be calculated from cluster-randomized designs

(Hedges, 2007, 2011). Baguley (2009) makes the broader point that all standardized effect sizes are sensitive to variations in study design.

This paper aims to provide a modicum of clarity regarding effect size conversion by addressing just two related types of conversions: from d to Pearson's product-moment correlation r and from r to z. I do not consider conversions in the opposite direction, from z to r or from r to d, and the methods presented here should not be interpreted as applying to those cases. In contrast to previous work, which presents conversion formulas as if they were applicable to any effect size of a given type, I focus narrowly on three types of study designs, considering distinct d-to-r conversion formulas and associated variance estimators in each case. These three designs, all likely to be encountered in practice, are: the controlled experimental design, the dichotomization design, and the extreme groups design. For context, I also discuss the bivariate sampling design, the most conventional method used to estimate the Pearson correlation between two continuous variables. In each of the three designs, results are commonly reported using d effect sizes, even though Pearson product-moment correlations may be of greater interest or utility; thus, researchers and meta-analysts will often need to convert from one metric to the other. Furthermore, some meta-analysts may also wish to convert from r to z for purposes of synthesizing effect sizes across studies.

Some of the methods discussed below are well-known, while others are novel. For the controlled experimental design, the d-to-r conversion formulas are closely related to the point-biserial correlation coefficient, as discussed in Hunter and Schmidt (2004, Chapter 7), but are to the best of my knowledge novel. For the dichotomization design, Hunter and Schmidt (2004) present formulas for converting a reported d statistic to a correlation coefficient r, though the variance estimators proposed for use with the converted effect size statistic are problematic. For the extreme groups design, Feldt (1961) presented a d-to-r conversion formula assuming groups of equal size. Though Feldt's work was cited in one early text on meta-analysis (Glass, McGaw, & Smith, 1981, Table 5.8), it seems to have

been largely ignored by meta-analysts and is not discussed or cited in more recent textbooks. Below, I provide a generalization of Feldt's result to unequal group sizes and study associated variance estimators in detail. In so doing, I also demonstrate that dichotomization is a special case of the extreme groups design; thus, conversion formulas for both designs can be studied by using a general formulation of the extreme groups design. Finally, I propose a novel solution to a technical problem that arises in dichotomization and extreme groups designs when converting from r to z.

In the next section, I explain each of the designs in detail and describe motivating examples. The following sections present models and formulas for converting from d to r to z effect sizes and address variance estimation. After that, a simulation study examines the small-sample properties of the estimators. I then return to the motivating examples, demonstrating how to use the appropriate conversion formulas for each design. A brief conclusion discusses limitations and future directions. Readers eager to begin converting effect sizes should consult Table 1, which displays the equation numbers of the appropriate conversion formulas and variance estimators for each type of design, as well as the online supplementary material, which includes a spreadsheet implementing all of the formulas.

Study designs and examples

This section describes four types of study designs that may be relevant to the estimation of a Pearson correlation: bivariate sampling, the controlled experiment, dichotomization, and the extreme groups design. For each design, I provide a general description of the procedures and statistics involved and then give an example.

Bivariate sampling

The bivariate sampling design is conventional and simple: procedurally, it involves nothing more than measuring two variables X and Y on a sample of n individuals and computing the Pearson correlation coefficient r based on the sample data. If (x_i, y_i) denote

the measured variables for individual $i = 1, ..., n$, the sample correlation is calculated as

$$r_p = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}, \quad (1)$$

where $\bar{x}$ and $\bar{y}$ are the means of $x_1, ..., x_n$ and $y_1, ..., y_n$, respectively. I denote this statistic r_p to distinguish it from other correlation statistics discussed in later sections. The statistic r_p provides a sample estimate of the correlation between X and Y in the population.

Example 1. Chatzisarantis, Hagger, and Smith (2007) studied the relationship between an individual's intentions to participate in vigorous physical activity and his or her level of perceived autonomy support (PAS). The authors collected Likert-scale measures of both intentions and PAS on a sample of 165 high school students. In a preliminary analysis, they reported that the Pearson correlation between intentions and PAS was $r_p = 0.58$ (Chatzisarantis et al., 2007, Table 3). This correlation coefficient was later used in a meta-analysis of the same constructs (Hagger & Chatzisarantis, 2009).

Controlled experiment

Along with the bivariate sampling design, the controlled experimental design is a primary method of psychological science. As oppose to measuring two variables that capture individual differences, the controlled experiment studies the effect on an outcome variable Y of an active manipulation by the researcher. A basic between-groups experiment involves assigning (often randomly) two groups of individuals to receive a treatment condition or a control condition, then taking measurements of an outcome variable on each individual. In an experiment with a total of n participants, suppose that the first n_1 individuals receive a control condition and have outcome measurements $y_1, ..., y_{n_1}$, while the remaining the remaining $n_2 = n - n_1$ individuals receive a treatment condition and have outcome measurements $y_{n_1+1}, ..., y_n$. The sample means ($\bar{y}_1$ and $\bar{y}_2$) and variances (s_1^2

and s_2^2) in each group are calculated as

$$\begin{aligned} \bar{y}_1 &= \frac{1}{n_1}\sum_{i=1}^{n_1} y_i, & s_1^2 &= \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(y_i - \bar{y}_1)^2, \\ \bar{y}_2 &= \frac{1}{n_2}\sum_{i=n_1+1}^{n} y_i, & s_2^2 &= \frac{1}{n_2 - 1}\sum_{i=n_1+1}^{n}(y_i - \bar{y}_1)^2. \end{aligned} \tag{2}$$

These statistics will often appear in reports of controlled experiments. Authors might describe the magnitude of the difference between treatment and control conditions using the standardized mean difference effect size d, which is calculated as

$$d = \frac{\bar{y}_2 - \bar{y}_1}{s_p}, \tag{3}$$

where s_p is the pooled sample standard deviation

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}.$$

The standardized mean difference measures the average causal effect on Y of assigning individuals to the treatment versus control condition.

Those who think of treatments as intervention packages containing many component pieces may wonder how the results of an experiment could relate to the correlation between continuous measurements. The circumstances where it is useful to combine correlational and causal evidence may be quite limited–perhaps only to situations where the researcher is interested in quantifying the contrast between observational and experimental results. Furthermore, converting an experimental d to the r scale involves making strong modeling assumptions, as I detail in the next section. In circumstances where the exercise is indeed relevant, one can imagine that individuals in the treatment and control conditions are induced to receive distinct, fixed levels of a variable X. Specifically, suppose that the individuals in the control condition all receive a certain fixed level (call it q) of X, so that $x_i = q$ for $i = 1, ..., n_1$, while the individuals in the treatment condition all receive a fixed level of X that is higher by w units, so that $x_i = q + w$ for $i = n_1 + 1, ..., n$. The standardized mean difference can then be interpreted as the causal effect on Y of increasing

X by the treatment-control differential w; thus, the results of the experiment may bear on the relationship between X and the outcome variable Y.

The substantive interpretation of the standardized mean difference d, as well as the formula for converting it to r, will depend on one's assumption about the size of the treatment-control differential. For all but the simplest interventions, it may be difficult to identify a reasonable value for w that captures the difference between treatment and control conditions in terms of a single quantitative variable. Still, without making some such assumption, one is left without any way of relating the experimental findings to the association between continuous variables. The model described in the following section provides this connection by relying on an assumption about the magnitude of w.

Example 2. Edmunds (2008) studied the effect of a teaching intervention that was designed to support participants' senses of autonomy, relative to a conventional teaching style, in the context of an exercise class. The outcome measure was a scale that measured individuals' intentions to continue exercising. The investigators reported that, after 10 weeks of treatment, $n_1 = 31$ individuals in the control group had a mean score of $\bar{y}_1 = 4.67$ ($s_1 = 1.52$), while $n_2 = 25$ individuals in the treatment group had a mean score of $\bar{y}_2 = 5.67$ ($s_2 = 1.33$) (Edmunds, 2008, Table 2). Using the reported statistics, the standardized mean difference can be calculated according to Equation (3): $d = 0.70$. Over the course of the experiment, the investigators also collected self-reports of autonomous support. I will use these data to inform an assumption about the treatment-control differential w. After 5 weeks of treatment, the difference between treatment and control groups was $w = 2.18$ SDs; after 10 weeks, the difference was reduced to 1.22 SDs (Edmunds, 2008, Table 1). Hagger and Chatzisarantis (2009) used the results of the study, along with the results described in Example 1, in a meta-analysis of the relationship between autonomous support and planned exercise behavior.

Dichotomization

Though I refer to it as a design, dichotomization is more precisely a data-analysis technique; it is a common practice in psychology (Decoster, Iselin, & Gallucci, 2009; MacCallum, Zhang, Preacher, & Rucker, 2002). Procedurally, data collection is identical to the bivariate sampling design: a researcher measures two continuous variables X and Y on a set of n individuals, producing data (x_i, y_i) for $i = 1, ..., n$. Rather than computing a correlation coefficient directly, the researcher creates an auxiliary variable for data analysis purposes. For some cutoff value c_1, the auxiliary variable is set equal to 0 for all cases with $x_i \leq c_1$ and equal to 1 for $x_i > c_1$. The researcher can choose a cutoff value in one of two ways: based on prior information about the population distribution of X or based on the sample data. For instance, a population-based cutoff might be chosen based on the scale median reported in a large norming study; in contrast, a sample-based cutoff might be chosen based on the median value of X in the sample (this is known colloquially as a "median split").

After creating the auxiliary variable, the researcher calculates means and standard deviations of the y_i's in the two groups defined by unique values of the auxiliary variable. For ease of notation, suppose that the first n_1 cases have $x_i \leq c_1$ and the remaining $n_2 = n - n_1$ cases have $x_i > c_1$. The sample means and variances in each group are then calculated just as in Equation (2). The researcher might report these four values directly, or might instead report a standardized mean difference effect size d, calculated as in Equation (3).

Example 3. Mussweiler, Gabriel, and Bodenhausen (2000) studied the relationship between self-esteem and the extent to which individuals focus on gender as part of their self-identity. In a preliminary study to assess the baseline focus on gender identity, the authors divided their sample of $n = 27$ individuals into two groups with low ($n_1 = 19$) and high ($n_2 = 8$) levels on a scale measuring self-esteem; the cutoff point was chosen based on the median of a larger study. The authors reported that individuals in the high self-esteem

group "focused more on their gender ($M = 2.16$) than did [individuals in the low self-esteem group] ($M = 3.09$), $t(25) = 2.71$, $p < .01$" (Mussweiler et al., 2000, p. 401). One can recover the sample d statistic from the reported t value using the algebraic relationship between the two statistics:

$$d = t\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 2.71\sqrt{\frac{1}{19} + \frac{1}{8}} = 1.14$$

(for details, see Borenstein, 2009, Table 12.1).

Extreme groups

The extreme groups design falls into a category of procedures sometimes described as "direct range enhancement" (Hunter & Schmidt, 2004, p. 104). Preacher, Rucker, MacCallum, and Nicewander (2005) review the use of and rationale for the design in psychological research. For studying the relationship between continuous two variables X and Y, an extreme groups design typically entails the following: To begin, the investigator measures X on an initial sample of m individuals. Based either on known population values or on sample quantiles, two cutoff values c_1 and $c_2 \geq c_1$ are chosen to define subsets of individuals with low and high values of X. From the full set of m individuals, Y is measured only on these two subsets: the n_1 individuals with $x_i \leq c_1$ and the n_2 individuals with $x_i > c_2$. No data on Y are collected from the remaining cases that have $c_1 < x_i \leq c_2$; measured X values from these cases are discarded from the analysis.

As with the dichotomization design, there are two slightly different methods of choosing cutoff values. First, an investigator might define population cutoffs based on known quantiles of the X distribution. For instance, if X is a widely used psychological scale, the experimenter might use as cutoffs the published values c_1 and c_2 corresponding to the lower and upper quartiles. Alternately, the investigator might define sample cutoffs by taking the n_1 lowest sample values and the n_2 highest sample values of X.

Studies using the extreme groups design often report the standardized mean difference between groups as a summary effect size and in a test of the hypothesis that X

and Y are uncorrelated.[1] As with the dichotomization design, suppose that cases are arranged so that $y_1, ..., y_{n_1}$ are in the lower group and $y_{n_1+1}, ..., y_{n_1+n_2}$ are in the higher group. Sample means and variances of Y are calculated as in Equation (2); based on these values, the standardized mean difference is then calculated as in Equation (3).

Example 4. Goldinger, Kleider, Azuma, and Beike (2003) conducted a study using an extreme groups design. These researchers used a memory span task to select participants with high and low levels of working memory. They then gave the selected participants the further task of reviewing hypothetical court cases and assigning monetary damages, the outcome of which can be interpreted as a measure of pro-social behavior. To begin, the researchers measured working memory on an initial sample of 138 participants. Of these, the bottom and top 25% of participants were assigned to the low and high working memory groups, respectively, resulting in 35 participants in each group (only these 70 participants reviewed the hypothetical court cases). The authors report an F-test for the difference in average monetary compensation between the low- and high-memory span groups: "A main effect of [memory] span, $F(1, 68) = 18.2$, $p < .001$, showed that high-span participants generally awarded more money than low-span participants" (Goldinger et al., 2003, p. 82). Using the F statistic and the reported sample sizes, the standardized mean difference can be determined using

$$d = \sqrt{F\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{18.2 \times \frac{2}{35}} = 1.02$$

(Borenstein, 2009, Table 12.1).

Other recent examples of the extreme groups design can be found in Bernichon, Cook, and Brown (2003); Cross, Morris, and Gore (2002); Deffenbacher, Huff, Lynch, Oetting, and Salvatore (2000); Eys, Hardy, Carron, and Beauchamp (2003); Pontari and Schlenker (2000); Skinner and Drake (2003); Verplanken and Holland (2002); and Stafford and Gonier (2004).

[1] This practice is common despite the fact that using the standardized mean difference is inferior to other analytic approaches (Alf & Abrahams, 1975).

Dichotomization as a special case. The description of the extreme groups design encompasses dichotomization as a special case. Note that if $c_1 = c_2$, then $m = n_1 + n_2$ and both X and Y are measured on the entire sample of cases. No data are discarded, and the sample means, variances, and d statistic are calculated for the lower and upper groups just as before. Due to this relationship between designs, I use the more general formulation of the extreme groups design to develop models and conversion formulas, then simply apply those formulas to the dichotomization design.

Converting from d to r

In principle, one might use any of the four designs to estimate the same correlation parameter. In order to maintain such comparability, I use a common set of modeling assumptions for all four designs, assuming throughout that the researcher is interested in studying the correlation between two continuous variables X and Y. For each design, I assume that variables of interest follow a bivariate normal distribution with population means $E(X) = 0$, $E(Y) = \mu_y$, population variances $\text{Var}(X) = 1$, $\text{Var}(Y) = \sigma_y^2$, and population correlation $\text{cor}(X, Y) = \rho$.[2] Thus, the target parameter to be estimated in each design is ρ.

In what follows, I employ some additional notation for describing normally distributed random variables with mean zero and variance one. I denote the probability density function for a standard normal random variable by $\phi(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$, the corresponding cumulative distribution function by $\Phi(x) = \int_{-\infty}^{x} \phi(x)dx$, and the inverse of the cumulative distribution (sometimes known as the quantile function) by $\Phi^{-1}(p)$.[3]

[2] Fixing the mean and variance of X does not reduce generality.

[3] Microsoft Excel (2010) users will know these functions by the names $\phi(x)$ = NORM.S.DIST(x, FALSE), $\Phi(x)$ = NORM.S.DIST(x, TRUE), and $\Phi^{-1}(p)$ = NORM.S.INV(p). Corresponding functions in earlier versions of Excel bear similar names.

Bivariate sampling

In the bivariate sampling design, the sample Pearson correlation r_p is calculated directly from the observations (x_i, y_i), $i = 1, ..., n$ following Equation (1). Assuming that the data represent independent observations from the bivariate normal distribution as described above, no conversion is necessary because r_p provides a direct estimate of ρ. Conversion from r to z will be addressed in the next section.

Controlled experiment

For the controlled experimental design, I assume that the observations in the control group all have $x_i = q$ for $i = 1, ..., n_1$, while the observations in the treatment group all have $x_i = q + w$ for $i = n_1 + 1, ..., n$, where q and w are known values. Since I have assumed that X follows a standard normal distribution, the treatment-control differential w should be understood as being measured in standard deviation units of X. For instance, $w = 2$ means that the treatment group receives a value of X that is 2 standard deviations higher than the value received by the control group.

As previously, I assume that X and Y are bivariate normally distributed in the population. It then follows that, for the fixed value $X = q$, the outcome observations in the control group $y_1, ..., y_{n_1}$ are normally distributed with mean $\mu_y + \rho\sigma_y q$ and variance $\sigma_y^2(1 - \rho^2)$. Similarly, for the fixed value $X = q + w$, the outcome observations in the treatment group $y_{n_1+1}, ..., y_n$ are normally distributed with mean $\mu_y + \rho\sigma_y(q + w)$ and identical variance $\sigma_y^2(1 - \rho^2)$. Under these assumptions, an estimate of ρ can be constructed as

$$r_{ce} = \frac{\bar{y}_2 - \bar{y}_1}{\sqrt{(\bar{y}_2 - \bar{y}_1)^2 + w^2 s_p^2}} = \frac{d}{\sqrt{d^2 + w^2}}. \tag{4}$$

I denote this estimator r_{ce} to distinguish it from r_p as given in Equation (1) and from the other conversion formula given below. The derivation of r_{ce} is discussed in Appendix A. Because r_{ce} is an estimate of the population correlation ρ between X and Y, it is on the same scale as a Pearson sample correlation, and thus suitable for comparison and meta-analysis with other correlation coefficients. However, the simulation studies below

reveal that r_{ce} has substantial biases when the treatment-control differential w is small in magnitude, even for large sample sizes.

It should be noted that this conversion formula differs from the methods discussed in Borenstein (2009, p. 234) and in Hunter and Schmidt (2004, pp. 277-282). Though they are similar in form to Equation (4), both of those conversion formulas replace w^2 with the factor $a = (n_1 + n_2)^2/(n_1 n_2)$, which has to do with the proportion of observations in the treatment and control groups. However, in controlled experiments, these proportions are arbitrary (and often equal) and do not provide any information about the magnitude of the treatment-control differential on the continuous variable X. The minimum possible value of the factor a is achieved when the two groups are of equal size. In this case, $a = 4$, which is equivalent to assuming that the absolute value of the treatment-control differential is $w = 2$. For treatment-control differentials less than 2, results based on the conversion formulas of Borenstein (2009) or Hunter and Schmidt (2004) will differ from results based on Equation (4).

Extreme groups (and dichotomization)

The data from an extreme groups design (and also that from a dichotomization design) can be modeled by treating the cutoff values c_1 and c_2 as fixed quantities. For modeling purposes, I assume that the lower cutoff c_1 corresponds to the p_1^{th} quantile of X, while the upper cutoff c_2 corresponds to the $(1 - p_2)^{th}$ quantile of X. For instance, if c_1 and c_2 correspond to the lower and upper tertiles of a scale, then $p_1 = \frac{1}{3}$ and $p_2 = 1 - \frac{2}{3} = \frac{1}{3}$. Alternately, if sample cutoffs are defined by taking the n_1 lowest values and the n_2 highest values from the m sample cases with measured values of X, then $p_1 = n_1/m$ and $p_2 = n_2/m$. Because the modeling assumptions maintain that X is drawn from a standard normal distribution, it follows that $c_1 = \Phi^{-1}(p_1)$ and $c_2 = \Phi^{-1}(1 - p_2)$.

Given p_1 and p_2, the values $y_1, ..., y_{n_1}$ from the lower group are independent draws from the distribution of Y, conditional on the fact that $X \leq \Phi^{-1}(p_1)$; similarly, the values $y_{n_1+1}, ..., y_n$ from the upper group are independent draws from the distribution of Y,

conditional on the fact that $X > \Phi^{-1}(p_1)$. The distributions of $(Y|X < \Phi^{-1}(p_1))$ and $(Y|X > \Phi^{-1}(1-p_2))$ are known as skew-normal distributions (Arnold, Beaver, Groeneveld, & Meeker, 1993; Azzalini, 1985). Known properties of the skew-normal distribution can be used to derive a formula for converting a reported d statistic into an r statistic; for details of the derivation, interested readers should consult Appendix B.

To construct an estimate of ρ, I first need to define the auxiliary constants $f = n_1/(n_1+n_2)$, $v_j = \phi(c_j)/p_j$ for $j = 1, 2$, and

$$a = \frac{(v_1+v_2)^2}{fv_1(v_1+c_1)+(1-f)v_2(v_2-c_2)},$$
$$b = \frac{1}{\sqrt{fv_1(v_1+c_1)+(1-f)v_2(v_2-c_2)}}. \tag{5}$$

Note that f, c_1, c_2, v_1, and v_2 are all known quantities, depending only on the sample sizes n_1, n_2 and the proportions p_1, p_2 that define the cutoff values used in the design. Therefore, a meta-analyst should be able to calculate the constants a and b based only on information in a published article or research report. Using the constants a and b, a reported value of d can be converted to a correlation using

$$r_{eg} = \frac{bd}{\sqrt{d^2+a}}. \tag{6}$$

I denote this estimate r_{eg} to distinguish it from the other correlation coefficients given in Equations (1) and (4). Because r_{eg} is an estimate of ρ, it is on the same scale as the other correlation coefficients r_p and r_{ce}; therefore, r_{eg} is suitable for comparison or meta-analysis with these other correlation coefficients.

The formula for converting from a d statistic to r_{eg} can be understood as a two-step process. First, d is converted to an estimate of the point-biserial correlation between Y and an indicator variable for the lower group versus higher group, using the formula

$$r_{pbs} = \frac{d}{\sqrt{d^2+a}}. \tag{7}$$

However, the target parameter is the correlation ρ between the continuous measures X and Y, rather than the correlation between Y and group membership. In order to recover an

estimate of ρ, it is therefore necessary to multiply r_{pbs} by the constant b, producing the estimator given in (6).

Figure 1 plots the values of the constants a and b as functions of the cutoff proportions p_1 and p_2. These constants can be expressed more simply in some common special cases. The first special case is the balanced extreme groups design, in which $p_1 = p_2$ and $n_1 = n_2$; this case was studied by Feldt (1961). For balanced extreme groups, $c_1 = -c_2$, $v_1 = v_2$, and $f = 1/2$; therefore, the constants reduce to $a = 4v_1/(v_1 + c_1)$ and $b = [v_1(v_1 + c_1)]^{-1/2}$. As seen in Figure 1a, a is always greater than or equal to 4 and is at minimum when $p_1 = p_2 = 1/2$. As seen in Figure 1b, b decreases towards 1 as p_1 decreases towards 0; thus, the bias-correction factor becomes less consequential for designs with more extreme cut-offs.

The dichotomization design is another special case, in which $p_1 = 1 - p_2$ and $f = p_1$. For the dichotomization design,

$$a = \frac{1}{p_1(1-p_1)}, \qquad b = \frac{1}{v_1}\sqrt{\frac{1-p_1}{p_1}}. \tag{8}$$

Here, r_{eg} is equivalent to the estimator proposed by Hunter and Schmidt (1990). As depicted in Figure 1b, the bias correction factor b is always greater than 1.25 for dichotomization designs, and increases as the design becomes more imbalanced. Therefore, bias correction is crucial with this design.

A final special case is the median split, which can be viewed as both a dichotomization and a balanced extreme groups design. If $p_1 = p_2 = \frac{1}{2}$ and $n_1 = n_2$, then $c_1 = 0$ and the constants simplify further to $a = 4$ and $b = 1/v_1 = \sqrt{\pi/2} \approx 1.253$.

It may occur to statistically inclined readers that the formula for r_{eg} could be improved by replacing constants a and b with other constants so that the numerator and squared denominator in (6) are both exactly unbiased. More precisely, one might use constants a_n and b_n such that (i) $E[(\bar{y}_2 - \bar{y}_1)^2 + a_n s_p^2]$ does not depend on ρ and (ii)

$b_n E(\bar{y}_2 - \bar{y}_1)/\sqrt{E[(\bar{y}_2 - \bar{y}_1)^2 + a_n s_p^2]} = \rho$. Constants satisfying these criteria are given by

$$a_n = (n_1 + n_2 - 2)\left(\frac{(v_1+v_2)^2 - \frac{1}{n_1}v_1(v_1+c_1) - \frac{1}{n_2}v_2(v_2-c_2)}{(n_1-1)v_1(v_1+c_1) + (n_2-1)v_2(v_2-c_2)}\right),$$

$$b_n = \frac{\sqrt{\frac{1}{n_1} + \frac{1}{n_2} + a_n}}{v_1 + v_2}.$$

The constants given in (5) are the limits of a_n and b_n as n_1, n_2 increase towards infinity in fixed proportion. Based on simulation results, it appears that using a_n and b_n in (6) reduces the estimator bias under certain circumstances, but also increases the mean-squared error. The differences are generally so slight as to be inconsequential; therefore this refinement is not considered further.

Converting from r to z

There is debate about whether correlation coefficients are best meta-analyzed on the natural scale or after applying Fisher's z-transformation, which is defined as

$$z(r) = \frac{1}{2}\log\frac{1+r}{1-r}, \tag{9}$$

where log denotes the natural logarithm. Note that the natural r scale ranges from -1 to 1, while $z(r)$ takes values of r in the interval $(-1, 1)$ and produces values in the range from $-\infty$ to ∞. Hunter and Schmidt (2004) advocate meta-analyzing correlations on the natural scale, while Hedges and Olkin (1985) propose applying the z-transformation. The latter authors recommend z-transformation because the resulting estimator is closer to being normally distributed and has a variance that does not depend as strongly on ρ, both of which are desirable properties for meta-analysis. Of course, the transformed estimator is no longer an estimate of ρ, but rather of $z(\rho)$. For random-effects meta-analysis, the debate over which method is preferable would seem to hinge on differences in maintained modeling assumptions, and specifically on whether the prior distribution of the random effects is on the scale of ρ or of $z(\rho)$ (Field, 2005; Hafdahl, 2009; Hafdahl & Williams, 2009; Schulze, 2004). Rather than weighing in on the debate regarding the merits of z-transformation, this section merely provides conversion formulas that can be applied if deemed appropriate.

Bivariate sampling. Fisher's z-transformation was developed for use with Pearson sample correlation coefficients, so Equation (9) is directly applicable to r_p effect sizes from bivariate sampling designs. For ease of notation, I denote $z(r_p)$ as z_p. For the other designs, use of Fisher's z-transformation warrants further consideration, because the transformation does not have the same normalizing and variance-stabilizing properties when applied to other correlation coefficients.

Controlled experiment. At least in terms of algebra, it is possible to use Equation (9) to transform a correlation estimate r_{ce} from a controlled experiment because the absolute value of r_{ce} will never exceed 1. I denote $z(r_{ce})$ as z_{ce}.

Extreme groups (and dichotomization)

Applying Fisher's z-transformation to r_{eg} from an extreme groups or dichotomization design presents further difficulties, even on the basic level of algebra, because the absolute value of r_{eg} may exceed 1. Since $z(r)$ is undefined for values $|r| \geq 1$, applying the transformation directly to r_{eg} is not always possible. To see that values of r_{eg} can fall outside of the domain of $z(r)$, first note that r_{pbs} can take values anywhere in the range (-1,1) because d can be arbitrarily large; r_{pbs} will never fall outside (-1,1), though, because a is positive and thus d (the numerator of r_{pbs}) will always be less than $\sqrt{d^2 + a}$ (the denominator or r_{pbs}). Next, note that bias-correcting r_{pbs} to get r_{eg} involves multiplication by a constant b that is always greater than one (as is apparent in Figure 1b). It follows that, for any values of p_1 and p_2, the absolute value of r_{eg} may be almost as large as b, and therefore may be larger than one. Due to this possibility, transforming r_{eg} to the z scale must involve more than just applying the formula.

As shown in the simulation studies discussed below, values of $|r_{eg}|$ greater than one have a non-negligible probability of occuring in small samples, especially when ρ is large in magnitude. If one is meta-analyzing correlation estimates on the natural scale, estimates falling outside the range are not necessarily problematic. In contrast, if using the z-transformation, such estimates must be handled in some fashion. Here I consider two

intuitive approaches to doing so.

The first approach is simply to truncate r_{eg} at some value close to 1. For some constant $s > 0$, define

$$z_s^S(r_{eg}) = z\left(\max[\min(r_{eg}, 1 - 10^{-s}), 10^{-s} - 1]\right). \tag{10}$$

For small or moderate values of ρ, truncation should have negligible impact on the distribution of z_s^S because few values of r_{eg} will fall outside the range of $(10^{-s} - 1, 1 - 10^{-s})$. Truncation will have greater consequences if the true value of ρ is very large. In the simulation study described below, a value of $s = 12$ is used, which corresponds to a maximum possible value of 14.16 for z_{12}^S.

The second approach is to use a Taylor series approximation to $z(\cdot)$ around the point $r_{pbs} = r_{eg}/b$. For a fixed number of terms t, define the function

$$z_t^T(r, b) = \sum_{k=0}^{t} \frac{(b-1)^k (r)^k}{k!} z^{(k)}(r), \tag{11}$$

where $z^{(k)}(\cdot)$ is the k^{th} derivative of $z(\cdot)$. Table 2 provides the first five derivatives of $z(\cdot)$. Recall that r_{pbs} will always fall within (-1,1); thus, $z^{(k)}(r_{pbs})$ and $z_t^T(r_{pbs}, b)$ will always be defined. The Taylor series estimator z_t^T will provide a close approximation to $z(r_{eg})$ when $|\rho|$ is not close to one and when b is small. For larger sample sizes, the number of terms must be increased to maintain the consistency of the transformed estimator. In the simulation study below, $t = 5$ is used.

Variance estimation

In this section, I present the large-sample variances of the transformed r and z estimators for each of the four designs.[4] Large-sample variances are approximations, but are used commonly in meta-analysis because exact results are seldom available. All of the

[4]By large-sample variance, I mean in technical terms that for a sequence of random variables $Q_1, Q_2, Q_3, ...$, if $[Q_n - E(Q_n)]/\sqrt{V_n^Q}$ converges in distribution to a standard normal random variable as n increases, then V_n^Q is a large-sample variance of Q_n.

large-sample variances presented here were derived using the delta method (for details, see Casella & Berger, 2002, p. 243).

Bivariate sampling

Hedges and Olkin (1985) detail the properties of r_p, assuming as here that X and Y are bivariate normally distributed. They note in particular that r_p has a small bias, the magnitude of which depends on sample size n, and that r_p has large-sample variance

$$\text{Var}(r_p) \approx V_p^r = \frac{\left(1 - r_p^2\right)^2}{n} \tag{12}$$

(Hedges & Olkin, 1985, p. 226).[5] Note that this variance depends on the value of the correlation, which is one motivation for converting from r to Fisher's z scale for purposes of meta-analysis. z-Transformation removes this dependence, so that the large-sample variance of z_p is

$$\text{Var}(z_p) \approx V_p^z = \frac{1}{n-3}. \tag{13}$$

Using $n - 3$ rather than n in the denominator provides a closer approximation to the exact variance of z_p in small samples (Hedges & Olkin, 1985, p. 227).

Controlled experiment

The large-sample variance of r_{ce} can be derived by first considering the variance of the d statistic from a controlled experimental design. Under the assumptions of this design, the large-sample variance of d is

$$V_{ce}^d = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)}, \tag{14}$$

(Borenstein, 2009, p. 226). Applying the delta method to the transformation formula given in Equation (4) yields the large sample variance of r_{ce}:

$$V_{ce}^r = \frac{w^4}{[d^2 + w^2]^3}\left(\frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)}\right). \tag{15}$$

[5] Some textbooks present large-sample variances in terms of the population parameters involved; for instance, Hedges and Olkin (1985) gives the large-sample variance of r_p as $(1-\rho^2)^2/n$. In all of the formulas in this section, I substitute sample estimators in place of the corresponding parameters, because this is how large-sample variance estimates will actually be computed.

Another application of the delta method, this time to the z-transformation function given in Equation (9), yields the large-sample variance of z_{ce}:

$$V_{ce}^{z} = \frac{1}{d^2 + w^2}\left(\frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)}\right), \tag{16}$$

which is somewhat simpler than V_{ce}^{r}.

Extreme groups (and dichotomization)

Large-sample variance expressions for effect size statistics from the extreme groups design (including the dichotomization design as a special case) can be developed along the same lines as for the controlled experimental design. I first give expressions for the large-sample variance of the d statistic from an extreme groups design, then use these to express the large-sample variances of r_{eg} and z_{eg}, where z_{eg} is based on either of the two transformation approaches discussed in the previous section. Unfortunately, a general expression for the large-sample variance of d involves some rather cumbersome algebra. Rather than present it here, I instead give expressions for the special cases of balanced extreme groups designs, dichotomization designs, and median splits, all of which are somewhat more tractable. Appendix B provides a general expression for the large-sample variance of d from an extreme groups design.

Consider first the balanced extreme groups design, where $p_2 = p_1$ and $n_1 = n_2$. With the auxiliary constants a and b as defined in Equation (5), the large-sample variance of d from this design is:

$$V_{eg}^{d} = \frac{1}{n}\left(4 + \frac{d^2}{2} - d^4\frac{a - ab^2 + 4}{a^2} - d^6\frac{64 + 64a + 4a^2 - 48ab^2 - 4a^2b^2}{64a^3}\right). \tag{17}$$

Next, consider the dichotomization design and recall that $c_1 = \Phi^{-1}(p_1)$ and $v_1 = \phi(c_1)/p_1$.

If X is dichotomized at the p_1^{th} quantile so that $p_1 = 1 - p_2 = n_1/n$, then

$$V_{eg}^d = \frac{1}{n}\left[\frac{1}{p_1(1-p_1)} - d^2\frac{v_1(9p_1^2 - 9p_1 + 2) + 2c_1(1-p_1)(1-2p_1)}{2v_1p_1(1-p_1)}\right.$$
$$- \frac{d^4}{v_1^2}\left(2v_1^2(3p_1^2 - 3p_1 + 1) + 3c_1v_1(1-p_1)(1-2p_1) - (1-c_1^2)(1-p_1)^2\right)$$
$$\left. - \frac{d^6p_1(1-p_1)}{4v_1^3}\left(2v_1^3(5p_1^2 - 5p_1 + 2) + 8c_1v_1^2(1-p_1)(1-2p_1) + (5c_1^2 - 4)v_1(1-p_1)^2\right)\right]. \tag{18}$$

Finally, if d is calculated using a median split so that $p_1 = p_2 = \frac{1}{2}$ and $n_1 = n_2$, a further simplification is obtained. In this case,

$$V_{eg}^d = \frac{1}{n}\left(4 + \frac{d^2}{2} - d^4\frac{4-\pi}{8} + d^6\frac{\pi - 3}{32}\right).$$

Applying the delta method to the transformation functions given in (6) and (9) produces large-sample variances for r_{eg} and z_{eg}, respectively. The large sample variance of r_{eg} can be expressed as a function of the auxiliary constants a and b and the large-sample variance of d:

$$V_{eg}^r = \frac{a^2b^2}{(d^2+a)^3}V_{eg}^d. \tag{19}$$

The large sample variance of z_{eg} can be expressed in similar fashion:

$$V_{eg}^z = \frac{a^2b^2}{(d^2+a)(d^2(1-b^2)+a)^2}V_{eg}^d. \tag{20}$$

For calculation purposes, one should first find V_{eg}^d using one of the formulas given in (17) or (18) (or the general formula in Appendix B), then use this computed value to evaluate Equation (19) or (20).

Even in the special cases considered here, the large-sample variance of d involves lengthy formulas. Consequently, the working meta-analyst may be inclined to use other, simpler formulas as rough approximations, and it is instructive to consider how accurate doing so would be. First, one might think to use the simpler formula for the variance of d from a controlled experiment V_{ce}^d, as given in Equation (14). Note that in a balanced

extreme groups design where $p_1 = p_2$ and $f = 1/2$, the variance under normality is equal to

$$V_{ce}^{d} = \frac{1}{n}\left(4 + \frac{d^2}{2}\right),$$

which is the same as the two leading terms in (17). Thus, the discrepancy between this approximation and the actual V_d will depend on the magnitude of d and on the value of p_1, which affects the auxiliary quantities a and b.

Figure 2 plots the ratio of V_{ce}^{d} to V_{eg}^{d} as a function of ρ (restricting to $\rho \geq 0$ because of symmetry). Figure 2a presents the balanced extreme groups design, assuming $p_2 = p_1$ and $n_2 = n_1$. In the balanced extreme groups design, for $p_1 = 1/3$ or more, the approximation V_{ce}^{d} exceeds the true variance, particularly for ρ greater than 0.6; however, for smaller values of ρ, the approximation is quite close. Furthermore, as p_1 decreases, the controlled experiment variance provides a very close approximation to the true variance over a wider and wider range of values for ρ. Thus, using V_{ce}^{d} in place of V_{eg}^{d} in Equations (19) or (20) might provide a reasonable approximation for the extreme groups design, particularly for more extreme cutoffs. Such is not the case in the dichotomization design, presented in Figure 2b. Here, the approximation becomes progressively worse for smaller values of p_1.

It is similarly instructive to compare the true large-sample variance V_{eg}^{r} to other possible approximations that are simpler to compute. For instance, Hunter and Schmidt (2004) propose to use the large-sample variance from a bivariate sampling design V_{p}^{r} as an estimate of $\mathrm{Var}(r_{eg})$ for dichotomization designs. Figure 3 plots the ratio of V_{p}^{r} to V_{eg}^{r} as a function of ρ, and is constructed in the same fashion as Figure 2. For both the balanced extreme groups design (Figure 3a) and the dichotomization design (Figure 3b), it is apparent that V_{p}^{r} provides a very poor approximation to the true large-sample variance.

Simulation study

The converted effect size estimators r_{ce} and r_{eg} are both based on large-sample approximations, rather than exact distribution theory. It is therefore important to study their statistical properties for sample sizes that are likely to be encountered in practice. To

do so, I conducted a set of simulations based on controlled experiments, dichotomization, and extreme groups designs. To provide points of reference, I also simulated the Pearson correlation coefficient r_p from a bivariate sampling design. Table 3 summarizes the structure of the simulation for each of the four study designs. For each combination of design parameters, the number of iterations was taken as $10^7/n$. All of the simulations and calculations were performed with R, version 2.15.0.

For each of the designs, I simulated several quantities. First, I computed the biases of the transformed r and z estimators.[6] For dichotomization and extreme groups designs, I also simulated the probability that $|r_{eg}| \geq 1$ in order to establish the extent to which alternative z-transformation methods are needed. Finally, I calculated the relative biases of the large-sample variance estimators, where the relative bias of an estimator for $\text{Var}(r)$ is defined as

$$Relative\ Bias(V^r) = \frac{E(V^r)}{\text{Var}(r)} - 1,$$

and is similarly defined for estimators of $\text{Var}(z)$. For the dichotomization and extreme groups designs, it is also of interest to compare the full large-sample variances to the approximations discussed in the previous section. This section reports only selected simulation results, focusing on those where a meta-analyst will face choices about how to perform an analysis. Full simulation results, with accompanying R code, are available upon request.

Bivariate sampling

For the bivariate sampling design, the simulation consisted of a 96×4 factorial structure; true correlations ρ varied from 0 to 0.95 in increments of 0.01, while total sample size n was set equal to 20, 40, 80, or 160 in order to represent studies ranging from small to large. Figure 4 presents simulation results for this study design, which are useful as reference points for interpreting the simulation results regarding the other study designs.

[6]Note that z estimators are compared to $z(\rho)$, rather than to ρ. That is, for a given true correlation ρ, the bias of z is defined as the expected value of z minus $z(\rho)$.

As seen in Figures 4a and 4b respectively, the Pearson sample correlation coefficient r_p and the z-transformed statistic z_p both have biases of negligible magnitude, even at the smallest sample size considered; when $n = 20$, the absolute bias of r_p is never more than 0.011, while the absolute bias of z_p is never more than 0.027. The large-sample variance of r_p has a negative bias that is most pronounced when $\rho = 0$ (Figure 4c), while the large-sample variance of z_p has a small, positive relative bias that increases with ρ (Figure 4d). The greater accuracy of V_p^z versus V_p^r provides further justification for using z-transformation in meta-analyses of correlations.

Controlled experiment

For the controlled experimental design, the simulation consisted of a $96 \times 4 \times 3$ factorial structure; in addition to varying the true correlation ρ and the sample size n, the simulations examined treatment-control differentials w of 0.5, 1.0 and 2.0. Figure 5 displays the simulation results, and is organized just like Figure 4. Based on Figure 5a, the bias of r_{ce} is larger than that of r_p and is more pronounced for smaller values of the treatment-control differential. For $w = 0.5$, r_{ce} has a large, negative bias, even at the largest sample size considered; unfortunately, this bias would be difficult to correct due to its non-linear relationship with ρ. A similar pattern of results holds for the bias of z_p (Figure 5b). As with the bivariate sampling design, the large-sample variance of r_{ce} (V_{ce}^r, Figure 5c) is in general less accurate than the large-sample variance on the z scale (V_{ce}^z, Figure 5d).

Dichotomization

The simulations for the dichotomization design varied the true correlation ρ, the sample size n, the cutoff percentiles p_1, and whether the cutoffs were population- or sample-based, producing a $96 \times 4 \times 4 \times 2$ factorial structure. Results for the smallest cutoff value $p_1 = \frac{1}{8}$ are not presented, as such an extreme dichotomization seems unlikely to be used in practice. In the data-generating model for the dichotomization design, I assumed that observations in each group are drawn from a skew-normal distribution with fixed truncation points (i.e., a population cutoff), but in practice the truncation points are often

based on sample percentiles (i.e., a sample cutoff). I therefore simulated both approaches.

Figure 6a displays the bias of r_{eg} versus ρ for the dichotomization design. Each panel of the figure shows a different combination of the value of p_1 and whether percentiles were fixed or sample-based. Using a sample cutoff rather than a population cutoff results in greater bias. The estimator bias is larger for more extreme cutoffs; for a cutoff $p_1 = \frac{1}{5}$, the most negative bias occurs when ρ is approximately 0.8. Note that if one were to ignore the bias-correction factor b when calculating r_{eg}, the bias of the resulting estimator would be more negative than the bias given here. Moreover, such bias would be present even in large samples.

The simulations also examined the extent to which r_{eg} falls outside of the domain of $z(\cdot)$. Figure 6b plots the probability that $|r_{eg}| < 1$ for $0.4 \leq \rho \leq 0.95$, with varying levels of p_1 and n. The figure is based on using sample cutoffs; results with population cutoffs are very similar. As one would expect, values of $|r_{eg}|$ greater than one occur mostly for large ρ and small sample sizes. For example, with a median split ($p_1 = \frac{1}{2}$) and $n = 20$, $\Pr(|r_{eg}| \geq 1)$ is less than 0.95 when $\rho \geq 0.8$. Values outside of (-1,1) occur more frequently for more extreme cutoffs. Overall, I would argue that having one in twenty estimates fall outside of the domain of $z(\cdot)$ presents a sufficiently common problem that it warrants further consideration; specifically, different approaches to z-transformation should be considered carefully.

In a previous section, I outlined two different methods of z-transformation that deal with the possibility that $|r_{eg}|$ may be greater than or equal to one. Figure 6c compares the bias of both of these, plotting the bias of the truncated estimator z_{12}^S and the Taylor series approximation z_5^T using sample cutoffs; results for population cutoffs are similar but with somewhat larger biases. The figure is restricted to $\rho \geq 0.5$ because the bias of each estimator is very small except at large values of ρ. Across levels of ρ, p_1, and n, z_5^T has smaller biases than z_{12}^S. This is perhaps surprising, considering that the Taylor series approximation about r_{pbs}/b always under-estimates $z(\cdot)$ for high values of ρ. It seems that

applying the z-transformation directly to r_{eg} introduces a positive bias, which is partially mitigated by using the Taylor series approximation. However, for the Taylor series with a fixed number of terms, it should be noted that z_5^T will become negatively biased as sample size grows. This can be observed in the lower panels of Figure 6c, where the bias of z_5^T bends below zero for larger sample sizes. As noted previously, the bias from the approximation can be removed by using more terms in the Taylor series.

Finally, the simulations examined the performance of several estimators for the variance of r_{eg} and z_{eg}, focusing on z_5^T due to its smaller bias. In particular, I compared the estimators using the full large-sample variance V_{eg}^d, as given in (18) to those that use the approximation based on a controlled experiment V_{ce}^d, as given in (14). Figures 7a and 7b display the relative bias of V_{eg}^r versus V_{ce}^r and V_{eg}^z versus V_{ce}^z, respectively, for varying sample sizes. For sake of space, the figures are restricted to sample cutoffs with $p_1 = \frac{1}{2}$; results for population cutoffs and other values of p_1 are qualitatively similar. Additionally, Figure 7b omits $n = 20$ because both variance estimators are highly biased for this sample size. Based on Figure 7a, the actual large-sample variance V_{eg}^r is close to unbiased for all but very high values of ρ, whereas the controlled experiment approximation V_{ce}^r tends to over-estimate for $\rho > 0.4$. A similar pattern holds in Figure 7b for the z_{eg} variance estimators. Thus, V_{eg}^r and V_{eg}^z are recommended for estimating $\text{Var}(r_{eg})$ and $\text{Var}(z_{eg})$ from a dichotomization design. However, it should be noted that V_{eg}^z performs poorly for samples sizes less than $n = 40$.[7]

Extreme groups

The final set of simulations were for balanced extreme groups designs, assuming that $p_1 = p_2$ and $n_1 = n_2$. The simulations were structured just as for the dichotomization design, as a 96 (true correlation ρ) $\times$ 4 (sample size n) $\times$ 4 (cutoff percentile p_1) $\times$ 2 (population versus sample cutoffs) factorial design. Results for $p_1 = \frac{1}{2}$ are not displayed

[7] As suggested by one anonymous reviewer, meta-analysts requiring an accurate variance estimate from a small sample may want to consider using re-sampling techniques such as bootstrapping rather than using V_{eg}^z.

because they are identical to those for the dichotomization design.

Figure 8a display the bias of r_{eg} versus ρ for the extreme groups design, and is constructed just like Figure 6a. As with the dichotomization design, using sample cutoffs results in somewhat larger biases than using population cutoffs. Still, r_{eg} has only small bias, comparable to that of r_p from bivariate sampling. For $p_1 = \frac{1}{3}$, the absolute bias is always less than 0.02 at the smallest sample size considered; furthermore, in designs using sample percentiles, the remaining bias decreases with more extreme cut-points.

Figure 6b plots the probability that $|r_{eg}|$ is less than one for $0.4 \leq \rho \leq 0.95$, varying levels of p_1 and n while using sample cutoffs; the results with population cutoffs are very similar. In contrast to the dichotomization design, values outside of the interval (-1,1) are somewhat less of a concern in the extreme groups design; furthermore, more extreme cut-points lead to lower probability of observing estimates outside of the interval.

Figure 8c compares the bias of the truncated estimator z_{12}^S and the Taylor-series approximation z_5^T using sample cutoffs, and is constructed just like Figure 6c. As with the dichotomization design, z_5^T always has smaller biases than z_{12}^S. Thus, even though values of r_{eg} will seldom fall outside the interval (-1,1), using the Taylor-series approximation to Fisher's z-transformation is still recommended.

Figures 9a and 9b display the relative bias of V_{eg}^r versus V_{ce}^r and V_{eg}^z versus V_{ce}^z, respectively, for varying sample sizes and for population and sample cutoffs. For sake of space, the figures are restricted to $p_1 = \frac{1}{3}$; results for other values of p_1 are qualitatively similar. As with the dichotomization design, Figure 9b excludes the smallest sample size because both estimators are badly biased when $n = 20$. For designs with population cutoffs, the actual large-sample variances V_{eg}^r and V_{eg}^z are close to unbiased for all but very high values of ρ, whereas the controlled experiment approximations V_{ce}^r and V_{ce}^z tend to over-estimate for $\rho > 0.4$. Therefore, in extreme groups designs with population cutoffs, the full large-sample variances should be used rather than the approximations based on controlled experiments. However, the opposite conclusion holds when sample cutoffs are

used. In this case, V^r_{ce} maintains smaller biases than V^r_{eg} across values of ρ; the comparison between V^z_{ce} and V^z_{eg} is similar. Therefore, if an extreme groups design uses sample cutoffs, the simpler approximations V^r_{ce} and V^z_{ce} should be used for estimating the variances of r_{eg} and z_{eg}, respectively. Finally, it should be noted that both V^z_{eg} and V^z_{ce} are badly biased for small sample sizes, just as in dichotomization designs.

Discussion

Based on the simulation study, there are several general recommendations for meta-analytic practice. First, in extreme groups and dichotomization designs, the Taylor series approximation given in Equation (11) should be preferred because using it produces a less biased transformed statistic than using the original transformation function. Second, for estimating the variance of r_{eg} and z_{eg}, the full large-sample variance formulas should be used rather than the approximations based on V^d_{ce} or V^r_p, except in extreme groups designs that use sample cutoffs; in the latter circumstance, V^d_{ce} should be used rather than V^d_{eg}. Finally, r- or z-type effect sizes that are converted based on standardized mean differences from controlled experiments can be badly biased when the treatment-control differential is small, and should be treated with caution.

These conclusions, and the simulation results in general, are limited in that only balanced designs have been considered. In practice, one will certainly encounter designs that have some degree of imbalance, such as controlled experiments with unequal sample sizes or extreme groups designs with sample sizes in each group that are not proportional to the cutoff percentiles. One may reasonably expect that the results and recommendations given here will hold in designs with small or moderate degrees of imbalance. In very imbalanced designs, such as dichotomizations that use $p_1 < \frac{1}{5}$ or extreme groups with very unequal sample sizes, one should proceed with more caution and check the sensitivity of conclusions to the various analytic approaches suggested here. The following section illustrates some sensitivity analyses that the analyst should consider.

Examples

This section revisits the motivating examples presented for each design, demonstrating how to convert reported d statistics into r- and z-type effect sizes and how to estimate the variance of those statistics. Readers may also want to consult the spreadsheet available in the online supplementary materials, which implements all of the formulas presented in previous sections and demonstrates the calculations used in the following examples.

Example 1. Using a bivariate sampling design, Chatzisarantis et al. (2007) reported a Pearson correlation of $r_p = 0.58$ between intention to exercise and perceive autonomy support. Inserting r_p and $n = 165$ into Equation (12), the sampling variance of this estimate is $V_p^r = (1 - 0.58^2)^2/165 = 0.0027$, corresponding to a standard error of 0.05. To convert the effect size and variance to the z-scale, I use the transformation formula given in Equation (9), finding that $z_p = [\log(1 + 0.58) - \log(1 - 0.58)]/2 = 0.66$, and the sampling variance from (13), finding that $V_p^z = 1/(165 - 3) = 0.0062$. Although z_p is more difficult to interpret than the effect size on the r scale, it is suitable for meta-analysis with other z-transformed correlations.

Example 2. Edmunds (2008) used a controlled experimental design to manipulate individuals' senses of autonomous support and measure the effect on their intentions to exercise. They reported statistics that correspond to a standardized mean difference effect size of $d = 0.70$. To convert this effect size into an r or z that estimates the correlation between autonomous support and intention to exercise, one must make an assumption about the treatment-control differential induced by the experimental manipulation. Here, I rely on the self-reported measures of autonomous support collected over the course of the experiment. I use the larger differential of $w = 2.18$ in initial calculations, then check the sensitivity of the estimates to a reduced treatment-control differential of $w = 1.22$.

Inserting $d = 0.70$ and $w = 2.18$ into Equation (4) leads to an estimated correlation of $r_{ce} = 0.70/\sqrt{0.70^2 + 2.18^2} = 0.30$. Using d and w with reported sample sizes of $n_1 = 31$ and

$n_2 = 25$, the estimated variance of r_{ce} from Equation (15) is $V^r_{ce} = 0.0121$, corresponding to a standard error of 0.11. Transforming r_{ce} to a z effect size by using Equation (9), I find that $z_{ce} = 0.31$, with a variance based on Equation (16) of $V^z_{ce} = 0.0146$.

Repeating the calculations with a smaller value of the treatment control differential $w = 1.22$ leads to $r_{ce} = 0.50$, $V^r_{ce} = 0.0222$, $z_{ce} = 0.54$, and $V^z_{ce} = 0.0390$. The estimated effect sizes are much larger than those based on the larger treatment control differential, with only slightly increased variances. This example illustrates that in controlled experiments, conversion from d to r is very sensitive to one's assumption regarding the treatment-control differential.

Example 3. Mussweiler et al. (2000) described the association between self-esteem and gender-focused self-identity by dichotomizing a measure of self-esteem and reporting the difference in gender-focus between low and high self-esteem groups. Based on reported statistics, the standardized mean difference between the groups is $d = 1.14$. The authors dichotomized their sample using a cutoff point based on the median of a larger study, which implies that $p_1 = 0.5$. Because this study used a median split, the auxiliary constants are particularly simple to calculate: $a = 4$ and $b = 1.253$. Using these auxiliary constants in Equation (6) leads to the converted effect size of $r_{eg} = 1.253 \times 1.14/\sqrt{1.14^2 + 4} = 0.62$. This is an estimate of the underlying Pearson correlation between the continuous measures of self-esteem and gender-focus. Note that if the bias-correction factor were ignored, one would find that $r_{pbs} = 0.50$, which is 20% smaller than r_{eg}.

To find the large-sample variance of r_{eg}, I first calculate the variance of the d statistic. Unfortunately, the simplified special case formulas for dichotomization and median split designs do not apply because the sample sizes of the two groups are not equal; using Equation (27) from Appendix B, I find that $V^d_{eg} = 0.1954$. Inserting this value together with $a = 4$, $b = 1.253$, and $d = 1.14$ into Equation (19), I find that $V^r_{eg} = 0.0329$. To transform the r-type effect size to the z scale, I use the Taylor-series approximation formula given in (11) and find that $z^T_5 = 0.73$ (using z^S_{12} produces a nearly identical estimate that

matches z_5^T to four decimal places). The large-sample variance of this effect size can be calculated by evaluating Equation (20) with $V_{eg}^d = 0.1954$ to produce $V_{eg}^z = 0.0874$.

Because the two groups created by the dichotomization were not of equal size ($n_1 = 19$ versus $n_2 = 8$), it is sensible to check the sensitivity of one's calculations to different values of the cutoff p_1. If the proportions in the two sample groups were representative of the entire population, one would have instead $p_1 = 19/(19 + 8) = 0.70$. Based on this cutoff, the auxiliary constants are calculated as $c_1 = \Phi^{-1}(0.70) = 0.535$ and $v_1 = \phi(0.535)/0.70 = 0.491$. Inserting $v_1 = 0.491$ and $p_1 = 0.70$ into Equation (8) leads to $a = 4.796$ and $b = 1.321$. Repeating the calculations with these auxiliary constants, I find $r_{eg} = 0.61$, $V_{eg}^r = 0.0332$, $z_5^T = 0.71$, and $V_{eg}^z = 0.0846$. These results are all quite similar to those based on $p_1 = 0.5$, implying that the transformed effect sizes are not sensitive to the assumption about how the dichotomization was created.

Example 4. Goldinger et al. (2003) studied the relationship between working memory and pro-social behavior using a balanced extreme groups design with sample-based cutoffs. The researchers used the lower and upper quartiles as cutoffs, corresponding to $p_1 = 0.25$ and $p_2 = 0.25$. The standardized mean difference between groups on a measure of pro-social behavior was $d = 1.02$. To convert this effect size to an r or z, I begin by calculating several auxiliary constants. First, based on the reported cutoffs $p_1 = p_2 = 0.25$ and sample sizes $n_1 = n_2 = 35$, I calculate that $c_1 = \Phi^{-1}(0.25) = -0.674$ and $v_1 = \phi(-0.674)/0.25 = 1.271$; because the design is balanced, it follows that $c_2 = -c_1 = 0.674$ and $v_2 = v_1 = 1.271$. Based on these constants and Equation (5), the further auxiliary constants are $a = 4 \times 1.271/(1.271 - 0.674) = 8.522$ and $b = [1.271 \times (1.271 - 0.674)]^{-1/2} = 1.148$. Inserting the calculated values of a, b, and d into the conversion formula given in (6) yields $r_{eg} = 0.38$. Note that ignoring the bias correction factor would produce the estimate $r_{pbs} = 0.33$, which is more than 10% smaller than r_{eg}. Applying the Taylor series approximation to Fisher's z-transformation given in (11), I find that $z_5^T = 0.40$. (The truncation approach produces an estimate z_{12}^S that differs from z_5^T by

less than 10^{-4}.)

Next, I calculate variance estimates for r_{eg} and z_{eg} using the actual large-sample variance formulas and compare these to the naive approximations based on V_{ce}^{d}. Using $d = 1.02$ along with the calculated values of $a = 8.522$ and $b = 1.271$, Equation (17) evaluates to $V_{eg}^{d} = 0.0643$. In this case, the naive approximation is quite close: $V_{ce}^{d} = 0.0646$. Together with calculated values of a, b, and d, inserting the value $V_{eg}^{d} = 0.0643$ into Expressions (19) and (20) yields $V_{eg}^{r} = 0.0070$ and $V_{eg}^{z} = 0.0096$, which correspond to estimated standard errors of 0.08 and 0.10, respectively. Finally, note that using the naive approximation V_{p}^{r} as an estimate of the variance of r_{eg} is not advised: in this example, Equation (12) evaluates to $V_{p}^{r} = 0.0105$, which is nearly 50% larger than V_{eg}^{r}.

General discussion

This paper has presented methods for converting effect sizes from d to r to z when the d statistics come from controlled experiments, dichotomizations, or extreme groups designs. For the latter two study designs, the general formulation of the data-generating model illustrates the connection between the conversion formula originally derived by Feldt (1961) for balanced extreme groups designs and the formula proposed by Hunter and Schmidt (1990) for dichotomization designs.

Currently, the literature on meta-analytic methods offers conflicting guidance regarding effect size conversion, with different schools of advice regarding the importance of bias correction and the appropriate formulas for variance estimation. The theoretical and simulation results presented in this paper suggest that each of these schools needs some modification, particularly regarding the following four aspects. First, some traditions of meta-analysis ignore the bias in the point-biserial correlation coefficient as an estimate of an underlying population correlation (e.g., Borenstein, 2009; Lipsey & Wilson, 2001). In contrast, the above results demonstrate that an approximately unbiased estimator can be obtained using Expression (6), which involves a bias-correction factor.

Second, though it has emphasized the importance of bias correction when converting

effect sizes in dichotomization designs, the psychometric tradition within meta-analysis has relied on variance formulas based on bivariate normal sampling, which are inappropriate for effect sizes from extreme groups designs, dichotomizations, and controlled experiments. The correct large-sample variance formulas provided in this paper are markedly different from the large-sample variance of the Pearson correlation coefficient under bivariate normality. Under limited circumstances, a simpler formula based on the variance from a controlled experimental design provides an adequate approximation to the variance. Based on the simulation study, V_{ce}^{d} should be used to estimate the variance of converted effect sizes when the data come from an extreme groups design with sample cutoffs. In other circumstances, and particularly for dichotomization designs, the actual asymptotic variance formulas are less biased and therefore preferable, despite being more tedious to calculate.

Third, I have proposed a model relating standardized mean differences from controlled experiments to the Pearson correlation coefficient between continuous variables. This model leads to a formula for converting effect sizes that differs from existing proposals (Borenstein, 2009; Hunter & Schmidt, 2004), and highlights that strong assumptions regarding the treatment-control differential are needed for the conversion to produce sensible results. Although correlational meta-analyses sometimes include converted standardized mean differences from controlled experiments, doing so seems to be of questionable value, and the converted effect sizes should be treated with skepticism. Even in the limited circumstances when a credible treatment-control differential can be estimated or assumed, the converted effect sizes are very sensitive to it; moreover, the results can be badly biased if the treatment-control differential is small.

Finally, there remains controversy regarding whether correlations should be analyzed in their natural r metric or after applying Fisher's z-transformation. The present results do not speak to this controversy directly, though they do address a complication that arises if bias-corrected correlation estimates are to be analyzed after z-transformation. The complication is created because bias-correction may lead to estimated correlations outside

of the interval (-1,1), making direct application of $z(\cdot)$ impossible. Rather than truncating the estimated correlation prior to transformation, I have proposed using a Taylor-series approximation to $z(\cdot)$. Mechanically, this produces an estimate of $z(\rho)$ even when r_{eg} falls outside the domain of $z(\cdot)$; more importantly, using this approximation produces a less-biased estimate than using a truncation approach. The simulation study used a Taylor-series approximation with five terms, but this choice is arbitrary; future work will need to provide guidance as to how many terms should be used for a given type of design and sample size.

It is interesting to note that the meta-analytic tradition that favors z-transformation has tended to ignore bias correction (Borenstein, 2009; Hedges & Olkin, 1985), while the tradition that emphasizes bias correction has argued against use of the z-transformation (Hunter & Schmidt, 2004). Of course, in psychometric meta-analysis, bias corrections arise from many other study features besides dichotomization or use of extreme groups. Further investigation is warranted regarding use of the Taylor-series approximate z-transformation after these other forms of bias correction, since the method may have wider application than that examined here. Another alternative to z-transformation would be to develop a generalized linear model for r-type effect sizes and estimate it via quasi-likelihood (see, e.g., McCullagh & Nelder, 1989).

As noted above, conversions among different types of effect sizes are needed for a variety of situations. This paper has addressed only a few study designs where results are typically reported using d-type effect sizes that one may want to convert to rs or zs. Readers and meta-analysts will no doubt encounter other study designs not addressed here, such as those using hierarchical linear models, a combination of dichotomization and extreme groups, extreme groups with multi-dimensional cutoffs (e.g., Cross et al., 2002), or dichotomization of latent traits (e.g., Séguin, Nagin, Assaad, & Tremblay, 2004). More broadly, further work is needed to examine and clarify conversion formulas for other types of effect sizes, such as from odds ratios to standardized mean differences and vice versa.

The present paper demonstrates how such future work might proceed. In studying conversions from d to r to z, I have relied on explicit and specific distributional assumptions to derive appropriate variance estimation procedures for the converted effect size estimates. In so doing, I have identified discrepancies between the actual large-sample variances and certain naive approximations. These discrepancies highlight the need to consider the sampling design and data-generating model of any particular study, rather than only the algebraic form of the effect size statistic, in determining appropriate variance estimation methods. Methodological texts could improve the quality of guidance regarding effect size conversions by emphasizing more directly these connections between design, model, and estimation method.

References

Alf, E., & Abrahams, N. (1975). The use of extreme groups in assessing relationships. *Psychometrika*, *40*(4), 563–572.

Arnold, B., Beaver, R., Groeneveld, R., & Meeker, W. (1993). The nontruncated marginal of a truncated bivariate normal distribution. *Psychometrika*, *58*(3), 471–488.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, *12*(2), 171–178.

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology*, *100*, 603–617.

Bernichon, T., Cook, K. E., & Brown, J. D. (2003). Seeking Self-Evaluative Feedback: The Interactive Role of Global Self-Esteem and Specific Self-Views. *Journal of Personality and Social Psychology*, *84*(1), 194–204.

Borenstein, M. (2009). Effect sizes for continuous data. In H. M. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 221–236). New York, NY: Russell Sage Foundation.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2009). *Introduction to meta-analysis*. West Sussex, United Kingdom: John Wiley & Sons.

Casella, G., & Berger, R. L. (2002). *Statistical Inference* (2nd ed.). Pacific Grove, CA: Duxbury.

Chatzisarantis, N. L. D., Hagger, M. S., & Smith, B. (2007). Influences of perceived autonomy support on physical activity within the theory of planned behavior. *European Journal of Social Psychology*, *37*, 934–954.

Cross, S. E., Morris, M. L., & Gore, J. S. (2002). Thinking about oneself and others: The relational-interdependent self-construal and social cognition. *Journal of Personality and Social Psychology*, *82*(3), 399–418.

Decoster, J., Iselin, A. R., & Gallucci, M. (2009). A Conceptual and Empirical Examination of Justifications for Dichotomization. *Psychological Methods*, *14*(4),

349–366.

Deffenbacher, J. L., Huff, M. E., Lynch, R. S., Oetting, E. R., & Salvatore, N. F. (2000). Characteristics and Treatment of High-Anger Drivers. *Journal of Counseling Psychology*, *47*(1), 5–17.

Edmunds, J. (2008). Testing a self-determination theory-based teaching style intervention in the exercise domain. *European Journal of Social Psychology*, *38*, 375–388.

Eys, M. A., Hardy, J., Carron, A. V., & Beauchamp, M. R. (2003). The relationship between task cohesion and competitive state anxiety. *Journal of Sport & Exercise Psychology*, *25*, 66–76.

Feldt, L. (1961). The use of extreme groups to test for the presence of a relationship. *Psychometrika*, *26*(3), 307–316.

Field, A. (2005). Is the meta-analysis of correlation coefficients accurate when population correlations vary? *Psychological Methods*, *10*(4), 444–467.

Glass, G., McGaw, B., & Smith, M. (1981). *Meta-analysis in social research* (Vol. 56). Beverly Hills, CA: Sage Publications.

Goldinger, S., Kleider, H., Azuma, T., & Beike, D. (2003). "Blaming the victim" under memory load. *Psychological Science*, *14*(1), 81.

Hafdahl, A. (2009). Improved Fisher z estimators for univariate random-effects meta-analysis of correlations. *British Journal of Mathematical and Statistical Psychology*, *62*(2), 233–261.

Hafdahl, A., & Williams, M. (2009). Meta-analysis of correlations revisited: Attempted replication and extension of Field's (2001) simulation studies. *Psychological methods*, *14*(1), 24–42.

Hagger, M. S., & Chatzisarantis, N. L. D. (2009). Integrating the theory of planned behavior and self-determination theory in health behaviour: A meta-analysis. *British Journal of Health Psychology*, *14*, 275–303.

Hedges, L. V. (2007, December). Effect sizes in cluster-randomized designs. *Journal of*

Educational and Behavioral Statistics, *32*(4), 341–370.

Hedges, L. V. (2011, June). Effect Sizes in Three-Level Cluster-Randomized Experiments. *Journal of Educational and Behavioral Statistics*, *36*(3), 346–380.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

Hunter, J., & Schmidt, F. (1990). Dichotomization of continuous variables: The implications for meta-analysis. *Journal of Applied Psychology*, *75*(3), 334–349.

Hunter, J., & Schmidt, F. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Thousand Oaks, CA: Sage Publications, Inc.

Lipsey, M., & Wilson, D. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage Publications, Inc.

Macbeth, A., & Gumley, A. (2012). Exploring compassion: A meta-analysis of the association between self-compassion and psychopathology. *Clinical psychology review*, *32*(6), 545–552.

MacCallum, R., Zhang, S., Preacher, K., & Rucker, D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*(1), 19–40.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London, UK: Chapman & Hall.

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, *7*(1), 105–125.

Mussweiler, T., Gabriel, S., & Bodenhausen, G. V. (2000). Shifting social identities as a strategy for deflecting threatening social comparisons. *Journal of personality and social psychology*, *79*(3), 398–409.

Pontari, B. A., & Schlenker, B. R. (2000). The influence of cognitive load on self-presentation: Can cognitive busyness help as well as harm social performance? *Journal of Personality and Social Psychology*, *78*(6), 1092–1108.

Preacher, K., Rucker, D., MacCallum, R., & Nicewander, W. (2005). Use of the extreme groups approach: a critical reexamination and new recommendations. *Psychological Methods*, *10*(2), 178–192.

Schulze, R. (2004). *Meta-analysis: A comparison of approaches.* Hogrefe & Huber.

Séguin, J. R., Nagin, D., Assaad, J.-M., & Tremblay, R. E. (2004). Cognitive-neuropsychological function in chronic physical aggression and hyperactivity. *Journal of abnormal psychology*, *113*(4), 603–13.

Skinner, N. F., & Drake, J. M. (2003). Behavioral implications of adaption-innovation: III. Adaption-innovation, achievement motivation, and academic performance. *Social Behavior and Personality*, *31*(1), 101–106.

Stafford, B. T. F., & Gonier, D. (2004). What Americans like about being online. *Communications of the ACM*, *47*(11), 107–112.

Verplanken, B., & Holland, R. W. (2002). Motivated decision making: Effects of activation and self-centrality of values on choices and behavior. *Journal of Personality and Social Psychology*, *82*(3), 434–447.

Table 1

Equation numbers for conversion and variance formulas

Design	d-to-r conversion	Variance of r	r-to-z conversion	Variance of z
Bivariate sampling	–	(12)	(9)	(13)
Controlled experiment	(4)	(15)	(9)	(16)
Dichotomization	(8) + (6)	(18) + (19)	(11)	(18) + (20)
Extreme groups	(5) + (6)	(17) + (19)	(11)	(17) + (20)

Table 2

Derivatives of $z(\cdot)$

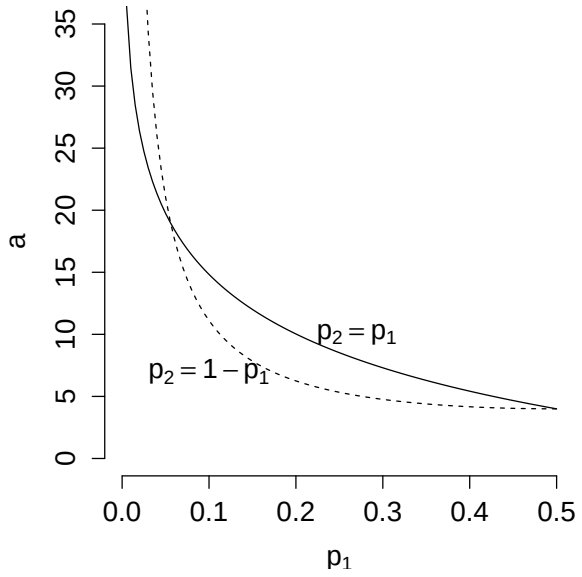
Derivative	Function
1^{st}	$z^{(1)}(r) = (1 - r^2)^{-1}$
2^{nd}	$z^{(2)}(r) = 2r(1 - r^2)^{-2}$
3^{rd}	$z^{(3)}(r) = (2 + 6r^2)(1 - r^2)^{-3}$
4^{th}	$z^{(4)}(r) = (24r + 24r^3)(1 - r^2)^{-4}$
5^{th}	$z^{(5)}(r) = (24 + 240r^2 + 120r^4)(1 - r^2)^{-5}$

Table 3

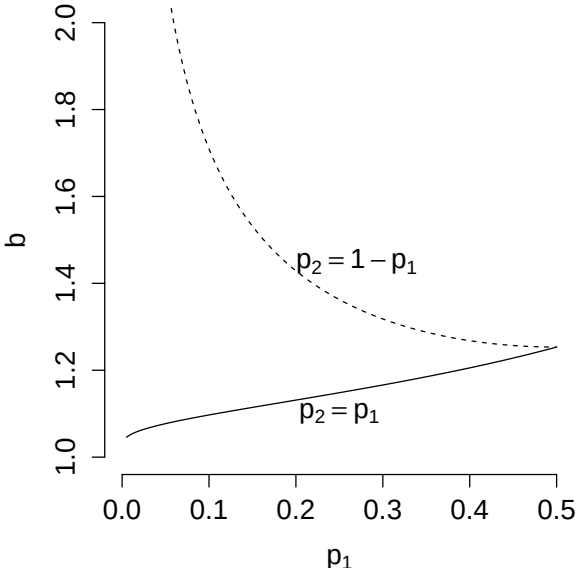
Simulation parameters

Design*	Parameter	Levels
All	Correlation ρ	0.00, 0.01,...,0.95
All	Total sample size n	20, 40, 80, 160
CE	T-C Differential w	0.5, 1.0, 2.0
DI, EG	Cutoff percentile p_1	$\frac{1}{2}, \frac{1}{3}, \frac{1}{5}, \frac{1}{8}$
DI, EG	Cutoff basis	population, sample

*CE controlled experimental design; DI dichotomization design; EG extreme groups design.

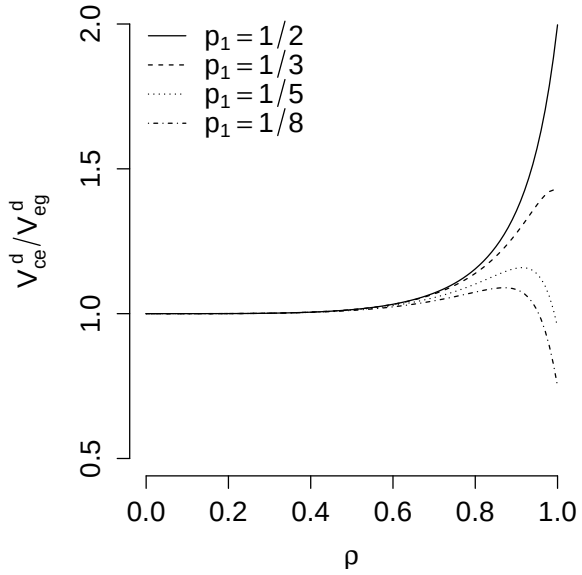


(a) Auxiliary constant a as a function of p_1

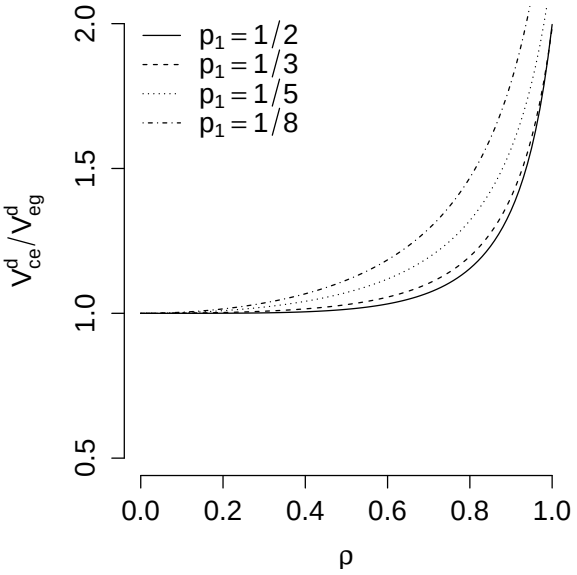


(b) Auxiliary constant b as a function of p_1

Figure 1. Auxiliary constants a and b. The solid line corresponds to a balanced extreme groups design with $p_2 = p_1$. The dashed line corresponds to a dichotomization design with $p_2 = 1 - p_1$.

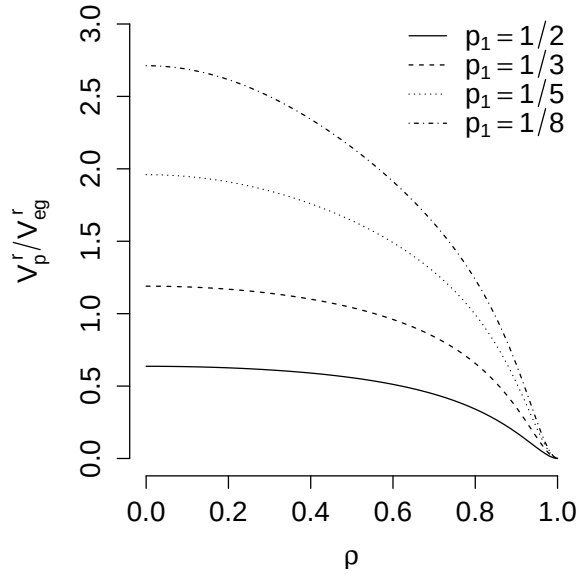


(a) Balanced extreme groups design, $p_2 = p_1$.

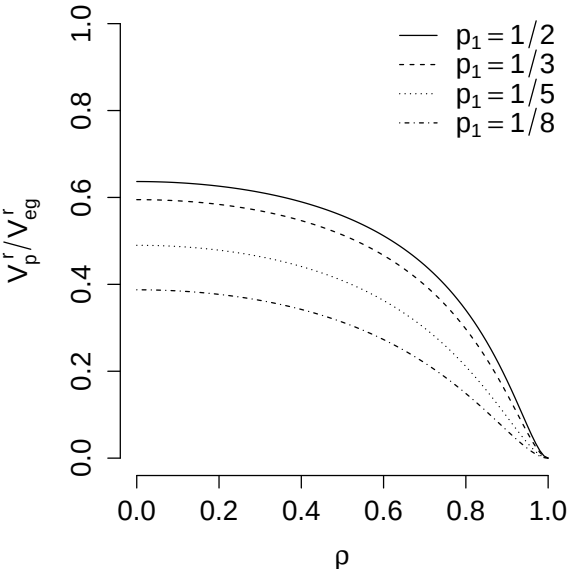


(b) Dichotomization design, $p_2 = 1 - p_1$.

Figure 2. Ratio of V_{ce}^d to V_{eg}^d versus ρ, for various values of p_1.

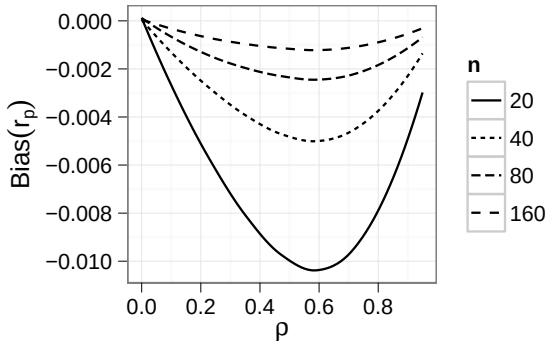


(a) Balanced extreme groups design, $p_2 = p_1$.

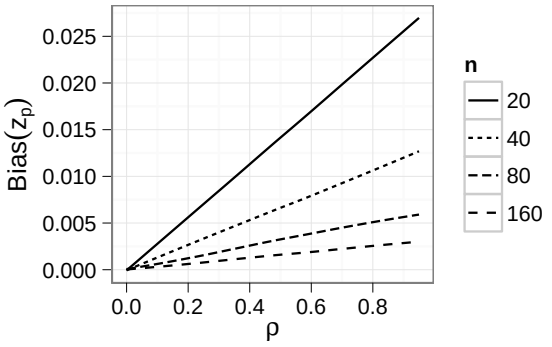


(b) Dichotomization design, $p_2 = 1 - p_1$.

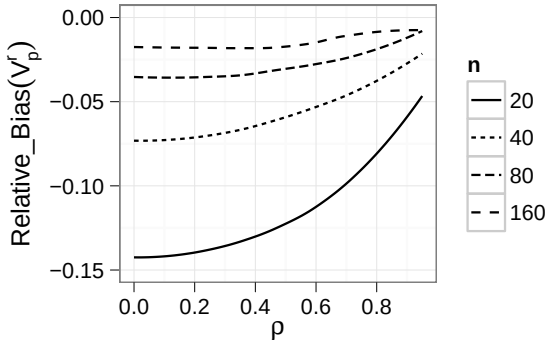
Figure 3. Ratio of V_p^r to V_{eg}^r versus ρ, for various values of p_1.



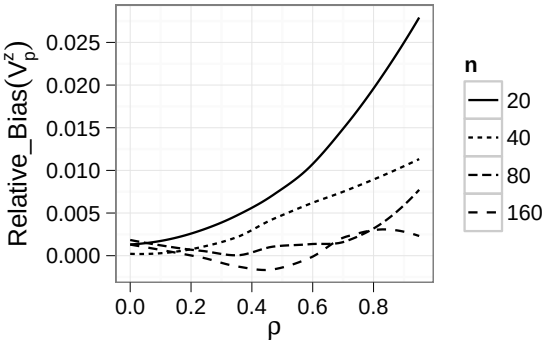
(a) Bias of r_p



(b) Bias of z_p

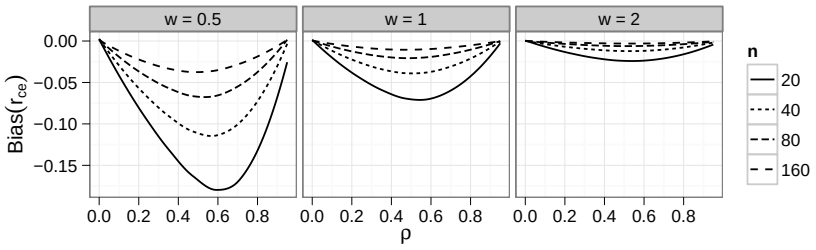


(c) Relative bias of V_p^r

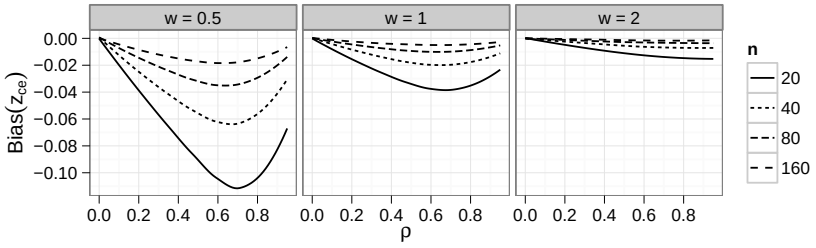


(d) Relative bias of V_p^z

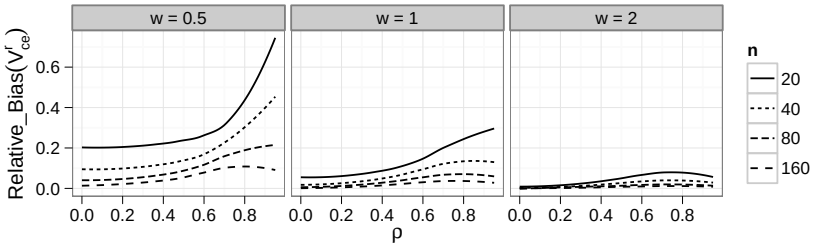
Figure 4. Simulation results for the bivariate sampling design



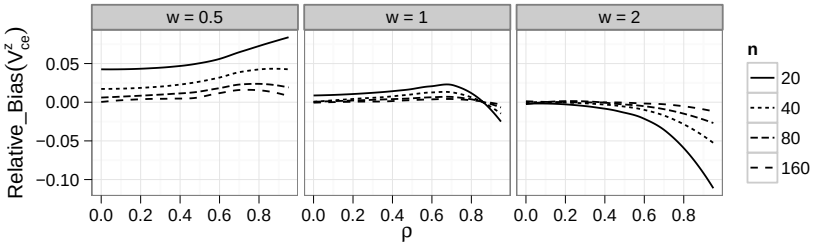
(a) Bias of r_{ce}



(b) Bias of z_{ce}

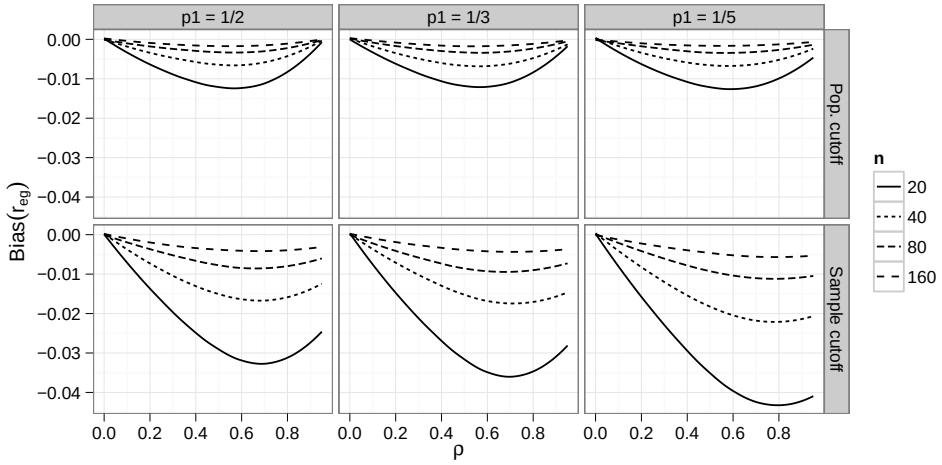


(c) Relative bias of V^r_{ce}

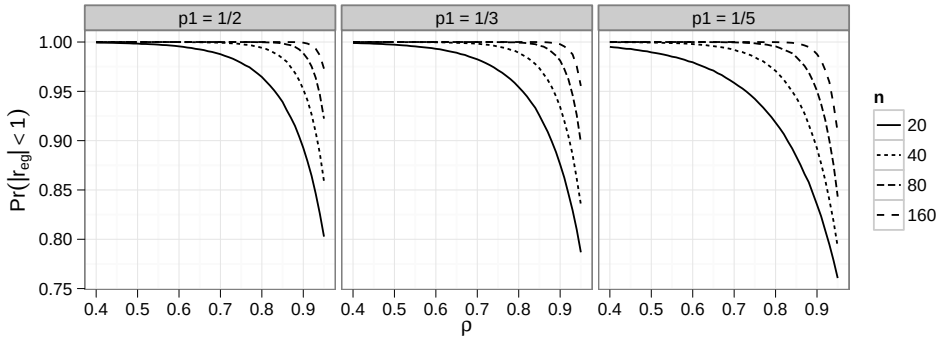


(d) Relative bias of V^z_{ce}

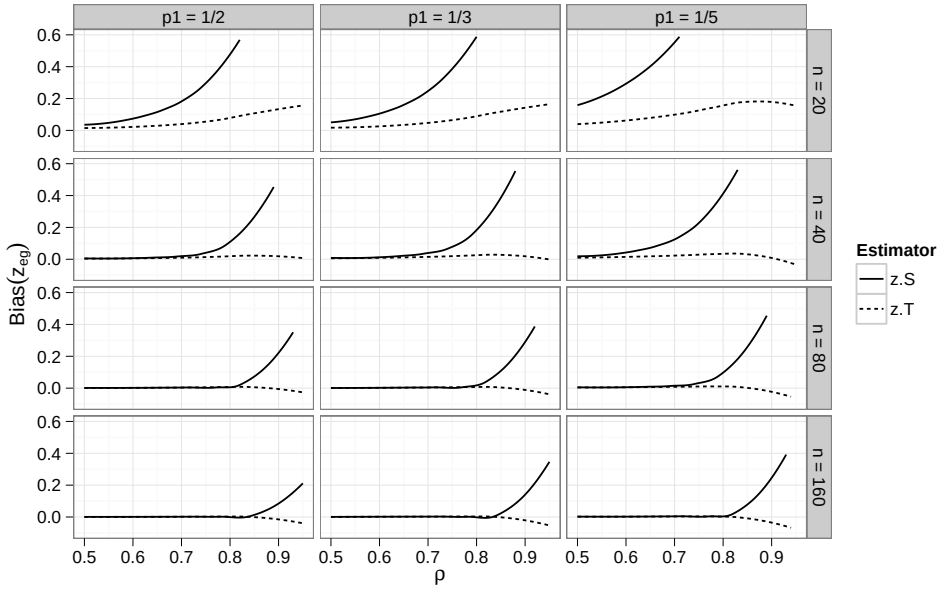
Figure 5. Simulation results for the controlled experimental design



(a) Bias of r_{eg}

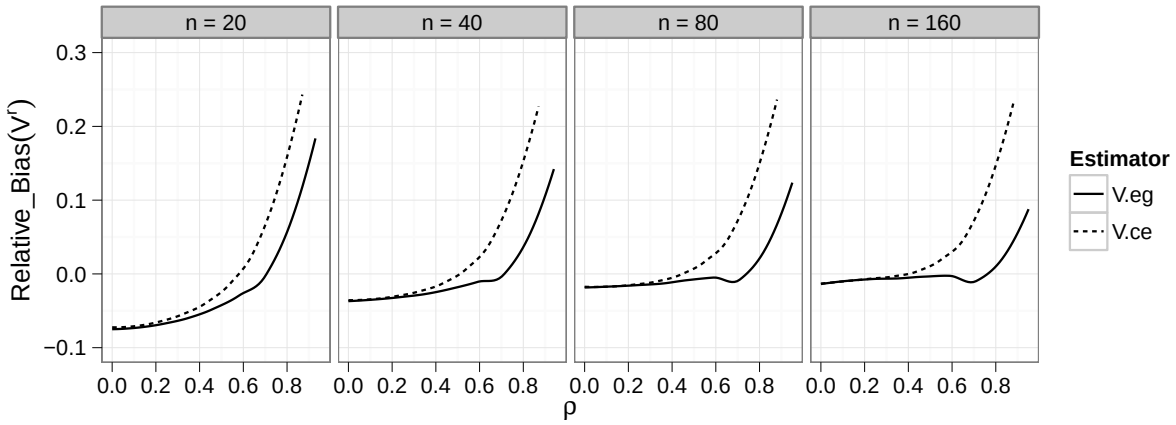


(b) $Pr(|r_{eg}| < 1)$ using sample cutoffs

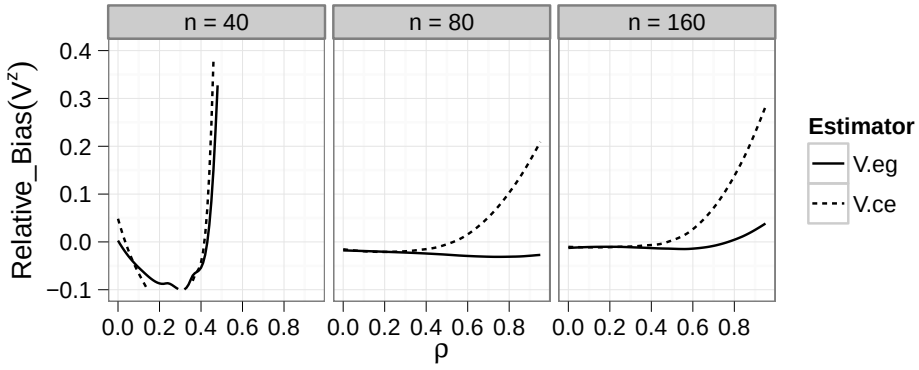


(c) Bias of z_{12}^S versus z_5^T using sample cutoffs

Figure 6. Simulation results for effect sizes from the dichotomization design

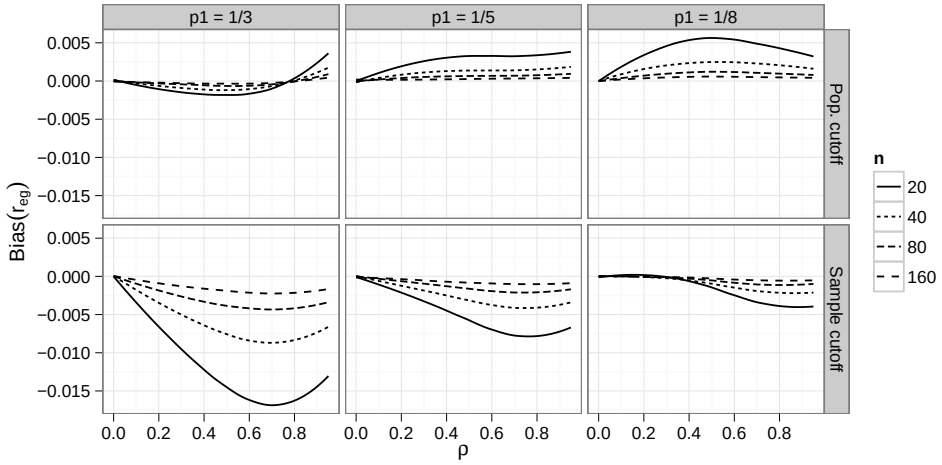


(a) Relative bias of V_{eg}^r versus V_{ce}^r using sample cutoffs and $p_1 = \frac{1}{2}$

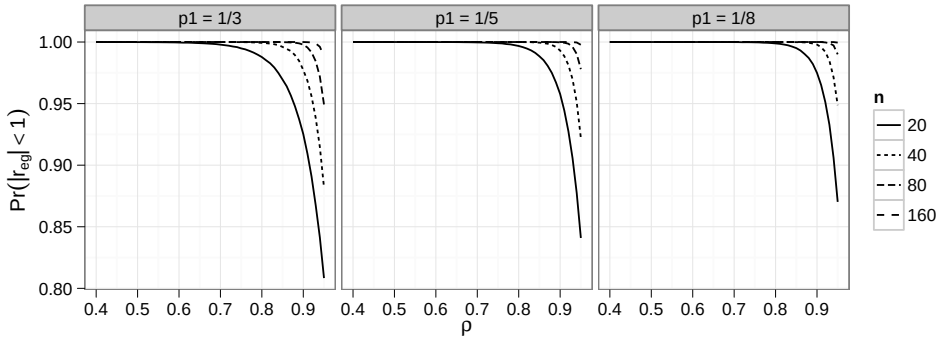


(b) Relative bias of V_{eg}^z versus V_{ce}^z using sample cutoffs and $p_1 = \frac{1}{2}$

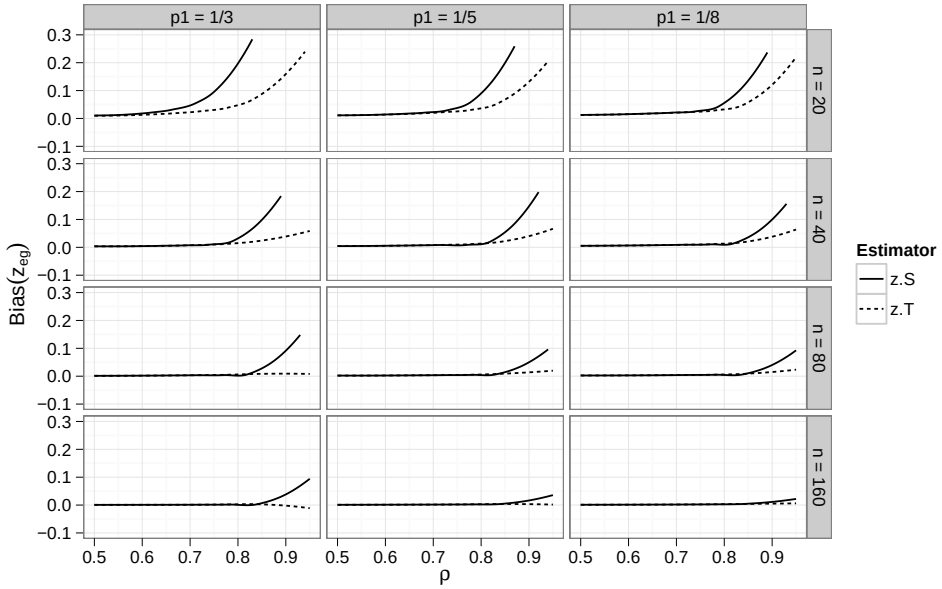
Figure 7. Simulation results for variance estimators from the dichotomization design



(a) Bias of r_{eg}

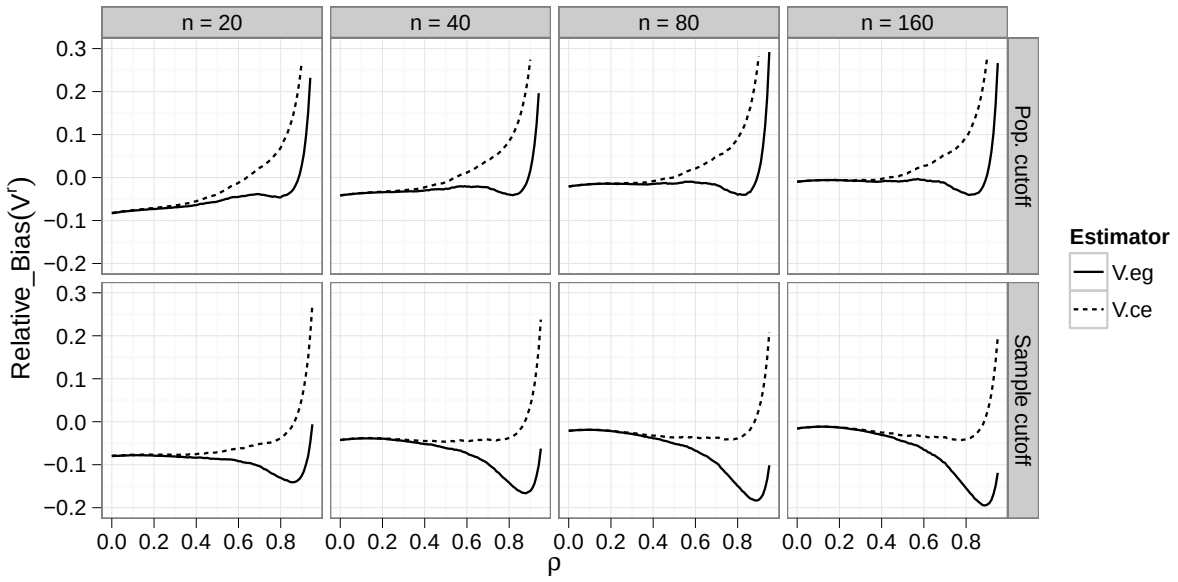


(b) $Pr(|r_{eg}| < 1)$ using sample cutoffs

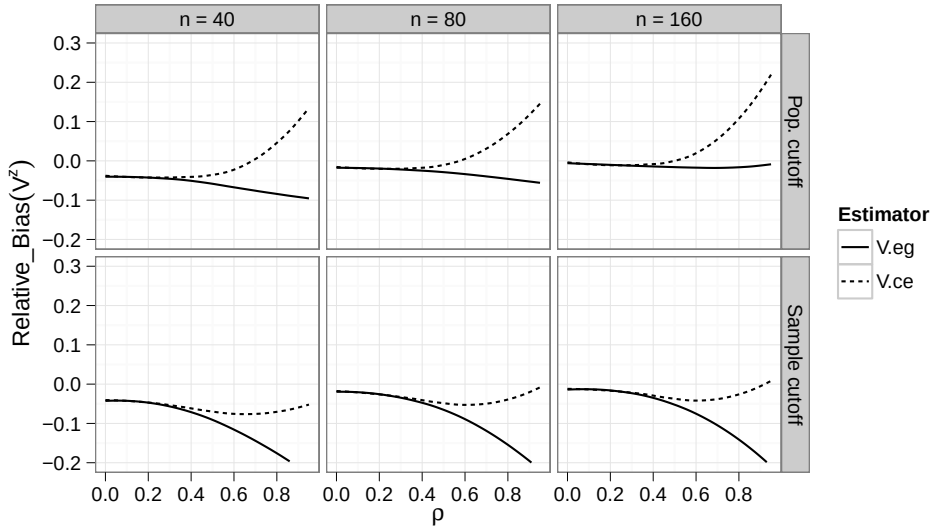


(c) Bias of z_{12}^S versus z_5^T using sample cutoffs

Figure 8. Simulation results for effect sizes from the extreme groups design



(a) Relative bias of V_{eg}^r versus V_{ce}^r for $p_1 = \frac{1}{3}$



(b) Relative bias of V_{eg}^z versus V_{ce}^z for $p_1 = \frac{1}{3}$

Figure 9. Simulation results for variance estimators from the extreme groups design

Appendix A

Controlled experimental design

Here I derive the conversion formula for the converted correlation coefficient r_{ce} as given in Equation (4). By the basic properties of the normal distribution, it follows that the expected value of $\bar{y}_2 - \bar{y}_1$ is $\rho\sigma_y w$ and the expected value of s_p^2 is $\sigma_y^2(1-\rho^2)$. From this, it is clear that d converges to the value $\delta_{ce} = w\rho/\sqrt{1-\rho^2}$. Solving this expression for ρ yields

$$\rho = \frac{\delta_{ce}}{\sqrt{\delta_{ce}^2 + w^2}}.$$

Substituting the sample estimate d in place of δ_{ce} produces r_{ce}, an estimator of ρ.

Appendix B

Extreme groups design

Here I derive the conversion formula for the converted correlation coefficient r_{eg} and the large-sample variance of the d statistic from an extreme groups design. The conversion formula is based on the limiting distribution of d, which can be derived from the first four central moments of $(Y|X < \Phi^{-1}(p_1))$ and $(Y|X \geq \Phi^{-1}(1-p_2))$. Arnold et al. (1993) give the moment-generating function for the variable $(U|\alpha < X < \beta)$, where (U, X) are bivariate normal with mean zero, unit variance, and correlation ρ. Setting $\beta = \Phi^{-1}(p_1), \alpha = -\infty$, this moment-generating function is

$$M_U(t) = \frac{1}{p_1}\Phi(\Phi^{-1}(p_1) - \rho t)\exp\left(\frac{t^2}{2}\right),$$

with corresponding cumulant-generating function

$$K_U(t) = \log\Phi(\Phi^{-1}(p_1) - \rho t) + \frac{t^2}{2} - \log p_1.$$

The cumulants of U can be found from the derivatives of $K_U(t)$ evaluated at $t = 0$. Let $c = c(p) = \Phi^{-1}(p)$ and $v = v(p) = \phi(c(p))/p$. The first four cumulants are then as follows:

$$\begin{aligned}
\kappa_1(p) &= -\rho v \\
\kappa_2(p) &= 1 - \rho^2 v[c + v] \\
\kappa_3(p) &= -\rho^3 v[(2v + c)(v + c) - 1] \\
\kappa_4(p) &= -\rho^4 v\left[(5v + c)(v + c)^2 + (v^2 - 3)(v + c) - v\right].
\end{aligned} \tag{21}$$

Since $Y = \mu_y + \sigma_y U$, it follows that $E[Y|X < \Phi^{-1}(p_1)] = m_{11} = \mu_y + \sigma_y E(U)$ and that $E[(Y - \mu_y)^k|X < \Phi^{-1}(p_1)] = \sigma_y^k E[(U - E(U))^k]$ for $k \geq 2$. The first four central moments of $(Y|X < \Phi^{-1}(p_1))$ are therefore

$$\begin{aligned}
m_{11} &= \mu_y + \sigma_y\kappa_1(p_1) = \mu_y - \sigma_y\rho v(p_1), \\
m_{12} &= \sigma_y^2\kappa_2(p_1), \\
m_{13} &= \sigma_y^3\kappa_3(p_1), \\
m_{14} &= \sigma_y^4[3\kappa_2^2(p_1) + \kappa_4(p_1)].
\end{aligned} \tag{22}$$

The central moments of $(Y|X \geq \Phi^{-1}(1-p_2))$ can be found through a property of skew-normal distributions: if $U \sim SN(\rho, \alpha, \beta)$ and $V \sim SN(-\rho, -\beta, -\alpha)$, then U and V follow the same distribution. Take $V \sim SN(-\rho, \Phi^{-1}(p_2), \infty)$, so that $(Y|X > \Phi^{-1}(1-p_2)) = \mu_y - \sigma_y V$. Then

$$\begin{aligned} m_{21} &= \mu_y - \sigma_y \kappa_1(p_2) = \mu_y + \sigma_y \rho \upsilon(p_2), \\ m_{22} &= \sigma_y^2 \kappa_2(p_2), \\ m_{23} &= -\sigma_y^3 \kappa_3(p_2), \\ m_{24} &= \sigma_y^4 [3\kappa_2^2(p_2) + \kappa_4(p_2)]. \end{aligned} \tag{23}$$

To derive the limiting distribution of d, note that $\mathrm{Var}(\bar{y}_i) = m_{i2}/n_i$, $\mathrm{Cov}(\bar{y}_i, s_i^2) = m_{i3}/n_i$, and

$$\mathrm{Var}(s_i^2) = \frac{(n_i-1)^2}{n_i^3} m_{i4} - \frac{(n_i-1)(n_i-3)}{n_i^3} m_{i2}^2.$$

Therefore,

$$\sqrt{n_i} \left(\begin{pmatrix} \bar{y}_i \\ s_i^2 \end{pmatrix} - \begin{pmatrix} m_{i1} \\ m_{i2} \end{pmatrix} \right) \xrightarrow{\mathcal{D}} N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} m_{i2} & m_{i3} \\ m_{i3} & m_{i4} - m_{i2}^2 \end{bmatrix} \right), \tag{24}$$

where $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution. Assume that $n_i = q_i n$, so that $f = q_1/(q_1+q_2)$ remains fixed as $n \to \infty$. Let

$$d' = \frac{\bar{y}_2 - \bar{y}_1}{\sqrt{f s_1^2 + (1-f) s_2^2}},$$

so that d' has the same limiting distribution as d. It then follows from (24) that d converges to the value

$$\begin{aligned} \delta_{eg} &= \frac{m_{21} - m_{11}}{\sqrt{f m_{12} + (1-f) m_{22}}} \\ &= \frac{k_1(p_2) + k_1(p_1)}{\sqrt{f k_2(p_1) + (1-f) k_2(p_2)}} \\ &= \frac{\rho(v_1 + v_2)}{\sqrt{1 - \rho^2 [f v_1 (v_1 + c_1) + (1-f) v_2 (v_2 - c_2)]}}, \end{aligned} \tag{25}$$

where $c_1 = \Phi^{-1}(p_1)$, $c_2 = \Phi^{-1}(1-p_2)$, $v_1 = \phi(c_1)/p_1$, and $v_2 = \phi(c_2)/p_2$. Working in the case that $p_1 = p_2$, Feldt (1961) derived an estimator for ρ by solving an expression similar to (25) for ρ and then substituting the sample quantity d for the parameter δ_{eg}. Following this approach in the more general case leads to the formula for r_{eg} given in (5) and (6).

Next, by applying the delta method to d', it follows from (24) that $(d - \delta_{eg})/\sqrt{V_d}$ converges in distribution to the standard normal, where $n = n_1 + n_2$ and

$$\begin{aligned} nV_d = \frac{m_{12}/f + m_{22}/(1-f)}{fm_{12} + (1-f)m_{22}} \\ - \frac{(m_{21} - m_{11})(m_{23} - m_{13})}{[fm_{12} + (1-f)m_{22}]^2} \\ + \frac{(m_{21} - m_{11})^2[fm_{14} + (1-f)m_{24}]}{4[fm_{12} + (1-f)m_{22}]^3} \\ - \frac{(m_{21} - m_{11})^2[fm_{12}^2 + (1-f)m_{22}^2]}{4[fm_{12} + (1-f)m_{22}]^3}. \end{aligned} \tag{26}$$

Substituting (22), (23), and (25) into (26) produces

$$\begin{aligned} nV_d = \frac{\kappa_2(p_1)/f + \kappa_2(p_2)/(1-f)}{f\kappa_2(p_1) + (1-f)\kappa_2(p_2)} \\ - \frac{[\kappa_1(p_1) + \kappa_1(p_2)][\kappa_3(p_1) + \kappa_3(p_2)]}{[f\kappa_2(p_1) + (1-f)\kappa_2(p_2)]^2} \\ + \frac{[\kappa_1(p_1) + \kappa_1(p_2)]^2[f\kappa_2^2(p_1) + (1-f)\kappa_2^2(p_2)]}{2[f\kappa_2(p_1) + (1-f)\kappa_2(p_2)]^3} \\ + \frac{[\kappa_1(p_1) + \kappa_1(p_2)]^2[f\kappa_4(p_1) + (1-f)\kappa_4(p_2)]}{4[f\kappa_2(p_1) + (1-f)\kappa_2(p_2)]^3}, \end{aligned} \tag{27}$$

in which the only unknown parameter is ρ. Thus, V_d may be estimated by evaluating the cumulants in (21) using $\rho = r_{eg}$, then calculating V_d using (27).