Using the Errors-in-Variables Method in Two-Group Pretest-Posttest Designs

Alyssa Counsell* & Robert A. Cribbie

Quantitative Methods Program

Department of Psychology

York University

* Correspondence should be addressed to Alyssa Counsell, Department of Psychology, York University, Toronto, ON, Canada M3J 1P3 (e-mail: counsell@yorku.ca)

Abstract

Culpepper and Aguinis (2011) highlighted the benefit of using the errors-in-variables (EIV) method to control for measurement error and obtain unbiased regression estimates. The current study investigated the EIV method and compared it to change scores and analysis of covariance (ANCOVA) in a two group pretest-posttest design. Results indicated that the EIV method's estimates were unbiased under many conditions, but the EIV method consistently demonstrated lower power than the change score method. An additional risk with using the EIV method is that one must enter the covariate reliability into the EIV model, and results highlighted that estimates are biased if a researcher chooses a value that differs from the true covariate reliability. Obtaining unbiased results also depended on sample size. Our conclusion is that there is no additional benefit to using the EIV method over change score or ANCOVA methods for comparing the amount of change in pretest-posttest designs.

Key words: analysis of covariance, change, errors-in-variables, posttest, pretest

Using the Errors-in-Variables Method in Two-Group Pretest-Posttest Designs

Measuring group differences across two time points has been a topic of debate for decades (e.g., Allison, 1990; Cribbie & Jamieson, 2000; Cronbach & Furby, 1970; Lord, 1967; Maris; 1998; Overall, 1989; Wright, 2006). The two most common contenders are change score models and analysis of covariance (ANCOVA). A recent article by Culpepper and Aguinis (2011) acknowledged that the issue with using ANCOVA is its assumption that the covariate has been measured without error. Since it is rarely the case that the types of covariates used in psychology are free of measurement error, using ANCOVA will produce biased (and often misleading) results. Based on their simulation study, Culpepper and Aguinis recommend using the errors-in-variables (EIV) method in lieu of ANCOVA. The EIV method is a modified ANCOVA procedure that takes the reliability of the covariate into account so that the regression model no longer produces biased estimates (Fuller, 1980; 1987; Warren, White, & Fuller, 1974). However, in their paper, Culpepper and Aguinis (2011) discussed the use of the EIV method for general covariates, not specifically for use with a pretest score as a covariate. Given the theoretical debate about when it is appropriate to use change scores and when it is appropriate to use ANCOVA, the EIV method should be investigated in such a context to determine whether it can provide a general solution to the problem of comparing groups in pretest-posttest designs.

ANCOVA or Change Scores?

Before delving into the EIV method it is important to discuss the similarities and differences between using change scores (also called difference scores or gain scores) versus using ANCOVA to compare the amount of change across two time points and two groups. The change score method involves running an independent samples ANOVA (or equivalently a *t*-test in cases with only two groups) to compare the amount of change from pretest to posttest between

the groups. It can also be expressed as a regression model: $(Y - X) = \beta_0 + \beta_1 G + \varepsilon$, where Y is the score at posttest, X is the score at pretest, G is the dummy coded grouping variable and ε represents the residual error. The other popular method for two time point-two group designs is to conduct an ANCOVA using the individual's pretest score as a covariate. ANCOVA, written as a regression model is: $Y = \beta_0 + \beta_1 G + \beta_2 X + \varepsilon$. One can see that the variables in the ANCOVA equation are the same as those in the change score model, and with some simple algebra, that the two models will be mathematically identical when $\beta_2 = 1$.In practice, however, $\beta_2$will not be equal to 1 and therefore the methods will not produce equivalent model coefficients. Specifically β_2 is the pooled within-group regression coefficient, which requires the assumption of homogeneity of regression (i.e., the slopes are equal across the groups) for unbiased estimates.

Although the methods may be used to address the same research design, researchers should be aware that the two methods are actually testing different null hypotheses. Specifically, the change score method is testing the null hypothesis that there is no raw difference between the groups in the amount of change from pretest to posttest, whereas the ANCOVA model is testing the null hypothesis that there is no difference in the groups' posttest scores, *had the groups started with the same pretest scores*. This theoretical difference has implications for the appropriateness of selecting one approach over another, particularly when different results are obtained using each of the two methods. When pretest differences between groups occur, a researcher can obtain radically different results regarding the differences between groups at posttest depending on the statistical approach used. Drawing conclusions from one method or the other could potentially be detrimental if the results were to be used for program implementation or policy change.

As a practical example, consider two different classes, one that receives a novel teaching method for improving vocabulary and another that continues with a previous teaching instruction method (control). However, the two classes differ at pretest, whereby the one that received the novel teaching method had, on average, lower vocabulary scores than the second class (control). If the means of each group remain the same from pre-intervention to post-intervention, the researcher could arrive at different conclusions about the effectiveness of the novel teaching method based on whether he or she used a change score or ANCOVA approach. More specifically, since the ANCOVA method assumes that the groups are from populations that are equivalent on pretest scores, it is possible for the ANCOVA method to conclude that the group that started with a higher score at pretest actually improved more, even though the mean differences are equivalent. This phenomenon is often termed Lord's paradox (Lord, 1967) as two apparently valid statistical methods provide contradictory results.

It is now clear that the ANCOVA approach can provide misleading results because it assumes that the mean pretest group differences are zero in the population, and therefore it is not the appropriate control for nontrivial pretest differences. In the example presented above, it would not be appropriate to assume that the classes with different pretest vocabulary scores actually started with the same pretest ability. Many researchers have examined this issue (Cribbie & Jamieson, 2004; Fitzmaurice, 2001; Jamieson, 1999, 2004; Linn & Slinde, 1977; Maris, 1998, Rogosa, 1988, 1995; Senn, 2006; Wright, 2006), and in general, the conclusions are that the ANCOVA model is slightly more powerful than the change score model for randomized experiments, but that ANCOVA should not be used when there are true population differences at baseline (unless participants were assigned to groups based on pretest scores, see Wright, 2006). In nonexperimental studies, baseline differences between the groups are often not trivial, so

regression-based control often results in erroneous conclusions (Cribbie & Jamieson, 2000; Miller & Chapman, 2001). Allison (1990) discusses situations where one method is more appropriate than the other. The authors would like to note that due to Cronbach and Furby's (1970) influential paper, there is a history of negative attitudes towards using change scores based on the argument that they are unreliable. We highly encourage researchers to read Rogosa (1995) for more information on this topic.

Errors-in-Variables Method

The issue that using ANCOVA may lead to biased results is a matter of having fallible covariates, i.e., covariates that contain measurement error are not perfectly reliable. In situations where there are nontrivial differences between the groups at pretest, measurement error (coupled with violating the homogeneity of regression assumption), is known to provide biased estimates (Culpepper & Aguinis, 2011; Porter & Raudenbush, 1987). The EIV method was developed to adjust the regression equation based on the covariate's degree of (un)reliability. Rather than using the raw covariance matrix of the independent variables to calculate the regression coefficients, it uses a corrected covariance matrix, which is adjusted to take into account the reliability of the covariate(s). After accounting for measurement error in the covariate (i.e., pretest score in the described research design), the EIV method creates an unbiased estimate. For specific details on how the covariance matrix is modified in the EIV formula see Fuller (1987). In Culpepper and Aguinis' simulation study, the EIV demonstrated unbiased estimates, good power, and accurate Type I error rates across a number of different conditions.

While the simulation results of Culpepper and Aguinis (2011) look promising, the EIV method was not utilized in the measurement of differences in pre-post change across groups. As

such, it is important to investigate whether the EIV method is the recommended method for comparing pre-post change with two groups. Since the EIV method is a modification of the ANCOVA method, one may raise the question about whether the EIV method is appropriate for a two time point-two group study when nontrivial pretest group differences occur. One would expect that the amount of bias would be equivalent to the change score method, given what is known about the regression formulae when ANCOVA's pooled regression coefficient is 1; in other words, if the pretest/posttest reliability is 1, then the slope of the relationship between pretest and posttest will be 1 (as in the change score model) . However, it also raises questions about the appropriateness of the EIV, given the previous discussion of the theoretical difference between ANCOVA and the change score approach. The current research will address the following questions:

1) Are the power, Type I error control, and estimates obtained by the EIV method superior to those of the change score method when baseline differences are nontrivial?
2) As part of the EIV method, one must use an estimate of the covariate's reliability. How precise must a researcher be in estimating the reliability of the covariate in order to obtain unbiased results?

We hypothesize that the EIV's estimates will be unbiased and more similar to the change score method than those of ANCOVA when there are nontrivial pretest differences. With trivial pretest differences (e.g., those that are due to randomization), all three procedures should produce similar model estimates. We also hypothesize that the EIV method will maintain accurate Type I error rates and similar power to the change score method across the range of conditions investigated.

Method

This study used computer simulations to examine the bias, Type I error control, and power of the EIV, ANCOVA, and change score methods for comparing the amount of change from pretest to posttest across two groups. Data generation involved creating a continuous underlying score labeled 'ability', which had a mean of 0 and a standard deviation of 1. Reliability of the pretest and posttest scores was fixed to be equal, although the reliability varied across conditions. Typically when researchers have a two group-two time point research design, the same instrument is used at both time points; this is the justification for why the reliability of the instrument did not change from pretest to posttest. The pretest score (*x*) was generated from the following model:

$$x = \sqrt{\rho_{xx}}\, X + \sqrt{1 - \rho_{xx}}\varepsilon_x$$

where ρ_{xx} is the reliability of the pretest score *x*, *X* is the underlying ability measure, and ε_x is a random error component for the pretest score. Group membership was determined by an allocation variable based on the correlation between group membership and pretest score. The simulation included four different conditions for group allocation. The first condition was random assignment to either the control or treatment group. This resulted in equal proportions across the groups. The next three allocation conditions were created such that the correlation between the underlying measure of ability and group membership was .2, .4, or .6. After the dummy coded group variable was created, a posttest score (*y*) was generated based on different effect sizes and group allocation with the following model:

$$y = \sqrt{\rho_{yy}}\, X + \delta G + \sqrt{1 - \rho_{yy}}\varepsilon_y$$

where ρ_{yy} is the reliability of the posttest score y, δ is the amount that the groups differ at pretest, *G* is the dummy coded grouping variable (0 or 1), and ε_y is a random error component for the posttest score.

The study included a total of 840 conditions. The following variables were manipulated: 1) δ, standardized population difference between groups at pretest (-.5, -.25, 0, .25, .5); 2) reliability of the pretest and posttest score (.5, .8); 3) sample size (20, 50, 100); 4) correlation between an underlying ability score and group membership (0, .2, .4, .6); 5) Difference between true reliability and researcher reported reliability of the covariate (-.2, -.1, -.05,05, .1, .2). This last condition is only relevant for the EIV method because users do not specify reliability estimates when using the change score or ANCOVA approaches. Five thousand replications were conducted for each condition using a nominal Type I error rate (α) of .05. The simulations were conducted using the open source statistical software R (R Development Core Team, 2013) and the EIV model was created using the function provided by Culpepper and Aguinis (2011).

Results

Relative bias was used as an indicator of the amount of bias present for each of the three methods. Specifically, relative bias is defined as the absolute difference between the observed model coefficient and the population treatment effect divided by the population treatment effect (e.g., effect size). When the population treatment effect was zero, bias was based on the raw difference between the model and population coefficients (to avoid division by zero). For the EIV model, bias was assessed under two separate conditions. In the first condition, the estimate of pretest (covariate) reliability entered in the model was exactly equal to the true reliability. A second bias condition examined the model coefficients when the reliability between the true

score and the researcher estimated reliability differed. This condition was investigated because we expected the EIV model coefficient to be unbiased in the first condition, but find it unlikely that applied researchers would be able to provide a value for their instrument's reliability that is exactly equal to its true value. Our Type I error was assessed by examining the proportion of rejections for each of the three models when the population treatment effect was zero. Power was defined as the proportion of rejections when the population treatment effect was non-zero.

Only a subset of the results is presented due to space constraints. Specifically, only the results for $N = 50$ are presented for the Type I error, power, and relative bias conditions for all three procedures. Since the amount of bias differed for the EIV method based on sample size, we also present a figure with all sample size conditions along with a large sample size ($N = 1000$) for the EIV method. The pattern of results for the change score and ANCOVA models remains the same regardless of sample size. If interested, the reader can request the full set of results from the authors.

Relative Bias Results

Table 1 presents the amount of bias present for the change score, ANCOVA, and EIV models when $\rho_{xx} = .5$ and $N = 50$. Table 2 presents the amount of bias present for the three models when $\rho_{xx} = .8$ and $N = 50$. Across all of the conditions investigated, the change score method demonstrated negligible bias. It was unbiased regardless of sample size, population treatment effect, pretest reliability, or correlation between group membership and underlying ability. Unsurprisingly, ANCOVA was found to be the most biased of the three methods across many conditions. Two conditions drastically affected the amount of bias in ANCOVA's estimates. The first was the correlation of group membership with the underlying ability score.

When group membership was not correlated with the underlying ability measure, the ANCOVA's estimates were unbiased. However, as expected, with small, medium, or large correlations between pretest scores and the grouping variable, bias was demonstrated for all of the effect sizes. The amount of bias increased as the correlation between group membership and underlying ability increased, where relative bias was as high as 217% in the condition with the highest correlation between group membership and ability score. The second condition that affected ANCOVA's estimates was the reliability of the pretest and posttest scores. Specifically bias in the regression estimates decreased as the reliability increased. This result is expected because if the pretest and posttest scores' reliability is exactly equal to 1, the ANCOVA model will be equivalent to the EIV model.

_____

INSERT TABLE 1 ABOUT HERE

_____

_____

INSERT TABLE 2 ABOUT HERE

_____

There were three conditions that interacted to affect the amount of bias for the EIV method. Sample size, reliability of the covariate (ρ_{xx}), and correlation between group membership and underlying ability. For smaller sample sizes (N = 20 or 50), the EIV method demonstrated little to no relative bias across the effect sizes when the correlation between group allocation and ability score was low. However, bias increased as the correlation increased when the reliability of the covariate was .5. In the largest correlation condition with pretest reliability of .5, the EIV demonstrated biased results up to 92% in the N = 50 condition. Given the

relationship with sample size and bias, we investigated bias under a large sample size (N = 1000), to see whether bias results approached zero. The interaction of sample size, ρ_{xx}, and correlation between group membership and underlying ability can be seen in Figure 1. Relative bias estimates at each correlation have been averaged over the four population treatment effect sizes (relative bias cannot be calculated with an effect size of 0). The figure demonstrates that in large sample sizes (e.g., N = 1000) there is no bias in any conditions, but for all of the other sample sizes, the amount of bias depends on the particular condition (i.e., combination of ρ_{xx} and correlation of group membership and ability).

INSERT FIGURE 1 ABOUT HERE

Bias when True Score Reliability Differs from Estimated Reliability (EIV only)

Given the importance of covariate reliability for the EIV method, we also investigated deviations of estimated covariate reliability from the true reliability to determine how accurate researchers must be for the EIV method to provide unbiased results. Table 3 presents a subset of the bias results for the EIV method when reliability estimates differ from true score reliability for N = 50. With random group assignment (correlation between group membership and underlying ability = 0), the results continued to show almost no bias regardless of degree of inaccuracy for reliability estimation. When the correlation was greater than zero, even small deviations (e.g., .05) from the true reliability resulted in bias across most of the conditions. More bias was present when the effect size was nonzero, and the amount of bias increased as the correlation between group membership and underlying ability increased. Increasing the correlation between underlying ability and group allocation resulted in bias regardless of the amount of deviation

from the true reliability across most of the conditions. In fact, at the highest correlation tested, underestimating or overestimating the true reliability of the covariate often resulted in more biased estimates than if one were to have used the ANCOVA method instead. This pattern of results occurred regardless of sample size.

INSERT TABLE 3 ABOUT HERE

Type I Error Rates

Table 4 displays the Type I error results across the presented conditions for $N = 50$. Here, ANCOVA's Type I error rates were close to the nominal level (.05) when random assignment was used for group membership. However, empirical Type I error rates were found to be highly inflated at the largest correlation between group and underlying ability (e.g., rates as high as .50 when ρ_{xx} was .5). Increasing the reliability to .8 decreased the amount of Type I error, but rates were still found to be much higher than the nominal level. Both the change score and EIV methods were found to accurately maintain the empirical Type I error rates at the nominal level across all of the conditions investigated.

INSERT TABLE 4 ABOUT HERE

Power Rates

Figure 2 displays four graphics presenting power results for several of the conditions with $N = 50$. The top left figure presents results for the ANCOVA, change scores, and EIV methods

when group membership was randomly assigned and $\rho_{xx}$was .5. Here, the ANCOVA was found to have the most power of the three procedures whereas the EIV method displayed the lowest power. The change score method consistently displayed higher power than the EIV, although the power advantage was relatively minor. In the top right figure, the correlation between group and ability remains at 0, but ρ_{xx} was .8. Increasing the reliability increased the power for each of the methods and lessened the power difference between ANCOVA and the other two methods.

The bottom left graph of Figure 2 examined power results when ρ_{xx} = .5 but the correlation between group membership and underlying ability was .6. This condition resulted in a different pattern of results from what was discussed above. When the effect size was negative, the change score method demonstrated the most power and the EIV and ANCOVA methods' power results were comparable to one another but significantly lower than the power of the change score approach. When the effect size was positive, the change score method continued to demonstrate higher power than the EIV method, but the ANCOVA's power exceeded that of the change score and EIV methods. It is important to note, however, that the ANCOVA's power results cannot be validly interpreted because the empirical Type I error results were almost five times the nominal level. The bottom right figure presents power results with a correlation of .6 and ρ_{xx} of .8. Increasing the reliability from .5 to .8 in the correlated group membership condition resulted in the same pattern of results as when $\rho_{xx}$was .5, although the power of the EIV and change score procedures increased.

INSERT FIGURE 2 ABOUT HERE

Discussion

Researchers are often interested in assessing a change in behaviour before and after some intervention. In order to provide valid claims about the effectiveness of the intervention, comparison to a control group is beneficial. While randomization is the only way to conclude that any differences between the groups were solely due to the effect of the intervention, in practice, some groups occur naturally or randomization may not be possible. Traditionally, when utilizing quasi-experimental designs, researchers were required to make a decision between two available statistical approaches: ANOVA (or *t*-test) on change scores or ANCOVA?

Culpepper and Aguinis (2011) suggested that the EIV method might be a viable alternative to ANCOVA. The benefit of using the EIV in lieu of change scores or ANCOVA is that estimates will be unbiased under many different conditions. Based on the current study, it was demonstrated that with larger sample sizes ($N > 100$) the EIV method showed little bias independent of pretest/posttest reliability, whether the groups displayed differences at pretest or not, and across differing correlations between group membership and underlying ability. However, several issues were demonstrated with using the EIV method instead of change scores or ANCOVA. With smaller sample sizes, there were conditions when the EIV's estimates were biased. Given the complexity of the EIV model, it is possible that with small sample sizes, there were slight differences between the empirical modification to the covariance matrix and the model-implied changes. Another potential downfall with using the EIV method for the two group-two time point design is that the EIV method's power was consistently lower than that of the change score model or ANCOVA across a wide range of conditions. While the power advantage of the change score or ANCOVA (when group membership was randomly allocated) may not always have been much larger, there were some instances in which the power difference

was marked. The last potential issue with using the EIV method in the current study was that bias was large when the researcher-estimated value of pretest/posttest reliability differed (even by small amounts) from the true reliability score. This bias was magnified when there was a relationship between group membership and one's underlying ability being measured at pretest and posttest. In practice, it is unlikely that researchers will have a precise estimate of reliability, and therefore may unknowingly bias their results by using the EIV method with covariate reliability estimates that are marginally different from their true value. Aside from the differences in power and bias using the EIV method, a theoretical conundrum arises by using the EIV method as a catch-all to reduce model bias and maintain Type I error rates. Using the EIV method takes away the necessity of thinking critically about the appropriateness of one's statistical analysis in this two group-two time point research design. After all, change scores and ANCOVA are testing two different null hypotheses, and this necessarily means that the language around interpretation and implication of results should be related back to the research hypotheses being tested.

Conclusion

Based on the results of the current research, we would not recommend using the EIV in place of the change score or ANCOVA methods with a two group, pretest-posttest research design. In smaller sample sizes, the EIV method is more biased and less powerful than the change score method when there are nontrivial pretest differences between groups. At larger sample sizes, the estimates are similar, but there is no advantage to using the EIV method in lieu of change scores. When trivial pretest differences exist, the ANCOVA is more powerful than the EIV (or change score) method and all of the methods are unbiased. Choosing the EIV method over change scores or ANCOVA also poses an additional risk. Researchers may obtain biased

parameter estimates if the researcher specifies a pretest reliability that is not exactly equal to the true population reliability or a researcher has a small sample size. As such, we recommend that researchers continue to think critically about their research design and hypotheses and choose the change score model with nontrivial pretest group differences or ANCOVA with trivial pretest differences, instead of implementing the EIV model for a research design with two time points and two groups.

References

Allison, P. D. (1990). Change scores as dependent variables in regression analysis. *Sociological methodology, 20*, 93-114.doi: 10.2307/271083

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*, 144-152.

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. doi: 10.1177/00131640021970970

Cribbie, R. A. & Jamieson, J. (2004). Decreases in posttest variance and the measurement of change. *Methods of Psychological Research, 9*, 37-55.

Cronbach. L., & Furby, L.(1970). How should we measure "change": Or should we? *Psychological Bulletin, 74*, 68-80. doi: 10.1037/h0029382

Culpepper, S. A., & Aguinis, H. (2011). Using analysis of covariance (ANCOVA) with fallible covariates. *Psychological Methods, 16*, 166-178. doi: 10.1037/a0023355

Fitzmaurice, G. (2001). A conundrum in the analysis of change. *Nutrition, 17*, 360-361. doi: 10.1016/S0899-9007(00)00593-1

Fuller, W. A. (1980). Properties of some estimators for the errors-in-variables model. *Annals of Statistics, 8*, 407–422. doi:10.1214/aos/1176344961

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

Jamieson, J. (1999). Dealing with baseline differences: Two principles and two dilemmas. *International Journal of Psychophysiology, 31*, 155-161.doi: 10.1016/S0167-8760(98)00048-8

Jamieson, J. (2004). Analysis of covariance (ANCOVA) with difference scores. *International Journal of Psychophysiology, 52*, 277-283. doi: 10.1016/j.ijpsycho.2003.12.009

Linn, R. L. & Slinde, J. A. (1977). The determination of the significance of change between pre- and posttesting periods. *Review of Educational Research, 47*, 212-150. doi: 10.1037/0022-006X.59.1.27

Lord, F. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin, 68*, 304-305.

Maris, E. (1998). Covariance adjustment versus gain scores-revisited. *Psychological Methods, 3*, 309-327. doi: 10.1037/1082-989X.3.3.309

Miller, G. M. & Chapman, J. P. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology, 110,* 40-48.doi: 10.1037/0021-843X.110.1.40

Porter, A. C., & Raudenbush, S. W.(1987). Analysis of covariance: Its model and use in psychological research. *Journal of Counseling Psychology, 34*, 383-392.doi: 10.1037/0022-0167.34.4.383

R Development Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org.

Rogosa, D. (1988). Myths about longitudinal research. In K. W. Schaie, R. T. Campbell, W. M. Meredith, & S. C. Rawlings (eds.), *Methodological issues in aging research* (p. 171-209). New York, NY: Springer.

Rogosa, D. R. (1995). Myths and methods: "Myths about longitudinal research" plus supplemental questions. In J. M. Gottman (Ed.), *The analysis of change*. Mahwah, NJ, Lawrence Erlbaum Associates, Inc.

Schafer, W. D. (1992). Analysis of pretest-posttest designs. *Measurement and Evaluation in Counseling and Development, 25,* 2-4.

Senn, S. (2006). Change from baseline and analysis of covariance revisited. *Statistics in Medicine, 25,* 4334-4344.doi: 10.1002/sim.2682

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.doi:10.1080/01621459.1974.10480223

Wright, D. B. (2006). Comparing groups in a before-after design: When *t* test and ANCOVA produce different results. *British Journal of Educational Psychology, 76,* 663-675. doi: 10.1348/000709905X52210

Table 1

Relative Bias (in percentages) for N =50 when ρ_{xx} = .5

	$\rho = 0$			ρ = .2			ρ = .4			ρ = .6		
δ	CS	A	EIV	CS	A	EIV	CS	A	EIV	CS	A	EIV
-.5	.300	.400	.500	1.30	27.3	2.80	.500	63.9	5.20	1.10	106	19.6
-.25	1.80	1.90	2.10	1.40	57.8	.300	.800	128	11.7	.100	214	33.6
0	**.003**	**.001**	**.004**	**.005**	**.147**	**.001**	**.002**	**.325**	**.025**	**.002**	**.534**	**-.025**
.25	.700	.200	1.10	2.00	54.5	6.00	2.20	126	16.1	2.60	217	44.8
.5	.400	.200	.400	.500	28.3	1.10	.300	63.7	5.40	.700	107	31.4

Note: ρ = the correlation between group membership and underlying ability; δ= effect size, the population treatment effect; CS = change score; A = ANCOVA, EIV= errors in variables; raw bias is used when $\delta = 0$ since we cannot calculate the relative bias with an effect of 0 (highlighted in bold)

Table 2

Relative Bias (in percentages) for N =50 when ρ_{xx} = .8

	$\rho = 0$			ρ = .2			ρ = .4			ρ = .6		
δ	CS	A	EIV	CS	A	EIV	CS	A	EIV	CS	A	EIV
-.5	.500	.400	.300	.300	14.6	.400	.500	34.0	1.40	.200	63.1	3.20
-.25	1.80	1.30	1.90	.500	29.9	.100	.800	68.0	2.80	.500	126	5.10
0	**.003**	**.002**	**.003**	**.004**	**.068**	**.006**	**.001**	**.170**	**.008**	**.002**	**.311**	**.019**
.25	.800	.900	.900	1.40	27.5	1.90	1.10	69.4	1.60	1.50	126	8.70
.5	.100	.200	.200	.500	13.9	1.00	1.20	35.0	.100	.200	63.4	2.50

Note: ρ = the correlation between group membership and underlying ability; δ= effect size, the population treatment effect; CS = change score; A = ANCOVA, EIV= errors in variables; raw bias is used when δ = 0 since we cannot calculate the relative bias with an effect of 0 (highlighted in bold)

Table 3

EIV's Relative Bias (percentages) when Estimated Reliability Differs from True Reliability when ρ_{xx} = .5 (N = 50)

	$\rho = 0$			$\rho = .2$			$\rho = .4$			$\rho = .6$		
Diff	$\delta = 0$	$\delta = .25$	$\delta = .50$	$\delta = 0$	$\delta = .25$	$\delta = .50$	$\delta = 0$	$\delta = .25$	$\delta = .50$	$\delta = 0$	$\delta = .25$	$\delta = .50$
-.20	**.003**	1.30	.400	**.270**	103	38.9	**.767**	198	58.1	**.533**	880	131
-.10	**.001**	1.60	1.00	**.085**	30.5	16.5	**.210**	89.9	61.3	**1.78**	216	149
-.05	**.001**	1.90	1.60	**.036**	16.6	8.70	**.133**	51.0	26.3	**.220**	5.4	75.1
.05	**.006**	.400	.400	**.021**	9.70	5.70	**.040**	17.0	7.70	**.062**	27.0	8.80
.10	**.001**	1.90	.900	**.047**	16.4	9.30	**.096**	37.1	20.2	**.174**	69.5	34.4
.20	**.001**	4.10	.200	**.080**	30.3	16.3	**.178**	72.6	35.2	**.329**	131	67.3

Note: ρ= correlation between group membership and underlying ability; Diff= the difference between true reliability and estimated reliability (negative difference is an underestimate of reliability); δ= effect size: population treatment effect; raw bias is used when $\delta = 0$ since we cannot calculate the relative bias with an effect of 0 (highlighted in bold)

Table 4

Type I Error Rates (N = 50)

	ρ_{xx} = .5			ρ_{xx} = .8		
ρ	Change Score	ANCOVA	EIV	Change Score	ANCOVA	EIV
0	.048	.045	.028	.050	.049	.046
.2	.050	**.088**	.032	.051	.068	.045
.4	.052	**.235**	.037	.053	**.149**	.045
.6	.049	**.502**	.045	.049	**.325**	.039

Note:ρ_{xx} is the reliability of the covariate;ρ= correlation between group membership and underlying ability. Bolded values are considered liberal if greater than .075 (e.g., Bradley, 1978)

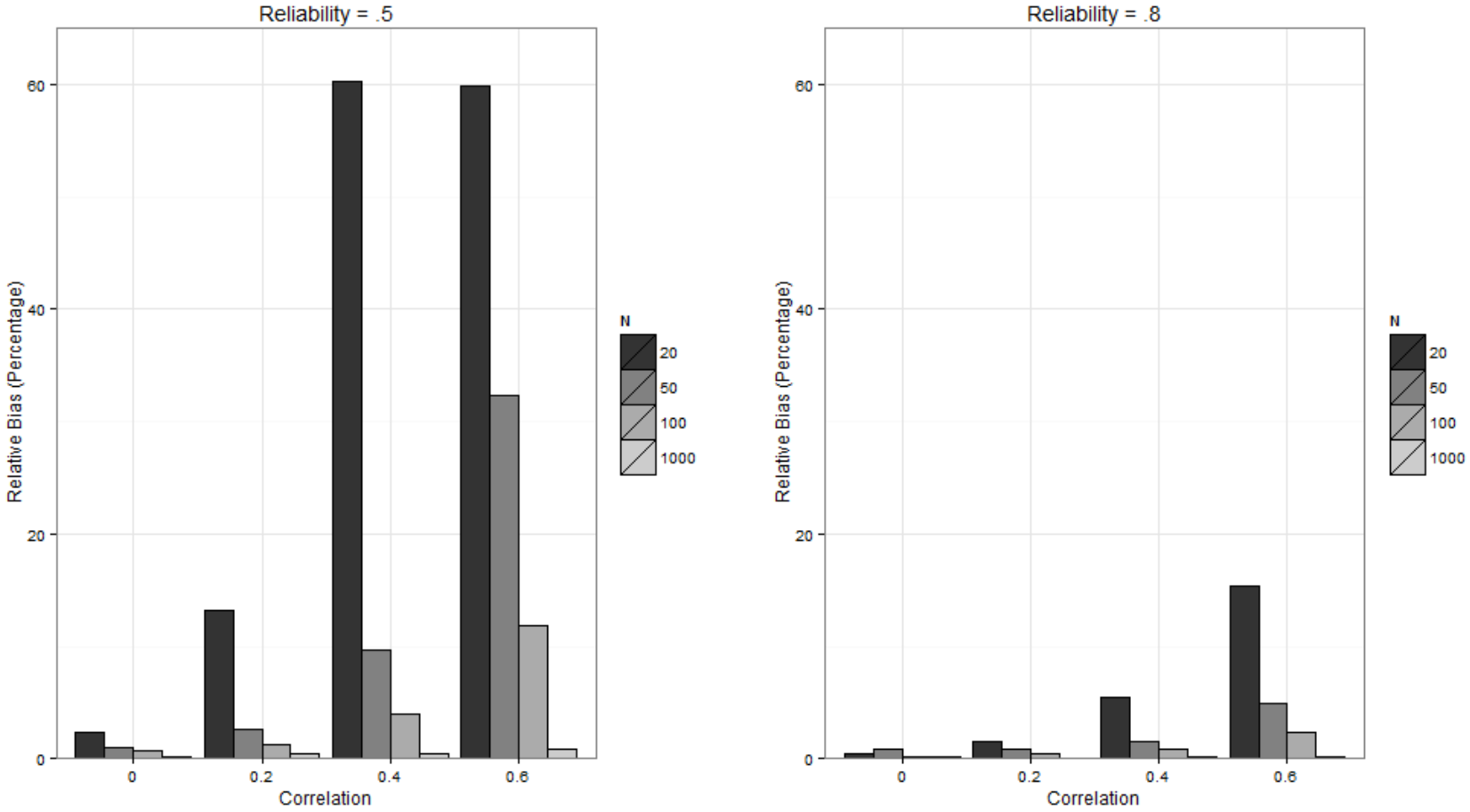

Figure 1: Amount of relative bias in the EIV method. Bias has been averaged across the population treatment effect size at each level of the correlation. The left figure displays the interaction of sample size and correlation between ability and group membership when ρ_{xx} = .5. The right figure displays the interaction of *N* and the correlation between group membership and ability when ρ_{xx} = .8.

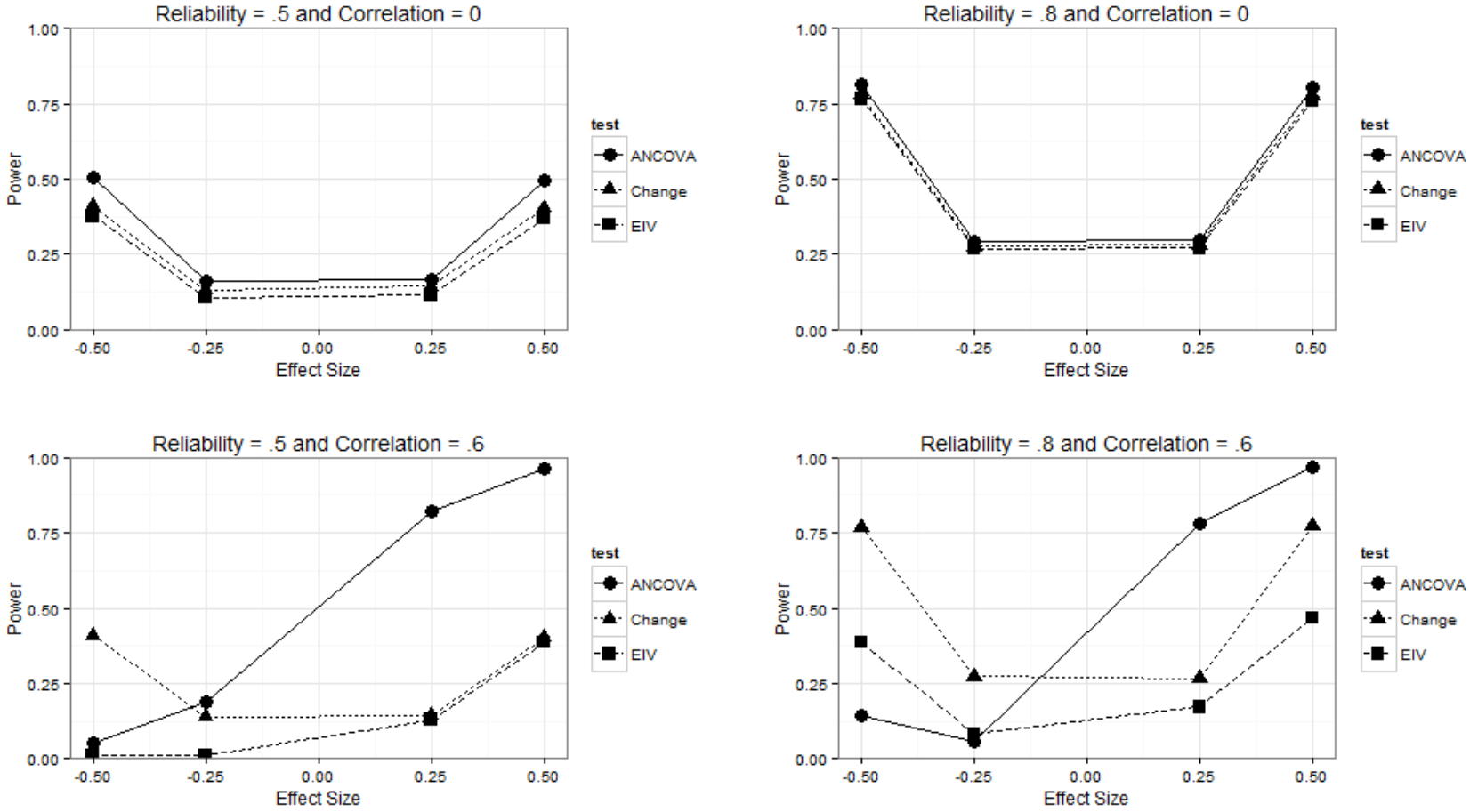

Figure 2. Four power plots when N = 50. The first row represents a correlation of group membership and ability of 0 whereas the second row is .6. The first column represents ρ_{xx} = .5 and the second column is ρ_{xx} = .8. Note that although ANCOVA appears to demonstrate a power advantage in the figures in the second row, the power results should not be interpreted as such, because its Type I error rates are extremely liberal.