British Journal of Mathematical and Statistical Psychology (2012), 65, 350–370
© 2011 The British Psychological Society

The British Psychological Society

www.wileyonlinelibrary.com

# Accuracy in parameter estimation for ANCOVA and ANOVA contrasts: Sample size planning via narrow confidence intervals

Keke Lai and Ken Kelley*

University of Notre Dame, Indiana, USA

Contrasts of means are often of interest because they describe the effect size among multiple treatments. High-quality inference of population effect sizes can be achieved through narrow confidence intervals (CIs). Given the close relation between CI width and sample size, we propose two methods to plan the sample size for an ANCOVA or ANOVA study, so that a sufficiently narrow CI for the population (standardized or unstandardized) contrast of interest will be obtained. The standard method plans the sample size so that the expected CI width is sufficiently small. Since CI width is a random variable, the expected width being sufficiently small does not guarantee that the width obtained in a particular study will be sufficiently small. An extended procedure ensures with some specified, high degree of assurance (e.g., 90% of the time) that the CI observed in a particular study will be sufficiently narrow. We also discuss the rationale and usefulness of two different ways to standardize an ANCOVA contrast, and compare three types of standardized contrast in the ANCOVA/ANOVA context. All of the methods we propose have been implemented in the freely available MBESS package in R so that they can be easily applied by researchers.

## 1. Introduction

The analysis of variance (ANOVA) and analysis of covariance (ANCOVA) are among the most popular statistical methods in psychology and related sciences. Most ANOVA or ANCOVA studies test the null hypothesis that the population group means are all equal. If the test result is statistically significant, the study reports that there is some difference among the groups. However, as has been echoed numerous times in the literature (e.g., Cohen, 1994; Meehl, 1997; Nickerson, 2000; Schmidt, 1996), there are serious limitations to null hypothesis significance testing (NHST). In the context of comparing group means, often before conducting an NHST, substantive theories already suggest that

This article has an online supplement that is accessible from https://repository.library.nd.edu/view/28/AIPE_for_Contrasts.pdf

*Correspondence should be addressed to Ken Kelley, Department of Management, Mendoza College of Business, University of Notre Dame, Notre Dame, IN 46556, USA (e-mail: kkelley@nd.edu).

DOI:10.1111/j.2044-8317.2011.02029.x

the null hypothesis of equal population means is almost certainly false. Correspondingly, it has been suggested (e.g., Harlow, Mulaik, & Steiger, 1997) that NHST does not facilitate the accumulation of knowledge because we are setting up a question that we essentially already know the answer to (e.g., there is some difference among the population group means).

Although NHST helps to understand whether or not an effect exists in the population, the result of an NHST should not be the endpoint of scientific inquiry. We believe that the method based on effect sizes and their corresponding confidence intervals (CIs) offers more and better information about the parameter(s) of interest than does NHST. In particular, we learn probabilistically not only what the parameter is not (i.e., those values excluded by the CI) but also a range of plausible values (i.e., those values contained within the CI), as opposed to NHST, which in and of itself does not provide information about the effect's magnitude. Due to the limitations of NHST, the emphasis in applied research has been moving from the dichotomous, reject or fail-to-reject NHST to the more informative approach based on CIs and effect sizes, which may well be the future of quantitative research (Thompson, 2002). For example, the American Psychological Association (2009, p. 33) states that NHST is 'but a starting point' and additional reporting elements such as effect sizes and CIs are necessary to 'convey the most complete meaning of the results' and are 'minimum expectations for all APA journals'.

An effect size can be *unstandardized*, if it is wedded to the particular scale of a measurement instrument, or *standardized*, if it is scaled in terms of the variability of the population from which the measure was taken. Both types are important, but one can be more useful than the other in certain circumstances, and the present paper focuses on standardized effects. Standardized effect sizes can be helpful when raw effects are difficult to interpret or raw effects from different studies cannot be directly compared, which is often the case in psychology and related sciences. Constructs in psychology and related disciplines are usually not observable and instruments used to measure those constructs do not generally have a natural scaling metric. Therefore, researchers often tend to adopt a metric that is believed to be 'reasonable' when measuring the latent construct, but such a metric is in fact simply chosen by the instrument developer and thus involves some degree of arbitrariness (Blanton & Jaccard, 2006; Embretson, 2006). As a result, effect measures based on raw scores are necessarily associated with the corresponding measurement unit, and multiple scaling schemes for the same phenomenon cannot be readily compared because there lacks a natural scale for the latent construct. To better interpret results in a study and synthesize results from different studies, a solution is to report standardized effect sizes (e.g., Glass, McGaw, & Smith, 1981; Hedges & Olkin, 1985; Hunter & Hamilton, 2002; Hunter & Schmidt, 2004). When the metric is well understood, however, raw effect sizes are much easier to interpret and more robust to factors such as instrument reliability and sample range, as compared to standardized effects (see Baguley, 2009, for a comprehensive review on the merits of raw effects and a detailed comparison between unstandardized and standardized effects).

To estimate a population effect of interest, one can construct a CI. All other things being equal, it is desirable to observe a narrow CI as compared to a wider one, because a narrow CI includes a narrower range of plausible parameter values and thus contains less uncertainty in the estimation. Since the width of a CI is closely related to the study's sample size, the goal to obtain a sufficiently narrow CI can be achieved by carefully planning the sample size for the study. Designing research with the goal to obtain a narrow CI dates back to at least Guenther (1965) and Mace (1964), and has been becoming popular recently due to the increasing emphasis on reporting effect sizes and

CIs. This approach to sample size planning, alternative to the power-analytic one, is termed *accuracy in parameter estimation* (AIPE; e.g., Kelley, 2007b, 2008; Kelley & Lai, 2011b; Kelley & Maxwell, 2003; Kelley & Rausch, 2006; Lai & Kelley, 2011), where the goal is to achieve a sufficiently narrow CI so that the parameter estimate will have a high degree of expected accuracy (see Maxwell, Kelley, & Rausch, 2008, for a review of the AIPE and power-analytic approaches to sample size planning). The goal to achieve a sufficiently narrow CI is operationally defined as obtaining a CI that is no wider than some desired value, denoted as $\omega$. The value of this desired width needs to be specified *a priori* by the researcher on a case-by-case basis according to the goals of the study.

In this paper we propose two procedures to plan sample size for ANCOVA and ANOVA, with the goal to obtain a sufficiently narrow CI for the population contrast of interest. We begin by briefly reviewing sample size planning for unstandardized contrasts, so as to familiarize readers with the general philosophy of AIPE; then we move on to standardized contrasts. An ANCOVA contrast can be standardized in at least two ways: (a) divided by ANCOVA's root mean square error; or (b) divided by the root mean square error of the ANOVA model excluding the covariate. We first develop CI formation and sample size planning for these two standardized effects, and then discuss their rationale and usefulness. The standard procedure calculates the necessary sample size so that the *expected* CI width is sufficiently small. However, since the CI width is a random variable, the expectation of the random width being sufficiently small does not guarantee that the CI width obtained in a particular study will be sufficiently small. An extended procedure is developed so that there is some specified, high degree of assurance (e.g., 90% of the time, 99% of the time, etc.) that the CI *observed* in a study will be sufficiently narrow. All of the CI formation and sample size planning methods we discuss in the present paper have been implemented in the MBESS package (Kelley, 2007a, 2007c; Kelley & Lai, 2011a) in R (R Development Core Team, 2011), so that they can be readily applied by researchers.[1]

## 2. Confidence interval and sample size for unstandardized contrasts

### 2.1 Notation and assumptions

Throughout this paper, we frame our discussion in the *randomized design* ANCOVA context, and treat ANOVA as a special case of ANCOVA. We assume that all of the ANCOVA and ANOVA model assumptions are satisfied (i.e., homogeneous variance, normality, independent observations). Let $j$ indicate the group status, $J$ the total number of groups ($j = 1, \ldots, J$), $n_j$ the number of individuals in the $j$th group, and $N$ the total number of individuals ($N = \sum_{j=1}^{J} n_j$). This paper is based on the ANCOVA model given by

$$Y_{ij} = \mu + \tau_j + \beta(X_{ij} - \mu_X) + \varepsilon_{ij}, \tag{1}$$

where $Y_{ij}$ is the response variable of individual $i$ in group $j$, $\mu$ is the population grand mean of the response, $\tau_j$ is the treatment effect of group $j$ (i.e., $\tau_j = \mu_j - \mu$, where $\mu_j$ is the population mean of group $j$), $\beta$ is the population regression coefficient of the

[1] For detailed software documentation and an empirical illustration, refer to this paper's online supplement at https://repository.library.nd.edu/view/28/AIPE_for_Contrasts.pdf

response on the covariate, $X_{ij}$ is the covariate score of individual $i$ in group $j$, $\mu_X$ is the population mean of the covariate, and $\varepsilon_{ij}$ is the error term following $N(0,\ \sigma^2_{\mathrm{ANCOVA}})$. Note that this model assumes there is no interaction between the treatment and the covariate. The ANOVA model that the present paper is based on is given by

$$Y_{ij} = \mu + \tau_j + \varepsilon'_{ij}, \tag{2}$$

where $\varepsilon'_{ij}$ is the error term following $N(0,\ \sigma^2_{\mathrm{ANOVA}})$. The relation between the population error variances for ANCOVA and for ANOVA can be described as (e.g., Maxwell & Delaney, 2004, p. 442; Rencher & Schaalje, 2007, p. 256)

$$\sigma^2_{\mathrm{ANCOVA}} = \sigma^2_{\mathrm{ANOVA}}(1 - \rho^2), \tag{3}$$

where $\rho$ is the population correlation coefficient between the response and the covariate. Equation (3) indicates that the population error variance of an ANCOVA model is always smaller than the population error variance of an ANOVA model, unless the covariate is uncorrelated with the response (in which case the two error variances are equal). Including a covariate in a randomized design can reduce the error variance and increase the precision of estimation.

### 2.2. *Unstandardized ANCOVA contrasts*

A sample contrast in the one-covariate ANCOVA context is defined as

$$\hat{\Psi}' = \sum_{j=1}^{J} c_j \overline{Y}'_j = \sum_{j=1}^{J} c_j \left[\overline{Y}_j - \hat{\beta}(\overline{X}_j - \overline{X}_{..})\right], \tag{4}$$

where $c_j$ is the contrast weight for group $j$ and restricted by $\sum_{j=1}^{J} c_j = 0$ and $\sum_{j=1}^{J} |c_j| = 2$, $\overline{Y}'_j$ is the adjusted mean of group $j$, $\overline{X}_j$ is the covariate mean of group $j$, and $\overline{X}_{..}$ is the covariate grand mean. The $t$-statistic associated with $\hat{\Psi}'$,

$$T = \frac{\hat{\Psi}' - \Psi'}{s_{\hat{\Psi}'}} \sim t_{(N-J-1)}, \tag{5}$$

follows a $t$-distribution with $N - J - 1$ degrees of freedom, where $s_{\hat{\Psi}'}$ is the standard error of $\hat{\Psi}'$. Given the covariate scores on the $N$ individuals, $s_{\hat{\Psi}'}$ equals $s_{\mathrm{ANCOVA}}\sqrt{(\sum_{j=1}^{J} c_j^2/n_j) + D}$, where $s_{\mathrm{ANCOVA}}$ is the root mean square error of the ANCOVA model, and $D = (\sum_{j=1}^{J} c_j \overline{X}_j)^2 / \sum_{j=1}^{J} \sum_{i=1}^{n_j} (X_{ij} - \overline{X}_j)^2$ (Cochran, 1957; Maxwell & Delaney, 2004, pp. 460–464).

Given the distribution of the $t$-statistic associated with $\hat{\Psi}'$, a $(1 - \alpha)100\%$ CI for $\Psi'$ can be formed:

$$\mathrm{CI}_{1-\alpha} = [\hat{\Psi}' - t_{(1-\alpha/2, N-J-1)} s_{\hat{\Psi}'} \leq \Psi' \leq \hat{\Psi}' + t_{(1-\alpha/2, N-J-1)} s_{\hat{\Psi}'}], \tag{6}$$

where $t_{(1-\alpha/2,\, N-J-1)}$ refers to the $(1 - \alpha/2)$th quantile of the $t$-distribution with $N - J - 1$ degrees of freedom. Note that the above CI is two-tailed with equal rejection probability on both tails. For symmetric unimodal distributions, an equal $\alpha$ split ensures the narrowest interval (Casella & Berger, 2002, Section 9.3.1); this paper assumes equal rejection probability on both tails hereafter. Based on equation (6), the full width of a CI for $\Psi'$ is

$$w = 2t_{(1-\alpha/2,N-J-1)}s_{\hat{\Psi}'}. \tag{7}$$

### 2.2.1. Sample size planning for unstandardized ANCOVA contrasts

In this section we briefly review the rationale for sample size planning from the AIPE perspective and explain how this approach applies to unstandardized ANCOVA contrasts. Detailed expositions on different kinds of unstandardized mean differences are available in Jiroutek, Muller, Kupper, & Stewart (2003), Kelley, Maxwell, and Rausch (2003), Kupper and Hafner (1989), and Pan and Kupper (1999). Their methods can be extended to the present unstandardized ANCOVA context.

Let $\omega$ denote the desired CI full width, which is specified by the researcher according to the goals of the study. The CI width obtained in a study is denoted by $w$ and calculated as in equation (7). Note that $\omega$ is a constant and known when planning the sample size, whereas $w$ is a random variable and remains unknown until the researcher constructs a CI based on a specific sample. The goal of AIPE is to find the *smallest* sample size such that $w \leq \omega$, and it can be achieved from two perspectives. The first perspective is to study the expectation of $w$ and find the sample size such that $\mathrm{E}[w] \leq \omega$. The expected CI width can be expressed as

$$\mathrm{E}[w] = \mathrm{E}[2t_{(1-\alpha/2,N-J-1)}s_{\hat{\Psi}'}] = 2t_{(1-\alpha/2,N-J-1)}\mathrm{E}[s_{\hat{\Psi}'}] \approx 2t_{(1-\alpha/2,N-J-1)}\sigma_{\hat{\Psi}'}. \tag{8}$$

The approximation follows because we use the population standard deviation as the expectation of the sample standard deviation.[2] Since ANCOVA and ANOVA designs with equal sample size per group are most robust to violations of the normality and homogeneous variance assumptions (e.g., Kirk, 1995, pp. 99–100), we develop sample size planning procedures by assuming equal sample size per group. In addition, using equal sample size per group reduces the amount of input information required when planning sample size. We assume equal sample size per group throughout the rest of the paper, and use $n$ to denote such group sample size.

Given the covariate scores in a design, $\sigma_{\hat{\Psi}'}$ equals $\sigma_{\mathrm{ANCOVA}}\sqrt{C/n + D}$, where $C = \sum_{j=1}^{J} c_j^2$. Since the covariate is usually considered as a random effect in psychology and related sciences, the interest is in the value of $\sigma_{\hat{\Psi}'}$ across all possible covariate values rather than conditional on certain specific covariate scores. In randomized designs, the covariate scores across different groups will be equal in the long run, and thus the numerator $(\sum_{j=1}^{J} c_j\overline{X}_j)^2$ in $D$ is zero in the population, reducing $\sigma_{\hat{\Psi}'}$ to $\sigma_{\mathrm{ANCOVA}}\sqrt{C/n}$. Moreover, in *randomized* designs the sample value of $D$ is usually close to zero and negligible, because its numerator is much smaller than the denominator. This is the case because (a) often the covariate means are not much different among groups as a result of randomization, making the numerator close to zero; and (b) the denominator

[2] Sample standard deviation is a biased estimator of its population counterpart, but the bias is negligible as long as the sample size is not too small. The expectation of sample standard deviation is $\mathrm{E}[s] = \sigma \cdot \phi(n)$, where $\phi(n) = \sqrt{n-1}\Gamma((n-1)/2)/[\sqrt{2}\Gamma(n/2)]$ and $\Gamma(\cdot)$ is the gamma function (e.g., Casella & Berger, 2002, p. 364). The quantity $\phi(n)$ depends on $n$ only and is a decreasing function of $n$. When $n = 50$, $\phi(n) = 1.0051$; when $n = 100$, $\phi(n) = 1.0025$.

is the error sum of squares of performing ANOVA on the covariate, whose value tends to be much larger compared to the numerator in randomized designs.[3] Therefore, in the planning stage, we use $\sigma_{\text{ANCOVA}}\sqrt{C/n}$ as the proxy population standard deviation of $\hat{\Psi}'$ to calculate the sample size. However, in the data analysis stage, the term $D$ is still involved in calculating $s_{\hat{\Psi}'}$, but its value is usually of little impact on $s_{\hat{\Psi}'}$ in randomized designs.

Based on the assumptions of randomized design and equal sample size per group, the goal $\text{E}[w] \leq \omega$ can be expressed formally as

$$2t_{(1-\alpha/2, nJ-J-1)}\sigma_{\text{ANCOVA}}\sqrt{C/n} \leq \omega, \tag{9}$$

where the left-hand side of the inequality follows from equation (8). Solving (9) for $n$ gives the (minimum) sample size per group that satisfies $\text{E}[w] \leq \omega$. Note that $n$ also plays a role in the $t$-distribution quantile, and thus $n$ needs to be solved for numerically rather than analytically. One way to solve for $n$ is implemented in the MBESS R package.[4]

Although the sample size obtained from the above process satisfies $\text{E}[w] \leq \omega$, it does not guarantee that in a particular study the observed CI width will be smaller than or equal to $\omega$. As is indicated by equations (7) and (8), $w$ is a random variable based on the random variable $s_{\text{ANCOVA}}$, whereas $\omega$ is a constant based on the constant $\sigma_{\text{ANCOVA}}$. When the sample size is not too small, $s_{\text{ANCOVA}}$ is smaller than $\sigma_{\text{ANCOVA}}$ about 50% of the time, making $w$ smaller than $\omega$ about 50% of the time (e.g., Liu, 2009). In order to obtain a sufficiently narrow CI in a study with high degree of assurance (e.g., 90% of the time, 99% of the time, etc.), a larger sample size is necessary. Let $\gamma$ ($.50 < \gamma < 1$) denote the desired degree of assurance, and the task now becomes one where we need to determine the (increased) sample size per group, $n^+$, such that $P(w \leq \omega) = \gamma$. When $s_{\text{ANCOVA}} < \sigma_{\text{ANCOVA}}$, $w$ will tend to be smaller than $\omega$, and when $s_{\text{ANCOVA}} > \sigma_{\text{ANCOVA}}$, $w$ will tend to be larger than $\omega$. Therefore, if a quantity $\sigma^*_{\text{ANCOVA}}$ can be found such that $P(s_{\text{ANCOVA}} \leq \sigma^*_{\text{ANCOVA}}) = \gamma$, substituting $\sigma^*_{\text{ANCOVA}}$ for $\sigma_{\text{ANCOVA}}$ in equation (9) will ensure $P(w \leq \omega) = \gamma$. Given the properties of a sample variance (e.g., Hays, 1994, pp. 355–358), the relation between $s^2_{\text{ANCOVA}}$ and $\sigma^2_{\text{ANCOVA}}$ can be described as

$$(n^+J - J - 1)\frac{s^2_{\text{ANCOVA}}}{\sigma^2_{\text{ANCOVA}}} \sim \chi^2_{(n^+J-J-1)}. \tag{10}$$

Based on this distribution, there is a quantity that $s^2_{\text{ANCOVA}}$ does not exceed $\gamma 100\%$ of the time:

$$P\left[s^2_{\text{ANCOVA}} \leq \sigma^2_{\text{ANCOVA}}\frac{\chi^2_{(\gamma,\, n^+J-J-1)}}{n^+J - J - 1}\right] = \gamma, \tag{11}$$

where $\chi^2_{(\gamma,\, n^+J-J-1)}$ is the $\gamma$th quantile of the chi-square distribution. Since taking the square root of both sides of the inequality in equation (11) does not change the

[3] Even when randomization is carefully implemented, the covariate means can still differ *statistically* significantly among groups about $\alpha 100\%$ of the time, where $\alpha$ is the Type I error rate used in comparing the covariate group means. This phenomenon is termed 'unhappy randomization' (Kenny, 1979, p. 217), and we will study its implications in more detail in Section 5.

[4] The function that fulfils this task is ss.aipe.c.ancova(). Please refer to its help file in the MBESS R package for detailed documentation. See also ci.c.ancova() for CI formation for $\Psi'$.

probability statement, substituting $\sigma_{\text{ANCOVA}}\sqrt{\chi^2_{(\gamma,\, n^+J-J-1)}/(n^+J-J-1)}$ for $\sigma_{\text{ANCOVA}}$ in (9) gives a way to calculate the sample size that ensures $P(w \le \omega) = \gamma$:

$$2t_{(1-\alpha/2,\, n^+J-J-1)}\sigma_{\text{ANCOVA}}\sqrt{\frac{\chi^2_{(\gamma,\, n^+J-J-1)}}{n^+J-J-1}}\sqrt{\frac{C}{n^+}} \le \omega. \tag{12}$$

All components in equation (12) except $n^+$ are either constants or specified *a priori* by the researcher, and thus solving the inequality for $n^+$ will give the necessary sample size per group that ensures $P(w \le \omega) = \gamma$. Since $n^+$ also plays a role in the distribution quantiles, this equation needs to be solved numerically with an iterative process.[5]

### 2.3. Unstandardized ANOVA contrasts

Let $\Psi$ denote the population ANOVA contrast; it can be estimated by $\hat{\Psi} = \sum_{j=1}^{J} c_j \overline{Y}_j$. The *t*-statistic associated with $\hat{\Psi}$, in the context of equal sample size per group, is

$$T = \frac{\hat{\Psi} - \Psi}{s_{\hat{\Psi}}} = \frac{\hat{\Psi} - \Psi}{s_{\text{ANOVA}}\sqrt{C/n}} \sim t_{(nJ-J)}, \tag{13}$$

and it follows a *t*-distribution with $nJ - J$ degrees of freedom. In the same manner as we formed a CI for $\Psi'$ in the ANCOVA context, a $(1-\alpha)100\%$ CI for $\Psi$ is

$$\text{CI}_{1-\alpha} = [\hat{\Psi} - t_{(1-\alpha/2, nJ-J)}s_{\hat{\Psi}} \le \Psi \le \hat{\Psi} + t_{(1-\alpha/2, nJ-J)}s_{\hat{\Psi}}]. \tag{14}$$

Regarding $\sigma_{\hat{\Psi}}$ as the approximate expectation of $s_{\hat{\Psi}}$, the expected width of the CI in equation (14) is approximately $2t_{(1-\alpha/2, nJ-J)}\sigma_{\hat{\Psi}}$, or equivalently

$$\text{E}[w] \approx 2t_{(1-\alpha/2, nJ-J)}\sigma_{\text{ANOVA}}\sqrt{C/n}. \tag{15}$$

Exactly as in the ANCOVA case, we can find the smallest sample size per group that satisfies $\text{E}[w] \le \omega$ with an iterative process. To ensure that the CI obtained in a particular study is no wider than desired with $\gamma100\%$ assurance, we can substitute $\sigma_{\text{ANCOVA}}\sqrt{\chi^2_{(\gamma,\, n^+J-J)}/(n^+J-J)}$ for $\sigma_{\text{ANCOVA}}$ in the standard sample size planning process.[6]

In summary, we have discussed CI formation and sample size planning for unstandardized contrasts. Since the CI width is a random variable, one approach is to focus on the expected width. An extended procedure ensures that the CI observed in a particular study is sufficiently narrow with a specified, high degree of assurance.

## 3. Confidence interval and sample size for standardized contrasts

### 3.1. Standardized ANCOVA contrasts

An ANCOVA contrast can be standardized in at least two ways: (a) divided by the root mean square error of the ANCOVA model, or (b) divided by the root mean square error

[5] The function that performs this task is ss.aipe.c.ancova().

[6] The functions for CI formation and sample size planning for $\Psi$ are ci.c() and ss.aipe.c(), respectively.

of the ANOVA model excluding the covariate. We first develop CI construction and sample size planning methods for these two standardized effect sizes, and then discuss their rationale and usefulness in a later section.

### *3.1.1. Root mean square error of ANCOVA as standardizer*

Let the observed standardized ANCOVA contrast be defined as $\hat{\psi}' = \hat{\Psi}'/s_{\text{ANCOVA}}$. The $t$-statistic associated with $\hat{\Psi}'$,

$$T' = \frac{\hat{\Psi}'}{s_{\text{ANCOVA}}\sqrt{C/n}} = \frac{\hat{\psi}'}{\sqrt{C/n}} \sim t'_{(nJ-J-1,\lambda)}, \tag{16}$$

follows a non-central $t$-distribution with $nJ - J - 1$ degrees of freedom and non-centrality parameter $\lambda$. Note that the difference between equations (16) and (5) is in the numerator: in equation (5) $\Psi'$ is subtracted from $\hat{\Psi}'$ and the resulting statistic follows a central $t$-distribution, whereas the numerator in equation (16) is $\hat{\Psi}'$ alone and the resulting statistic follows a non-central $t$-distribution (see Johnson, Kotz, & Balakrishnan, 1995; Johnson & Welch, 1940, for technical discussions of the non-central $t$-distribution). The population non-centrality parameter is defined as

$$\lambda = \psi'/\sqrt{C/n}, \tag{17}$$

and can be estimated by

$$\hat{\lambda} = \hat{\psi}'/\sqrt{C/n}. \tag{18}$$

Therefore, $\hat{\lambda}$ is equivalent to $T'$ in equation (16).

Although $\hat{\psi}'$ multiplied by some constant follows a non-central $t$-distribution as equation (16) indicates, the CI for $\psi'$ cannot be readily constructed based on the non-central $t$-distribution, because $\lambda$ is unknown. Given the monotonic relation between $\lambda$ and $\psi'$ (see equation (17)) and the relative ease of forming a CI for $\lambda$, the CI formation for $\psi'$ can be accomplished indirectly. First we form a CI for $\lambda$, and then we transform the CI limits of $\lambda$ onto the scale of $\psi'$, so that those limits become the CI limits of $\psi'$. This indirect method shares the same rationale with the CI formation for many other effect sizes in the general linear model context (Steiger & Fouladi, 1997; Steiger, 2004; see also Cumming & Finch, 2001; Fleishman, 1980; Kelley, 2007a; Smithson, 2003), as well as the CI formation for root mean square error of approximation (RMSEA, Browne & Cudeck, 1992; Steiger & Lind, 1980) of structural equation models.

To form a CI for $\lambda$, first let $t'_{(q,\nu,\lambda)}$ denote the $q$th quantile of the non-central $t$-distribution with $\nu$ degrees of freedom ($\nu = nJ - J - 1$ in the present context) and non-centrality parameter $\lambda$, and let $\lambda_U$ and $\lambda_L$ denote the $(1 - \alpha)100\%$ upper and lower confidence limits for $\lambda$, respectively. Then $\lambda_U$ and $\lambda_L$ are values that satisfy

$$t'_{(\alpha/2,\,\nu,\lambda_U)} = T' \text{and } t'_{(1-\alpha/2,\,\nu,\lambda_L)} = T',$$

where $T'$ is the observed non-central $t$-statistic defined in equation (16) (Casella & Berger, 2002, p. 432). The same $T'$ value can equal two different quantiles because those quantiles are from distributions of different non-centrality parameters. Also note that $\lambda_U$ and $\lambda_L$ have different values given different $T'$ values observed, and thus they are random variables. We denote a $(1-\alpha)100\%$ CI for $\lambda$ as $\mathrm{CI}_{1-\alpha} = [\lambda_L \leq \lambda \leq \lambda_U | T']$. Given the monotonically increasing relation between $\lambda$ and $\psi'$ (see equation (17)), the CI for $\psi'$ can be formed based on the CI for $\lambda$:

$$\mathrm{CI}_{1-\alpha} = \left[\lambda_L \sqrt{C/n} \leq \psi' \leq \lambda_U \sqrt{C/n}\ |\hat{\psi}'\right]. \tag{19}$$

The CI width obtained is thus

$$w = (\lambda_U - \lambda_L)\sqrt{C/n}\ |\hat{\psi}'. \tag{20}$$

Note that the observed CI width is a random variable depending only on $\hat{\psi}'$.

Similar to the case of unstandardized ANCOVA contrasts, we first develop methods to plan for sample size so that $\mathrm{E}[w] \leq \omega$. Due to the complexity and non-linear nature of the CI formation for $\psi'$, the exact expected value of $\mathrm{E}[w]$ is analytically intractable at present; nevertheless its expectation is approximately equal to the CI width evaluated at $\psi'$.[7] That is,

$$\mathrm{E}[w] \approx (\lambda_U - \lambda_L)\sqrt{C/n}\ |\psi'. \tag{21}$$

Therefore, the necessary sample size per group that satisfies $\mathrm{E}[w] \leq \omega$ is the smallest $n$ that satisfies $(\lambda_U - \lambda_L)\sqrt{C/n} \leq \omega|\psi'$. The quantity $n$ cannot be directly solved for from this inequality, because $n$ is involved in the degrees of freedom in $t'_{(\alpha/2,\nu,\lambda_U)}$ and $t'_{(1-\alpha/2,\nu,\lambda_L)}$, which in turn affect the value of $(\lambda_U - \lambda_L)|\psi'$. One needs to obtain the value of $n$ numerically with an iterative process.[8]

Although sometimes the goal is to plan the sample size so that $\mathrm{E}[w] \leq \omega$, in many situations the desire is to obtain a sufficiently narrow CI in the study the researcher is going to carry out. An extended procedure can be performed so that there is $\gamma 100\%$ assurance that the $w$ obtained in a particular study will be no larger than $\omega$. As can be seen from equation (20), when $n$ and $C$ remain constant, the observed CI width is affected only by the difference between $\lambda_U$ and $\lambda_L$, which are in turn affected by the observed $t$-value $T'$ (since $T'$ is simply a constant times $\hat{\psi}'$). Due to the asymmetric nature of non-central $t$-distributions, the farther the $t$-value is from zero, the more positively skewed (when $T' > 0$, which implies $\hat{\lambda} > 0$) or more negatively skewed (when $T' < 0$, which implies $\hat{\lambda} < 0$) the distribution becomes, all other things being equal. Moreover, as the skewness of the distribution increases, the distance between $\lambda_U$ and $\lambda_L$ increases, implying a larger value of $\lambda_U - \lambda_L$, all other things being equal. Therefore, if a quantity $t'_\gamma$ can be found such that

$$P\left(|T'| \leq |t'_\gamma|\right) = \gamma, \tag{22}$$

[7] Given $\alpha$, $\nu$, $C$, and $n$, the value of $w$ is solely determined by $\hat{\psi}'$, and we use $w = h(\hat{\psi}')$ to denote such a relation, where $h$ refers to the non-linear confidence interval formation function (i.e., the CI formation discussed previously). Applying the Taylor expansion to $h(\hat{\psi}')$ at $\psi'$, we obtain $h(\hat{\psi}') = h(\psi') + h'(\psi')(\hat{\psi}' - \psi') +$ Remainder, where $h'()$ refers to the first derivative of $h()$ (e.g., Casella & Berger, 2002, p. 241). Taking expectation on both sides of the equality and omitting the remainder leads to $\mathrm{E}[h(\hat{\psi}')] \approx \mathrm{E}[h(\psi')] + h'(\psi')\mathrm{E}[\hat{\psi}' - \psi']$. The left-hand side $\mathrm{E}[h(\hat{\psi}')]$ is $\mathrm{E}[w]$ according to the definition of $\mathrm{E}[w]$. The first term on the right-hand side equals $h(\psi')$, and the second term equals zero. Thus it follows that $\mathrm{E}[h(\hat{\psi}')] \approx h(\psi')$ or equivalently $\mathrm{E}[w] \approx h(\psi')$.

[8] The functions that perform CI formation and sample size planning for $\psi'$ are ci.sc.ancova() and ss.aipe.sc.ancova(), respectively.

then, $\gamma 100\%$ of the time, the CI width based on the random value $T'$ will be no wider than the one based on $t'_\gamma$ (see Kelley & Rausch, 2006, for an application in the context of two-group standardized mean difference). The algorithm to numerically obtain $t'_\gamma$ in equation (22) is a generalization of the method proposed by Kelley and Rausch (2006), and is implemented in a specialized function in the MBESS R package.[9] After $t'_\gamma$ is found, it is transformed into the population effect size $\psi'_\gamma$; planning the sample size based on $\psi'_\gamma$ instead of $\psi'$ ensures a particular CI will be no wider than desired $\gamma 100\%$ of the time. This is the case because $|\hat{\psi}'|$ does not exceed $|\psi'_\gamma|$, $\gamma 100\%$ of the time, causing $w$ (which is based on $|\hat{\psi}'|$) not to exceed $\omega$ (which is based on $|\psi'_\gamma|$) $\gamma 100\%$ of the time.

### *3.1.2. Root mean square error of ANOVA as standardizer*

Let the ANCOVA contrast standardized with ANOVA root mean square error be defined as $\psi'' = \Psi'/\sigma_{\text{ANOVA}}$; it can be estimated by $\hat{\psi}'' = \hat{\Psi}'/s_{\text{ANOVA}}$, which is equivalent to $\hat{\psi}' s_{\text{ANCOVA}}/s_{\text{ANOVA}}$. To form a CI for $\psi''$, equations (16) and (17) need to be rewritten in a form that contains the root mean square error of ANOVA. This can be achieved by rearranging the equations as follows:

$$T' = \frac{\hat{\Psi}'}{s_{\text{ANCOVA}}\sqrt{C/n}} \cdot \frac{1/s_{\text{ANOVA}}}{1/s_{\text{ANOVA}}} = \frac{\hat{\psi}''}{u\sqrt{C/n}} \sim t'_{(nJ-J-1,\lambda)} \tag{23}$$

and

$$\lambda = \frac{\Psi'}{\sigma_{\text{ANCOVA}}\sqrt{C/n}} \cdot \frac{1/\sigma_{\text{ANOVA}}}{1/\sigma_{\text{ANOVA}}} = \frac{\psi''}{\upsilon\sqrt{C/n}}, \tag{24}$$

where $u = s_{\text{ANCOVA}}/s_{\text{ANOVA}}$ and $\upsilon = \sigma_{\text{ANCOVA}}/\sigma_{\text{ANOVA}}$.

The CI for $\lambda$ depends on the value of $T'$ only and thus is not affected by the rearrangement in equation (23). The difference between the CIs for $\psi'$ and $\psi''$ is in how the confidence limits of $\lambda$ are transformed back to the scale of their respective standardized contrasts. After finding the $(1-\alpha)100\%$ CI for $\lambda$, the confidence limits for $\psi''$ can be obtained as follows:

$$\text{CI}_{1-\alpha} = \left[\lambda_L \upsilon\sqrt{C/n} \leq \psi'' \leq \lambda_U \upsilon\sqrt{C/n}\ |T'\right], \tag{25}$$

when $\upsilon$ is known. However, in practice $\upsilon$ is almost always unknown and thus it requires an estimate for this population quantity to construct the CI. A reasonable estimator of $\upsilon$ is $u$. Applying the delta method, it can be shown that the expectation of $s_{\text{ANCOVA}}/s_{\text{ANOVA}}$ is approximately $\sigma_{\text{ANCOVA}}/\sigma_{\text{ANOVA}}$ (e.g., Casella & Berger, 2002, p. 245).

Using $u$ to estimate its corresponding value, therefore, returns the CI for $\psi''$:

$$\text{CI}_{1-\alpha} \approx \left[\lambda_L u\sqrt{C/n} \leq \psi'' \leq \lambda_U u\sqrt{C/n}\ |T'\right]. \tag{26}$$

The full CI width is

$$w = (\lambda_U - \lambda_L)u\sqrt{C/n}\ |\hat{\psi}'', u. \tag{27}$$

[9] The function that fulfils this task is ss.aipe.sc.ancova(). See the online appendix to the present paper or the help file in the MBESS R package for illustration and documentation.

Although the CI for $\lambda$ is exact, the one for $\psi''$ is only approximate, because the confidence limits for $\lambda$ are later multiplied by $u$, which is an estimate of $\upsilon$. Nevertheless, extensive Monte Carlo simulations we conducted indicate that $u$ is a very good estimator of $\upsilon$, and that using $u$ instead of $\upsilon$ to construct CI for $\psi''$ has no substantive impact on the confidence level (e.g., the empirical Type I error rate is not affected).[10]

Similar to the case of sample size planning for $\psi'$, we first develop methods to find $n$ so that $\text{E}[w] \leq \omega$. The expected CI width is approximately equal to the width evaluated at the population effect sizes:

$$\text{E}[w] \approx (\lambda_U - \lambda_L)\upsilon\sqrt{C/n}\,|\psi'', \upsilon. \tag{28}$$

To calculate $n$, one can first substitute $\psi''$ and $\upsilon$ for $\hat{\psi}''$ and $u$ respectively in the CI formation to obtain $\text{E}[w]$, and then solve $\text{E}[w] \leq \omega$ for $n$. Since $n$ also plays a role in the values of $\lambda_U$ and $\lambda_L$, it needs to be solved for numerically with an iterative process. Since the sample size planning process requires only the ratio $\upsilon$ instead of the respective values of $\sigma_{\text{ANOVA}}$ and $\sigma_{\text{ANCOVA}}$, sometimes it is easier to conceptualize $\upsilon$ in terms of the correlation between the response and the covariate. Based on equation (3), $\upsilon = \sqrt{1-\rho^2}$, and thus one can specify $\rho$ instead of $\upsilon$ if the knowledge about the correlation is easier to obtain.

The extended method to ensure that a CI for $\psi'$ is sufficiently narrow with $\gamma$100% assurance is not appropriate in the present context of CI for $\psi''$. A comparison between equations (20) and (27) helps explain this problem. As can be seen from equation (20), $\lambda_U - \lambda_L$ is the only factor of $w$ that varies, and when $|T'| \leq |t'_\gamma|$, the CI width for $\psi'$ will be no larger than desired. On the other hand, the observed CI width for $\psi''$ is influenced by two random variables, $(\lambda_U - \lambda_L)$ and $u$; even when $|T'| \leq |t'_\gamma|$, the CI width for $\psi''$ would still be larger than desired if $u$ exceeds $\upsilon$ by a non-trivial amount. To develop a satisfactory procedure, it needs to restrict the variation of both $T'$ and $u$, but the distribution of $u$ is an issue that has received little attention in the literature and remains unknown.

To achieve $\gamma$100% assurance that the CI for $\psi''$ observed in a particular study is sufficiently narrow, currently there are two practical solutions. The first of these is to use a normal distribution to approximate the distribution of $\psi''$. Instead of relying on the exact non-central $t$-distribution of $\psi''$, Bonett (2008, 2009) proposed using a normal distribution to approximately construct CI and plan sample size for $\psi''$ and other kinds of standardized contrast. This is reasonable because as the sample size approaches infinity, the distribution of $\psi''$ approaches normal. However, it is not known how well Bonett's methods perform at common, finite sample sizes. In addition, at the time of writing, there is no software available that readily implements his methods.

The second solution is to perform *a priori* Monte Carlo simulations. In an *a priori* Monte Carlo simulation study, the researcher specifies both the population parameters

[10] The simulations were conducted in the four-group context with the covariate being considered as random. The specifications were $\rho = 0, .10, .30, .50$, and $.80$; $\omega = 0.1\ (0.1)\ 0.6$; $\psi'' = 0.1, 0.2, 0.3, 0.5, 0.7$, and $1.0$, $\alpha = .05$. Each condition was replicated 10,000 times, and there were $5 \times 6 \times 6 = 180$ conditions in total. Empirical Type I error rates typically ranged from 4.7% to 5.3%; the maximum was 5.6% and the minimum was 4.4%. The complete set of simulation results were given to the anonymous reviewers as part of the assessment of our paper. Also note that the sample size planning methods for all three of the standardized effect sizes (i.e., $\psi$, $\psi'$, and $\psi''$) developed in the present paper were tested by extensive Monte Carlo simulations, and the complete sets of results were given to the reviewers. The simulation results are included in this article's online supplement accessible from https://repository.library.nd.edu/view/28/AIPE_for_Contrasts.pdf

and the sample size $n^*$, and generates a large number (e.g., 10,000) of random samples of size $n^*$ under the specified population parameters. Then one can empirically evaluate the properties of the CIs obtained. With *a priori* Monte Carlo simulations, one can continue to increase the sample size until the empirical percentage of $w \leq \omega$ is equal to the desired degree of assurance.

To illustrate sample size planning for $\psi''$ using *a priori* Monte Carlo simulation, let $n_\gamma$ denote the sample size per group that satisfies $P(|T'| \leq |t'_\gamma|) = \gamma$. Then one can insert $n_\gamma + 1$ as the selected sample size, along with other necessary specifications, generate a large number of random samples, calculate the CI for $\psi''$ based on each random sample, and compute the empirical percentage of $w \leq \omega$. If the percentage is still less than $\gamma$, increase the sample size by 1 and repeat the simulation; otherwise stop the process and take the current sample size as the one that satisfies $w \leq \omega$ with $\gamma$100% assurance.[11]

### 3.2. Standardized ANOVA contrasts

Let the population standardized ANOVA contrast be defined as $\psi = \Psi/\sigma_{\text{ANOVA}}$, which can be estimated by $\hat{\psi} = \hat{\Psi}/s_{\text{ANOVA}}$. Note that the numerator in the present context is an unadjusted contrast because the model does not include a covariate. The $t$-statistic associated with $\hat{\psi}$,

$$T' = \frac{\hat{\Psi}}{s_{\text{ANOVA}}\sqrt{C/n}} = \frac{\hat{\psi}}{\sqrt{C/n}} \sim t'_{(nJ-J,\lambda)}, \tag{29}$$

follows the non-central $t$-distribution with $nJ - J$ degrees of freedom. Although equation (29) has a similar form to equation (16), the $T'$ value, the degrees of freedom, and the non-centrality parameter are all different. The population non-centrality parameter is $\lambda = \psi/\sqrt{C/n}$ and can be estimated by $\hat{\lambda} = \hat{\psi}/\sqrt{C/n}$, which is equivalent to $T'$. CI formation and sample size planning for $\psi$ are analogous to those for $\psi'$ discussed previously. An exact CI for $\psi$ can be constructed with the CI formation process discussed previously in the $\psi'$ case. Sample size can be planned with the standard method so that $\text{E}[w] \leq \omega$, or with the extended method so that $P(w \leq \omega) = \gamma$.[12]

In summary, we have discussed CI formation and sample size planning methods for three types of standardized contrasts. The methods are all based on non-central $t$-distributions. In the ANCOVA context a contrast can be standardized in two ways, and in the ANOVA context there is one standardized contrast. These three standardized effect sizes have great similarity as well as important distinctions, and we will discuss them in more detail in the following section.

## 4. Comparing three types of standardized contrasts

In randomized designs, the population adjusted contrast $\psi'$, which is calculated in an ANCOVA context, is equal to the population unadjusted contrast $\Psi$, which is calculated in an ANOVA context. This is the case because the expected value of the covariate is the same across all groups in randomized designs, and the adjustment in contrast due to the covariate is zero in the population. Therefore, the population standardized ANCOVA contrast using $\sigma_{\text{ANOVA}}$ as divisor, $\psi'' = \Psi'/\sigma_{\text{ANOVA}}$, is equal to the population standardized

[11] The function ss.aipe.sc.ancova.sensitivity() implements such *a priori* Monte Carlo simulation.

[12] The functions for CI formation and sample size planning for $\psi$ are ci.sc() and ss.aipe.sc(), respectively.

ANOVA contrast $\psi = \Psi/\sigma_{\text{ANOVA}}$. When the interest lies in $\psi$, the researcher can also use the CI for $\psi''$ to estimate $\psi$. Although these two quantities and their corresponding CIs usually have different observed values in the same sample, the interest is the population parameter instead of any specific sample value. When it is relatively easy to obtain information about a covariate, even if the interest is in $\psi$, the researcher can include a covariate in the model and use the CI for $\psi''$ to estimate $\psi$. The benefit is that in the same sample the CI for $\psi''$ is narrower than that for $\psi$ . As can be seen from equations (27) and (29), when $C$ and $n$ remain the same, the CI width for $\psi''$ is affected by $T'$ and $u$, while that for $\psi$ is affected by $T'$ only. The value of $T'$ in the CI formation for $\psi''$ is roughly $u$ times as large as the value of $T'$ in the CI formation for $\psi$, where $u$ falls between 0 and 1. Moreover, the value of $\lambda_U - \lambda_L$, which is a function of $T'$, changes much more slowly than does $u$. Both $(\lambda_U - \lambda_L)$ and $u$ are in turn affected by $\rho$, the correlation between the response and the covariate. When $\rho$ is small, both the difference in the $(\lambda_U - \lambda_L)$ values and the magnitude of the factor $u$ are negligible, and thus the CI widths for $\psi''$ and for $\psi$ are close. As $\rho$ increases, $u$ increases more quickly than does $\lambda_U - \lambda_L$, making the CI for $\psi''$ narrower than that for $\psi$. The benefit gained in the CI width by reporting $\psi''$ instead of $\psi$ becomes more obvious as $\rho$ increases. Therefore, in randomized designs, if it is relatively easy to obtain information about a covariate, even when the interest is in $\psi$, it is still beneficial to construct and report the CI for $\psi''$.

The present paper has discussed three types of standardized contrasts: $\psi$ , $\psi'$, and $\psi''$. Table 1 summarizes these three effect sizes. The $t$-statistics associated with them are of different forms and follow different non-central $t$-distributions. The CI formation and sample size planning methods for $\psi$ are in the ANOVA context, and those for $\psi'$ and $\psi''$ are in the ANCOVA context. Although the expected CI widths for $\psi$ and $\psi'$ are of the same form, they should not be confused with each other because they necessarily have different confidence limits for the population non-centrality parameter. Different notations (i.e., $[\lambda_U - \lambda_L]$ and $[\lambda'_U - \lambda'_L]$) are used in the table to emphasize the different values in the confidence limits for the non-centrality parameter. A comparison between equations (20) and (27) indicates the relation between the CIs for $\psi'$ and $\psi''$. The CI for $\psi''$ has an extra term $u$, and because $0 < u < 1$, the CI for $\psi''$ is narrower than that for $\psi'$, other things being equal. The notations $w_\psi$ , $w_{\psi'}$, and $w_{\psi''}$ in the last column of Table 1 denote the CI widths for $\psi$ , $\psi'$, and $\psi''$, respectively.

### 4.1. Which standardized ANCOVA contrast should we use?

When the metric of the response variable is not fully understood, standardized measures can help interpret an effect of interest. An ANCOVA contrast $\Psi'$ can be standardized with $\sigma_{\text{ANOVA}}$ or $\sigma_{\text{ANCOVA}}$. Both $\Psi'/\sigma_{\text{ANOVA}}$ and $\Psi'/\sigma_{\text{ANCOVA}}$ can be regarded as reasonable measures of the effect $\Psi'$, yet they describe the magnitude from different perspectives. In the ANOVA context, $\sigma_{\text{ANOVA}}$ describes the variability of $Y$ within a group; that is, $sd(Y_{\cdot j}) = \sigma_{\text{ANOVA}}$. In the ANCOVA context, $\sigma_{\text{ANCOVA}}$ describes the conditional variability of $Y$ within a group given a covariate value; that is, $sd(Y_{\cdot j}|X) = \sigma_{\text{ANCOVA}}$. Since the conditional variability of $Y$ is assumed to be homogeneous across all $X$ values, one can focus on the variability of $Y$ conditional on $\mu_X$ without loss of generality. Then the ANCOVA root mean square error can be interpreted as $sd(Y_{\cdot j} \mid \mu_X) = \sigma_{\text{ANCOVA}}$, and it ties in well with an important question ANCOVA addresses: how would the response of different groups compare if the groups are equivalent on the covariate?

To better illustrate the connection between $\psi'$ and $\psi''$, consider a simple case of two-group randomized ANCOVA as depicted in Figure 1. The effect size of interest is the adjusted difference between two group means, as graphically represented by the bold line segment of length $\tau_2 - \tau_1$. The quantity $\sigma_{\text{ANOVA}}$ is the unconditional within-group

**Table 1.** *Comparing* $\psi$, $\psi'$, and $\psi''$

| Population effect size | Definition | Associated $t$-statistic | Degrees of freedom | Model context | Expected confidence interval width | Obtained confidence interval width |
|---|---|---|---|---|---|---|
| $\psi$ | $\Psi/\sigma_{\text{ANOVA}}$: unadjusted contrast divided by ANOVA root mean square error | $\frac{\hat{\psi}}{\sqrt{\sum_{j=1}^{J} c_j^2 /n_j}}$ | $N-J$ | ANOVA | $(\lambda_U - \lambda_L)\sqrt{\sum_{j=1}^{J} c_j^2 /n_j} \mid \psi$ | exact; $w_{\psi} > w_{\psi''}$ |
| $\psi'$ | $\Psi'/\sigma_{\text{ANCOVA}}$: adjusted contrast divided by ANCOVA root mean square error | $\frac{\hat{\psi}'}{\sqrt{\sum_{j=1}^{J} c_j^2 /n_j}}$ | $N-J-1$ | ANCOVA | $(\lambda'_U - \lambda'_L)\sqrt{\sum_{j=1}^{J} c_j^2/n_j} \mid \psi'$ | exact; $w_{\psi'} > w_{\psi''}$ |
| $\psi''$ | $\Psi'/\sigma_{\text{ANOVA}}$: adjusted contrast divided by ANOVA root mean square error | $\frac{s_{\text{ANOVA}}}{s_{\text{ANCOVA}}} \cdot \frac{\hat{\psi}''}{\sqrt{\sum_{j=1}^{J} c_j^2/n_j}}$ | $N-J-1$ | ANCOVA | $(\lambda'_U - \lambda'_L)\frac{\sigma_{\text{ANCOVA}}}{\sigma_{\text{ANOVA}}}\sqrt{\sum_{j=1}^{J} c_j^2/n_j} \mid \psi'$ | usually approximate; $w_{\psi''} < w_{\psi'}$; $w_{\psi''} < w_{\psi}$ |

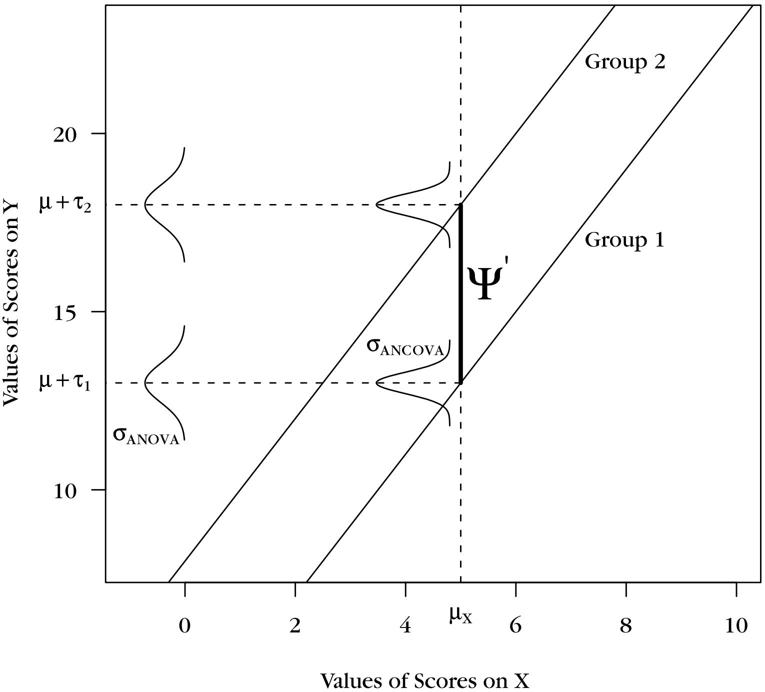

**Figure 1.** Graphical presentation of two types of standardized ANCOVA contrasts in the context of randomized one-way ANCOVA with two groups. The two oblique lines represent the population scores of the response and the covariate in the two groups. The bold line segment represents the adjusted mean difference of the two groups (i.e., $\Psi'$). There are two standardized measures of this raw effect: (a) in terms of unconditional within-group standard deviation (i.e., $\sigma_{\text{ANOVA}}$ as represented by the flatter bell curves); and (b) in terms of the within-group standard deviation conditional on the mean of the covariate (i.e., $\sigma_{\text{ANCOVA}}$ as represented by the sharper bell curves).

standard deviation of $Y$, which takes all values of $Y$ into account, whereas the quantity $\sigma_{\text{ANCOVA}}$ is the conditional within-group standard deviation of $Y$ evaluated at a specific value, say $X = \mu_X$, without loss of generality. The standardized contrast $\Psi'/\sigma_{\text{ANOVA}}$ measures the length of interest (i.e., the magnitude of $\Psi'$) in $\sigma_{\text{ANOVA}}$ units, and $\Psi'/\sigma_{\text{ANCOVA}}$ is another measure of the same length in $\sigma_{\text{ANCOVA}}$ units. Therefore, if the purpose is to interpret an adjusted ANCOVA (raw) contrast in a study, both ways of standardization are reasonable.

Another important use of standardized effect sizes is to compare results from different studies, and depending on the occasions, $\psi'$ or $\psi''$ can be more reasonable. An instance where $\Psi'/\sigma_{\text{ANOVA}}$ is more reasonable is when one compares $\Psi$ from an ANOVA study to $\Psi'$ from an ANCOVA study. Suppose two randomized studies investigate the same treatment effect, but study 1 is an ANOVA design and study 2 ANCOVA. If all assumptions hold and all other things are equal, at the population level the unadjusted contrast $\Psi$ in study 1 will equal the adjusted contrast $\Psi'$ in study 2, since the adjustment due to covariate is zero in a randomized design in the long run. That is, studies 1 and 2 would have the same expected treatment effect (though they would observe different values due to sampling variation). However, if one uses $\sigma_{\text{ANCOVA}}$ to standardize the contrast in study 2, $\Psi'/\sigma_{\text{ANCOVA}}$ from study 2 will be larger than $\Psi/\sigma_{\text{ANOVA}}$ from study 1 since study 2 has a smaller denominator, and thus it creates an illusion that study 2 observes a larger effect. In this case, comparing $\Psi'/\sigma_{\text{ANOVA}}$ from an ANCOVA study to $\Psi/\sigma_{\text{ANOVA}}$ from an ANOVA study is more appropriate (e.g., Glass *et al.*, 1981; Olejnik & Algina, 2000, 2003).

Consider another instance where one compares two randomized ANCOVA studies that investigate the same treatment effect. Suppose studies 1 and 2 are the same in all aspects except that study 1 samples adolescents only and study 2 samples people at any ages. Due to the restricted range of sampling, the within-group variability in study 1 will be smaller, and comparing $\Psi'/\sigma_{\text{ANOVA}}$ in the two studies leads to the biased conclusion that study 1 observes a larger effect. Based on equation (1), the within-group variance is $\text{Var}(Y_{\cdot j}) = \beta^2\text{Var}(X_{\cdot j}) + \sigma^2_{\text{ANCOVA}}$, and one source of the within-group variation is in the covariate. Since the two studies sample from different populations, the variances of the covariate in these two studies are also different. Therefore, the differential covariate variation contributes in part to the differential within-group variation. In this case, comparing $\Psi'/\sigma_{\text{ANCOVA}}$ is more appropriate, because $\sigma_{\text{ANCOVA}}$ removes the variation due to the covariate and makes the groups more homogeneous across the two studies. Nevertheless, note that besides the covariate, other components of the within-group variation may as well be different across the two studies. But it is not possible to single them out because they are all accounted for by the error term; consequently, $\sigma_{\text{ANCOVA}}$ in the two studies may still be different.

Both $\psi'$ and $\psi''$ can be regarded as reasonable and have their own advantages. If the primary purpose is to plan sample size and interpret the results within a study, we recommend using $\psi'$, because the extended sample size planning procedure for $\psi'$ is easier to implement. Comparison of standardized contrasts across different studies needs to proceed on a case-by-case basis. In fact, given the relation between $\sigma_{\text{ANOVA}}$ and $\sigma_{\text{ANCOVA}}$, there is a one-to-one mapping between the standardized contrasts $\Psi'/\sigma_{\text{ANOVA}}$ and $\Psi'/\sigma_{\text{ANCOVA}}$, which is $\psi' = \psi''/\sqrt{1-\rho^2}$. Therefore, it will facilitate the task to compare $\psi'$ and $\psi''$ among different studies if the researcher reports the denominator used in standardization, the correlation coefficient between $X$ and $Y$, as well as an explanation for choosing the particular measure of effect.

## 5. Discussion

Researchers should be reminded that sample size is not the only controllable factor in CI width, and that other factors such as confidence level and effect size can also influence the CI width (e.g., Baguley, 2004; Lenth, 2001). Other things being equal, a lower coverage gives a narrower CI. Although reducing the confidence level is generally discouraged, it can be an option in some circumstances (e.g., when the cost of Type I errors is not too high). An example in this regard is the CI for RMSEA: In the structural equation model context, conventionally a 90% CI for RMSEA is reported, yet a 95% CI is usually reported for other effects.[13] Another example is the two one-sided test (TOST) of (bio)equivalence. A TOST null hypothesis consists of two parts: $H_{0L}: \mu < \theta_1$ and $H_{0R}: \mu > \theta_2$. A TOST with $\alpha = .05$ rejects the null hypothesis only if both $H_{0L}$ is statistically

[13] The reason why a 90% CI for RMSEA is generally reported is simply due to convention: in their seminal paper Browne and Cudeck (1992) used .05 for both the upper and lower Type I error rate when illustrating their CI formation method for RMSEA, resulting a 90% confidence level. Mainstream structural equation modelling software also uses 90% as the default confidence level for RMSEA. We are not arguing that a 90% CI for RMSEA is more reasonable than a 95% one; we use this example simply to illustrate that the Type I error rate is flexible in some situations.

significant at .05 and $H_{0R}$ is statistically significant at .05. This is equivalent to forming a 90% two-sided CI for $\mu$ (Schuirmann, 1981; Westlake; 1981). Besides confidence level, in experimental studies, sometimes the researcher can manipulate the amount of treatment administered and thus influence the effect size, which in turn affects the CI width. A third way is to reduce standard error by increasing the reliability in measurement. In the ANCOVA or ANOVA context, one component of the root mean square error is the variation due to measurement error in the response. By reducing the measurement error, one can reduce the standard error in estimating contrasts. For more detailed discussions on various controllable factors other than $N$ on the CI width, interested readers are referred to Allison, Allison, Faith, Paultre, and Pi-Sunyer (1997), Baguley (2004), Lenth (2001), and McClelland (1997).

Although standardized effect sizes can facilitate the interpretation of results and comparison of different studies, they need to be used carefully. If the original scale of the (raw) effect is not meaningful or the standardizer is not meaningful, the mere act of standardizing will not make the resulting standardized effect any more meaningful (Bond, Wiitala, & Richard, 2003; Morris & DeShon, 2002; Tukey, 1969). Since the standardizer is usually a measure of variability (e.g., standard deviation of some kind), it is influenced by at least (a) the instrument's reliability, (b) the range of the sample, and (c) the design of the study (Baguley, 2009). All other things being equal, scores measured with an instrument that has high reliability have a smaller variance compared to those measured with lower reliability, and scores from a truncated distribution have a smaller variance compared to those from the complete population. Different designs usually have different error variances of their respective models, and thus analysing the same data with different models may lead to different estimates of error variance. The above three factors influence the estimation of standardized effect sizes to different extents in different studies, and therefore researchers need to be cautious when comparing studies that differ dramatically in those three factors.

The ANCOVA model on which we base our methods does not include an interaction between the treatment and the covariate, and it assumes that the treatment and the covariate are statistically independent. When this independence assumption is violated, the estimation of treatment effect will be affected. Maxwell and Delaney (2004, pp. 422–427) discussed four cases where lack of independence can arise and their possible solutions. Out of those four cases, the situation that can happen in the context of randomized studies is *unhappy randomization* (Kenny, 1979, p. 217), which refers to the fact that the researcher collects the covariate scores before the experiment, performs random assignment carefully, and only then finds that the groups differ statistically significantly on the covariate. In this case, the term $D$ used to calculate $s_{\hat{\Psi}'}$ is not close to zero and may not be negligible. However, the CI formation discussed previously is still valid because $D$ is always involved in $s_{\hat{\Psi}'}$ in data analysis (see equation (6)). The discrepancy is in the sample size planning procedure because it omits $D$. Therefore, when unhappy randomization happens, the planned $N$ may fail to help achieve the desired CI width, because the unusually large $D$ becomes not ignorable yet is ignored when calculating $\sigma_{\hat{\Psi}'}$ and planning for $N$. Nevertheless, note that such type of lack of independence between the treatment and the covariate is simply a Type I error: when the groups are equal on the covariate in the population, comparing the sample covariate means will still yield statistical significance about $\alpha 100\%$ of the time.[14] A second implication of unhappy randomization is that the large and non-ignorable $D$ term can

[14] Some people suggest performing randomization again in this situation (e.g., Rubin, 2008).

cause $s_{\hat{\Psi}'}$ to be larger than $s_{\hat{\Psi}}$; that is, the standard error of adjusted (raw) ANCOVA contrast can be larger than the standard error of unadjusted (raw) ANOVA contrast, and thus including covariates in the model can sometimes reduce the precision (Liu, 2011). This can happen when the sample size is especially small (e.g., $n = 10$), or the difference in the covariate among groups is large.[15]

Since equal sample size per group is usually beneficial, we have assumed equal $n$ to plan the sample size and reduce the amount of input information required. However, in practice unequal $n_j$ values can happen, for example, if the researcher uses smaller group sizes for more expensive treatments due to a limited budget. The sample size planning methods discussed previously can be extended to situations like this, and the researcher needs to plan the sample size for each group and use the full formula $s_{\text{ANCOVA}}\sqrt{(\sum_{j=1}^{J} c_j^2/n_j) + D}$ to calculate $s_{\hat{\Psi}'}$ instead of the simplified one $s_{\text{ANCOVA}}\sqrt{(C/n) + D}$. One way to achieve this is to define $n_j$ as $m_j \cdot \tilde{n}$, where $m_j$ is some measure of the cost per participant and $\tilde{n}$ is a baseline sample size being constant across groups (Hsu, 1994; see also Cochran, 1983; Liu, 2009). The sample size planning process requires $m_j$ values as input information and returns $\tilde{n}$, which in turn leads to $n_j$ values. Unequal $n_j$ values can also happen if some treatments are more disagreeable than others and cause more people to drop out. In the case of differential attrition, besides using the full formula to calculate $s_{\hat{\Psi}'}$ and form CIs, the research may also need to study the missingness pattern of the data and apply appropriate missing-data treatments. Analysing incomplete data is outside the scope of the present paper; interested readers are referred to sources such as Little and Rubin (2002). Note that if unequal $n_j$ values happen, the observed CI will tend to be wider than desired, because the full formula leads to a larger $s_{\hat{\Psi}'}$ than does the simplified one, and the study is planned assuming equal $n$ but analysed based on unequal $n_j$ values.

ANOVA and ANCOVA are among the most popular methods in psychology and related sciences. Contrasts of means are often of ultimate interest, because those targeted effects indicate to what extent groups differ from each other after administering treatments. CIs for contrasts provide more information than do hypothesis tests or point estimates, because from a CI we know not only what the population parameter is not, but also plausible values of that parameter. All other things being equal, a wider CI contains more uncertainty about the estimation, and therefore a narrow CI is usually desirable. Given the close relation between CI width and sample size, the issue of wider CIs can be addressed if the researcher plans the sample size from the AIPE perspective. Therefore, it is our hope that researchers consider the AIPE approach when planning an ANCOVA or ANOVA design.

## References

Allison, D. B., Allison, R. L., Faith, M. S., Paultre, F., & Pi-Sunyer, F. X. (1997). Power and money: Designing statistically powerful studies while minimizing financial costs. *Psychological Methods*, *2*, 20–33.

American Psychological Association. (2009). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

Baguley, T. (2004). Understanding statistical power in the context of applied research. *Applied Ergonomics*, *35*, 73–80.

[15] Based on the simulation results in Liu (2011), to ensure $s_{\hat{\Psi}'} < s_{\hat{\Psi}}$ about 80% of the time, a rule of thumb is to have at least 20 participants per group.

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

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist*, *64*, 27-41.

Bond, C. F., Jr, Wiitala, W. L., & Richard, F. D. (2003). Meta-analysis of raw mean differences. *Psychological Methods*, *8*, 406-418.

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

Bonett, D. G. (2009). Estimating standardized linear contrasts of means with desired precision. *Psychological Methods*, *14*, 1-5.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*, 230-258.

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

Cochran, W. G. (1957). Analysis of covariance: Its nature and uses. *Biometrics*, *13*, 261-281.

Cochran, W. G. (1983). *Planning and analysis of observational studies*. New York: Wiley.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997-1003.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, *61*, 532-574.

Embretson, S. E. (2006). The continued search for nonarbitrary metrics in psychology. *American Psychologist*, *61*, 50-55.

Fleishman, A. E. (1980). Confidence intervals for correlation ratios. *Educational and Psychological Measurement*, *40*, 659-670.

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

Guenther, W. C. (1965). *Concepts of statistical inference*. New York: McGraw-Hill.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum Associates.

Hays, W. L. (1994). *Statistics* (5th ed.). New York: Harcourt Brace College Publishers.

Hedges, L., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Hsu, L. (1994). Unbalanced designs to maximize statistical power in psychotherapy efficacy studies. *Psychotherapy Research*, *4*, 95-106.

Hunter, J. E., & Hamilton, M. A. (2002). The advantage of using standardized scores in causal analysis. *Human Communication Research*, *28*, 552-561.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.

Jiroutek, M. R., Muller, K. E., Kupper, L. L., & Stewart, P. W. (2003). A new method for choosing sample size for confidence interval-based inferences. *Biometrics*, *59*, 580-590.

Johnson, N. K., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions* (2nd ed., Vol. 2). New York: Wiley.

Johnson, N. K., & Welch, B. L. (1940). Applications of the noncentral *t*-distribution. *Biometrika*, *31*, 362-389.

Kelley, K. (2007a). Confidence intervals for standardized effect sizes: Theory, application, and implementation. *Journal of Statistical Software*, *20*, 1-24.

Kelley, K. (2007b). Sample size planning for the coefficient of variation: Accuracy in parameter estimation via narrow confidence intervals. *Behavior Research Methods*, *39*, 755-766.

Kelley, K. (2007c). Methods for the Behavioral, Educational, and Social Sciences: An R package. *Behavior Research Methods*, *39*, 979-984.

Kelley, K. (2008). Sample size planning for the squared multiple correlation coefficient: Accuracy in parameter estimation via narrow confidence intervals. *Multivariate Behavioral Research*, *43*, 524-555.

Kelley, K., & Lai, K. (2011a). MBESS (Version 3.2.1 or higher) [Computer software and manual]. Retrieved from http://www.cran.r-project.org/

Kelley, K., & Lai, K. (2011b). Accuracy in parameter estimation for the root mean square error of approximation: Sample size planning for narrow confidence intervals. *Multivariate Behavioral Research*, *46*, 1–32.

Kelley, K., & Maxwell, S. E. (2003). Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods*, *8*, 305–321.

Kelley, K., Maxwell, S. E., & Rausch, J. R. (2003). Obtaining power or obtaining precision: Delineating methods of sample-size planning. *Evaluation and the Health Professions*, *26*, 258–287.

Kelley, K., & Rausch, J. R. (2006). Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods*, *11*, 363–385.

Kenny, D. A. (1979). *Correlation and causality*. New York: Wiley-Interscience.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Belmont, CA: Brooks/Cole.

Kupper, L. L., & Hafner, K. B. (1989). How appropriate are popular sample size formulas? *American Statistician*, *43*, 101–105.

Lai, K., & Kelley, K. (2011). Accuracy in parameter estimation for targeted effects in structural equation modeling: Sample size planning for narrow confidence intervals. *Psychological Methods*, *16*, 127–148.

Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *American Statistician*, *55*, 187–193.

Little, R. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley.

Liu, X. S. (2009). Sample size and the width of the confidence interval for mean difference. *British Journal of Mathematical and Statistical Psychology*, *62*, 201–215.

Liu, X. S. (2011). The effect of covariate mean differences on the standard error and confidence interval for the comparison of treatment means. *British Journal of Mathematical and Statistical Psychology*, *64*, 310–319.

Mace, A. E. (1964). *Sample-size determination*. New York: Reinhold.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, *59*, 537–563.

McClelland, G. H. (1997). Optimal design in psychological research. *Psychological Methods*, *2*, 3–19.

Meehl, P. E. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 393–426). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Nickerson, R. S. (2000). Null hypothesis testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241–301.

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

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

Pan, Z., & Kupper, L. L. (1999). Sample size determination for multiple comparison studies treating confidence interval width as random. *Statistics in Medicine*, *18*, 1475–1488.

R Development Core Team. (2011). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Rencher, A. C., & Schaalje, G. B. (2007). *Linear models in statistics* (2nd ed.). Hoboken, NJ: Wiley.

Rubin, D. (2008). Comment: The design and analysis of gold standard randomized experiments. *Journal of the American Statistical Association*, *103*, 1350–1353.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, *1*, 115–129.

Schuirmann, D. L. (1981). On hypothesis testing to determine if the mean of a normal distribution is contained in a known interval. *Biometrics*, *37*, 617.

Smithson, M. (2003). *Confidence intervals*. Thousand Oaks, CA: Sage.

Steiger, J. H. (2004). Beyond the *F* test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods*, *9*, 164–182.

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

Steiger, J. H., & Lind, A. (1980). *Statistically based tests for the number of common factors*. Paper presented at the annual meeting of the Psychometric Society, Iowa City.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, *31*, 25–32.

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist*, *24*, 83–91.

Westlake, W. J. (1981). Response to T. B. L. Kirkwood: Bioequivalence testing – A need to rethink. *Biometrics*, *37*, 589–594.

*Received 9 November 2010; revised version received 30 July 2011*