Organizational Research Methods
Volume 11 Number 2
April 2008 364-386
© 2008 Sage Publications
10.1177/1094428106291059
http://orm.sagepub.com
hosted at
http://online.sagepub.com

Estimating Effect Sizes From Pretest-Posttest-Control Group Designs

Scott B. Morris
Illinois Institute of Technology

Previous research has recommended several measures of effect size for studies with repeated measurements in both treatment and control groups. Three alternate effect size estimates were compared in terms of bias, precision, and robustness to heterogeneity of variance. The results favored an effect size based on the mean pre-post change in the treatment group minus the mean pre-post change in the control group, divided by the pooled pretest standard deviation.

Keywords: *meta-analysis; effect size; repeated measures; experimental design*

Meta-analysis has become a critical tool for assessing the effectiveness of organizational interventions. In meta-analysis, researchers are able to combine data from multiple samples, thereby increasing the precision of treatment effect estimates. In addition, by analyzing the variability of results across studies, researchers can identify substantive or methodological moderators of treatment effectiveness. Meta-analysis has been used to evaluate the effectiveness of many types of organizational interventions, including employee training programs (Arthur, Bennett, Edens, & Bell, 2003; Arthur, Bennett, Stanush, & McNelly, 1998; Callahan, Kiker, & Cross, 2003; Carlson & Schmidt, 1999; Driskell, Willis, & Copper, 1992; M. A. Morris & Robie, 2001; Morrow, Jarrett, & Rupinski, 1997; Scott, Leritz, & Mumford, 2004; P. J. Taylor, Russ-Eft, & Chan, 2005; Woehr & Huffcutt, 1994), leadership development programs (Collins & Holton, 2004; Eden et al., 2000; McNatt, 2000; Niemiec, Sikorski, Clark, & Walberg, 1992), performance management systems (Guzzo, Jette, & Katzell, 1985; Kluger & DeNisi, 1996; Smither, London, & Reilly, 2005), and organizational development interventions (Neuman, Edwards, & Raju, 1989; Roberts & Robertson, 1992).

Despite advances in the statistical models available, researchers are still faced with a number of operational challenges when conducting a meta-analysis. One of these challenges is dealing with data from varying research designs. Program evaluation studies differ in the use of control groups, covariates, and between-groups versus repeated-measures designs. It is important for researchers to understand how results from alternate designs can best be integrated into a meta-analysis.

Although primary research studies utilize a variety of research designs, these differences are often ignored when conducting a meta-analysis. The majority of the meta-analyses listed above do not specify how they dealt with design differences across studies. Methods

Author's Note: Portions of the article were presented at the 18th annual conference of the Society for Industrial and Organizational Psychology, Orlando, Florida, April 2003.

for meta-analysis have been developed primarily for assessing the difference between means for two independent groups. Although methods exist to deal with more complex designs (Cortina & Nouri, 2000; Olejnik & Algina, 2000; Rosenthal, 1994), the focus of past research has been on obtaining effect size estimates that are comparable to those from a simple two-group study. Little attention has been given to the relative precision of effect size estimates from alternate designs and the impact of these differences on meta-analytic research.

The current research will focus on the pretest-posttest-control (PPC) design, a useful framework for evaluating organizational interventions. In the PPC design, research participants are assigned to treatment or control conditions, and each participant is measured both before and after the treatment has been administered. The PPC design can provide useful estimates of treatment effects as either an experimental design, where participants are randomly assigned to treatment conditions, or as a quasiexperimental design, where randomization is not feasible (Cook & Campbell, 1979).

The PPC design has a number of advantages over other common designs in evaluation research. The posttest only with control design (POWC) has participants assigned to treatment and control conditions, but participants are measured only after administration of the treatment. In quasiexperimental designs, preexisting differences between groups could artificially inflate or obscure differences at posttest, casting suspicion on results from the POWC design. In contrast, the PPC design allows researchers to control for preexisting differences, allowing estimates of treatment effectiveness even when treatment and control groups are nonequivalent (Cook & Campbell, 1979; S. B. Morris & DeShon, 2002). Even in experimental designs, where preexisting group differences are controlled through random assignment, there are advantages to the PPC design. The use of repeated measurements in the PPC design allows each individual to be used as his or her own control, which typically increases the power and precision of statistical tests (Hunter & Schmidt, 2004).

Another common approach to evaluating organizational interventions is the single-group pre-post (SGPP) design. As in the PPC design, participants are measured before and after treatment, allowing direct estimation of the amount of learning or change that occurs within an individual. However, the SGPP design does not include a no-treatment control group. Consequently, estimates of treatment effects from the SGPP design may be confounded with factors other than the intervention that produce changes in the outcome variable. For example, new hires are likely to improve in performance simply because of experience with the job, even if no training is provided. In addition, an intervention may correspond to other changes occurring in the work setting that could affect outcome measures. The PPC design appropriately controls for such factors by comparing the pre-post change in the treatment group to the amount of change in a group that experiences the same organizational context but did not receive the intervention.

Because of these factors, the PPC design is often recommended for program evaluation research in areas such as employee training (Goldstein & Ford, 2002; Quinones & Tonidandel, 2003) and occupational safety and health (Robson, Shannon, Goldenhar, & Hale, 2001). Consequently, the research base in these areas is likely to include studies that utilize the PPC design. Indeed, several recent training meta-analyses have identified a substantial number of studies reporting PPC designs (Carlson & Schmidt, 1999; Collins & Holton, 2004; P. J. Taylor et al., 2005).

Researchers analyzing data from PPC designs are faced with the challenge of choosing among a variety of data analysis strategies. Researchers might compute pre-post gain scores for each individual and then test for differences in the mean gain between treatment and control groups using a *t* test. Alternatively, the data could be analyzed using a mixed effects analysis of variance with treatment condition as a between-groups factor and pre- versus posttest scores as a within-subject factor. Another approach would be to conduct an analysis of covariance treating pretest scores as a covariate or, similarly, to test for group difference using residualized gain scores. Each of these approaches has advantages and disadvantages, and the choice of the best approach continues to be a matter of some debate (Arvey & Cole, 1989; Cook & Campbell, 1979; Maris, 1998).

A similar array of choices face researchers when defining effect sizes for a meta-analysis. To most effectively incorporate studies with the PPC design into a meta-analysis, researchers should utilize a definition of the effect size that takes full advantage of the relevant information available in the study. That is, the effect size should be defined using both pretest and posttest information. In addition, the meta-analytic procedures must be adapted to reflect the precision of effect sizes from alternate designs. Previous work on this topic offers several distinct estimates of effect size for the PPC design. This article compares these alternatives in terms of their precision and the practical challenges that limit their use in meta-analysis.

When choosing among alternate effect size estimates, several factors should be considered. First, the effect size estimate should be unbiased. Second, among unbiased estimates, the most precise effect size should be selected. In general, estimates with smaller sampling variance will provide more precise estimates of the mean effect size, particularly when the number of studies in the meta-analysis is small. Even in large meta-analyses, moderator analysis often requires the examination of subgroups with a relatively small number of studies. Therefore, the selection of a more precise effect size estimate can improve the accuracy of meta-analytic estimates and enhance the power of moderator tests.

A third consideration is that the distribution of the effect size must be known. Characteristics of the sampling distribution, such as the degree of bias or the sampling variance, are needed to conduct a meta-analysis. For example, estimates of sampling variance are used in several meta-analysis procedures. When computing the precision-weighted mean effect size, the weights are computed from the inverse of the sampling variance (Hedges & Olkin, 1985). Estimates of sampling variance are also needed to build confidence intervals around the mean effect size estimate, to test for homogeneity of effect size, and to estimate the between-study variance component in random effects models.

A forth factor that can be used to choose among alternate effect size estimates is robustness to violations of model assumptions. Standard meta-analysis procedures make many assumptions about the nature of the data (e.g., normality, homogeneity of variance) that may be inappropriate in many situations. Some effect size estimates may be more resistant than others to the effects of violating these assumptions.

The current article will consider violations of the homogeneity of variance assumption. All of the effect sizes to be compared assume that pre- and posttest scores have equal variance. However, when the effect of treatment is not the same for each individual, the treatment will tend to increase the variance of scores. Therefore, posttest variances are often larger than pretest variances, resulting in smaller effect size estimates for alternatives that use the posttest standard deviations (Bryk & Raudenbush, 1988; Carlson & Schmidt, 1999).

The following section will define an effect size for the PPC design and present three alternate estimates of this effect size. The distribution of each effect size will be discussed, and the results of a Monte Carlo simulation will be used to compare the relative efficiency of the alternatives. Next, the effect of violating the homogeneity of variance assumption will be examined.

Effect Size for the PPC Design

The data are assumed to be randomly sampled from two populations, corresponding to treatment and control conditions. Pretest and posttest scores in each population have a bivariate normal distribution with common variance σ^2 and common correlation ρ, but distinct means, indicated by $\mu_{\mathrm{T,pre}}$ for the treatment population pretest, $\mu_{\mathrm{T,post}}$ for the treatment population posttest, $\mu_{\mathrm{C,pre}}$ for the control group pretest, and $\mu_{\mathrm{C,post}}$ for the control group posttest.

The standardized mean change in each population is defined as the mean difference between posttest and pretest scores, divided by the common standard deviation. The standardized mean change for the treatment group (δ_{T}) is

$$\delta_T = \frac{\mu_{T,post} - \mu_{T,pre}}{\sigma}. \tag{1}$$

The standardized mean change for the control group (δ_{C}) is

$$\delta_C = \frac{\mu_{C,post} - \mu_{C,pre}}{\sigma}. \tag{2}$$

The effect size for the PPC design is defined as the difference between the standardized mean change for the treatment and control groups,

$$\Delta = \delta_T - \delta_C = \frac{(\mu_{T,post} - \mu_{T,pre}) - (\mu_{C,post} - \mu_{C,pre})}{\sigma}. \tag{3}$$

Alternate Effect Size Estimates

To illustrate the choices involved in the calculation of effect size estimates, Table 1 presents the descriptive statistics for five studies taken from a meta-analysis on training effectiveness by Carlson & Schmidt (1999). An individual study consists of n_{T} participants receiving treatment and n_{C} participants in the control group. The pretest and posttest means for the treatment group are indicated by $M_{\mathrm{pre,T}}$ and $M_{\mathrm{post,T}}$, respectively. The pretest and posttest means for the control group are indicated by $M_{\mathrm{pre,C}}$ and $M_{\mathrm{post,C}}$, respectively. A separate estimate of the standard deviation can be obtained for the treatment groups at pretest ($SD_{\mathrm{pre,T}}$) and posttest ($SD_{\mathrm{post,T}}$) and for the control group at pretest ($SD_{\mathrm{pre,C}}$) and posttest ($SD_{\mathrm{post,C}}$). These standard deviations can be combined in several different ways to derive different estimates of the effect size Δ.

Table 1
Descriptive Statistics and Effect Size Estimates for Studies Included in Meta-Analysis

	Treatment Group					Control Group								
		Pretest		Posttest			Pretest		Posttest			Effect Size Estimate		
Study	n	M	SD	M	SD	n	M	SD	M	SD	Pre-Post r	d_{ppc1}	d_{ppc2}	d_{ppc3}
Blicksenderfer, Cannon-Bowers, and Salas (1997)	20	30.6	(15.0)	38.5	(11.6)	20	23.1	(13.8)	19.7	(14.8)	.47	0.74	0.77	0.81
Ivancevich and Smith (1981)	50	23.5	(3.1)	26.8	(4.1)	42	24.9	(4.1)	25.3	(3.3)	.64	0.95	0.80	0.79
Katz and Schwebel (1976)	9	0.5	(0.1)	0.7	(0.1)	9	0.6	(0.2)	0.6	(0.2)	.77	1.81	1.20	1.24
McGehee and Gardner (1955)	10	53.4	(14.5)	75.9	(4.4)	11	55.7	(17.3)	60.7	(17.9)	.89	1.15	1.05	1.16
Miraglia (1963)	14	35.6	(4.7)	36.0	(4.6)	14	34.8	(3.1)	33.4	(6.9)	.44	0.51	0.44	0.35

Effect Size Estimate Using Separate Pretest *SD*s

Becker (1988) described an effect size measure for the PPC design, referred to here as g_{ppc1},

$$g_{ppc1} = \frac{M_{post,T} - M_{pre,T}}{SD_{pre,T}} - \frac{M_{post,C} - M_{pre,C}}{SD_{pre,C}}. \tag{4}$$

Becker showed that g_{ppc1} is a biased estimator of the population effect size, although the bias is quite small when the sample size in each group is greater than 10. An approximately unbiased estimate can be obtained using

$$d_{ppc1} = c_T\left(\frac{M_{post,T} - M_{pre,T}}{SD_{pre,T}}\right) - c_C\left(\frac{M_{post,C} - M_{pre,C}}{SD_{pre,C}}\right), \tag{5}$$

where the bias adjustments c_{T} and c_{C} can be approximated by

$$c_j = 1 - \frac{3}{4(n_j - 1) - 1}. \tag{6}$$

Applying this to the first study in Table 1 yields the following estimate,

$$d_{ppc1} = \left(1 - \frac{3}{4(20-1)-1}\right)\left(\frac{38.5 - 30.6}{15.0}\right) - \left(1 - \frac{3}{4(20-1)-1}\right)\left(\frac{19.7 - 23.1}{13.8}\right) = 0.74. \tag{7}$$

Effect Size Estimate Using Pooled Pretest *SD*

A limitation of d_{ppc1} is that separate estimates of the sample standard deviation are used ($SD_{\mathrm{pre,T}}$ and $SD_{\mathrm{pre,C}}$), despite the assumption that the population variances are homogeneous. Under this assumption, a better estimate of the population standard deviation could be obtained by pooling the data from the treatment and control groups. This suggests an alternative effect size estimate, which will provide a more precise estimate of the population treatment effect,

$$d_{ppc2} = c_P\left[\frac{(M_{post,T} - M_{pre,T}) - (M_{post,C} - M_{pre,C})}{SD_{pre}}\right] \tag{8}$$

where the pooled standard deviation is defined as

$$SD_{pre} = \sqrt{\frac{(n_T - 1)SD^2_{pre,T} + (n_C - 1)SD^2_{pre,C}}{n_T + n_C - 2}} \tag{9}$$

and

$$c_P = 1 - \frac{3}{4(n_T + n_C - 2) - 1}. \tag{10}$$

Except for the bias correction, d_{ppc2} is the same as the effect size estimate (ES_{PPWC}) recommended by Carlson and Schmidt (1999). Applying Equation 8 to the first study in Table 1 yields a pooled *SD* of 14.4 and an effect size estimate of

$$d_{ppc2} = \left(1 - \frac{3}{4(20+20-2)-1}\right)\left[\frac{(38.5-30.6)-(19.7-23.1)}{14.4}\right] = 0.77. \qquad (11)$$

Effect Size Based on the Pooled Pre- and Posttest *SD*

Both of the preceding estimates consider only the pretest standard deviations. Under the assumed model, pretest and posttest variances are homogeneous. Therefore, a more precise estimate (d_{ppc3}) can be obtained by pooling estimates across both pretest and posttest measurements for both treatment and control conditions (Dunlap, Cortina, Vaslow, & Burke, 1996; M. J. Taylor & White, 1992). Specifically,

$$d_{ppc3} = c_{PP}\left[\frac{(M_{post,T} - M_{pre,T}) - (M_{post,C} - M_{pre,C})}{SD_{pre+post}}\right]. \qquad (12)$$

where the pooled standard deviation is defined as

$$SD_{pre+post} = \sqrt{\frac{(n_T-1)SD^2_{pre,T} + (n_C-1)SD^2_{pre,C} + (n_T-1)SD^2_{post,T} + (n_C-1)SD^2_{post,C}}{2(n_T+n_C-2)}}. \qquad (13)$$

For d_{ppc3}, the exact value of the bias correction is not known. If the four standard deviations were from independent groups, the bias correction would be

$$c_{PP} = 1 - \frac{3}{4(2n_T+2n_C-4)-1}. \qquad (14)$$

However, because pre- and posttest scores are not independent, the amount of information gained by adding the posttest standard deviations will be less than if the groups were independent (Kish, 1965). Therefore, the degree of bias is likely to be greater than indicated by Equation 14 (i.e., c_{PP} will be too large), particularly when there is a substantial correlation between pre- and posttest scores. Consequently, it was expected that the use of Equation 14 would result in a slight overestimate of the population effect size for large values of ρ.

Applying Equation 12 to the first study in Table 1 would yield a pooled *SD* of 13.9 and an effect size estimate of

$$d_{ppc3} = \left(1 - \frac{3}{4(2*20+2*20-4)-1}\right)\left[\frac{(38.5-30.6)-(19.7-23.1)}{13.9}\right] = 0.81. \qquad (15)$$

Comparison of Alternate Effect Size Estimates

The alternate effect sizes estimated in Table 1 illustrate that the choice of an estimate can make a difference in the value obtained. Although the alternative methods will often provide similar values, substantial differences can occur. For example, the effect size for Katz and Schwebel (1976) ranged from 1.20 for d_{ppc2} to 1.81 for d_{ppc1}. Given the potential differences, it is important to determine which method of computing the effect size is most effective.

The two alternatives that involve pooling of standard deviations across groups should allow more precise estimates of effect size from studies using the PPC design. However, factors other than precision should also influence the choice of an effect size estimate. To use the alternatives in a meta-analysis, it is necessary to first specify their sampling distribution. Estimates of the sampling variance are needed for meta-analytic procedures, such as computing the weighted mean or estimating variance of effect sizes in a random effect model. In addition, it is important to consider the behavior of each alternative under violations of the model assumptions. The following sections will discuss the theoretical sampling distribution of each effect size estimate, followed by a Monte Carlo simulation comparing the mean and variance of each estimate under a variety of conditions. A second simulation compares the performance of each statistic when the assumption of homogeneity of variance is violated.

Distribution of Alternate Effect Size Estimates

Distribution of d_{ppc1}. Becker (1988; cf. S. B. Morris, 2000; S. B. Morris & DeShon, 2002) derived the asymptotic distribution of d_{ppc1}. When sample size is large, d_{ppc1} is approximately normally distributed, with a mean equal to the population effect size Δ and variance,

$$\sigma^2(d_{ppc1}) = c_T^2\left(\frac{2(1-\rho)}{n_T}\right)\left(\frac{n_T-1}{n_T-3}\right)\left(1+\frac{n_T\delta_T^2}{2(1-\rho)}\right)-\delta_T^2$$
$$+c_C^2\left(\frac{2(1-\rho)}{n_C}\right)\left(\frac{n_C-1}{n_C-3}\right)\left(1+\frac{n_C\delta_C^2}{2(1-\rho)}\right)-\delta_C^2\,. \qquad (16)$$

Distribution of d_{ppc2}. The asymptotic distribution of d_{ppc2} can be derived using the approach developed by Becker (1988). Following Hedges (1981), the distribution of d_{ppc2} is derived by relating the effect size to the noncentral *t* distribution.

The effect size estimate without the correction for small sample bias will be referred to as g_{ppc2}, which is defined,

$$g_{ppc2} = \frac{(M_{post,T}-M_{pre,T})-(M_{post,C}-M_{pre,C})}{SD_{pre}}. \qquad (17)$$

The numerator is normally distributed with a mean of σΔ and a standard error of

$$\sigma\sqrt{2(1-\rho)\left(\frac{n_T+n_C}{n_T\,n_C}\right)} \qquad (18)$$

Therefore,

$$\frac{g_{ppc2}}{\sqrt{2(1-\rho)\left(\frac{n_T+n_C}{n_T\,n_C}\right)}} = \frac{\left[\frac{(M_{post,T}-M_{pre,T})-(M_{post,C}-M_{pre,C})}{\sigma\sqrt{2(1-\rho)\left(\frac{n_T+n_C}{n_T\,n_C}\right)}}\right]}{\left[\frac{SD_{pre,P}}{\sigma}\right]} \tag{19}$$

is distributed as a noncentral *t*, with $df = n_{\mathrm{T}} + n_{\mathrm{C}} - 2$ and noncentrality parameter,

$$\phi = \frac{\Delta}{\sqrt{2(1-\rho)\left(\frac{n_T+n_C}{n_T\,n_C}\right)}}. \tag{20}$$

As a result, g_{ppc2} is approximately normally distributed with expected value

$$E(g_{ppc2}) = \left(\sqrt{2(1-\rho)\left(\frac{n_T+n_C}{n_T\,n_C}\right)}\right)\left(\frac{\Delta}{c_P\left(\sqrt{2(1-\rho)\left(\frac{n_E+n_C}{n_T\,n_C}\right)}\right)}\right) = \frac{\Delta}{c_P}. \tag{21}$$

c_P is a function indicating the degree of bias in g_{ppc2},

$$c_P = \sqrt{\frac{2}{df}}\left(\frac{\Gamma[df/2]}{\Gamma[(df-1)/2]}\right), \tag{22}$$

where Γ is the gamma function (Johnson & Kotz, 1970). An unbiased estimate (d_{ppc2}) can be obtained by multiplying the effect size by c_{P}, as in Equation 7. A close approximation to c_{P} is provided in Equation 10.

The variance of *g* is $2(1-\rho)\,(n_T + n_C)/(n_T n_C)$ times the variance of a noncentral *t*. The variance of a noncentral *t* is defined by Johnson and Kotz (1970) as

$$\sigma^2(t) = \left(\frac{df}{df-2}\right)(1+\phi^2) - \frac{\phi^2}{[c(df)]^2}. \tag{23}$$

Substituting $df = n_{\mathrm{T}} + n_{\mathrm{C}} - 2$ and the noncentrality parameter specified in Equation 17, the variance of g_{ppc2} is

$$\sigma^2(g_{ppc2}) = 2(1-\rho)\left(\frac{n_T+n_C}{n_T\,n_C}\right)\left(\frac{n_T+n_C-2}{n_T+n_C-4}\right)\left(1+\frac{\Delta^2}{2(1-\rho)\left(\frac{n_T+n_C}{n_T\,n_C}\right)}\right) - \left(\frac{\Delta^2}{c_p^2}\right). \tag{24}$$

The variance of d_{ppc2} is $c_{\text{p}}^{\,2}$ times the variance of g_{ppc2}, or

$$\sigma^2(d_{ppc2}) = 2(c_P^2)(1-\rho)\left(\frac{n_T+n_C}{n_T\,n_C}\right)\left(\frac{n_T+n_C-2}{n_T+n_C-4}\right)\left(1+\frac{\Delta^2}{2(1-\rho)\left(\frac{n_T+n_C}{n_T\,n_C}\right)}\right)-\Delta^2. \quad (25)$$

Distribution of d_{ppc3}. The third alternative (d_{ppc3}) is problematic because the sampling distribution is unknown. For d_{ppc3}, the standard deviation is pooled across nonindependent scores (i.e., pretest and posttest scores). As a result, the gain in precision will be less than if standard deviations were pooled across independent groups. The exact sampling variance of d_{ppc3} is currently unknown but is expected to be smaller than the other alternatives and to approach d_{ppc2} as ρ approaches 1.

Bias and Precision of Alternate Effect Size Estimates

A Monte Carlo simulation was used to investigate the mean and sampling variance of each effect size. Monte Carlo analyses involve the use of computer simulations to generate artificial data sets that can be subjected to statistical analyses. Such simulations are advantageous for studying the behavior of statistical procedures because large amounts of data can be generated with known characteristics. By computing the effect size estimates on a large number of randomly generated samples, it is possible to obtain an empirical approximation of the actual sampling distributions under a variety of controlled conditions. These empirical sampling distributions can then be compared to theoretically derived values to determine the conditions under which alternative procedures will yield accurate results.

The performance of each estimate was investigated across varying levels of Δ, ρ, and *n*. The population effect size, Δ, was manipulated by varying the standardized mean change in the treatment and control groups. The standardized mean change in the treatment group, δ_{T}, was set at 0.0, 0.5, or 1.0 to reflect the range of effect sizes typically observed in organizational intervention research. For example, a recent large-scale meta-analyses of training evaluation studies (Arthur et al., 2003) found an average effect size of about 0.6, with effect sizes for specific subgroups varying widely (0.3-1.9). The standardized mean change in the control group, δ_{C}, was set at 0.0 or 0.2, which corresponds to the range of control-group effect sizes obtained in training meta-analyses (Carlson & Schmidt, 1999). The correlation between pre- and posttest scores, ρ, was always the same for the treatment and control groups and was manipulated to reflect a wide range of possible situations, 0.0, 0.45, or 0.9. The latter two values are consistent with the range of values obtained for different criteria in a meta-analysis by P. J. Taylor et al. (2005). The condition with $\rho = 0.0$ was included for the sake of completeness. The sample size was always equal across groups, and the within-group *n* was 10, 25, or 50. The corresponding total sample sizes (20, 50, or 100) provide a good representation of the typical values found in organizational intervention research. The meta-analysis by Arthur et al. (2003) reported average samples sizes for different outcome variables ranging from 62 to 128, with average sample size for some subgroups analyses as small as 23.

A Fortran program was used to conduct the simulation. For each combination of the parameters, the program generated 10,000 samples with *n* observations from both the treatment and control groups. For each observation, pre- and posttest scores were generated from a bivariate normal distribution with a pre-post correlation of ρ. The data were generated using the multivariate random number generator (DRNMVN) of the International Mathematical and Statistical Library (IMSL, 1984). The effect of treatment was manipulated by adding δ_T or δ_C to the posttest scores, depending on the group.

For each sample, the three effect size estimates (d_{ppc1}, d_{ppc2}, and d_{ppc3}) were computed using the formulas given above. The mean and variance of each effect size was computed across the 10,000 random samples. The mean effect sizes were evaluated based on how closely they matched the population effect size specified in the simulation (i.e., bias). Sampling variances were compared across the three alternatives in terms of relative efficiency (i.e., the ratio of the variance of one estimator to the variance of another estimator) and how closely the observed variance matched the theoretical variance, when known. The results of the simulation are summarized in Tables 2 and 3.

As expected, d_{ppc1} and d_{ppc2} were nearly unbiased. Across all conditions, the mean effect size was always within .008 of the population parameter. Generally, d_{ppc3} provided a good estimate, except when n was small and ρ was large, in which case there was a slight positive bias. For example, with $\delta_T = 1.0$, $\delta_C = 0.0$, $\rho = 0.9$, and $n = 10$, the mean effect size was 1.013. As expected, when ρ was large, the bias correction, which was based on the degrees of freedom from four independent groups, underestimated the degree of bias, resulting in an overestimate of the population effect size. However, the degree of bias was not large.

The results also verify the accuracy of the sampling variance formulas for d_{ppc1} and d_{ppc2}. Under most conditions, the theoretical variance differed from the observed variance by less than 3%, and in no case was the difference greater than 7% of the observed variance.

The analysis of relative efficiency confirmed the expectation that d_{ppc2}, where the pretest standard deviations were pooled across treatment and control groups, would provide a more precise estimate of the sampling variance than d_{ppc1}. In general, d_{ppc2} had a smaller sampling variance than d_{ppc1}, as indicated by relative efficiency less than 1.0. Figure 1 displays the relative efficiency for conditions with no change in the control group. When both $\delta_E = 0.0$ and $\delta_C = 0.0$, there was little difference between the two estimates. However, as δ_T increased, the superiority of d_{ppc2} over d_{ppc1} became more apparent. The difference became larger as sample size decreased and as ρ increased. The pattern of results was similar but more extreme when $\delta_C = 0.2$ (see Figure 2). Under the most extreme conditions ($\delta_T = 1.0$, $\delta_C = 0.2$, $\rho = 0.9$, and $n = 10$), the variance of d_{ppc2} was as much as 50% lower than the variance of d_{ppc1}.

Contrary to expectations, pooling both pre- and posttest standard deviations in d_{ppc3} did not substantially reduce the sampling variance relative to d_{ppc2}. Across conditions, the variance of d_{ppc3} was generally lower than the variance of d_{ppc2}, but the difference was quite small. The relative efficiency of d_{ppc3} to d_{ppc2} was generally greater than .96 and never fell below .94.

Heterogeneity of Variance

A common concern with the PPC design is that the posttest variance in the treatment group may be larger than the variance of the untreated population, violating the homogeneity of variance assumption. If there are individual differences in the effectiveness of a treatment (a subject by treatment interaction), some individuals in the treatment condition will improve more than others, and the distribution of scores at posttest will be more spread out than at pretest (Bryk & Raudenbush, 1988; Cook & Campbell, 1979). It is

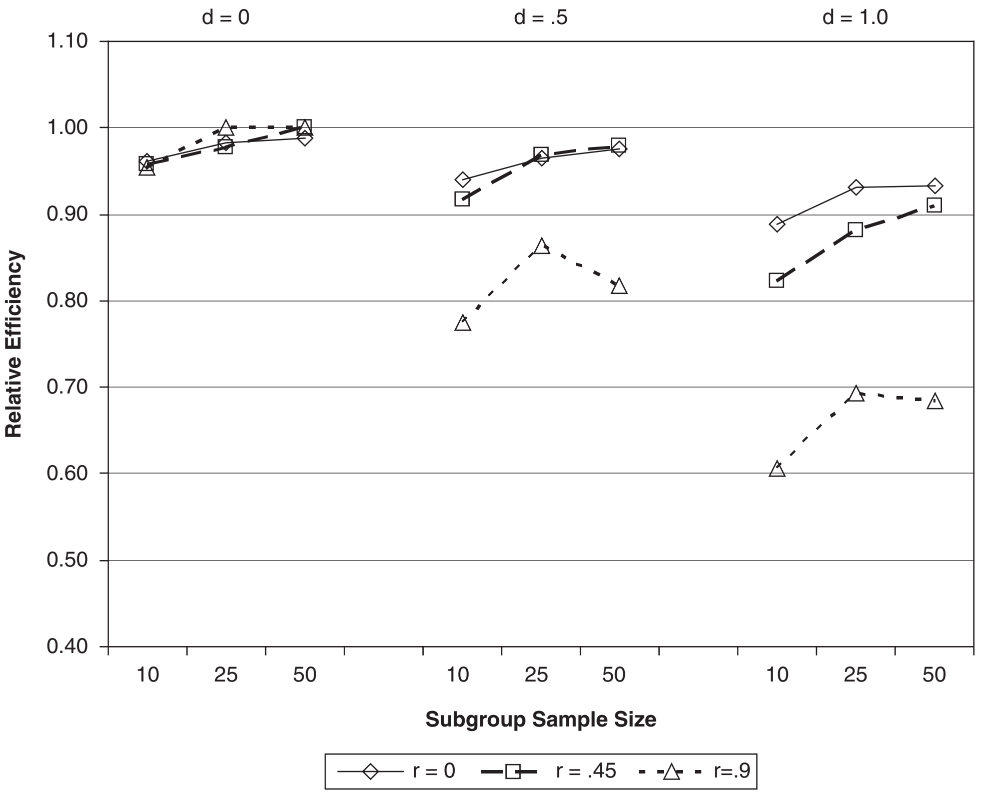
Table 2
Mean and Variance of Alternate Effect Size Estimates When δc = 0.0

			d_{ppc1}			d_{ppc2}			d_{ppc3}		Relative Efficiency	
δ_T	ρ	*n*	*M*	Observed Variance	Theoretical Variance	*M*	Observed Variance	Theoretical Variance	*M*	Observed Variance	d_{ppc2}/d_{ppc1}	d_{ppc3}/d_{ppc2}
0.0	0.00	10	0.003	0.425	0.430	0.002	0.409	0.413	0.000	0.404	0.96	0.99
0.0	0.00	25	0.004	0.163	0.164	0.004	0.161	0.162	0.004	0.159	0.99	0.99
0.0	0.00	50	0.003	0.079	0.081	0.004	0.078	0.080	0.003	0.078	0.99	1.00
0.0	0.45	10	–0.002	0.230	0.236	–0.004	0.221	0.227	–0.003	0.218	0.96	0.99
0.0	0.45	25	–0.001	0.091	0.090	–0.001	0.090	0.089	–0.001	0.090	0.99	1.00
0.0	0.45	50	0.001	0.044	0.044	0.001	0.044	0.044	0.001	0.044	1.00	1.00
0.0	0.90	10	–0.001	0.044	0.043	–0.001	0.042	0.041	–0.001	0.043	0.96	1.03
0.0	0.90	25	0.001	0.016	0.016	0.001	0.016	0.016	0.001	0.016	0.99	1.01
0.0	0.90	50	–0.001	0.008	0.008	–0.001	0.008	0.008	–0.001	0.008	1.00	1.01
0.5	0.00	10	0.507	0.451	0.448	0.508	0.417	0.421	0.505	0.405	0.92	0.97
0.5	0.00	25	0.499	0.168	0.169	0.499	0.163	0.164	0.500	0.161	0.97	0.99
0.5	0.00	50	0.495	0.082	0.084	0.496	0.081	0.082	0.496	0.080	0.98	0.99
0.5	0.45	10	0.497	0.253	0.255	0.497	0.234	0.235	0.499	0.230	0.93	0.98
0.5	0.45	25	0.507	0.094	0.096	0.506	0.090	0.092	0.508	0.089	0.96	0.99
0.5	0.45	50	0.499	0.048	0.047	0.499	0.046	0.046	0.500	0.045	0.97	0.99
0.5	0.90	10	0.499	0.060	0.061	0.500	0.049	0.049	0.508	0.049	0.81	1.01
0.5	0.90	25	0.495	0.022	0.022	0.495	0.019	0.019	0.498	0.019	0.85	1.00
0.5	0.90	50	0.500	0.011	0.011	0.499	0.009	0.009	0.501	0.009	0.87	0.99
1.0	0.00	10	0.998	0.499	0.503	0.995	0.440	0.444	0.996	0.418	0.88	0.95
1.0	0.00	25	0.999	0.187	0.187	0.999	0.171	0.173	0.999	0.164	0.92	0.96
1.0	0.00	50	0.989	0.091	0.092	0.989	0.085	0.086	0.990	0.082	0.93	0.97
1.0	0.45	10	1.000	0.308	0.310	1.003	0.254	0.259	1.008	0.239	0.83	0.94
1.0	0.45	25	1.003	0.113	0.113	1.002	0.101	0.100	1.004	0.096	0.89	0.95
1.0	0.45	50	0.999	0.055	0.055	1.000	0.049	0.049	1.000	0.047	0.90	0.96
1.0	0.90	10	0.994	0.116	0.117	0.996	0.073	0.073	1.013	0.071	0.63	0.97
1.0	0.90	25	1.000	0.038	0.039	1.000	0.027	0.027	1.006	0.026	0.69	0.96
1.0	0.90	50	1.000	0.019	0.019	1.000	0.013	0.013	1.003	0.013	0.72	0.96

Table 3
Mean and Variance of Alternate Effect Size Estimates When δc = 0.2

			d_{ppc1}			d_{ppc2}			d_{ppc3}		Relative Efficiency	
δ_T	ρ	*n*	*M*	Observed Variance	Theoretical Variance	*M*	Observed Variance	Theoretical Variance	*M*	Observed Variance	d_{ppc2}/d_{ppc1}	d_{ppc3}/d_{ppc2}
0.0	0.00	10	–0.195	0.439	0.432	–0.197	0.419	0.414	–0.197	0.412	0.96	0.98
0.0	0.00	25	–0.203	0.165	0.165	–0.203	0.162	0.162	–0.203	0.162	0.98	1.00
0.0	0.00	50	–0.200	0.081	0.081	–0.200	0.080	0.081	–0.199	0.080	0.99	0.99
0.0	0.45	10	–0.201	0.243	0.239	–0.201	0.231	0.228	–0.202	0.229	0.95	0.99
0.0	0.45	25	–0.201	0.092	0.091	–0.201	0.090	0.089	–0.201	0.090	0.98	1.00
0.0	0.45	50	–0.200	0.045	0.045	–0.199	0.045	0.044	–0.199	0.045	0.99	1.00
0.0	0.90	10	–0.198	0.045	0.046	–0.198	0.042	0.043	–0.201	0.043	0.93	1.03
0.0	0.90	25	–0.202	0.017	0.017	–0.201	0.017	0.017	–0.203	0.017	0.96	1.01
0.0	0.90	50	–0.200	0.009	0.009	–0.199	0.008	0.008	–0.200	0.008	0.98	1.00
0.5	0.00	10	0.296	0.438	0.451	0.298	0.410	0.416	0.300	0.403	0.94	0.98
0.5	0.00	25	0.300	0.173	0.170	0.300	0.164	0.163	0.301	0.164	0.95	1.00
0.5	0.00	50	0.296	0.083	0.084	0.296	0.080	0.081	0.296	0.080	0.97	1.00
0.5	0.45	10	0.305	0.253	0.258	0.302	0.227	0.230	0.303	0.224	0.90	0.99
0.5	0.45	25	0.299	0.098	0.097	0.299	0.091	0.090	0.299	0.091	0.94	0.99
0.5	0.45	50	0.302	0.048	0.048	0.302	0.046	0.045	0.302	0.045	0.94	0.99
0.5	0.90	10	0.296	0.063	0.064	0.298	0.045	0.044	0.303	0.046	0.72	1.02
0.5	0.90	25	0.300	0.023	0.023	0.301	0.017	0.017	0.303	0.017	0.75	1.01
0.5	0.90	50	0.301	0.011	0.011	0.301	0.008	0.009	0.302	0.008	0.77	1.00
1.0	0.00	10	0.806	0.512	0.506	0.805	0.438	0.433	0.804	0.419	0.86	0.96
1.0	0.00	25	0.805	0.188	0.188	0.804	0.168	0.169	0.805	0.164	0.89	0.98
1.0	0.00	50	0.802	0.092	0.092	0.803	0.084	0.084	0.803	0.082	0.91	0.98
1.0	0.45	10	0.806	0.308	0.313	0.808	0.241	0.247	0.810	0.233	0.78	0.96
1.0	0.45	25	0.804	0.118	0.114	0.804	0.098	0.096	0.805	0.095	0.83	0.97
1.0	0.45	50	0.799	0.056	0.056	0.797	0.048	0.048	0.798	0.046	0.85	0.97
1.0	0.90	10	0.803	0.123	0.120	0.799	0.063	0.062	0.813	0.062	0.51	0.99
1.0	0.90	25	0.804	0.040	0.040	0.801	0.023	0.023	0.806	0.023	0.58	0.99
1.0	0.90	50	0.800	0.019	0.019	0.800	0.011	0.011	0.803	0.011	0.60	0.98

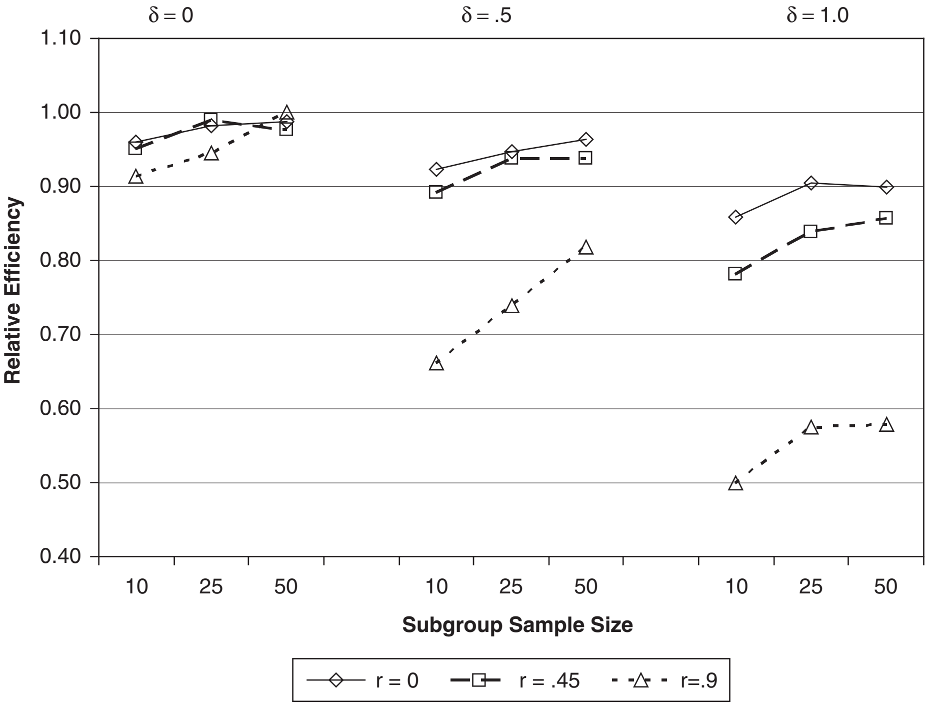
Figure 1
Relative Efficiency of Effect Size Estimates Using Pooled Pretest *SD* Versus Separate Pretest *SD* With No Control Group Change



likely that the pretest scores in the treatment group and both pre- and posttest scores in the control group will reflect the untreated population, and, therefore, homogeneity of variance is reasonable for these conditions. However, the variance of posttest scores in the treatment group might be inflated relative to the other conditions. This pattern has been shown in research on training effectiveness (Carlson & Schmidt, 1999); therefore, it is important to examine the performance of the alternate effect size estimates under this pattern of heterogeneity.

Heterogeneity of variance raises the question of which standard deviation to use in the definition of the population effect size. This study followed the recommendations of Becker (1988) and Carlson and Schmidt (1999) by defining the population effect size using the standard deviation of the untreated population. Because the larger posttest standard deviation in the treatment group results from a Subject × Treatment interaction, it may depend on the magnitude of the treatment effect, which may not be the same across studies. The standard deviation of the untreated population is more likely to be comparable across studies.

Figure 2
Relative Efficiency of Effect Size Estimates Using Pooled Pretest *SD* Versus Separate Pretest *SD* With Control Group δ = 0.2



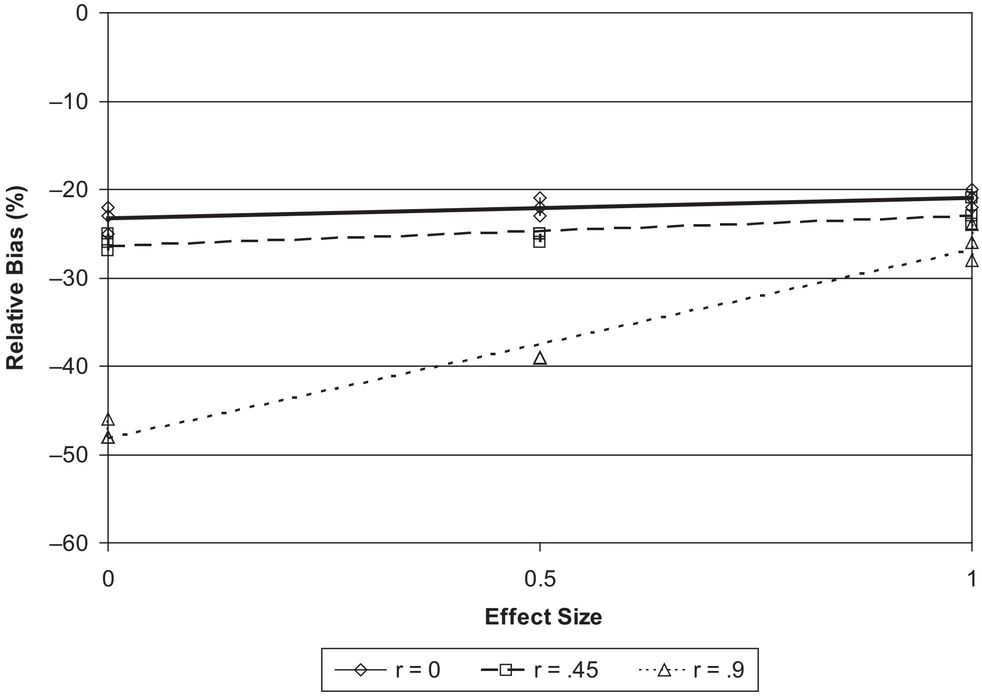
It was expected that heterogeneity of variance would be most problematic for d_{ppc3}. Although homogeneity of variance is assumed by all three estimates, only d_{ppc3} uses the standard deviation of posttest scores when computing the effect size estimate. The inflated posttest standard deviation in the treatment group will tend to increase the pooled standard deviation, resulting in an effect size estimate that is too small. Using only the pretest standard deviations, d_{ppc1} and d_{ppc2} should provide an unbiased estimate of the population effect size, even when posttest scores are heterogeneous.

A Monte Carlo simulation was conducted to examine the performance of the alternate effect size estimates when the variance of posttest scores in the treatment group was inflated. The posttest standard deviation in the treatment groups was set at 1.5 times the standard deviation of the untreated population. Scores in each group were initially generated from a bivariate normal distribution with all means equal to 0, variances equal to 1, and correlation ρ. Then, posttest scores were computed by first multiplying each score by the posttest standard deviation (1.5) and then adding the population standardized mean change, δ_T or δ_C, depending on the group. Otherwise the simulation was the same as the first simulation, except that the standardized mean change in the control group was always $\delta_C = 0.0$. The results of the simulation are presented in Table 4.

Table 4
Mean and Variance of Alternate Effect Size Estimates When δc = 0.0 and Heterogeneous Variance

			d_{ppc1}			d_{ppc2}			d_{ppc3}	
δ_T	ρ	*n*	*M*	Observed Variance	Theoretical Variance	*M*	Observed Variance	Theoretical Variance	*M*	Observed Variance
0.0	0.00	10	–0.013	0.572	0.430	–0.012	0.552	0.413	–0.010	0.417
0.0	0.00	25	0.000	0.208	0.164	0.001	0.205	0.162	0.001	0.156
0.0	0.00	50	0.003	0.107	0.081	0.003	0.107	0.080	0.002	0.081
0.0	0.45	10	–0.006	0.324	0.236	–0.006	0.311	0.227	–0.005	0.237
0.0	0.45	25	–0.001	0.121	0.090	0.000	0.120	0.089	0.000	0.092
0.0	0.45	50	0.001	0.060	0.044	0.001	0.059	0.044	0.000	0.045
0.0	0.90	10	0.001	0.083	0.043	0.002	0.079	0.041	0.002	0.062
0.0	0.90	25	0.002	0.031	0.016	0.002	0.030	0.016	0.002	0.024
0.0	0.90	50	0.000	0.015	0.008	0.000	0.015	0.008	0.000	0.011
0.5	0.00	10	0.508	0.579	0.448	0.507	0.547	0.421	0.442	0.410
0.5	0.00	25	0.501	0.220	0.169	0.502	0.215	0.164	0.438	0.162
0.5	0.00	50	0.499	0.110	0.084	0.499	0.109	0.082	0.436	0.082
0.5	0.45	10	0.508	0.335	0.255	0.507	0.315	0.235	0.446	0.241
0.5	0.45	25	0.501	0.129	0.096	0.500	0.124	0.092	0.438	0.094
0.5	0.45	50	0.501	0.064	0.047	0.501	0.062	0.046	0.438	0.047
0.5	0.90	10	0.501	0.097	0.061	0.501	0.084	0.049	0.446	0.066
0.5	0.90	25	0.499	0.036	0.022	0.498	0.033	0.019	0.438	0.025
0.5	0.90	50	0.498	0.017	0.011	0.497	0.016	0.009	0.436	0.012
1.0	0.00	10	1.004	0.644	0.503	1.006	0.582	0.444	0.880	0.425
1.0	0.00	25	1.000	0.234	0.187	1.001	0.221	0.173	0.874	0.163
1.0	0.00	50	0.999	0.117	0.092	0.997	0.111	0.086	0.873	0.083
1.0	0.45	10	0.997	0.393	0.310	0.998	0.347	0.259	0.879	0.258
1.0	0.45	25	1.002	0.144	0.113	1.003	0.131	0.100	0.878	0.097
1.0	0.45	50	0.997	0.072	0.055	0.997	0.066	0.049	0.871	0.049
1.0	0.90	10	0.996	0.158	0.117	0.995	0.111	0.073	0.885	0.086
1.0	0.90	25	0.997	0.053	0.039	0.998	0.042	0.027	0.877	0.032
1.0	0.90	50	1.000	0.027	0.019	1.000	0.021	0.013	0.876	0.016

Figure 3
Relative Bias in Variance Estimate for Pooled Pretest Method When Variances Are Unequal



As expected, d_{ppc3} underestimated the population effect size, except when $\delta_T = 0$. In general, the mean effect size tended to underestimate the population parameter by about 12%, regardless of the sample size or the magnitude of ρ. Providing mean estimates comparable to those when the homogeneity assumption was satisfied, d_{ppc1} and d_{ppc2} were nearly unbiased.

Although d_{ppc1} and d_{ppc2} were unbiased estimates, introducing heterogeneity of variance had a negative impact on their precision. The sampling variance of both estimates was substantially larger than in the previous simulation. In addition, the inflated sampling variance was systematically larger than the theoretical variance estimated using Equations 16 and 25. As shown in Figure 3, the theoretical variance underestimated the actual variance by between 21% and 48%. More extreme errors occurred for large ρ, particularly when the effect size was small.

It should be noted that individual differences in treatment effectiveness might create other problems beyond variance heterogeneity. Specifically, these differences may also cause the pre-post correlation to differ across treatment and control groups. Under such circumstances, the theoretical variance estimates, which assume a common correlation across groups, may be even less accurate than indicated in this simulation.

Impact on Meta-Analytic Results

To illustrate how the choice of effect size might affect research conclusions, meta-analyses were conducted separately using d_{ppc1} and d_{ppc2} for the five studies in Table 1. Because the meta-analytic procedures require estimates of sampling variance, which is unknown for d_{ppc3}, d_{ppc3} was not included in this illustration.

When comparing meta-analytic methods, it is important to distinguish two types of research questions typically addressed through meta-analysis. One goal of most meta-analyses is to estimate the magnitude of the intervention under investigation. The primary statistic is the mean effect size estimate and the associated confidence interval. A second common goal of meta-analysis is to evaluate the extent to which the effect is consistent across studies and to identify potential moderators of the effect size.

Given a set of k studies, let the estimated effect size from the jth study be indicated by d_{j}. Hedges and Olkin (1985) show that the optimal estimate of the effect size utilizes a weighted mean, where the weights are defined using the inverse of the sampling variance of each effect size,

$$\overline{d} = \frac{\sum_{j=1}^{k} \frac{d_j}{\hat{\sigma}^2(d_j)}}{\sum_{j=1}^{k} \frac{1}{\hat{\sigma}^2(d_j)}}. \tag{26}$$

Alternate definitions of the effect size will lead not only to different values of d for each study but also to different estimates of the sampling variance. Applying this formula to the five studies in the example yields a mean effect size of 1.00 using d_{ppc1} and 0.87 using d_{ppc2} (see Table 5). The sampling variance of the mean effect size can be estimated by

$$VAR(\overline{d}) = \frac{1}{\sum_{j=1}^{k} \frac{1}{\hat{\sigma}^2(d_j)}}. \tag{27}$$

This can be used to compute a 95% confidence interval around the mean effect size using $CI = \bar{d} \pm 1.96\sqrt{VAR(\bar{d})}$, as illustrated in Table 5.

The homogeneity of effect size across studies can be evaluated using the Q test developed by Hedges (1981). The null hypothesis is that a set of k studies each have a common effect size. The test statistics is

$$Q = \sum_{j=1}^{k} \frac{(d_j - \overline{d})^2}{\hat{\sigma}^2(d_j)}. \tag{28}$$

The Q test is evaluated against a chi-square distribution with k-1 df. For both meta-analyses in Table 5, the Q test was nonsignificant. However, the magnitude of the test statistic differed substantially, with d_{ppc1} producing a Q value about twice as large as d_{ppc2}.

Although this example illustrates how the choice of effect size can lead to different meta-analytic results, the simulation results provide a basis for a more systematic examination of the issue. With regard to estimates of the mean effect size, both d_{ppc1} and d_{ppc2} were found

Table 5
Comparison of Meta-Analytic Results for Alternate Definitions of Effect Size (d_{ppc1} and d_{ppc2})

Study	d_{ppc1}	$\hat{\sigma}^2(d_{ppc1})$	d_{ppc2}	$\hat{\sigma}^2(d_{ppc2})$
Blicksenderfer, Cannon-Bowers, and Salas (1997)	0.74	0.14	0.77	0.12
Ivancevich and Smith (1981)	0.95	0.04	0.80	0.04
Katz and Schwebel (1976)	1.81	0.19	1.20	0.13
McGehee and Gardner (1955)	1.15	0.12	1.05	0.07
Miraglia (1963)	0.51	0.21	0.44	0.18
M	1.00		0.87	
95% confidence interval	0.72–1.28		0.63-1.11	
Q test ($df = 4$)	5.24, ns		2.59, ns	

to be unbiased across all conditions examined in this study, meaning that they would provide equally good estimates of the mean effect size. On the other hand, d_{ppc3} was biased under a common violation of the homogeneity of variance assumption (i.e., when posttest variance in the treatment group was inflated). Inflation of the posttest variance resulted in values of d_{ppc3} that tended to underestimate the population effect size.

The magnitude of bias in d_{ppc3} depended on the values of the parameters in the simulation. Based on a review of past meta-analyses on training interventions (Arthur et al., 2003; Carlson & Schmidt, 1999; P. J. Taylor et al., 2005), the most typical conditions for a meta-analyses would be $\delta_T = 0.5$, $\rho = 0.45$, and $n = 25$. Under these conditions, and with a posttest *SD* 1.5 times larger than the pretest *SD* in the treatment group, d_{ppc3} underestimated the population effect size by 12%. Although the difference between effect sizes of .5 and .44 is not likely to change the conclusion about whether an intervention is effective, a difference of this magnitude might be large enough to suggest that one type of intervention is more effective than another.

The smaller sampling variance for d_{ppc2}, relative to d_{ppc1}, will increase the precision of the mean effect size estimate, which will be reflected in a smaller confidence interval. The sampling variance of the mean effect size (Equation 27) is approximately equal to $1/k$ times the average sampling variance of the individual effect sizes. Consequently, the relative efficiency (RE) of the two estimators provides an approximation of the relative size of the variances of the mean effect sizes. Because the confidence interval is computed using the square root of the sampling variance, a confidence interval around d_{ppc2} will be smaller than a confidence interval around d_{ppc1} by a factor of approximately $1-\sqrt{RE}$. Under typical conditions ($\delta_T = 0.5$, $\delta_C = 0.2$, $\rho = 0.45$, and $n = 25$), the relative efficiency of d_{ppc2} to d_{ppc1} was found to be .94, and the corresponding confidence interval would be about 3% smaller using d_{ppc2} than d_{ppc1}.

Although it is always advisable to use the most precise estimator available, the choice of effect size will not make much difference in the precision of the mean effect size estimate. In addition, because the aggregated sample sizes in meta-analysis are typically quite large, confidence intervals on the mean effect size are generally quite small, regardless of how the effect size is defined. Therefore, the choice between d_{ppc1} and d_{ppc2} is unlikely to have much impact on conclusions concerning the magnitude of the treatment effect.

The differential precision of the alternate estimates is more likely to influence conclusions regarding the variability of effect sizes across studies. Methods for evaluating the homogeneity of effect sizes (e.g., Hedges *Q* test or the Hunter-Schmidt 75% rule) are based on a comparison of the observed variance of effect sizes to the variance predicted based on the theoretical estimates of sampling variance. Because the theoretical sampling variance serves as the standard of comparison for these procedures, methods that have smaller sampling variance will have greater sensitivity to variances across studies. The magnitude of this improvement will depend not only on the relative efficiency of the estimators but also on the degree to which there is true variability in effect sizes across studies.

The *Q* test for homogeneity (Equation 28) is approximately equal to *k* times the ratio of the observed variance of effect sizes to the average theoretical variance, where *k* is the number of studies in the meta-analysis. Using an effect size with a smaller theoretical variance will decrease the denominator of Equation 28 by a factor approximately equal to the relative efficiency. However, changing the effect size will also reduce the numerator. The observed variance includes both true variance and sampling variance, and the sampling variance will decrease when a more precise estimate is used. When the null hypothesis is true (i.e., when the true effect size is constant across studies), the value of *Q* will be unaffected by the choice of effect size. However, when there are true differences in the magnitude of the effect size across studies, using a more precise estimate of effect size will decrease the observed variance (the numerator) less than the theoretical variance (the denominator), resulting in a larger value for the *Q* test. In general, the use of d_{ppc2} will lead to larger *Q* values than d_{ppc1}. The difference between the methods will increase with the extent of true variance, but will generally be less than 1/RE.

Under typical conditions ($\delta_T = 0.5$, $\delta_C = 0.2$, $\rho = 0.45$, $n = 25$), the sampling variance does not differ substantially between d_{ppc1} and d_{ppc2} (RE = .94), and, therefore, use of d_{ppc2} would increase the obtained *Q* values only slightly (less than 6%). Although any increase in statistical power would be beneficial, the improvement under many conditions is fairly small. On the other hand, when the pre-post correlation is high ($\rho = 0.9$), a *Q* test based on d_{ppc2} could be substantially larger (up to 30% larger) than a test computed using d_{ppc1}.

When the homogeneity of variance assumption was violated, the sampling variances of both d_{ppc1} and d_{ppc2} were inflated. Thus, the theoretical estimates of variance because of sampling error (which assume homogeneity of variance) tended to underestimate the observed variance. An underestimate of the sampling variance will artificially inflate the *Q* test for homogeneity of variance (Equation 28). Under common conditions ($\delta_T = 0.5$, $\rho = 0.45$, and $n = 25$) and with treatment group posttest *SD* inflated 1.5 times the pretest *SD*, the theoretical estimate of sampling variance for both d_{ppc1} and d_{ppc2} underestimated the observed variance by about 26%. The *Q* test would be inflated by a factor of 1.0/0.26, or 35%. If there were no true differences between studies, this inflation of the *Q* statistic would result in an increased chance of Type I error.

Conclusion

The research base on the effectiveness of organizational interventions includes studies with a variety of research designs. The relative strengths and weaknesses of alternative designs are widely recognized when conducting primary research studies, but these differences are often overlooked when conducting a meta-analysis.

The PPC design provides an effective framework for program evaluation research. Compared to studies with only posttest measures or with no control group, the PPC design can yield increased precision in the estimation of treatment effects and can provide superior control over threats to internal validity. By applying the procedures outlined in this article, results from the PPC design can be incorporated into a meta-analysis in a way that takes full advantage of the strengths of this design.

The choice of an effect size depends on a number of considerations, such as bias, precision, having a known sampling distribution, and robustness to violations of assumptions. For the PPC design, three effect size estimates have been recommended in the literature. Becker (1988) suggested a measure of effect size using separate estimates of the pretest standard deviation in the treatment and control groups (d_{ppc1}). Carlson and Schmidt (1999) used an effect size based on the pooled pretest standard deviation (d_{ppc2}). Others (Dunlap et al., 1996; M. J. Taylor & White, 1992) have recommend pooling standard deviations across both pretest and posttest scores (d_{ppc3}).

The results support d_{ppc2} as the best choice because it provides an unbiased estimate of the population effect size and has a known sampling variance that is smaller than the sampling variance of d_{ppc1}. Pooling standard deviations across both pre- and posttest scores, as in d_{ppc3}, is problematic. Because pre- and posttest scores are not independent, the sampling variance of d_{ppc3} is more complex than the other two estimates and is currently unknown. In addition, d_{ppc3} demonstrated very little improvement over d_{ppc2} in terms of sampling variance. Because meta-analytic procedures require estimates of sampling variance, the use of d_{ppc3} is not recommended.

An additional limitation of d_{ppc3} is its tendency to underestimate the true effect size when the homogeneity of variance assumption was violated. When the intervention is not equally effective for all individuals, some individuals will tend to show greater change than others. This will tend to inflate the variance of posttest scores for those who have received the intervention. Measures of effect size that are based on the posttest standard deviations, such as d_{ppc3}, will have a downward bias. It should be noted that this bias in not particular to the PPC design but would occur any time posttest standard deviations are used. Therefore, the common practice of defining the effect size based only on posttest scores will produce a similar bias (Carlson & Schmidt, 1999).

Violation of the homogeneity of variance assumption also poses a problem when assessing the homogeneity of effect sizes across studies, and this bias is likely to occur regardless of how the effect size is defined. Formulas for estimating the sampling variance have been derived under the assumption that variances are equal across conditions. When variances were heterogeneous, sampling variance formulas for both d_{ppc1} and d_{ppc2} underestimated the true sampling variance. Because the estimated variances are too small, tests for homogeneity of effect size will be inflated. Consequently, a meta-analyst might incorrectly conclude that effect sizes are heterogeneous and potentially search for moderators of the effect size where none exist. Additional work is therefore needed to better estimate the sampling variance of effect sizes when there are individual differences in treatment effectiveness, not only for the PPC design but also for estimates of effect size from other designs.

References

Arthur, W., Bennett, W., Edens, P. S., & Bell, S. T. (2003). Effectiveness of training in organizations: A meta-analysis of design and evaluation features. *Journal of Applied Psychology, 88*, 234-245.

Arthur, W., Bennett, W., Stanush, P. L., & McNelly, T. L. (1998). Factors that influence skill decay and retention: A quantitative review and analysis. *Human Performance, 11*, 57-101.

Arvey, R. D., & Cole, D. A. (1989). Evaluating change due to training. In I. L. Goldstein (Ed.), *Training and development in organizations* (pp. 89-117). San Francisco: Jossey-Bass.

Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology, 41*, 257-278.

Blicksenderfer, E., Cannon-Bowers, J. A., & Salas, E. (1997, April). *Training teams to self-correct: An empirical investigation.* Paper presented at the 12th Annual Meeting of the Society for Industrial and Organizational Psychology, St. Louis, MO.

Bryk, A. S., & Raudenbush, S. W. (1988). Heterogeneity of variance in experimental studies: A challenge to conventional interpretations. *Psychological Bulletin, 104*, 396-404.

Callahan, J. S., Kiker, D. S., & Cross, T. (2003). Does method matter? A meta-analysis of the effects of training method on older learner training performance. *Journal of Management, 29*, 663-680.

Carlson, K. D., & Schmidt, F. L. (1999). Impact of experimental design on effect size: Findings from the research literature on training. *Journal of Applied Psychology, 84*, 851-862.

Collins, D. B., & Holton, E. F. (2004). The effectiveness of managerial leadership development programs: A meta-analysis of studies from 1982 to 2001. *Human Resource Development Quarterly, 15*, 217-248.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton Mifflin.

Cortina, J. M., & Nouri, H. (2000). *Effect size for ANOVA designs.* Thousand Oaks, CA: Sage.

Driskell, J. E., Willis, R. P., & Copper, C., (1992). Effect of overlearning on retention. *Journal of Applied Psychology, 77*, 615-622.

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods, 1*, 170-177.

Eden, D., Geller, D., Gewirtz, A., Gordon-Terner, R., Inbar, I., Liberman, M., et al. (2000). Implanting pygmalion leadership style through workshop training: Seven field experiments. *Leadership Quarterly, 11*, 171-210.

Goldstein, I. L., & Ford, J. K. (2002). *Training in organizations* (4th ed.). Belmont, CA: Wadsworth.

Guzzo, R. A., Jette, R. D., & Katzell, R. A. (1985). The effects of psychologically based interventions programs on worker productivity: A meta-analysis. *Personnel Psychology, 38*, 275-291.

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

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

Hunter, J. E., & Schmidt, F. L. (2004). *Method of meta-analysis: Correcting error and bias in research findings* (2nd ed). Thousand Oaks, CA: Sage.

International Mathematical and Statistical Library. (1984). *User's manual: IMSL Library, problem-solving software system for mathematical and statistical FORTRAN programming* (Vol. 3, 9.2 ed.). Houston, TX: Author.

Ivancevich, J. M., & Smith, S. V. (1981). Goal setting interview skills training: Simulated and on the job analysis. *Journal of Applied Psychology, 66*, 697-705.

Johnson, N. L., & Kotz, S. (1970). *Continuous univariate distributions*. New York: John Wiley.

Katz, S. I., & Schwebel, A. I. (1976). The transfer of laboratory training. *Small Group Behavior, 7*, 271-285.

Kish, L. (1965). *Survey sampling*. New York: John Wiley.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*, 254-284.

Maris, E. (1998). Covariance adjustment versus gain scores—Revisited. *Psychological Methods, 3*, 309-327.

McGehee, W., & Gardner, L. E. (1955). Supervisory training and attitude change. *Personnel Psychology, 8*, 449-460.

McNatt, D. B. (2000). Ancient Pygmalion joins contemporary management: A meta-analysis of the result. *Journal of Applied Psychology, 85*, 314-322.

Miraglia, J. F. (1963). *An experiential study of the effects of communication training upon perceived job performance of nursing supervisors in two urban hospitals.* Unpublished doctoral dissertation, Purdue University, West Lafayette, Indiana.

Morris, M. A., & Robie, C. (2001). A meta-analysis of the effects of cross-cultural training on expatriate performance and adjustment. *International Journal of Training & Development, 5*, 112-125.

Morris, S. B. (2000). Distribution of the standardized mean change effect size for meta-analysis on repeated measures. *British Journal of Mathematical and Statistical Psychology, 53,* 17-29.

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.

Morrow, C. C., Jarrett, M. Q., & Rupinski, M. T. (1997). An investigation of the effect and economic utility of corporate-wide training. *Personnel Psychology, 50*, 91-119.

Neuman, G. A., Edwards, J. E., & Raju, N. S. (1989). Organizational development interventions: A meta-analysis of their effects on satisfaction and other attitudes. *Personnel Psychology, 42*, 461-489.

Niemiec, R. P., Sikorski, M. F., Clark, G., & Walberg, H. J. (1992). Effects of management education: A quantitative synthesis. *Evaluation & Program Planning, 15*, 297-302.

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

Quinones, M. A., & Tonidandel, S. (2003). Conducting training evaluation. In J. E. Edwards, J. C. Scott, & N. S. Raju (Eds.), *The human resources program-evaluation handbook* (pp. 225-243). Thousand Oaks, CA: Sage.

Roberts, D. R., & Robertson, P. J. (1992). Positive-findings bias, and measuring methodological rigor, in evaluations of organization development. *Journal of Applied Psychology, 77*, 918-925.

Robson, L. S., Shannon, H. S., Goldenhar, L. M., & Hale, A. R. (2001). *Guide to evaluating the effectiveness of strategies for preventing work injuries* (DHHS NIOSH 2001-119). Washington, DC: Department of Health & Human Services, Public Health Service, Centers for Disease Control and Prevention, National Institute for Occupational Safety and Health.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231-244). New York: Russell Sage.

Scott, G., Leritz, L. E., & Mumford, M. D. (2004). The effectiveness of creativity training: A quantitative review. *Creativity Research Journal, 16*, 361-388.

Smither, J. W., London, M., & Reilly, R. R. (2005). Does performance improve following multisource feedback? A theoretical model, meta-analysis and review of empirical findings. *Personnel Psychology, 58*, 33-66.

Taylor, M. J., & White, K. R. (1992). An evaluation of alternative methods for computing standardized mean difference effect size. *Journal of Experimental Education, 61*, 63-72.

Taylor, P. J., Russ-Eft, D. F., & Chan, D. W. L. (2005). A meta-analytic review of behavior modeling training. *Journal of Applied Psychology, 90*, 692-709.

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational & Organizational Psychology, 67*, 189-205.

Scott B. Morris, PhD, is an associate professor in the Institute of Psychology at Illinois Institute of Technology. His research interests include employee selection, the evaluation of employment discrimination, and statistical issues in meta-analysis.