Psychological Methods
2011, Vol. 16, No. 2, 166–178

1082-989X/11/$12.00 DOI: 10.1037/a0023355

Using Analysis of Covariance (ANCOVA) With Fallible Covariates

Steven Andrew Culpepper
University of Colorado Denver

Herman Aguinis
Kelley School of Business, Indiana University

Analysis of covariance (ANCOVA) is used widely in psychological research implementing nonexperimental designs. However, when covariates are fallible (i.e., measured with error), which is the norm, researchers must choose from among 3 inadequate courses of action: (a) know that the assumption that covariates are perfectly reliable is violated but use ANCOVA anyway (and, most likely, report misleading results); (b) attempt to employ 1 of several measurement error models with the understanding that no research has examined their relative performance and with the added practical difficulty that several of these models are not available in commonly used statistical software; or (c) not use ANCOVA at all. First, we discuss analytic evidence to explain why using ANCOVA with fallible covariates produces bias and a systematic inflation of Type I error rates that may lead to the incorrect conclusion that treatment effects exist. Second, to provide a solution for this problem, we conduct 2 Monte Carlo studies to compare 4 existing approaches for adjusting treatment effects in the presence of covariate measurement error: errors-in-variables (EIV; Warren, White, & Fuller, 1974), Lord's (1960) method, Raaijmakers and Pieters's (1987) method (R&P), and structural equation modeling methods proposed by Sörbom (1978) and Hayduk (1996). Results show that EIV models are superior in terms of parameter accuracy, statistical power, and keeping Type I error close to the nominal value. Finally, we offer a program written in R that performs all needed computations for implementing EIV models so that ANCOVA can be used to obtain accurate results even when covariates are measured with error.

Keywords: measurement error, analysis of covariance, structural equation modeling, research design

Researchers use analysis of covariance (ANCOVA) to answer research questions, test theories, and evaluate treatments while implementing nonexperimental research designs. Adjusting treatment effects for confounding variables in nonexperimental designs is important for accurately determining the value and practical usefulness of treatments, interventions, and programs (Arvey, Cole, Hazucha, & Hartanto, 1985; Grant & Wall, 2009; Harwell, 2003; Maris, 1998; Schafer & Kang, 2008).

Equation 1 shows an ANCOVA model with one treatment effect, α_j, and a single covariate, x_{ij}, centered by the average covariate value, $\bar{x}$:

$$y_{ij} = \mu + \alpha_j + \beta(x_{ij} - \bar{x}) + e_{ij} \quad (1)$$

where y_{ij} is the dependent variable of interest for subject i in group j, μ represents the grand mean, e_{ij} is a residual, β measures the effect of x_{ij} on y_{ij}, and using effect coding requires $\Sigma_j \alpha_j = 0$. In this article, we consider the case where $j = 0$ for the control group and $j = 1$ for the treatment group and $y_{ij} \sim N[\mu + \alpha_j + \beta(x_{ij} - \bar{x}), \sigma^2]$. Additionally, let $\mu_{1x} - \mu_{0x}$ represent the degree of covariate mean differences, or nonequivalence, between the treatment and control groups on x_{ij}, and let ρ_{xx} denote the reliability of x_{ij}. In general, ANCOVA makes the following assumptions: e_{ij} are identically and independently normally distributed; the slope, β, is equal across treatment and control groups; the relationship between y_{ij} and x_{ij} is linear conditioned on group membership (note that a more general polynomial function of x_{ij} could be modeled as well, as long as the shape of the curve is the same across groups); and homogeneity of variance is satisfied across groups.

Another important assumption of ANCOVA is that covariates are measured without error. In fact, controlling for fallible covariates leads to biased treatment effects (Bartlett, 1949; Cochran, 1968; Elashoff, 1969; Fuller, 1987; Kahneman, 1965; Linn & Werts, 1971; Lord, 1960; Madansky, 1959; Porter & Chibucos, 1975; Porter & Raudenbush, 1987; Raaijmakers & Pieters, 1987; Ree & Carretta, 2006; Stanley & Robinson, 1990). Note that covariate measurement error is only a problem for nonexperimental designs with groups that differ in average covariate values. More precisely, covariate measurement error (i.e., $\rho_{xx} < 1$) coupled with group average differences on the covariate (i.e., $\mu_{1x} - \mu_{0x} \neq 0$), which arises in nonexperimental designs (Porter & Raudenbush, 1987), leads to biased treatment effects.

Appendix A includes a derivation of the following equation for computing the exact treatment effect bias when x_{ij} is fallible and the null hypothesis of no treatment effect (i.e., H_o: $\alpha_j = 0$) is true:

$$\Delta R^2_\alpha = \frac{\rho^2_{xy}\rho_{yy}(\mu_{1x} - \mu_{0x})^2 p(1-p)(1-\rho_{xx})^2}{\sigma^2_x - \rho_{xx}(\mu_{1x} - \mu_{0x})^2 p(1-p)}. \quad (2)$$

Specifically, Equation 2 represents the change in R^2 associated with the null hypothesis of no treatment effect after controlling for the covariate when H_0 is true. Also, note in Equation 2 that ρ_{xy} is

This article was published Online First April 25, 2011.

Steven Andrew Culpepper, Department of Mathematical and Statistical Sciences, University of Colorado Denver; Herman Aguinis, Department of Management and Entrepreneurship, Kelley School of Business, Indiana University.

Correspondence concerning this article should be addressed to Steven Andrew Culpepper, Department of Mathematical and Statistical Sciences, University of Colorado Denver, Campus Box 170, P.O. Box 173364, Denver, CO 80217-3364. E-mail: steve.culpepper@ucdenver.edu

the true score correlation between x_{ij} and y_{ij}, σ_x^2 is the variance of x_{ij}, p is the proportion of subjects in the treatment group, and ρ_{yy} is the reliability of the dependent variable scores. Equation 2 shows that ΔR_α^2 will be unbiased (i.e., $\Delta R_\alpha^2 = 0$) when either $\rho_{xx} = 1$ or $\mu_{1x} - \mu_{0x} = 0$ and reaffirms concerns about employing ANCOVA in nonexperimental settings when x_{ij} is measured with error and groups differ in covariate averages. Stated differently, testing the null hypothesis of no treatment effect with standard F critical values (i.e., F^*) is inappropriate because F^* does not account for the biased effects when $\rho_{xx} < 1$ and $\mu_{1x} - \mu_{0x} \neq 0$ (Raaijmakers & Pieters, 1987). Moreover, using F^* to test treatment effects will frequently lead to incorrect statistical inferences and inflated Type I error rates, which can lead to incorrect substantive conclusions such as concluding that a certain treatment works when it actually may not. Appendix A also includes expressions using Equation 2 to compute the real Type I error rate when $\rho_{xx} < 1$ and $\mu_{1x} - \mu_{0x} \neq 0$.

We used equations in Appendix A to create Figure 1 to show the exact degree of Type I error inflation across a set of illustrative conditions. In Figure 1, a value of .05 represents a situation where α_r (i.e., real Type I error rate) equals α (i.e., nominal Type I error rate). Figure 1 includes four panels with different values of ρ_{xy} and illustrates the degree of inflated Type I error rates as a function of $\mu_{1x} - \mu_{0x}$ and ρ_{xx} and holding the sample size constant at 500. All four panels in Figure 1 show that Type I error rates are severely inflated as ρ_{xy} and $\mu_{1x} - \mu_{0x}$ increase and ρ_{xx} decreases. For instance, Panel B shows that Type I error rates are nearly four times larger than the nominal level when $\rho_{xy} = 0.5$, $\mu_{1x} - \mu_{0x} = 0.5$, and $\rho_{xx} = 0.7$. The problem becomes even more severe for larger values of ρ_{xy}. For instance, as shown in Panel D, even a small $\mu_{1x} - \mu_{0x}$ value and small amounts of covariate measurement error distort Type I errors when $\rho_{xy} = 0.9$.

Previous research has proposed methods for correcting biased treatment effects for covariate measurement error in nonexperimental designs, and at least four methods have been developed in the statistics and econometrics literatures: errors-in-variables (EIV; Warren et al., 1974), Lord's (1960) method, Raaijmakers and Pieters's (1987) method (R&P), and structural equation modeling (SEM) methods proposed by Sörbom (1978) and Hayduk (1996). However, we are not aware of any research that has evaluated the relative merits of these approaches. Therefore, researchers interested in using ANCOVA do not have guidelines regarding which approach works best and under what conditions. Accordingly, in the present study we implement Monte Carlo simulations to evaluate the relative performance of existing approaches for adjusting treatment effects in nonexperimental designs when the covariate is fallible.

The remainder of our article is organized into four primary sections. The first section describes four competing methods for addressing covariate measurement error. The second and third sections describe results from two Monte Carlo simulations that assess the relative accuracy (i.e., bias in estimating treatment effects), statistical power, and Type I error rates of these competing methods. Specifically, the first simulation compares the performance of the EIV, Lord method, R&P method, and sparse SEM models (Hayduk, 1996) in cases where a single measure (i.e., observed indicator) of the covariate and dependent variable are available. The second simulation study compares the EIV and the SEM (i.e., Sörbom, 1978) approaches in cases where multiple measures are available for the covariate and dependent variable.

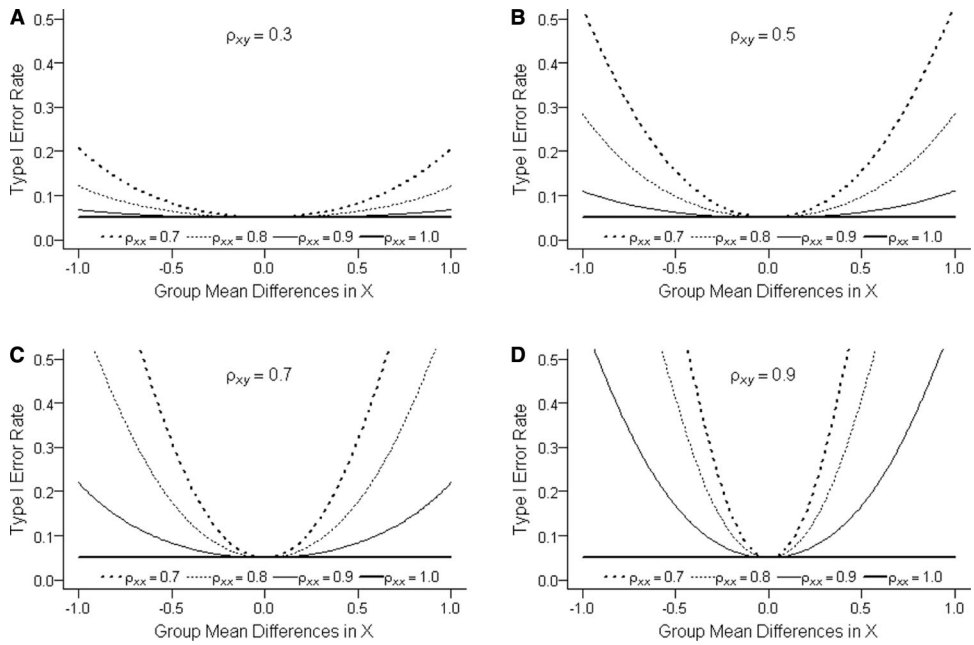

Figure 1. Effects of group mean covariate differences ($\mu_{1x} - \mu_{0x}$), covariate reliability (ρ_{xx}), and covariate correlation with dependent variable (ρ_{xy}) on real Type I error rate for the null hypothesis of no treatment effect for a sample size of 500, nominal rejection level of 0.05, and equal proportions of subjects in treatment and control groups ($p = .5$). Real Type I error rates were computed using equations in Appendix A.

The last section includes a discussion of our results and implications for research and practice.

Summary of Methods for Addressing Biased ANCOVA Results

Researchers have three inadequate options to deal with covariate measurement error when implementing nonexperimental designs: (a) pretend the problem does not exist, use ordinary least squares (OLS) regression, and hope that results will be unbiased; (b) employ a correction method without knowing which method leads to more accurate results under which conditions; or (c) as some have suggested (Porter & Raudenbush, 1987; Wicherts, 2005), not use ANCOVA at all. In this section, we discuss four approaches that have been proposed in the context of adjusting ANCOVA effects in the presence of covariate measurement error. Specifically, we address EIV models (Fuller, 1980, 1987; Fuller & Hidiroglou, 1978; Warren et al., 1974), Lord's (1960) method, the functional R&P method described by Raaijmakers and Pieters (1987), and SEM methods proposed by Sörbom (1978) and Hayduk (1996).

EIV Models

EIV models were developed and popularized by Fuller (1980, 1987). The attenuated OLS unstandardized coefficients are defined by $\mathbf{b} = \mathbf{\Sigma}_{\mathbf{XX}}^{-1}\mathbf{\Sigma}_{\mathbf{XY}}$, where $\mathbf{\Sigma}_{\mathbf{XX}}$ is the estimated covariance matrix among the independent variables and $\mathbf{\Sigma}_{\mathbf{XY}}$ is a vector of covariances between the predictors and dependent variable. The disattenuated EIV coefficients are estimated by $\tilde{\boldsymbol{\beta}} = \mathbf{\Sigma}_{\boldsymbol{xx}}^{-1}\mathbf{\Sigma}_{\mathbf{XY}}$, where $\mathbf{\Sigma}_{\boldsymbol{xx}}$ is the corrected covariance matrix among the independent variables defined by $\mathbf{\Sigma}_{\boldsymbol{xx}} = \mathbf{\Sigma}_{\mathbf{XX}} - [1 - (K + 1)/n]\mathbf{S}_{\mathbf{uu}}$, and $\mathbf{S}_{\mathbf{uu}}$ is the error covariance matrix and K is the total number of predictors. If an estimate for the reliability of x_i is available (the subscript j is dropped from x_{ij} and y_{ij} in the remainder of the article for simplicity), the error covariance matrix for the model in Equation 1 is defined by

$$\mathbf{S}_{\mathbf{uu}} = \begin{bmatrix} (1 - \rho_{xx})\sigma_x^2 & 0 \\ 0 & 0 \end{bmatrix}. \tag{3}$$

Fuller (1987) noted that the variance–covariance matrix of the disattenuated effects (i.e., $\mathbf{\Sigma}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}}$) is defined by

$$\mathbf{\Sigma}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}} = \frac{s_v^2}{n}\mathbf{\Sigma}_{\boldsymbol{xx}}^{-1} + \frac{1}{n}\mathbf{\Sigma}_{\boldsymbol{xx}}^{-1}(\mathbf{S}_{\mathbf{uu}}s_v^2 + \mathbf{S}_{\mathbf{uu}}\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}'\mathbf{S}_{\mathbf{uu}} + 2\hat{\mathbf{R}})\mathbf{\Sigma}_{\boldsymbol{xx}}^{-1}, \tag{4}$$

where $s_v^2 = (\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}})/(n - K)$, which is a measure of the conditional variance of the dependent variable, and $\hat{\mathbf{R}} = \text{diag}(\tilde{\boldsymbol{\beta}}'\mathbf{S}_{\mathbf{uu}} \odot \tilde{\boldsymbol{\beta}}'\mathbf{S}_{\mathbf{uu}})$, where $\odot$ is a Hadamard operator that represents elementwise multiplication.

The Lord Model

Lord (1960) developed a large-sample method for correcting ANCOVA treatment effects for covariate measurement error. Specifically, let W_{xx}, W_{yy}, and W_{xy} represent the pooled within-group sums of squares and cross-products. The Lord model assumes researchers have an estimate of covariate reliability derived from parallel measurements. In this case, Lord wrote the error variance of x_i as $\sigma_\delta^2 = n^{-1}(1 - \rho_{xx})W_{xx}$ and the true score variance of x_i as $S_T^2 = n^{-1}\rho W_{xx}$, where ρ (see Lord, 1960, for more details about computing ρ) is the pooled within-group reliability coefficient. Lord derived approximations for the disattenuated slope coefficient (β_{Lord}) and treatment effect (α_{Lord}) as follows:

$$\beta_{\text{Lord}} = \frac{W_{xy}}{nS_T^2\left[1 + \dfrac{2S_T^2\sigma_\delta^2}{n(S_T^2 + \sigma_\delta^2)^2}\right]} \text{ and}$$

$$\alpha_{\text{Lord}} = (\bar{Y}_1 - \bar{Y}_0) - \beta_{\text{Lord}}(\bar{x}_1 - \bar{x}_0). \tag{5}$$

Lord also defined the error variance of Y_i as $\sigma_\varepsilon^2 = n^{-1}W_{xx} - \beta_{\text{Lord}}^2 S_T^2$ and approximated the variance of the slope and treatment effect as follows:

$$\sigma^2\{\beta_{\text{Lord}}\} = \frac{S_T^2(\sigma_\varepsilon^2 + \beta_{\text{Lord}}^2\sigma_\delta^2) + \sigma_\varepsilon^2\sigma_\delta^2}{nS_T^4} - \frac{2\beta_{\text{Lord}}^2\sigma_\delta^4}{n(S_T^2 + \sigma_\delta^2)^2} \text{ and}$$

$$\sigma^2\{\alpha_{\text{Lord}}\} = \frac{(\sigma_\delta^2 + \beta_{\text{Lord}}^2\sigma_\delta^2)n}{n_1 n_2} + \sigma^2\{\beta_{\text{Lord}}\}(\bar{x}_1 - \bar{x}_0)^2. \tag{6}$$

Lord's procedure uses a first-order approximation of the disattenuated estimates and standard errors.

R&P Model

Raaijmakers and Pieters (1987) described a correction method that is a functional measurement error model (Gleser, 1981; Moran, 1971) that does not assume a distribution for error variances. Their R&P model is related to orthogonal regression and/or total least squares in the statistics literature (Carroll & Ruppert, 1996; DeGracie & Fuller, 1972). Raaijmakers and Pieters noted that their model is more restrictive than the Lord model because the functional model assumes that the measurement error variance for x_i is equivalent to the variance of the error in Y_i. It is important to note that the assumption in the functional model is violated every time the reliability of x_i differs from the proportion of variance accounted for in the dependent variable, Y_i. Consequently, we expect the functional method to yield the least accurate corrections of the models we evaluate. Although the R&P method has not been examined empirically, we hypothesize that the R&P method will be relatively more biased as the assumption of equal variances in x_i and Y_i deviate, which occurs in our simulation as the covariate becomes more reliable (i.e., ρ_{xx} increases) and the treatment effect size increases.

In the R&P model, the error variance is defined as

$$\sigma_\varepsilon^2 = \frac{\beta_{\text{RP}}^2 W_{xx} - 2\beta_{\text{RP}}W_{xy} + W_{yy}}{2n(1 + \beta_{\text{RP}}^2)}, \tag{7}$$

where the formulas for the disattenuated slope coefficient and treatment effect are

$$\beta = \frac{W_{yy} - W_{xx} + \sqrt{(W_{yy} - W_{xx})^2 + 4W_{xy}^2}}{2W_{xy}},$$

$$\beta_{RP} = \frac{\beta}{1 + \dfrac{2\sigma_\varepsilon^2[(1 + \beta^2) + 2\sigma_\varepsilon^2]}{n(1 + \beta^2)S_T^4}},$$

$$\alpha_{RP} = (\bar{Y}_1\bar{Y}) - \beta_{RP}(\bar{x}_1 - \bar{x}), \text{ and} \quad (8)$$

$$S_T^4 = n^{-1}W_{xx} - 2\sigma_\varepsilon^2.$$

Approximate estimates for the variances of the slope and treatment effects are

$$\sigma^2\{\beta_{RP}\} = \frac{\sigma_\varepsilon^2[(1 + \beta_{RP}^2)S_T^2 + \sigma_\varepsilon^2]}{nS_T^4}$$

$$\sigma^2\{\alpha_{RP}\} = n^{-1}\sigma_\varepsilon^2(1 + \beta_{RP}^2) + 0.25\sigma^2\{\beta_{RP}\}(\bar{x}_1 - \bar{x}_0)^2. \quad (9)$$

SEM Approaches: Sörbom's (1978) Method and Hayduk's (1996) Sparse SEM Model

Sörbom (1978) developed a multigroup SEM to adjust ANCOVA treatment effects for covariate and dependent variable measurement error. Specifically, let g index the gth group and $\mathbf{y}^{(g)}$ and $\mathbf{x}^{(g)}$ represent matrices of observed measures of latent variables $\eta^{(g)}$ and $\xi^{(g)}$. Sörbom discussed an approach for estimating the disattenuated effects as

$$\mathbf{y}^{(g)} = \boldsymbol{\mu}_y + \boldsymbol{\Lambda}_y\eta^{(g)} + \boldsymbol{\varepsilon}_y^{(g)},$$

$$\mathbf{x}^{(g)} = \boldsymbol{\mu}_x + \boldsymbol{\Lambda}_x\xi^{(g)} + \boldsymbol{\varepsilon}_x^{(g)}, \text{ and}$$

$$\eta^{(g)} = \alpha^{(g)} + \Gamma^{(g)}\xi^{(g)} + \zeta^{(g)}, \quad (10)$$

where the equations for $\mathbf{y}^{(g)}$ and $\mathbf{x}^{(g)}$ represent measurement models where $\boldsymbol{\mu}_y$ and $\boldsymbol{\mu}_x$ are vectors of means and $\boldsymbol{\Lambda}_y$ and $\boldsymbol{\Lambda}_x$ are loadings. Additionally, $\alpha^{(g)}$ is the adjusted mean for group g, $\Gamma^{(g)}$ is the relationship between the latent covariate and dependent variable for group g, and $\zeta^{(g)}$ is the error term for group g. Sörbom noted that in the case of two groups, the treatment effect is computed as $\alpha^{(1)} - \alpha^{(2)}$. Sörbom also presented equations for finding maximum likelihood estimates and noted that standard SEM software can be used to estimate the effects in Equation 10.

Hayduk (1996) discussed the use of sparse SEM models that include single rather than multiple indicators. Sparse SEM models require researchers to fix parameters in the measurement model. In Study 1, we estimated Sörbom's (1978) model as described in existing SEM software documentation (Arbuckle, 2008) with fixed parameters in the measurement model to understand the performance of sparse SEM models for estimating treatment effects. Specifically, we used the simulated covariate reliability to fix the covariate factor loading, or theta, as described by Hayduk.

Previous Comparative Research on the EIV, OLS, Lord Method, R&P, and SEM

Previous research has used simulations to understand the performance of the OLS for conducting ANCOVA. For example, Overall and Woodward (1977) and Cappelleri, Trochim, Stanley, and Reichardt (1991) conducted several simulations to understand the role of fallible covariates on estimated treatment effects. Overall and Woodward and Cappelleri et al. found inflated Type I error rates for tests of treatment effects and substantial bias in the OLS estimated treatment effect when groups were not equivalent ($\mu_{1x} - \mu_{0x} \neq 0$), and the covariate was measured with error ($\rho_{xx} < 1$).

Some previous research studied the relative performance of the R&P (i.e., orthogonal regression), SEM, and OLS, as well. For example, previous research found evidence to prefer the R&P over the OLS in linear and nonlinear models (Boggs, Spiegelman, Donaldson, & Schnabel, 1988) and models that include the EIV (Ketellapper, 1983). Whereas previous research examined orthogonal regression, we are not aware of any research that specifically studied the viability of R&P for estimating treatment effects. Similarly, Cribbie and Jamieson (2000) examined the relative bias of OLS and SEM treatment effect estimates but did not compare these methods with other estimators, such as the Lord method, R&P, or EIV. Moreover, we were unable to find any simulations comparing the OLS with the Lord method. Consequently, the studies described in the present article are the most extensive and comprehensive simulations on ANCOVA and measurement error in terms of the number of competing models and parameters investigated.

Study 1

Our first Monte Carlo simulation examined the effects of covariate measurement error on the accuracy of the OLS regression, EIV, Lord model, R&P model, and sparse SEM model. Factors manipulated included sample size, treatment effect size, covariate reliability, group mean differences on the covariate, and proportion of the sample in the treatment group. We investigated a wide range of values for each of these factors and their effects on the bias of treatment effects, statistical power, and Type I error rates. The parameter values included in our simulation were specifically chosen to include the range of values observed by applied researchers. Next, we describe the manipulated parameters and their values, data generation procedures, dependent variables, key accuracy checks, and simulation results.

Manipulated Parameters and Data Generation Procedures

The simulation estimated 3,500 unique combinations of the parameter values with 5,000 replications for each combination. Table 1 includes the values for each manipulated parameter, including four values of ρ_{xx} (ranging from .6 to .9) and five values of n (ranging from 100 to 1,000). The unique effect of the treatment beyond the covariate (denoted by $\Delta\psi_G^2$) was manipulated and took on seven different values. That is, $\Delta\psi_G^2$ equaled 0 to .03 in increments of .005 and represents the change in R^2 associated with adding the treatment effect to a model with the covariate. The simulation also manipulated the proportion of subjects in the treatment group (p) and true group mean differences on X_i ($\mu_{1X} - \mu_{0X}$). Table 1 shows that we included five values of p (ranging from .1 to .9) and five values of $\mu_{1X} - \mu_{0X}$ (ranging from –1 to 1).

We used the following equation to generate observed covariate scores (X_i) with true group mean differences between the control and treatment groups represented by $\mu_{1X} - \mu_{0X}$:

Table 1
Summary of Simulation Parameters and Parameter Values

Parameter	Study 1 values	Study 2 values
$\Delta\psi_G^2$	0, .005, .010, .015, .020, .025, .030	0, .005, .010, .015, .020, .025, .030
ρ_{xx}	.6, .7, .8, .9	
n	100, 250, 500, 750, 1,000	250, 500, 750
$\mu_{1X} - \mu_{0X}$	−1.0, −.5, 0, .5, 1.0	−1.0, −.5, 0, .5, 1.0
p	.1, .3, .5, .7, .9	.1, .3, .5, .7, .9
λ_x		.3, .4, .5, .6
λ_y		.3, .4, .5, .6
K		2, 4, 6, 8
L		2, 4, 6, 8

Note. In Study 1, ρ_{yy} (i.e., criterion reliability) was held constant at 1.0 (ρ_{yy} was determined by λ_y and L in Study 2). In both Studies 1 and 2, ρ_{xy} (i.e., covariate–dependent variable correlation) was held constant at .5. Study 1 included 3,500 unique permutations of parameter values that were each replicated 5,000 times. Study 2 included 134,400 unique permutations that were replicated 1,000 times. $\Delta\psi_G^2$ = unique effect of the treatment beyond the covariate; ρ_{xx} = covariate reliability; $\mu_{1X} - \mu_{0X}$ = average difference in standard deviation units between treatment group and control group (i.e., the predictor, X_i, was standardized with a mean of zero and variance of one); p = proportion of the sample in the treatment group; λ_x and λ_y = loadings for observed measures of X_i and Y_i, respectively; K and L = number of observed measures (i.e., indicators) of X_i and Y_i, respectively.

$$X_i = (\mu_{1X} - \mu_{0X})(G_i - p) + \sqrt{1 - \rho_{XG}^2}\, e_{Xi}, \tag{11}$$

where ρ_{XG} is the point biserial correlation between X_i and G_i (where G_i is 1 for the treatment and 0 for the control group) defined by $\rho_{XG} = \sqrt{p(1-p)}(\mu_{1X} - \mu_{0X})$ and e_{Xi} is a standard normal random variable. Observed covariate scores (x_i) were created by introducing random measurement error (e_{xi}) using the following equation:

$$x_i = \sqrt{\rho_{xx}}\, X_i + \sqrt{1 - \rho_{xx}}\, e_{xi}. \tag{12}$$

The dependent variable, Y_i, was generated as a standard normal random variable using the following equation:

$$Y_i = \rho_{xy} X_i + \Delta\psi_G \frac{G_i - p}{\sqrt{p(1-p)}} + \sqrt{1 - \rho_{xy}^2 - \Delta\psi_G^2}\, e_{Yi}, \tag{13}$$

where $\Delta\psi_G^2$ is the unique treatment effect, ρ_{xy} is the true correlation between X_i and Y_i, and e_{Yi} is a standard normal random variable. Note that the G_i is centered by p (the proportion of subjects in the treatment group) and divided by the standard deviation to ensure Y_i has a mean of zero and variance of one.

We conducted the Monte Carlo simulation using Indiana University's Big Red supercomputer, which is a distributed shared-memory cluster consisting of 1,024 IBM JS21 Blades, each with two dual-core PowerPC 970 MP processors, 8 GB of memory, and a PCI-X Myrinet 2000 adapter for high-bandwidth, low-latency message-passing interface applications. Big Red has a theoretical peak performance of more than 40 teraflops (i.e., more than 40 thousand billion floating-point operations a second) and uses a SuSE Linux Enterprise Server operating system. We wrote all programs in R (Culpepper & Aguinis, in press; R Development Core Team, 2008), and they are available from the authors upon request.

Dependent Variables

The dependent variables were relative bias for the treatment effect and Type I error and statistical power rates for the null hypothesis of no effect (i.e., H_o: $\alpha_j = 0$). Relative bias was computed as the average of the absolute value difference between the observed (i.e., sample-based) and true (i.e., population) treatment effect. Note that the population treatment effect was generated as $\Delta\psi_G/\sqrt{p(1-p)}$ in Equation 13. Additionally, Type I error and power rates were computed as the proportion of statistically significant estimates out of the 5,000 simulated replications for different combinations of parameter values. Specifically, Type I error rates were proportions for design cells for which $\Delta\psi_G^2 = 0$ and statistical power rates were proportions for design cells for which $\Delta\psi_G^2 > 0$.

Key Accuracy Checks

We computed theoretical values for ΔR_α^2 and compared them with the empirical values using the OLS regression from the Monte Carlo simulation, and differences were not larger than expected due to sampling error alone. Specifically, the median absolute valued difference between the empirical and theoretical ΔR_α^2 values was .000003, and in no cell in the simulation design were differences larger than .0003. Thus, these results provide evidence in support of the validity of the data generation procedures.

Results

We conducted three analyses of variance (ANOVA) for the comparison of the relative bias, power, and Type I error among five methods we evaluated in Study 1: the OLS regression, EIV, Lord method, R&P, and sparse SEM. Specifically, each of these three ANOVAs included main effects for the following factors:

statistical approach (i.e., the OLS, EIV, Lord method, R&P, or SEM), sample size (n), proportion of sample in treatment group (p), true group mean differences on the covariate ($\mu_{1X} - \mu_{0X}$), and covariate reliability (ρ_{xx}), in addition to interactions between the statistical approach factor and each of the other aforementioned design characteristics. This section includes three subsections devoted to the results for bias, power, and Type I errors. Full ANOVA tables for each of the three sets of analysis can be obtained from the authors upon request.

Comparison of relative bias. Figure 2 includes five panels with results for absolute-value relative bias (i.e., absolute valued difference between estimated and true treatment effects divided by true effects) of each approach by study design characteristic. The ANOVA using bias as the dependent variable provided evidence that all of the manipulated factors contributed to differences in bias except for sample size and the interaction between statistical approach and sample size (e.g., Panel A of Figure 2 displays results showing no main or interaction effect associated with sample size). Additionally, with the exception of sample size, all of the main effects and interactions were statistically significant at the .001 level. However, some of the effects were substantively small (e.g., the η^2 associated with the main effects for p, ρ_{xx}, and $\Delta\psi_G/\sqrt{p(1-p)}$ were less than 2%).

ANOVA results provided evidence that statistical approach had the largest effect on treatment effect bias out of all the manipulated factors (i.e., $\eta^2 = .56$). In fact, the five panels in Figure 2 show that, overall and across all conditions, the OLS and R&P produced the most biased estimates and the EIV and Lord method produced the least biased estimates. OLS performed just as expected on the basis of the derivations included in Appendix A. Specifically, the OLS estimates were more biased as the covariate became less reliable (see Panel C of Figure 2), group differences deviated from zero (see Panel D), and p approached .5 (see Panel E). The R&P estimates were the least accurate of the methods. This result is likely due to the violation of the assumption of equal error variances in x_i and Y_i, as is the case in typical psychological research. For instance, the R&P method was more biased as the error in Y_i decreased due to increases in the treatment effect (see Panel B) and as measurement error in x_i decreased (see Panel C). Although the sparse SEM model produced less biased treatment effects than did the OLS and R&P, it was less accurate than the EIV and Lord method. In fact, the SEM produced estimates that were biased by 9.4% on average across the 3,500 conditions, and Figure 2 shows that the SEM yielded treatment effect estimates that differed from true effects by 15.8%, 11.3%, 7.2%, and 3.5% when covariate reliability was 0.6, 0.7, 0.8, and 0.9, respectively. In contrast, the relative bias for the EIV was 2.1%, 1.9%, 1.9%, and 1.8% when covariate reliability was 0.6, 0.7, 0.8, and 0.9, respectively, which suggests that the SEM was more biased than the EIV when $\rho_{xx} < 0.9$.

Because results regarding relative bias demonstrate the superiority of the EIV, Lord method, and SEM, next we present results regarding statistical power and Type I error rates for these three statistical approaches only (and not for the OLS and R&P).

Comparison of relative statistical power. This section examines the simulation results for all combinations of the parameter values for conditions where ΔR^2_α (i.e., unique effect of the treatment beyond the covariate) was greater than zero. ANOVA results using statistical power as the dependent variable indicated that all of the main effects and interactions with statistical approach (i.e.,

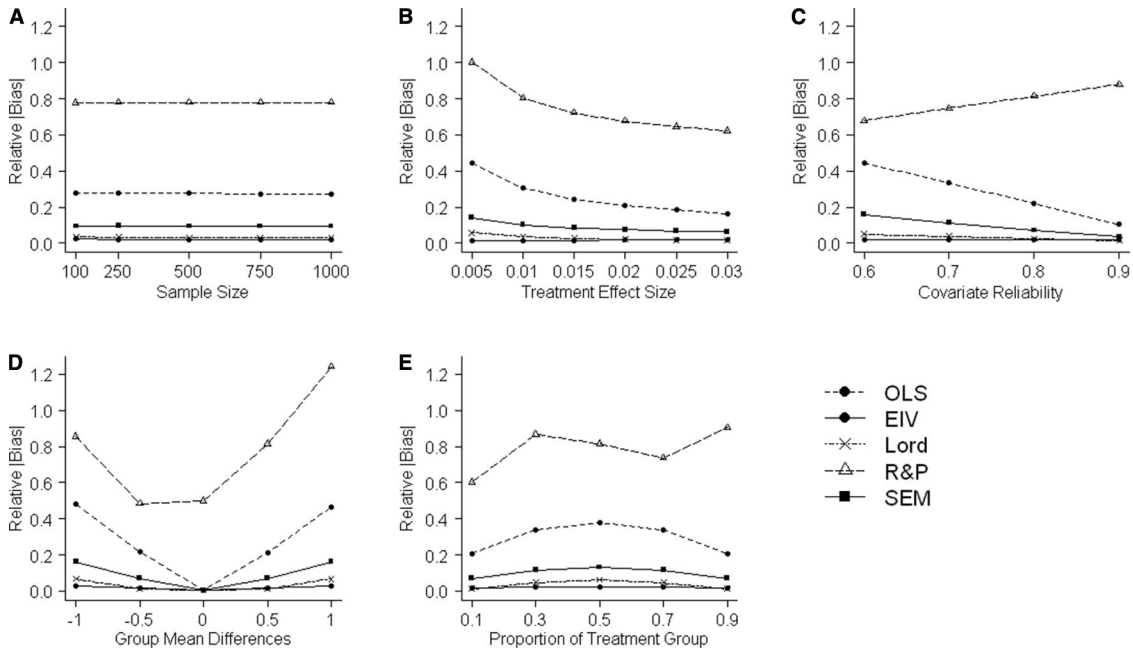

Figure 2. Simulation results for bias of ordinary least squares (OLS), errors-in-variables (EIV), Lord's (1960) method, Raaijmakers and Pieters's (1987) method (R&P), and sparse structural equation modeling (SEM) as a function of sample size, treatment effect size, covariate reliability, group mean differences, and proportion of subjects in the treatment group.

the Lord method, EIV, or SEM) affected statistical power. In fact, statistical approach and sample size (e.g., see Panel A of Figure 3) had the greatest effects on statistical power. That is, the EIV offered a significantly more powerful test of treatment effects than did the Lord method, although, as expected, power increased as sample size increased for both approaches. Additionally, the EIV also had a statistical power advantage over the SEM. Specifically, the EIV was more powerful than the SEM across sample sizes, treatment effect sizes, covariate reliability, and proportion of the sample in the treatment group. However, the SEM had a slight advantage in power over the EIV when groups differed by one standard deviation in covariate means.

Additionally, the main effects for $\mu_{1X} - \mu_{0X}$ (see Panel D of Figure 3) and ρ_{xx} (see Panel C) were substantively small, and statistical approach only slightly interacted with n, $\mu_{1X} - \mu_{0X}$, ρ_{xx}, and $\Delta\psi_G/\sqrt{p(1-p)}$ (see Panel B). In contrast, statistical approach interacted with the proportion of subjects in the treatment group. Specifically, statistical power associated with the Lord method declined as p increased, whereas the power of the EIV remained fairly constant (see Panel E).

In summary, results indicate that the EIV is superior to the Lord method and the SEM in terms of statistical power despite the fact that the three approaches produced estimates with relatively similar accuracy, as shown in the previous section. The discrepancy in power between the EIV and the other two methods is likely attributed to the fact that the approximated standard errors for treatment effects in the Lord method and the SEM are inflated, whereas the standard errors for the EIV derived by Fuller (1980, 1987) are more accurate. In fact, the average statistical power for the EIV across all of the values of the manipulated parameters was nearly twice as large (i.e., .67) as the power for the Lord method (i.e., .28). The EIV was often more powerful than the SEM but never by more than 10% in typical cases (e.g., the average power of the SEM across all manipulated parameters was .612), so the SEM appears to yield more accurate standard errors than does the Lord method.

Comparison of Type I error rates. This section compares the Type I error rates of the EIV, SEM, and Lord method for conditions where ΔR_α^2 (i.e., unique effect of the treatment beyond the covariate) equaled zero. ANOVA results using Type I error rates as the dependent variable indicated that statistical approach had the greatest effect on Type I error rates, followed by the main effect of p and the interaction between statistical approach and p (see Panel D of Figure 4). The four panels in Figure 4 suggest that the EIV effectively controlled Type I error rates at the nominal level set at .05. In contrast, the Lord method offered a conservative test. Furthermore, whereas the SEM did a better job of controlling the Type I error rate than did the Lord method, the Type I error rate of the SEM was affected by sample size, covariate reliability, group mean differences on the covariate, and the proportion of sample in the treatment group.

In summary, the simulation results regarding Type I error rates indicate that the Lord method offers a test of treatment effects that is overly conservative because Type I error rates are biased downwardly due to less-accurate standard errors. For example, Panel D of Figure 4 shows that the Lord test is more conservative for larger values of p, which, coupled with the results shown in Panel E of Figure 3, suggests that the standard errors associated with the Lord estimated treatment effect become increasingly inaccurate as p increases. Additionally, across all conditions, the average Type I error rate was .048 for the EIV, .007 for the Lord method, and .033 for the SEM, which suggests that the Lord method suffers from

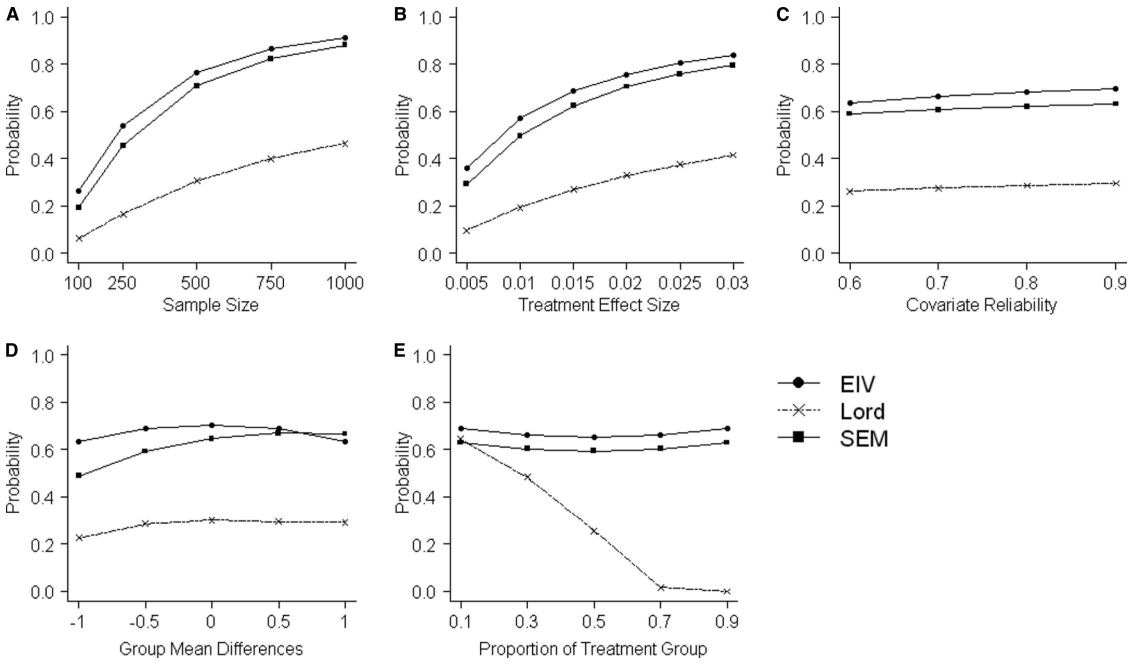

Figure 3. Simulation results for statistical power (i.e., probability of detecting an existing effect) of errors-in-variables (EIV), Lord's (1960) method, and structural equation modeling (SEM) as a function of sample size, treatment effect size, covariate reliability, group mean differences, and proportion of subjects in the treatment group.

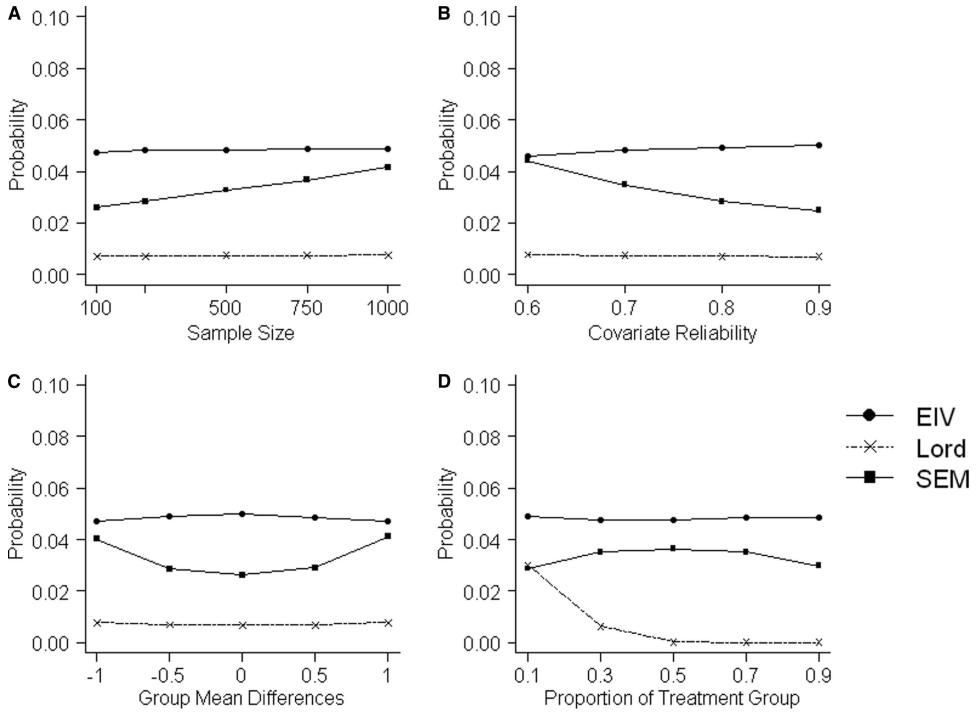

Figure 4. Simulation results for Type I error rates (i.e., probability of rejecting a true null hypothesis of no effect) of errors-in-variables (EIV), Lord's (1960) method, and structural equation modeling (SEM) as a function of sample size, covariate reliability, group mean differences, and proportion of subjects in the treatment group.

deflated Type I error rates, whereas the EIV is able to control Type I error at the prespecified nominal level, and the SEM tended to be slightly more conservative.

Study 2

The goal of the second study was to compare the performance of the EIV and SEM in terms of bias, control of Type I error rates, and statistical power. This study focuses on the EIV because this was the approach that outperformed all others on the basis of Study 1. Also, this study focuses on the SEM because of its ability to incorporate multiple indicators for both the covariate and the dependent variable and, hence, its ability to potentially address identification issues. Study 2 did not examine the performance of the Lord method in the presence of multiple indicators because this approach is limited to pre- and postdesigns, where researchers control for only a single fallible covariate. In other words, the Lord method is not capable of adjusting treatment effects when more than one fallible covariate is used to control for group differences. In contrast, both the EIV and the SEM can be used in the presence of multiple covariates, and therefore they are more applicable in circumstances more typical in research in psychology and related fields.

Manipulated Parameters and Data Generation Procedures

The true covariate and dependent variable scores were generated using Equations 11 and 13. We conducted key accuracy checks that were similar to those in Study 1 and also generated the data using Indiana University's Big Red supercomputer. In contrast to the procedure in Study 1, in Study 2 we manipulated the number of observed covariate and dependent variable measures (i.e., indicators). Specifically, the observed measures for X_i and Y_i were generated using the following equation:

$$x_{ik} = \lambda_x X_i + \sqrt{1 - \lambda_{xk}^2}\, e_{x1k} \text{ and}$$

$$y_{il} = \lambda_y Y_i + \sqrt{1 - \lambda_{yl}^2}\, e_{yil}, \tag{14}$$

where x_{ik} and y_{il} are the kth and lth observed measures of X_i and Y_i, respectively, and $k = 1, \ldots, K$ and $l = 1, \ldots, L$. Additionally, λ_x and λ_y are loadings, and e_{xik} and e_{yil} are standard normal error terms. λ_x and λ_y were constant across the K and L observed measures for X_i and Y_i, respectively.

Table 1 includes the parameter values used in Study 2. Specifically, Study 2 examined 134,400 unique combinations of parameter values and generated 1,000 replications for each combination. Also, Table 1 shows that λ_x and λ_y equaled one of four values (i.e., .3, .4, .5, or .6) and that the number of observed measures equaled one of four values (i.e., 2, 4, 6, or 8).

We implemented the SEM approach using Sörbom's (1978) model as described earlier. The EIV was implemented by creating total scores of the indicators for X_i and Y_i. Cronbach's alpha was computed from the K x_{ik} and L y_{il}. The estimated Cronbach's alpha for the K x_{ik} was used to compute $\boldsymbol{\Sigma}_{\boldsymbol{xx}}$, and the square root of the estimate of internal consistency was used to disattenuate the covariance between the total score for the covariate and dependent variable.

Results

As in Study 1, we conducted an ANOVA using bias, statistical power, and Type I error rates as the dependent variables. ANOVA results indicated a high degree of consistency regarding each of these dependent variables across the two approaches in that the EIV demonstrated superior performance. So, given that the results regarding power and Type I error rates were consistent, we report only results related to relative bias of the two procedures. Additional figures illustrating results regarding power and Type I error rates are available from the authors upon request.

Figure 5 includes six panels that plot the relative bias of the EIV and SEM in the case where more than one covariate and response indicators are available. Note that to simplify the graphs, we created Panel C with only cases where the loadings for X_i and Y_i were equal; similarly, we created Panel F using conditions where the number of indicators for X_i and Y_i were equal. The full ANOVA table and additional graphs are available from the authors upon request.

Figure 5 illustrates the ANOVA results that only the magnitude of factor loadings (see Panel C) and the extent of group mean differences (see Panel D) affected the relative bias of the EIV and SEM estimates. In contrast, the main effects for sample size, treatment effect size, proportion of treatment group, and number of indicators were not statistically significant. Panels A, B, E, and F show that the SEM produced treatment effect estimates that were smaller than the true value by approximately 30% across values of sample size, treatment effect size, proportion of treatment group, and number of indicators for X_i and Y_i. In contrast, Panels A, B, E, and F show that the EIV estimates were essentially unbiased. Panel C shows that the EIV estimates were unbiased across factor loading values whereas the SEM produced less-biased estimates as the factor loadings increased. Panel D shows that the EIV estimates were less biased in situations where group mean differences on the covariate were small and more biased when groups differed in covariate mean differences. In summary, results of Study 2 provide evidence that the EIV yields more accurate treatment effect estimates compared with the SEM.

General Discussion

Given the pervasive use of ANCOVA to address important theoretical and practical issues in psychology and related fields, the present study makes both methodological and substantive contributions. From a methodological perspective, previous research has documented biases in ANCOVA treatment effects when fallible covariates are included in the model in nonexperimental designs. The present article provides new formulas describing the exact degree of bias under various conditions and the underlying mechanisms that lead to inflation in Type I error rates and subsequent erroneous substantive conclusions. This analytic material can be used by future researchers to attempt to replicate past studies (either for single studies or in a meta-analytic fashion) that may have committed Type I errors and reported possibly nonexistent treatment effects. Specifically, researchers would need to collect the necessary information on the sample (e.g., group mean difference, sample size, covariate reliability, proportion in the treatment group) and apply the equations included in Appendix A.

Additionally, the Monte Carlo simulations offer new and comprehensive knowledge about the relative performance of existing methods for disattenuating parameter estimates in the presence of fallible covariates. An important implication for substantive researchers and for practitioners is that EIV models are superior compared with their competitors (namely, the SEM, Lord, and

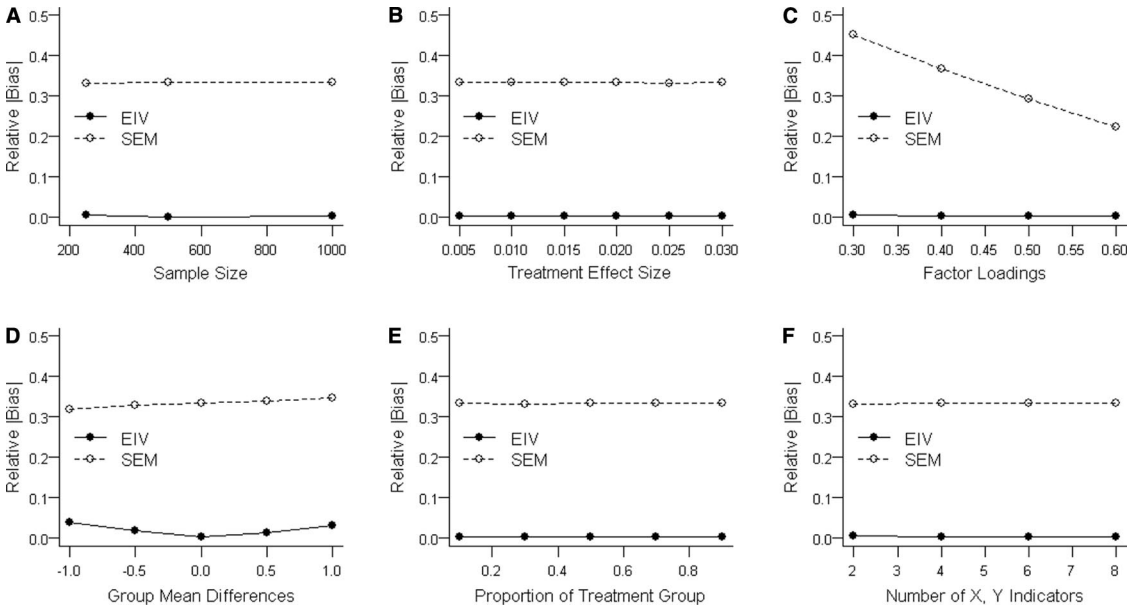

Figure 5. Simulation results for relative bias of errors-in-variables (EIV) and structural equation modeling (SEM) for multiple covariate and response variable indicators and different sample size, treatment effect size, factor loadings, group mean differences, proportion of treatment group, and number of X, Y indicators.

R&P models). In particular, EIV methods produce accurate estimates of the true treatment effects, have greater levels of statistical power, and provide a better control of Type I error rates. Unfortunately, researchers may not have access to statistical software to implement EIV methods; in fact, Stata (StataCorp, 2009), which is primarily used by applied economists, is the only widespread program that to our knowledge is able to conduct EIV analysis. Thus, Appendix B includes a program written for R, which is an open source statistical software package (Culpepper & Aguinis, in press), that researchers can download for free and use to accurately assess treatment effects. Also, to avoid the need to retype it, this program can be downloaded from http://math.ucdenver.edu/~sculpeppe/EIV.R or http://mypage.iu.edu/~haguinis/eiv.html

We acknowledge some limitations of our study that also serve as impetus for future research. First, our simulation examined loadings that were constant for all covariate and dependent variable indicators. It is possible that the SEM is superior to the EIV in cases where loadings differ. Consequently, additional research is needed to understand the impact of nonconstant loadings on the relative performance of the methods examined in this study. Second, we used Cronbach's alpha to correct EIV estimates in the case of several indicator variables. However, internal consistency, or lack thereof, is only one of several sources of measurement error (Aguinis, Pierce, & Culpepper, 2009; Le, Schmidt, & Putka, 2009). Thus, it is possible that the SEM performs better than the EIV in situations where alpha provides a less-accurate estimate of reliability, and this issue can be investigated in future research.

Concluding Remarks

In conclusion, the primary goal of this study is to assist researchers in addressing substantive questions by implementing more accurate data-analytic procedures. In addition to its value in terms of basic research, ANCOVA is often used to assess treatment effects that may consist of determining the value and/or merit of programs, interventions, or organizational and social initiatives (Arvey et al., 1985). Current applications of ANCOVA make it difficult to accurately assess and evaluate treatment effects given the normative presence of covariate measurement error. The present study provides researchers with new knowledge that the EIV is the best available procedure for estimating treatment effects accurately when using ANCOVA with nonexperimental designs. Using the EIV minimizes bias, maximizes statistical power, and keeps the Type I error rate close to its nominal level. In addition, we provide a practical tool (i.e., computer program) that allows researchers to implement the EIV in the future, with the goal of yielding more accurate ANCOVA results that, in turn, are likely to lead to more accurate assessments regarding the size of treatment effects and better decisions in terms of interventions, practices, and policy making.

References

Aguinis, H., Pierce, C. A., & Culpepper, S. A. (2009). Scale coarseness as a methodological artifact: Correcting correlation coefficients attenuated from using coarse scales. *Organizational Research Methods, 12,* 623–652.

Arbuckle, J. L. (2008). *Amos 17.0 user's guide.* Chicago, IL: SPSS.

Arvey, R. D., Cole, D. A., Hazucha, J. F., & Hartanto, F. M. (1985). Statistical power of training evaluation designs. *Personnel Psychology, 38,* 493–507.

Bartlett, M. S. (1949). Fitting a straight line when both variables are subject to error. *Biometrics, 5,* 207–212.

Boggs, P. T., Spiegelman, C. H., Donaldson, J. R., & Schnabel, R. B. (1988). A computation examination of orthogonal distance regression. *Journal of Econometrics, 38,* 169–201.

Cappelleri, J. C., Trochim, W. M., Stanley, T. D., & Reichardt, C. S. (1991). Random measurement error does not bias the treatment effect estimate in the regression-discontinuity design: I. The case of no interaction. *Evaluation Review, 15,* 395–419.

Carroll, R. J., & Ruppert, D. (1996). The use and misuse of orthogonal regression in linear errors-in-variables models. *American Statistician, 50,* 1–6.

Cochran, W. G. (1968). Errors of measurement in statistics. *Technometrics, 10,* 637–666.

Cramer, E. M. (1974). The distribution of partial correlations and generalizations. *Multivariate Behavioral Research, 9,* 119–122.

Cribbie, R. A., & Jamieson, J. (2000). Structural equation models and the regression bias for measuring correlates of change. *Educational and Psychological Measurement, 60,* 893–907.

Culpepper, S. A., & Aguinis, H. (in press). R is for revolution: A cutting-edge, free, open source statistical package. *Organizational Research Methods.*

DeGracie, J. S., & Fuller, W. A. (1972). Estimation of the slope and analysis of covariance when the concomitant variable is measured with error. *Journal of the American Statistical Association, 67,* 930–937.

Elashoff, J. D. (1969). Analysis of covariance: A delicate instrument. *American Educational Research Journal, 6,* 383–401.

Fuller, W. A. (1980). Properties of some estimators for the errors-in-variables model. *Annals of Statistics, 8,* 407–422.

Fuller, W. A. (1987). *Measurement error models.* New York, NY: Wiley.

Fuller, W. A., & Hidiroglou, M. A. (1978). Regression estimation after correcting for attenuation. *Journal of the American Statistical Association, 73,* 99–104.

Gleser, L. J. (1981). Estimation in a multivariate "errors in variables" regression model: Large sample results. *Annals of Statistics, 9,* 24–44.

Grant, A. M., & Wall, T. D. (2009). The neglected science and art of quasi-experimentation: Why-to, when-to, and how-to advice for organizational researchers. *Organizational Research Methods, 12,* 653–686.

Harwell, M. (2003). Summarizing Monte Carlo results in methodological research: The single-factor, fixed-effects ANCOVA case. *Journal of Educational and Behavioral Statistics, 28,* 45–70.

Hayduk, L. A. (1996). *LISREL issues, debates, and strategies.* Baltimore, MD: Johns Hopkins University Press.

Jennings, E. (1965). Matrix formulas for part and partial correlation. *Psychometrika, 30,* 353–356.

Kahneman, D. (1965). Control of spurious association and the reliability of the controlled variable. *Psychological Bulletin, 64,* 326–329.

Ketellapper, R. H. (1983). On estimating parameters in a simple linear errors-in-variables model. *Technometrics, 25,* 43–47.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th ed.). Boston, MA: McGraw–Hill Irwin.

Le, H., Schmidt, F. L., & Putka, D. J. (2009). The multifaceted nature of measurement artifacts and its implications for estimating construct-level relationships. *Organizational Research Methods, 12,* 165–200.

Linn, R. L., & Werts, C. E. (1971). Considerations for studies of test bias. *Journal of Educational Measurement, 8,* 1–4.

Lord, F. M. (1960). Large-sample covariance analysis when the control variable is fallible. *Journal of the American Statistical Association, 55,* 307–321.

Madansky, A. (1959). The fitting of straight lines when both variables are subject to error. *Journal of the American Statistical Association, 54,* 173–205.

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

Maxwell, S. E., Delaney, H. D., & Manheimer, J. M. (1985). ANOVA of

residuals and ANCOVA: Correcting an illusion by using model comparisons and graphs. *Journal of Educational and Behavioral Statistics, 10,* 197–209.

Moran, P. A. P. (1971). Estimating structural and functional relationships. *Journal of Multivariate Analysis, 1,* 232–255.

Mudholkar, G. S., Chaubey, Y. P., & Lin, C. (1976). Some approximations for the noncentral-F distribution. *Technometrics, 18,* 351–358.

Overall, J., & Woodward, J. A. (1977). Common misconceptions concerning the analysis of covariance. *Multivariate Behavioral Research, 12,* 171–186.

Pearson, E. S., & Hartley, H. O. (1951). Charts of the power function for analysis of variance tests, derived from the non-central F-distribution. *Biometrika, 38,* 112–130.

Porter, A. C., & Chibucos, T. R. (1975). Common problems of design and analysis in evaluative research. *Sociological Methods Research, 3,* 235–257.

Porter, A. C., & Raudenbush, S. W. (1987). Analysis of covariance: Its model and use in psychological research. *Journal of Counseling Psychology, 34,* 383–392.

R Development Core Team. (2008). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Raaijmakers, J. G. W., & Pieters, L. P. M. (1987). Measurement error and ANCOVA: Functional and structural relationship approaches. *Psychometrika, 52,* 521–538.

Ree, M. J., & Carretta, T. R. (2006). The role of measurement error in familiar statistics. *Organizational Research Methods, 9,* 99–112.

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandom studies: A practical guide and simulated example. *Psychological Methods, 13,* 279–313.

Sörbom, D. (1978). An alternative to the methodology for analysis of covariance. *Psychometrika, 43,* 381–396.

Stanley, T. D., & Robinson, A. (1990). Sifting statistical significance from the artifact of regression-discontinuity design. *Evaluation Review, 14,* 166–181.

StataCorp. (2009). *Stata statistical software: Release 11.* College Station, TX: Author.

Warren, R. D., White, J. K., & Fuller, W. A. (1974). An errors-in-variables analysis of managerial role performance. *Journal of the American Statistical Association, 69,* 886–893.

Wicherts, J. M. (2005). Stereotype threat research and the assumptions underlying analysis of covariance. *American Psychologist, 60,* 267–269.

Appendix A

Analytic Derivations

The purpose of Appendix A is twofold. First, we show the derivation of Equation 2, which provides an explanation for the various factors that bias treatment-effect estimates. Second, we show equations for computing real Type I error rates associated with tests for treatment effects in the presence of fallible covariates.

Derivation of Equation 2

The analysis of covariance (ANCOVA) estimates in Equation 1 are depicted in matrix terms as follows:

$$\begin{bmatrix} \alpha_1 \\ \beta \end{bmatrix} = \frac{1}{1 - \rho_{x\alpha}^2} \begin{bmatrix} 1 & -\rho_{x\alpha} \\ -\rho_{x\alpha} & 1 \end{bmatrix} \begin{bmatrix} \rho_{y\alpha} \\ \rho_{xy} \end{bmatrix}, \tag{A1}$$

where ρ_{xy} and $\rho_{y\alpha}$ represent correlations between the dependent variable and the covariate and treatment, respectively, and $\rho_{x\alpha}$ is the correlation between treatment and covariate. Equation A1 can be expressed as $\mathbf{b} = \mathbf{R}_{x\alpha}^{-1}\,\boldsymbol{\rho}_y$, where $\mathbf{b}$ is a vector containing the estimates in Equation A1, the product of the fraction and matrix represents the inverse of the correlation matrix between treatment and covariate, $\mathbf{R}_{x\alpha}^{-1}$, and $\boldsymbol{\rho}_y$ is a vector of correlations between the independent variables and y_{ij}. Equation A1 can be further modified by accounting for measurement error in x_{ij} and y_{ij}. That is, the unattenuated, true score correlations for $\rho_{y\alpha}$, ρ_{xy}, and $\rho_{x\alpha}$ can be adjusted for measurement error by substituting $\rho_{ya}\sqrt{\rho_{yy}}$, $\rho_{xy}\sqrt{\rho_{xx}}\sqrt{\rho_{yy}}$, and $\rho_{x\alpha}\sqrt{\rho_{xx}}$, respectively. The matrix expression can be updated to $\mathbf{b} = [\mathbf{R}_{x\alpha} \odot \mathbf{R}_{xx}]^{-1}\rho_y\sqrt{\rho_{yy}}$, where $\odot$ denotes the Hadamard product (i.e., element-wise multiplication) between $\mathbf{R}_{x\alpha}$ and a reliability matrix, $\mathbf{R}_{x x\alpha}$, which contains $\sqrt{\rho_{xx}}$ in the upper and lower triangles and ones in the diagonal. Also note that $\boldsymbol{\rho}_y$ includes $\rho_{xy}\sqrt{\rho_{xx}}$ in the second row.

First, we demonstrate that null hypothesis tests (i.e., H_o: $\alpha_j = 0$) are incorrect when $\rho_{xx} < 1$ and $\mu_{1x} - \mu_{0x} \neq 0$. The formula for squared part correlations in the following equation is useful for deriving an expression for $\boldsymbol{\rho}_{y\alpha}$ in the presence of an error-free covariate when H_o is true:

$$\boldsymbol{\Delta}R_{\alpha}^2 = \frac{(\rho_{y\alpha} - \rho_{xy}\rho_{x\alpha})^2}{1 - \rho_{x\alpha}^2}. \tag{A2}$$

Specifically, if H_0 is true (i.e., $\Delta R_{\alpha}^2 = 0$), then Equation A2 simplifies to $\rho_{y\alpha} = \rho_{xy}\rho_{x\alpha}$ and $\mathbf{b} = [\mathbf{R}_{x\alpha} \odot \mathbf{R}_{xx}]^{-1}[\rho_{x\alpha}, \sqrt{\rho_{xx}}]'\,\rho_{xy}\sqrt{\rho_{yy}}$, where $[\rho_{x\alpha}, \sqrt{\rho_{xx}}]'$ is a 2×1 vector with the elements separated by a comma.

If both x_{ij} and y_{ij} are error-free, then the unique contribution of the treatment beyond the covariate is $R_{x\alpha}^2 = \Delta R_{\alpha}^2 + \rho_{xy}^2$, where $R_{x\alpha}^2$ is the variance accounted for in y_{ij} by the covariate and treatment and ρ_{xy}^2 is the proportion of variance in y_{ij} accounted for by x_{ij} alone. Previous research showed that $R_{x\alpha}^2 = [\rho_{x\alpha}, \sqrt{\rho_{xx}}][\mathbf{R}_{x\alpha} \odot \mathbf{R}_{xx}]^{-1}[\rho_{x\alpha}, \sqrt{\rho_{xx}}]'\rho_{xy}^2\rho_{yy}$ (Cramer, 1974; Jennings, 1965), which suggests that

(*Appendices continue*)

$$\Delta R^2_\alpha = [\rho_{x\alpha},\ \sqrt{\rho_{xx}}][\mathbf{R}_{x\alpha} \odot \mathbf{R}_{xx}]^{-1}[\rho_{x\alpha},\ \sqrt{\rho_{xx}}]'\rho^2_{xy}\rho_{yy} - \rho^2_{xy}\rho_{xx}\rho_{yy}. \quad \text{(A3)}$$

We represent Equation A3 using matrix algebra as follows:

$$\Delta R^2_\alpha = \rho^2_{xy}\rho_{yy}\left\{\frac{1}{1-\rho_{xx}\rho^2_{x\alpha}}\begin{bmatrix}\rho_{x\alpha}\\ \sqrt{\rho_{xx}}\end{bmatrix}\begin{bmatrix}1 & -\rho_{x\alpha}\sqrt{\rho_{xx}}\\ -\rho_{x\alpha}\sqrt{\rho_{xx}} & 1\end{bmatrix} \times \begin{bmatrix}\rho_{x\alpha}\\ \sqrt{\rho_{xx}}\end{bmatrix} - \rho_{xx}\right\}. \quad \text{(A4)}$$

Pre- and postmultiplying the matrix by the vectors yields

$$\Delta R^2_\alpha = \rho^2_{xy}\rho_{yy}\left[\frac{\rho^2_{x\alpha}(1-\rho_{xx}) + \rho_{xx}(1-\rho^2_{x\alpha})}{1-\rho_{xx}\rho^2_{x\alpha}} - \rho_{xx}\right], \quad \text{(A5)}$$

which simplifies to

$$\Delta R^2_\alpha = \rho^2_{xy}\rho_{yy}\left[\frac{\rho^2_{xa}(1-\rho_{xx}) + \rho_{xx}(1-\rho^2_{xa}) - \rho_{xx}(1-\rho_{xx}\rho^2_{xa})}{1-\rho_{xx}\rho^2_{xa}}\right]. \quad \text{(A6)}$$

Rearranging terms yields the following desired solution:

$$\Delta R^2_\alpha = \rho^2_{xy}\rho_{yy}\rho^2_{xa}\left[\frac{1-2\rho_{xx}+\rho^2_{xx}}{1-\rho_{xx}\rho^2_{xa}}\right] = \frac{\rho^2_{xy}\rho_{yy}\rho^2_{xa}(1-\rho_{xx})^2}{1-\rho_{xx}\rho^2_{xa}}. \quad \text{(A7)}$$

The definition of point biserial correlations can be substituted into Equation A1 for $\rho_{x\alpha}$ and $\rho_{y\alpha}$ as

$$\rho_{x\alpha} = \frac{(\mu_{1x}-\mu_{0x})\sqrt{p(1-p)}}{\sigma_x} \text{ and}$$

$$\rho_{y\alpha} = \frac{(\mu_{1y}-\mu_{0y})\sqrt{p(1-p)}}{\sigma_y}, \quad \text{(A8)}$$

where p represents the proportion of subjects in the treatment group, σ_x is the standard deviation of x_{ij} across the treatment and control groups, σ_y is the standard deviation of y_{ij} across the groups, and $\mu_{1y} - \mu_{0y}$ are unadjusted mean differences between the groups on y_{ij}. The following equation includes the updated expression, which is identical to Equation 2 in the article:

$$\Delta R^2_\alpha = \frac{\rho^2_{xy}\rho_{yy}(\mu_{1x}-\mu_{0x})^2 p(1-p)(1-\rho_{xx})^2}{\sigma^2_x - \rho_{xx}(\mu_{1x}-\mu_{0x})^2 p(1-p)}. \quad \text{(A9)}$$

Equation A9 shows that ΔR^2_α will be unbiased when either $\rho_{xx} = 1$ or $\mu_{1x} - \mu_{0x} = 0$.

Computing Type I Errors in the Presence of Fallible Covariates

The standard F statistic for testing treatment effects in an ANCOVA model is $F = (R^2_{x\alpha} - \rho^2_{xy})(n-3)/(1-R^2_{x\alpha})$, where $R^2_{x\alpha}$ is the proportion of variance explained by the covariate, x_{ij}, and the treatment effect, α_j (Kutner, Nachtsheim, Neter, & Li, 2005; Maxwell, Delaney, & Manheimer, 1985). Consider the case where there is no treatment effect in the presence of a fallible covariate. Given that $R^2_{x\alpha} = \Delta R^2_\alpha + \rho^2_{xy}$, the F statistic can be rewritten as $F_\alpha = \Delta R^2_\alpha(n-3)/(1-R^2_{x\alpha})$, which suggests that the standard F statistic is biased in the amount of F_α when H_o is true. Consequently, the central F distribution is inappropriate and the noncentral F distribution (Mudholkar, Chaubey, & Lin, 1976; Pearson & Hartley, 1951) can be used to compute Type I error rates by specifying a noncentrality parameter. The noncentral F distribution is defined as

$$f(F,\ \nu_1,\ \nu_2,\ \lambda) = \frac{\chi^2_n(\nu_1,\ \lambda)\nu_2}{\chi^2(\nu_2)\nu_1}, \quad \text{(A10)}$$

where χ^2_n is the noncentral chi-square distribution with ν_1 degrees of freedom and noncentrality parameter $\lambda = F_\alpha$. Type I error rates are computed from the integral in the following equation:

$$P(F > F^*_{1-\alpha,1,N-3}) = \int_{F^*_{1-\alpha,1,N-3}}^{\infty} f\left(F,1,N-3,\ \frac{\Delta R^2_\alpha(n-3)}{1-R^2_{x\alpha}}\right)dF, \quad \text{(A11)}$$

where $F^*_{1-\alpha,1,N-3}$ is the standard critical value from the central F distribution.

(*Appendices continue*)

Appendix B

R Program for Errors-in-Variables (EIV) Regression

The following R code defines a function called "eiv":

```
eiv<-function(formula,reliability,data){
mfx<-model.matrix(formula,data=data)
p<-length(mfx[1,])-1;n<-length(mfx[,1])
  mf <- match.call(expand.dots = FALSE)
  m <- match(c("formula", "data", "subset", "weights", "na.action",
    "offset"), names(mf), 0L)
  mf <- mf[c(1L, m)]
  mf$drop.unused.levels <- TRUE
  mf[[1L]] <- as.name("model.frame")
  mf <- eval(mf, parent.frame())
mf<-data.frame(mf)
MXX<-cov(mfx[,c(2:(p+1))]);MXY<-cov(mfx[,c(2:(p+1))],mf[,1])
Suu<-matrix(0,p,p);diag(Suu)<-(1-reliability)*diag(MXX)
Mxx<-MXX-(1-p/n)*Suu;Btilde<-solve(Mxx)%*%MXY
MSEtilde<-as.numeric(n*(1-2*t(Btilde)%*%MXY+t(Btilde)%*%MXX%*%Btilde)/(n-3))
Rhat<-matrix(0,p,p);diag(Rhat)<-(t(Btilde)%*%Suu)^2
VCtilde<-MSEtilde*(1/n)*solve(Mxx)+(1/n)*solve(Mxx)%*%(Suu*MSEtilde+Suu%*
  %Btilde%*%t(Btilde)%*%Suu+2*Rhat)%*%solve(Mxx)
ttilde<-Btilde/sqrt(diag(VCtilde))
output<-cbind(reliability,Btilde,sqrt(diag(VCtilde)),ttilde,2*(1-pt(ttilde,n-p)))
colnames(output)<-c('Reliability','Est.','S.E.','t','Prob.(>|t|)')
output
}
```

Users can implement the eiv function by first submitting the aforementioned code to R. Once the eiv function is entered into R, users need to specify a statistical model, a vector of reliability coefficients for the predictors, and the name of the data set. For example, the following code would compute an EIV analysis with a dependent variable (y), two covariates (x_1 and x_2), and a dichotomously coded treatment effect (treat):

```
eiv(y~x1+x2+treat,reliability=c(.8,.9,1),data=eivdata)
```

Also, note that the option `reliability=` allows users to specify the reliability of the three predictors (in this case, the reliability coefficients for x_1, x_2, and treat are .8, .9, and 1.0, respectively). Finally, the option denoted by `data=` specifies the name of the data set containing the dependent and independent variables (the name of the data set in this example is `eivdata`). Submitting the eiv command will produce a table of regression output with disattenuated estimates, parameter standard errors based upon Equation 4 of the article, t values, and p values. This code is also available at http://math.ucdenver.edu/~sculpeppe/EIV.R or http://mypage.iu.edu/~haguinis/eiv.html

Received September 18, 2009
Revision received November 17, 2010
Accepted November 30, 2010 ■