Published in final edited form as:

Educ Psychol Meas. 2013 February ; 73(1): 47–62. doi:10.1177/0013164412450574.

A Monte Carlo Comparison Study of the Power of the Analysis of Covariance, Simple Difference, and Residual Change Scores in Testing Two-Wave Data

Yasemin Kisbu-Sakarya, David P. MacKinnon, and Leona S. Aiken

Arizona State University

Behavioral research often investigates two wave data in which a treatment group and a control group are assessed at baseline (i.e., before treatment) and after the treatment is given to the treatment group. However a controversial issue arises for the researcher when analyzing the data--choosing the statistical approach that would allow him/her to accurately test the relevant null hypotheses of no treatment effect. The three most common statistical approaches used by researchers to test hypotheses in the analysis of change are analysis of covariance (ANCOVA), difference score, and residual change score methods. Even though the differences among the three methods have been investigated by statisticians, there is still need for more simulation studies as pointed out by Rogosa (1995) and Zumbo (1999). For instance, which method is more sensitive to detect the treatment effect under which conditions is not yet fully addressed with simulation studies. A recent study by Petscher and Schatschneider (2011) compared simple difference score method of measuring differential change with ANCOVA by examining how statistical power to detect treatment effects was influenced by stability between pre- and posttest measurements, non-normality and variances of pretest and posttest scores in a randomized experiment. The authors reported that ANCOVA always had greater statistical power than the difference score method; and in the case of equal pretest and posttest variances, the discrepancy in statistical power between the methods decreased as stability increased. In this study, we extended the Petscher and Schatschneider paper to compare the three methods (i.e., simple difference score, ANCOVA and residual change score) in terms of statistical power and Type 1 error rate by studying the effect of reliability of scores and its interaction with stability between pre- and posttest measurement, and non-randomization using a Monte Carlo simulation.

Different Approaches

Suppose a researcher wants to examine the effect of a program (variable G) on reading ability. Reading ability is first measured at baseline to obtain a pretest score (variable X). Then the treatment is given to half the participants; and one year later, the posttest score on the outcome variable (variable Y) is measured. The *simple difference score* approach takes the difference between the pretest and posttest scores and then regresses that difference score () on the binary treatment variable G (0=control, 1=treatment):

$$\Delta = \beta_1 G + \varepsilon_1 \quad (1)$$

In contrast to the difference score, *ANCOVA* treats the pretest score as a covariate which may be an uncontrolled source of variation and may influence the posttest score. Thus, the posttest score Y is regressed on both the treatment variable G and pretest score X as in equation 2. In *ANCOVA*, the within class regression coefficient is estimated within each group and then the pooled within class regression coefficient ($\beta_{2pooled\ within}$) is computed by pooling the regression coefficients in the two groups assuming that the data exhibit homogeneity of within class regression. *ANCOVA* partials out the effect of the pretest score on the within and between class variance on the posttest score. By removing the within class variance that is due to differences among participants at pretest, *ANCOVA* increases the power to detect a treatment effect because of a smaller error term if the baseline covariate has a substantial association with the posttest score. However, the between class adjustment of *ANCOVA* may have different effects on power depending on the design. In a randomized trial, the expected baseline difference on any covariate is zero, and there is no expected adjustment of between class variation. However, in a non-randomized design, the *ANCOVA* between class adjustment can lead to an increase or decrease in statistical power depending on which group has a higher score at pretest (Cohen, Cohen, West, & Aiken, 2003; Bonate, 2000).

$$Y = \beta_1 + G + \beta_{2pooled\ within} X + \varepsilon_2 \quad (2)$$

where $X = \beta_3 G + \varepsilon_3$, reflecting the association between X and group membership.

Additionally, it should be noted that the difference score model tests the null hypothesis of no difference across groups in the raw change from pretest to posttest (i.e., $H_0: \Delta = 0$). However, *ANCOVA* model tests the null hypothesis of no difference between the treatment and control posttest scores, conditional on the pretest scores. Another difference between the two methods is that the difference score approach requires the use of the same measurement instrument so the baseline and posttest scores have the same units in order for the difference score to have the same metric.

In the *residual change score* approach, first the predicted posttest scores are estimated by regressing the posttest score Y on the pretest score X , ignoring group membership. Thus the regression coefficient for predictor X is the total regression coefficient of Y on X . Then the residual change scores are computed by subtracting the predicted posttest scores from the observed posttest scores. The residual change scores are then regressed on the binary treatment variable G . Theoretically the residual change score approach is similar to *ANCOVA* since both analyses adjust for pretest measurement. However, the statistical adjustment that generates the residuals in the residual change score method uses the *total regression coefficient* for Y on X , whereas *ANCOVA* adjustment is based on the regressions of Y on X within each group pooled across groups, the *pooled within class regression coefficient* of Y on X .

Reliability and Stability

The reliability of the pretest and posttest measures has been the most problematic issue for its use in two wave models, specifically for the analysis of difference score. The discussion is based primarily on Lord's (1956) assertion that "differences between scores tend to be much more unreliable than the scores themselves". Using the reliability definition in classical test theory, where the observed score X is the sum of true score T and error score ε , reliability of X is $\rho_{XX} = \sigma^2_T / \sigma^2_X$. Then it can be shown that the reliability of difference score () is equal to (Zimmerman & Williams, 1982):

$$\rho_{\Delta\Delta} = \frac{\lambda\rho_{XX} + \lambda^{-1}\rho_{YY} - 2\rho_{XY} + 2\rho(\varepsilon_X, \varepsilon_Y)[(1 - \rho_{XX})(1 - \rho_{YY})]^{1/2}}{\lambda + \lambda^{-1} - 2\rho_{XY}} \quad (3)$$

where ρ_{XY} represents the correlation between the first and second measurement; and $\lambda = \sigma_X / \sigma_Y$. If we assume that $\rho(\varepsilon_X, \varepsilon_Y) = 0$ then equation 3 reduces to:

$$\rho_{\Delta\Delta} = \frac{\lambda\rho_{XX} + \lambda^{-1}\rho_{YY} - 2\rho_{XY}}{\lambda + \lambda^{-1} - 2\rho_{XY}} \quad (4)$$

If, in addition, one assumes that the pretest and posttest measures are parallel forms of a test and that $\sigma_X = \sigma_Y$ and $\rho_{XX} = \rho_{YY}$, then (4) reduces to:

$$\rho_{\Delta\Delta} = \frac{\rho_{XX} - \rho_{XY}}{1 - \rho_{XY}} \quad (5)$$

Based on equation 5, Rogosa et al. (1982) argued that if measurement stability (i.e., the correlation between the pretest and posttest scores) is low then the difference score will be reliable. For example, using the assumptions for equation 5, if the reliability of pretest and posttest scores is equal to .7 and their correlation ρ_{XY} is .5, the reliability of their difference score will be .4. If their correlation ρ_{XY} is .6, then the reliability of difference score decreases to .25 (Rogosa, 1988). Therefore, it is argued that for a difference score to be reliable (i.e., able to distinguish among individuals on a particular score), there should be individual differences in true change scores. If everyone changes by the same amount (i.e., high stability case), all variation in that measure is a function of measurement error resulting in almost parallel growth curves (Rogosa et al., 1982; Zumbo, 1999). Thus, it is concluded and mathematically shown that difference scores can be reliable and able to distinguish among individuals if the pretest and posttest scores are not highly correlated. However, it is demonstrated that this case is not relevant when testing for treatment differences using aggregated data (Thomas & Zumbo, in press). When the focus is not the individual but rather the group, the relevant measure is the divergence between the sampling distributions of the test statistic under the null and alternative hypotheses (i.e., noncentrality parameter) rather than classical reliability (see Thomas & Zumbo, in press for the mathematical illustration and discussion). Therefore, the analysis of difference score may lead to good statistical power even when the reliability of the difference score is low when testing aggregated data. In this study, we used a Monte Carlo simulation to investigate how the statistical power and Type 1 error rate of the difference score method is influenced by

reliability of measures and stability in comparison to ANCOVA and residual change score approaches in a pretest-posttest control group design.

Baseline imbalance

Researchers generally do not pay attention to the existence of imbalance between groups at baseline when choosing which analysis method to conduct. However, different methods produce different results in case of baseline imbalance. For instance, ANCOVA is often perceived as being equivalent to the residual change score approach since both analyses statistically adjust for the covariate pretest measure, albeit with different adjustment models. However, it has been shown that the two approaches differ mathematically (Maxwell, Delaney, & Manheimer, 1985). This difference arises from the different use of regression slopes. As described above the most commonly used version of the residual change score method obtains the residual score by using the regression coefficient for the total sample combined into one group (that is, the regression coefficient of Y on X across all cases, ignoring group membership) whereas ANCOVA uses the pooled within slope across groups. Consequently, if the groups differ at baseline, then ANCOVA and residual change score lead to different results. Use of the total slope in residual change score analysis leads to an underestimated treatment effect, thus leading to incorrect estimates of change. Because the test becomes too conservative, it also produces invalid confidence intervals. The bias increases as the baseline imbalance increases and sample size decreases (Maxwell, Delaney, & Manheimer, 1985; Forbes & Carlin, 2005).

Moreover, it has been also shown that ANCOVA and difference score would perform differently in case of baseline imbalance (e.g., Lord's ANCOVA Paradox). In case of no baseline imbalance, ANCOVA has good statistical power since it assumes the groups are equivalent at baseline (Fitzmaurice, Laird, & Ware, 2004). It is illustrated that ANCOVA will approximately be 30% more precise than the difference score method in case of no baseline imbalance and no measurement error (Oakes & Feldman, 2001; Fitzmaurice et al., 2004). However, as baseline imbalance increases, ANCOVA may lead to biased results. The baseline imbalance may influence the statistical performance of ANCOVA depending on which group has a higher a pretest score and whether the scores are expected to increase or decrease at posttest (Cribbie & Jamieson, 2000; Jamieson, 1999). The reason ANCOVA can lead to different conclusions under these different circumstances is due to ANCOVA's approach for adjusting for baseline scores. Whenever the pooled within slope across groups is less than 1, ANCOVA expects the posttest means to come closer together (i.e., regress to the mean). Thus, in case the treatment group has a higher pretest mean than the control group, ANCOVA is more likely to detect a change if there is an increase from pretest to posttest and less likely to detect the change if there is a decrease from pretest to posttest, since ANCOVA is biased towards finding a significant effect when the means behave contrary to its expectation of regression to the mean (readers are referred to Figure 4 in Jamieson, 1999 for a graphical illustration of this directional bias). Furthermore, Oakes and Feldman (2001) argues that difference score method should yield better statistical power in case of non-randomization since its power calculations (shown by the minimum detectable difference formula) is not affected by baseline non-equivalence.

This study compares the difference score, analysis of covariance, and residual change score methods in testing the group effect for two-wave data in terms of power and Type 1 error rates. Following the literature, we concentrate on addressing the most commonly encountered issues related to statistical performance of the tests – reliability, stability and baseline imbalance. Although, previous studies mathematically illustrated and discussed the effects on outcomes of reliability, stability, and non-randomization, there is still need for addressing the statistical power performance across the three methods simultaneously in a simulation study. Therefore, we conducted a Monte Carlo simulation study to test the empirical statistical power and Type 1 error rate performance of the three methods under the above-mentioned circumstances.

Method

The SAS 9.2 programming software was used for the simulation. Variables were generated from the normal distribution using the RANNOR function. The treatment condition variable G was simulated to be binary for an intervention design (0 = control, 1 = treatment). Pretest scores (variable X) and posttest scores (variable Y) were always simulated to be continuous. The regression parameter reflecting the relation between G and Y (i.e., treatment effect) was varied as $-.59, -.39, -.14, 0, .14, .39$, and $.59$ corresponding to Cohen's criteria for zero, small (2% of the variance), medium (13% of the variance), and large (26% of the variance) effect sizes, respectively (Cohen, 1988, pp. 412–414). The regression parameter (β_3) reflecting the relation between G and X was varied as $0, -.14, .14, -.59$, and $.59$ to investigate an experimental design with perfect randomization (i.e., no baseline imbalance), small baseline imbalance where the control group has higher pretest score and vice versa, and large baseline imbalance where the control group has higher pretest score and vice versa conditions. Stability between the pretest score X and posttest score Y was varied by setting the regression coefficient to predict Y from X to $.3, .5, .7$ and 1 for low, medium, high, and perfect stability conditions. Reliabilities of X and Y are set to be equal, and they varied as $.5, .8$, and 1 to simulate low, high, and perfect reliability. The simulation was conducted for sample sizes of 50, 100, 200, and 500 per condition to be comparable to the commonly used sample sizes in the social sciences. We note that because the simulation mimics an ANCOVA model, it may intrinsically favor the ANCOVA approach. To summarize, a 4 (sample size) $\times 7$ (treatment effect) $\times 4$ (stability) $\times 3$ (reliability) $\times 5$ (baseline imbalance) factorial design was used in the simulation study. A total of 1000 replications of each condition were conducted.

For the ANCOVA method, the posttest score Y was regressed on both X and G. For the difference score model, the difference score Y-X was computed and then was regressed on G. For the residual change score method, the posttest score Y was regressed on the pretest score X; then the residual change score, which was obtained by subtracting the predicted Y scores from the observed Y scores, was regressed on G.

Empirical power and type 1 error rates were calculated using 5% level of significance, since it is the most common value used in most research areas. For each condition, the proportion of times that the treatment group variable G was statistically significant in predicting change in 1000 replications was computed. When the regression parameter reflecting the relation

between G and Y (i.e., β_I) was equal to zero, the proportion of replications in which the null hypothesis of no treatment effect was rejected was an estimate of Type I error rate. When the regression parameter was greater than zero, the proportion of times across the replications in a cell that each method yielded a significant treatment effect presented the estimate of statistical power.

Results

Reliability and stability

Table 1 presents the results for the effect of reliability and stability on statistical power and Type 1 error rates for the ANCOVA, difference score, and residual change score methods in the case of zero baseline imbalance (i.e., perfect randomization). To ease the interpretation, we only included the positive values of treatment effect into the Table 1. In terms of Type 1 error rates, all three methods were similar with rates between the nominal values .04 and .07 across all conditions. There was no effect of reliability or stability on Type 1 error rates for either method.

As expected, ANCOVA and residual change score methods produced nearly equivalent results when the groups were balanced at baseline. All three methods had low power for small treatment effect sizes and small samples. In the case of perfect reliability, all three methods attained power of .80 or greater for large effects with more than 100 participants or medium effects at 200 participants. When the reliability of measures was less than perfect (.5 or .8), all three methods approached .80 power to detect a large effect with more than 200 participants or a medium effect for a sample size of 500. Furthermore, ANCOVA and residual change score methods always produced higher statistical power than the difference score method in all conditions when stability was less than 1. Whereas when stability was equal to 1, the methods did not differ in terms of statistical power.

For low reliability of .5, the power of the ANCOVA method declined with increasing stability in case of medium and large effect sizes. When reliability was equal to .8 or 1, there was no consistent effect of stability on power. In contrast, for the difference score method, when the reliability was .5, there was no effect of stability on power; whereas when reliability was .8 or 1, power increased as stability increased for medium and large effect sizes. The finding that, in contrast to ANCOVA, statistical power of the difference score method increased as stability increased was a result of the correlation of pretest scores with difference scores (r_x) (Bonate, 2000). The difference score method is less powerful as r_x becomes more negative, but ANCOVA is not influenced by r_x (Jamieson, 1995). In the current study, in the low stability case, r_x was more negative compared to high stability case leading to a loss of power. Figure 1 illustrates the pair-link plots for two randomly selected data sets from the simulated data sets. Figure 1a presents a data set from the low stability condition and Figure 1b presents a data set from the high stability condition. The r_x in Figure 1a was more negative leading to a loss of power to measure the change. The formula showing r_x as a function of the correlation between pretest and posttest scores is provided below (Cohen, Cohen, West, & Aiken, 2003, p. 59).

$$r_{x\Delta} = \frac{r_{xy}sd_y - sd_x}{\sqrt{sd_y^2 + sd_x^2 - 2r_{xy}sd_ysd_x}} \quad (6)$$

If X and Y are standard normal variables with a mean 0 and standard deviation 1, then equation (7) reduces to:

$$r_{x\Delta} = \frac{r_{xy} - 1}{\sqrt{2 - 2r_{xy}}} \quad (7)$$

Baseline imbalance

The effect of baseline imbalance on Type 1 error rates and statistical power of the three methods were examined for different levels of program effect size and reliability. To ease the interpretation, the levels of treatment effect were categorized to three categories: no effect (i.e, regression parameter estimate for treatment effect = 0), increase in the outcome variable from pretest to posttest (i.e, regression parameter estimates for treatment effect = .14, .39, and .59 for small, medium and large effects), and decrease in the outcome variable from pretest to posttest (i.e, regression parameter estimates for treatment effect = -.14, -.39, and -.59 for small, medium and large effects). We compared ANCOVA and residual change score methods with the classical model of the difference score method which assumes $\beta_2=1$. Figures 2 and 3 show the results for Type 1 error and statistical power rates of all three methods across the levels of baseline imbalance, program effect, and reliability averaged over all sample size conditions. As can be seen in Figure 2, the Type 1 error rates for all three methods were approximately .05 when there was no baseline imbalance across all conditions of reliability. However, in case of baseline imbalance, ANCOVA and residual change score methods produced large Type 1 error rates when reliability was less than perfect. When the reliability was perfect, ANCOVA and residual change score methods produced Type 1 error rates of .05. On the other hand, Type 1 error rates for the difference score method were not influenced by neither baseline imbalance nor reliability. Figure 3 shows that the statistical power of the difference score method increased as reliability increased. The ANCOVA and residual change score methods had similar power rates across all conditions, with ANCOVA having slightly more power than the residual change score method. When reliability was less than perfect, ANCOVA had more power than the difference score method when there was an increase from pretest to posttest and a positive baseline imbalance (i.e., treatment group had higher pretest scores than the control group), or when there was a decrease from pretest to posttest and a negative baseline imbalance, and vice versa. In case of perfect reliability, the statistical power of ANCOVA did not differ from the difference score method.

Discussion

The pretest-posttest control group design is a common research design. In this study, we addressed several issues relevant to three methods used to analyze change: analysis of covariance, difference scores, and residual change scores. First, we demonstrated how stability and measurement reliability affected the statistical power performance of difference

scores compared with ANCOVA and residual change score methods. Similar to Petscher and Schatschneider (2011), we found that the ANCOVA and difference score methods did not differ in terms of Type 1 error rate performance. However, Petscher and Schatschneider (2011) found that the statistical power of the difference score method decreased when stability increased in the case of equal pretest and posttest variances, which was contrary to the results of the current study. According to the current study results for the case of high reliability, the power of the difference score method increased as stability increased. Our finding was explained by the change in the correlation between baseline and difference scores (r_x), specifically that this correlation became increasingly negative as the pretest-posttest correlation r_{xy} decreased. However, an important difference between the Petscher and Schatschneider (2011) and the current study is that in Petscher and Schatschneider (2011) study, the r_x , rather than being a function of r_{xy} as shown by equation 7, was simulated by using the pretest and posttest variances; and accordingly r_x was presumed to be zero when the pretest and posttest variances were equal. Estimating r_x independent of r_{xy} may lead to inaccurate conclusions.

The current simulation study furthermore showed how different methods for analysis of change may lead to dissimilar results under baseline imbalance. The difference score method had similar statistical power performance compared to ANCOVA and residual change score methods when there was no baseline imbalance; and its statistical power performance was not influenced by baseline imbalance, yet increased as reliability increased. On the other hand, ANCOVA and residual change score methods produced differential power under baseline imbalance when reliability was less than perfect. One of the main goals of this paper was to inform researchers how the properties of their data will influence the statistical power of their method of choice and allow them to make a more informed choice and interpret their results cautiously. Nevertheless, we should note that the method researchers will choose to analyze their data mainly depends on the research questions they want to investigate. If a researcher is interested in answering the question “which group changed more?”, then the difference score method will be selected as the mean of analysis. However, if the researcher is interested in a conditional question such as “if the groups have had the same pretest score, which would have changed more?”, then ANCOVA will be the appropriate method. Another controversy raised by Holland and Rubin (1983) is that when there is no randomization at baseline, there is no true control condition; thus, the control condition should be specified by the researcher in order to infer causality. And the different specifications of the control condition may then be the basis of getting different results. The researcher may for instance specify and estimate the control condition as the pretest score and use the difference score method or as the posttest score adjusted by pretest and use the ANCOVA method.

We covered specific topics in the analysis of pretest-posttest data. However, more topics merit investigation. Some of those topics include the effect of reliability in non-normal distributions, and binary outcomes. It appears that the discrepancies of power for different methods are the result of many interacting characteristics of both the design and the data.

Acknowledgments

This research was supported by the National Institute on Drug Abuse Grant R01DA009757.

References

- Bonate, PL. Analysis of Pretest-Posttest Designs. New York, USA: Chapman & Hall/CRC; 2000.
- Cohen, J. Statistical power for the behavioral sciences. Hillsdale, NJ: Erlbaum; 1988.
- Cohen, J.; Cohen, P.; West, SG.; Aiken, LS. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. New Jersey: Lawrence Erlbaum Associates; 2003.
- Cribbie RA, Jamieson J. Structural equation models and the regression bias for measuring correlates of change. *Educational and Psychological Measurement*. 2000; 60:893–907.
- Fitzmaurice, GM.; Laird, NM.; Ware, JH. Applied Longitudinal Analysis. Hoboken, NJ: Wiley; 2004.
- Forbes AB, Carlin JB. “Residual change” analysis is not equivalent to analysis of covariance. *Journal of Clinical Epidemiology*. 2005; 58:540–541. [PubMed: 15845342]
- Holland, PW.; Rubin, DB. On Lord’s paradox. In: Wainer, H.; Messick, S., editors. *Principals of modern psychological measurement*. Hillsdale, NJ: Erlbaum; 1983.
- Jamieson J. Measurement of change and the law of initial values: A computer simulation study. *Educational and Psychological Measurement*. 1995; 55:38–46.
- Jamieson J. Dealing with baseline differences: two principles and two dilemmas. *International Journal of Psychophysiology*. 1999; 31:155–161. [PubMed: 9987061]
- Lord FM. The measurement of growth. *Educational and Psychological Measurement*. 1956; 18:437–454.
- Maxwell SE, Delaney HD, Manheimer JM. ANOVA of residuals and ANCOVA: Correcting an illusion by using model comparisons and graphs. *Journal of Educational Statistics*. 1985; 10:197–209.
- Oakes JM, Feldman HA. Statistical power for nonequivalent pretest-posttest designs. *Evaluation Review*. 2001; 25:3–28. [PubMed: 11205523]
- Petscher Y, Schatschneider C. A simulation study on the performance of the simple difference and covariance-adjusted scores in randomized experimental designs. *Journal of Educational Measurement*. 2011; 48:31–43. [PubMed: 26379310]
- Rogosa, D. Myths about longitudinal research. In: Schaie, KW.; Campbell, RT.; Meredith, W.; Rawlings, SC., editors. *Methodological issues in aging research*. New York: Springer Publishing Company; 1988. p. 171-209.
- Rogosa, DR. Myths and methods: Myths about longitudinal research plus supplemental questions. In: Gottman, JM., editor. *The analysis of change*. Mahwah, NJ: Lawrence Erlbaum; 1995. p. 4-66.
- Rogosa D, Brandt D, Zimowski M. A growth curve approach to the measurement of change. *Psychological Bulletin*. 1982; 92:726–748.
- Thomas DR, Zumbo BD. Difference scores from the point of view of reliability and repeated measures ANOVA: In defense of difference scores for data analysis. *Educational and Psychological Measurement*. (in press).
- Zimmerman DW, Williams RH. Gain scores in research can be highly reliable. *Journal of Educational Measurement*. 1982; 19:149–154.
- Zumbo, BD. The simple difference score as an inherently poor measure of change: Some reality, much mythology. In: Bruce, Thompson, editor. *Advances in Social Science Methodology*. Greenwich, CT: JAI Press; 1999. p. 269-304.

(a)

(b)

Figure 1. Pair-link plots

(a) Individual time paths for low stability case with a sample size of 100. (For these time paths, the correlation between pretest and posttest scores is .15 and the r_x is $-.63$).

(b) Individual time paths for high stability case with a sample size of 100. (For these time paths, the correlation between pretest and posttest scores is .56 and the r_x is $-.41$).

Figure 2.
Mean Type I error rates within reliability conditions.

(a) posttest score > pretest score

(b) posttest score < pretest score

Figure 3.
Mean statistical power within reliability conditions.

Table 1

Type I Error Rates and Statistical Power for ANCOVA, Residual Change and Difference Score Methods in Case of No Baseline Imbalance

N=50		Residual change score				ANCOVA				Difference score			
		Reliability (ρ_{xx}, ρ_{yy})											
<i>Treatment β_1</i>	<i>Stability β_2</i>	.5	.8	1	.5	.8	1	.5	.8	1	.5	.8	1
Type I Error													
Zero	.3	.060	.065	.056	.059	.065	.056	.048	.053	.059			
	.5	.052	.057	.052	.052	.059	.053	.058	.052	.055			
	.7	.051	.055	.070	.052	.054	.071	.044	.058	.072			
	1.0	.056	.054	.052	.053	.052	.048	.052	.056	.049			
Power													
Small	.3	.070	.083	.095	.073	.083	.092	.071	.072	.068			
	.5	.066	.095	.090	.066	.096	.085	.057	.072	.085			
	.7	.063	.080	.099	.066	.078	.100	.072	.074	.089			
	1.0	.070	.074	.090	.066	.072	.087	.057	.065	.083			
Medium	.3	.173	.229	.264	.172	.224	.264	.114	.157	.193			
	.5	.140	.212	.283	.138	.213	.284	.136	.175	.239			
	.7	.124	.198	.290	.123	.199	.293	.111	.193	.269			
	1.0	.120	.202	.274	.120	.202	.271	.107	.196	.273			
Large	.3	.286	.434	.530	.288	.432	.529	.205	.318	.411			
	.5	.225	.400	.548	.226	.400	.552	.177	.335	.471			
	.7	.196	.379	.522	.201	.380	.520	.177	.351	.487			
	1.0	.185	.317	.537	.181	.316	.541	.163	.325	.550			
N=100		Residual change score				ANCOVA				Difference score			
		Reliability (ρ_{xx}, ρ_{yy})											
<i>Treatment β_1</i>	<i>Stability β_2</i>	.5	.8	1	.5	.8	1	.5	.8	1	.5	.8	1
Type I Error													
Zero	.3	.057	.061	.057	.057	.060	.056	.056	.058	.049			
	.5	.053	.059	.049	.053	.058	.048	.067	.041	.058			

N=50		Residual change score			ANCOVA			Difference score		
		Reliability (ρ_{xx}, ρ_{yy})								
$Treatment \beta_1$	$Stability \beta_2$	.5	.8	1	.5	.8	1	.5	.8	1
Small	.7	.052	.060	.043	.052	.061	.044	.052	.048	.041
	1.0	.054	.058	.057	.056	.060	.055	.047	.061	.055
	.3	.086	.072	.105	.084	.073	.102	.066	.078	.081
	.5	.083	.083	.114	.082	.082	.116	.082	.084	.111
Medium	.7	.082	.091	.105	.082	.090	.106	.066	.078	.106
	1.0	.060	.077	.115	.060	.076	.112	.052	.075	.105
	.3	.291	.377	.500	.289	.377	.501	.185	.267	.378
	.5	.195	.395	.494	.194	.396	.491	.151	.325	.418
Large	.7	.233	.349	.479	.230	.348	.479	.191	.319	.456
	1.0	.181	.312	.517	.183	.312	.519	.146	.312	.515
	.3	.444	.687	.831	.443	.687	.829	.322	.517	.692
	.5	.432	.673	.828	.432	.673	.829	.316	.573	.760
N=200	.7	.387	.652	.820	.388	.654	.822	.342	.584	.791
	1.0	.331	.592	.835	.330	.595	.834	.298	.594	.833
		Residual change score			ANCOVA			Difference score		
		Reliability (ρ_{xx}, ρ_{yy})								
$Treatment \beta_1$	$Stability \beta_2$	.5	.8	1	.5	.8	1	.5	.8	1
Zero	.3	.052	.070	.066	.052	.070	.066	.044	.054	.056
	.5	.048	.066	.043	.048	.066	.043	.052	.052	.044
	.7	.053	.051	.050	.053	.050	.049	.048	.044	.042
	1.0	.046	.052	.049	.048	.052	.049	.043	.048	.049
Small	.3	.105	.144	.162	.105	.144	.161	.081	.119	.117
	.5	.093	.140	.170	.094	.139	.169	.089	.096	.154
	.7	.095	.138	.159	.095	.138	.159	.086	.130	.155
	1.0	.095	.138	.159	.095	.138	.159	.086	.130	.155

N=50		Residual change score			ANCOVA			Difference score		
		Reliability (ρ_{xx}, ρ_{yy})								
<i>Treatment β_1</i>	<i>Stability β_2</i>	.5	.8	1	.5	.8	1	.5	.8	1
Medium	1.0	.075	.122	.160	.076	.121	.160	.076	.113	.154
	.3	.440	.635	.783	.439	.634	.783	.292	.486	.599
	.5	.420	.633	.787	.419	.631	.788	.296	.531	.672
	.7	.373	.594	.775	.373	.593	.774	.313	.549	.741
Large	1.0	.296	.574	.799	.297	.575	.799	.281	.557	.799
	.3	.759	.931	.986	.760	.931	.987	.589	.816	.934
	.5	.709	.942	.993	.709	.940	.993	.560	.852	.962
	.7	.658	.921	.988	.658	.921	.987	.571	.876	.977
N=500		.561	.857	.989	.560	.855	.990	.507	.857	.991
		Residual change score			ANCOVA			Difference score		
		Reliability (ρ_{xx}, ρ_{yy})								
<i>Treatment β_1</i>	<i>Stability β_2</i>	.5	.8	1	.5	.8	1	.5	.8	1
Zero	.3	.056	.042	.047	.056	.042	.046	.054	.050	.042
	.5	.060	.038	.045	.058	.038	.045	.050	.048	.058
	.7	.052	.062	.071	.052	.062	.070	.048	.050	.066
	1.0	.054	.051	.046	.052	.049	.045	.059	.047	.050
Small	.3	.204	.262	.355	.203	.262	.356	.142	.176	.252
	.5	.175	.262	.340	.175	.263	.340	.132	.225	.288
	.7	.147	.244	.351	.148	.244	.351	.130	.206	.327
	1.0	.133	.219	.341	.133	.219	.339	.123	.215	.335
Medium	.3	.832	.966	.991	.832	.966	.992	.604	.849	.939
	.5	.784	.962	.995	.784	.962	.995	.614	.899	.978
	.7	.713	.946	.995	.715	.946	.995	.579	.901	.989
	1.0	.610	.912	.995	.610	.912	.995	.551	.896	.996
Large	.3	.988	1	1	.988	1	1	.903	.995	1

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Treatment β_1	Stability β_2	N=50				Residual change score				ANCOVA				Difference score			
		Reliability (ρ_{XX} , ρ_{YY})															
		.5	.8	1	1	.5	.8	1	1	.5	.8	1	1	.5	.8	1	1
	.5	.980	1	1	1	.980	1	1	1	.905	.995	1	1	.905	.995	1	1
	.7	.966	1	1	1	.966	1	1	1	.924	1	1	1	.924	1	1	1
	1.0	.926	.996	1	1	.926	.996	1	1	.887	.995	1	1	.887	.995	1	1

Note. For Treatment β_I small value = .14, medium value = .39, and large value = .59. Tests are two-tailed $p = .05$. For each method, when β_I is equal to zero, values in each row are empirical estimates of the Type I error rate. Sample sizes (N) indicated are per cell.