Journal of Data Science 14(2016), 205-230

Comparing Pre-Post Change Across Groups: Guidelines for Choosing between Difference Scores, ANCOVA, and Residual Change Scores

Megan A. Jennings & Robert A. Cribbie*

Department of Psychology, York University

Abstract: Psychological researchers often investigate differences between groups in the amount of change from pre-test to post-test. For example, researchers treating a group of depressed students may wish to compare the amount of improvement from pre-intervention to post-intervention for males and females, or the researchers may randomly assign participants to groups and compare their improvement across two treatments. In the first case, there are likely to be pre-test differences between the groups, whereas in the second case no pre-test group differences are expected. Three of the most popular methods for comparing independent groups with regard to the amount of change are difference scores, ANCOVA, and residualized change scores. Although the choice between these methods is sometimes clear, in most instances this is not the case. In this research, a simulation study was used to determine the effect of many common issues on the bias, Type I error rates, and power of the difference score, ANCOVA, and residual change score. These issues included: sample size, reliability of the measure, floor/ceiling effects, effect size, and baseline group differences. Results from the study are used to provide specific recommendations with regards to applying each of these three methods.

Key words: ANCOVA, difference scores, change scores, residualized change, pretest-posttest design.

1. Introduction to the Problem

Psychological researchers often investigate differences between groups in the amount of change from pre-test to post-test over a period of time. For example, a study conducted by Nochajski, Stasiewixz, and Patterson (2013) measured the effect of a substance abuse program by measuring the amount of change between a treatment group and a control group using an ANCOVA. The authors used the baseline and follow-up measures of depression and readiness to change and found that individuals in the treatment group had greater improvements on both

* Corresponding author.

variables. Another study conducted by Koch, Morlinghaus, and Fuchs (2007) compared changes in depression between a dance group, a music group and an ergometer group. The authors found that the individuals in the dance group benefited more from the intervention than those in the music or ergometer group via difference scores. Although selecting a procedure for comparing the amount of change over two time points across groups may sound relatively simple, this issue is surprisingly complex and numerous articles have debated which statistical approach should be used for analyzing these designs. Given the complexity of these issues, researchers struggle when deciding which statistical approach to use to determine if there has been a significant difference in the amount of change across groups.

The three statistical methods frequently adopted by researchers to test whether groups differ in the amount of change from pre-test to post-test are a *t*-test on the difference scores (also referred to as change or gain scores), analysis of covariance (ANCOVA), and residual change score analysis. Pre-post designs are extremely common in the field of psychology, and it is well known that these methods often lead to different conclusions (Petscher, 2009; Smith & Beaton, 2008; Wright, 2006). Even though there has been a substantial amount of research on these topics over the past few decades, there is still a need for specific guidelines for the recommended use of each approach.

Thus, the purpose of this paper is to expand on previous research that examines the appropriateness of the popular methods for comparing the amount of change across groups. This is accomplished by investigating common data issues that affect these measures of change, and quantifying the relative impact of these issues and how their interactions with one another affect the performance of the difference score, ANCOVA and residual change score analysis. The end goal is to be able to provide specific recommendations with regard to the use of each of the available procedures. This paper will begin by briefly outlining and comparing the available approaches, as well as discussing the various issues that affect them, before conducting a simulation study to examine the impact of these common data issues on each statistical approach.

2. The Difference Score

The difference score was one of the earliest statistical methods created to analyze data across multiple time points. This approach provides the raw gain observed by individuals as an index of change over time or the difference between two measures using the same sampling unit (Thomas & Zumbo, 2012). Discussions surrounding the difference score are often associated with classical test theory, in which an individual's observed score is a function of their true score plus some error (Petscher, 2009).

One of the most common applications of the difference score is to compare the amount of change from pre-test to post-test across groups. The simple difference score model takes the difference between the pre-test and post-test scores and regresses this difference on the grouping variable, denoted for the two-group case by:

$$post_i - pre_i = \beta_0 + \beta_1 group_i + \varepsilon_i ,$$

where β_0 represents the difference score for individuals in group = 0, β_1 represents the difference in difference scores between group = 0 and group = 1, and ε_i represents the error of estimation. Of primary interest is β_1, which can be interpreted as the difference between the average difference scores of the groups. H_0: $\mu_{D1} = \mu_{D2}$ is rejected if $|t_{\beta 1}| \geq t_{\alpha/2,\nu}$ where:

$$t_{\beta_1} = \frac{\beta_1}{se(\beta_1)}$$

and se(β_1) is the standard error associated with β_1. μ_{D1} and μ_{D2} are the population difference scores for group 1 and 2 respectively, and $t_{1-\alpha/2,\nu}$ is the one-sided *t* critical value with a nominal Type I error rate of α and ν degrees of freedom. This approach tests the null hypothesis of no difference across groups in the amount of raw change from pre-test to post-test.

3. The Analysis of Covariance

Analysis of covariance treats the pre-test value as a covariate that can be a source of variation that may influence post-test scores, and accordingly the post-test score is regressed on both the pre-test score and the grouping variable (Kisbu-Sakarya et al., 2013). ANCOVA adjusts the pre-test scores, increasing the power to determine whether or not there has been a treatment effect and is expressed as:

$$post_i = \beta_0 + \beta_1 group_i + \beta_2 pre_i + \varepsilon_i \text{ , or}$$

$$post_i - pre_i = \beta_0 + \beta_1 group_i + (\beta_2 - 1) pre_i + \varepsilon_i \text{ in difference score form.}$$

This approach tests the null hypothesis of no difference between the control and treatment post-test scores, conditional on the pre-test scores.

4. Lord's Paradox

In 1967, Frederic Lord wrote a paper entitled *A Paradox in the Interpretation of Group Comparisons*, in which he presented a problem now known as Lord's Paradox. In his paper, Lord mentions that researchers often rely on ANCOVA instead of the difference score when investigating group differences, even when individuals are not randomly assigned to groups. Lord provides a hypothetical example of preexisting groups and the problem that occurs when trying to interpret the data. The example involves a large university that wants to know if the new diet in the dining hall has any effect on the students and if there are any sex differences. The weight of each student is measured at the beginning and end of the school year. Two statisticians then examined the data. The first statistician used the difference scores for males and females and found that the mean weight for the girls and boys at the beginning and end of

the school year was the same. She concluded that there was no mean gain for either of the sexes and the diet had no effect. The second statistician used ANCOVA to interpret the data. Since the ANCOVA assumes that the groups are equal at baseline, which they are not, the second statistician concluded that the boys gained significantly more weight than the girls. Thus, the apparent paradox: both approaches seem viable but they lead to different conclusions. Lord concluded that neither approach can properly account for uncontrollable preexisting group differences. Wright (2006) mentions that gain scores and ANCOVA can lead to different results because they ask different questions. The t-test asks if the average gain is different for the two groups and the ANCOVA asks if the average gain, while partialling out pre-test scores, is different between the two groups. It is important to determine which question is of importance when deciding on which approach to use, however in many cases it is difficult to determine which hypothesis is more appropriate for a given problem. In other words, often the research hypothesis is only expressed in terms of comparing change across groups. Lord's Paradox has sparked a heated debate about the reliability and usefulness of the different approaches, especially when pre-test group differences are present.

5. The Residual Change Score

The residual change score method initially estimates the predicted post-test scores by regressing the post-test scores on the pre-test scores, ignoring group assignment. Residual change is then calculated by subtracting the predicted post-test scores from the observed post-test scores, which is then regressed on the grouping variable (Kisbu-Sakarya et al., 2013). This can be expressed as:

$$post\_adj_i = \beta_0 + \beta_1 group_i + \varepsilon_i \text{, where}$$

$$post\_adj_i = post_i - post_i' \quad ,$$

and $post_i'$ is the predicted score from the regression of post-test on pre-test (i.e., $post_i' = \beta_0 + \beta_1 pre_i$). This method uses the regression coefficient for post-test on pre-test in order to adjust for the pre-test, whereas ANCOVA uses the pooled within-group coefficient. Thus, the residual change score approach can result in slightly different statistical power when compared to the ANCOVA. This approach tests the null hypothesis of no difference between the control group and treatment group post-test scores, conditional on the pre-test scores where the conditioning occurs in the absence of group membership.

6. Factors Affecting the Difference Score, ANCOVA, and Residual Change Score

6.1. Group Allocation and Baseline Differences

It is important to determine which statistical approach is better at detecting significant effects between the pre-test and post-test data. If assignment to the treatment group is random,

the baseline difference on any measure is expected to be zero. Petscher and Schatschneider (2011) state that the difference score is just as likely to be used as the ANCOVA in a randomized experiment. However, it has been shown that when randomization to groups occurs, the ANCOVA has slightly greater statistical power than the difference score. According to Kisbu-Sakarya et al. (2013), the residual change score method is comparable to the ANCOVA when groups are randomly assigned, with both producing higher statistical power than the difference score. If baseline imbalances occur during random assignment, Jamieson (1999) argues that the ANCOVA should be used because these random baseline differences will be affected by regression towards the mean. In other words, if baseline imbalances occur "by accident", it is important to control for these difference by using an approach that controls for pre-test.

With non-random assignment to groups (e.g., naturally occurring groups), these baseline imbalances can lead to an increase or decrease in statistical power depending on which group has the higher pre-test score, and which group changes more (Jamieson, 1999). In the presence of nontrivial baseline differences, the ANCOVA and residual change score methods may lead to biased results; further, because these two methods differ in their method of controlling for pre-test scores, they can produce different results (Kisbu-Sakarya et al., 2013). The difference score is less influenced by baseline differences, and Thomas and Zumbo (2012) have shown that the difference score has merit when baseline imbalances exist because it is unbiased and has good statistical power even when the reliability is low. However, it is important to distinguish between nontrivial pretest differences due to a priori differences (e.g., males vs females, young vs old) and those where subjects are "grouped" based on the pretest scores (e.g., give subjects an IQ test and then group them based on those scoring high and low). In this latter case, measurement error in the pretest means that 'regression to the mean' is likely to occur. Since the difference score does not control for 'regression to the mean' it would not be appropriate in this sense. To summarize, if pre-test differences are random or due to scores on the pretest, it is important to control for these differences, however if pre-test differences are based on pre-existing abilities, then controlling for these differences will bias the results.

6.2. Floor and Ceiling Effects

Floor and ceiling effects can affect the amount of change because of baseline imbalance. If a test or measure is relatively easy, participants may answer every question correctly and thus their score is not a good indication of their ability. According to Wang et al. (2008) the ceiling threshold (CT) is not a valid data value but is a proxy value for some larger true value. Ceiling effects can cause underestimation of means and standard deviations, weakening the validity and reliability of the measure. When ceiling or floor effects are present, a negative correlation between initial status and change can occur due to decreasing variance from the pre-test to post-test. Smaller changes are produced by the same external stimulus as scores approach a ceiling or floor (Jamieson, 1995). Higher scores will show a smaller increase when approaching a ceiling, and lower scores will show a smaller decrease when approaching a floor; both resulting in a

negative correlation known as the law of initial values. Cribbie and Jamieson (2004) and Jamieson (1995) found that as post-test variance decreased as a result of a floor or ceiling effect, the difference score, but not the ANCOVA, was negatively affected.

6.3. Unreliability of the Difference Score

The unreliability of the difference score was initially highlighted by Cronbach and Furby (1970), who argued for the abandonment of this method for measuring change. From the classical test theory perspective, where X represents the pre-test score and Y represents the post-test score, the reliability of the difference score is defined by:

$$\rho_{\Delta\Delta'} = \frac{\lambda\rho_{XX'} + \lambda^{-1}\rho_{YY'} - 2\rho_{XY}}{\lambda + \lambda^{-1} - 2\rho_{XY}}$$

where $\lambda = \sigma_X/\sigma_Y$ is the ratio of the standard deviations, $\rho_{XX'}$ and $\rho_{YY'}$ are the reliability coefficients, and ρ_{XY} is the correlation between X and Y (Kisbu-Sakarya et al., 2013; Zimmerman & Williams, 1982b, 1998). Therefore, the ratio of the standard deviations is important in determining the reliability of difference scores. The difference score gets a bad reputation because when there is no variability between individuals (i.e., the pre-test, post-test correlation is high), this equation shows that the reliability of the difference score will be equal to zero (Petscher, 2009). However, when the variance and reliability of the post-test scores exceeds that of the pre-test scores, the reliability of the difference score can be substantial (Zimmerman & Williams, 1998).

An alternate expression has been proposed for the previous equation which can be further reduced if one assumes that the pre-test and post-test measures are parallel forms of a test and that $\rho_{XX'} = \rho_{YY'}$ and $\sigma_X = \sigma_Y$:

$$\rho_{\Delta\Delta'} = \frac{\rho_{XX'} - \rho_{XY}}{1 - \rho_{XY}}$$

As described above, Rogosa et al. (1982) argue that the difference score will be reliable if the correlation between the pre-test and post-test scores is low. Another way to look at this is that even when the difference score is unreliable it can still have adequate power for testing hypotheses about change since the standard error will often be low. As Webb, Shavelson, and Haertel put it, *the reliability coefficient of the difference score may not be an important factor to consider*.

The perceived unreliability of the difference score has also been associated with unrealistic assumptions of the classical test theory formula (Chiou & Spreng, 1996). According to Chiou and Spreng, measurement error can promote the reliability of the post-test score relative to the pre-test score, producing a condition in which the equal variance assumption is not met. This significantly improves the reliability of the difference score. Thus, the claims by Cronbach and Furby (1970), Ling and Slinde (1977), and Overall and Woodward (1975) asserting that the

difference score is an unreliable measure are not necessarily true; the reliability can be high if more realistic assumptions are made.

To conclude, the unreliability of the difference score has been used to discount the use of the procedure, however under realistic conditions it may not be so unreliable. Further, even if it has low reliability, that may not be a factor that should be used in deciding whether or not to use the difference score.

6.4. Stability

The stability of a test is normally measured by the correlation between pre-test and post-test; the higher the correlation, the higher the stability. The correlation between pre-test and post-test scores is linked to the reliability of the different measures of change. Thus, the difference score is influenced differently by reliability and stability than the ANCOVA and residual change score methods. The stability of the measure is higher when the variance of the individual true change score is smaller; if the variance is kept small, the reliability of the difference score will also be very small (Chiou & Spreng, 1996). Petscher and Schatschneider (2011), following Rogosa (1995), used the variances between the pre-test and post-test scores to determine the functional form of change over time, but Kisbu-Sakarya et al. (2013) argued that this is an inaccurate measure and it instead should be a function of the pre-test and post-test correlation. However, it is important to recall that the correlation between the pre-test and post-test scores depends on the reliability of these scores (Chiou & Spreng, 1996; Zimmerman et al. 1993). It is also important to highlight that lower pre-test post-test correlations can reduce the power of the difference score method (Zhang et al., 2014). In this study, the stability is a function of the reliability.

6.5. Correlation between Change and Initial Status

The correlation between difference scores and initial status has been studied extensively due to the troublesome problems of reliability and validity of gains (Zimmerman & Williams, 1982a). With regard to the correlation between baseline and change, there are three types of correlation discussed in the literature (Petscher, 2009). First is the law of initial values, which occurs when a negative correlation between change and initial status is observed. The second type is fan-spread change, where a positive correlation is observed between initial status and change. Finally, the overlap hypothesis refers to the existence of no relationship between change and initial status. The simple difference score typically has a negative correlation with the pre-test scores, which is argued to be one of the major disadvantages and reasons for the abandonment of the difference score.

Rogosa et al. (1982) discussed that many researchers feel that the law of initial values makes difference scores an inappropriate method to evaluate individual change when

individuals have different pre-test scores and/or certain pre-test scores provide some individuals with an advantage. However, as Rogosa et al. defend, this seems confusing considering that difference scores are an unbiased estimate of true change. It is true, as discussed in this paper and elsewhere, that there can be a relationship between pre-test and change; however, this does not invalidate the use of the difference score as a measure of individual change. Further, one reason a negative correlation between change and initial status can occur is if the ratio of the pre-test and post-test standard deviations exceeds the correlation between the pre-test and post-test scores (Zimmerman & Williams, 1982b). This can occur as a result of regression toward the mean or measurement error.

In two-wave data, regression toward the mean is possible because pretest and posttest scores represent the same operational variable (Furby, 1973). When an individual has a low pretest score, they are likely to have a large positive post-pre difference score, and those with a high pretest score will typically have a large negative post-pre difference score (Linn & Slinde, 1977). In other words, if an individual scores high on the pre-test, they are more likely to score lower on the post-test, and an individual who scores low on the pre-test will tend to score higher on the post-test. Rogosa et al. (1982) note that formal statements of regression towards the mean standardize the variables by defining these statements in standard deviation units; they further argue that this standardization of regression toward the mean is not useful and that the inequality of change should be viewed in the metric of the measure. In terms of the true scores this definition is:

$$E[T_2 \mid T_1 = C] - \mu_{T2} < C - \mu_{T1},$$

where T_1 is the true score at pre-test, T_2 is the true score a post-test and C is any value on the pretest greater than the mean. For this equation to be satisfied, there must be a negative correlation between initial status and change (indicating again that regression toward the mean is observed when the pre-test to change correlation is negative). Most importantly, regression to the mean, and negative correlations between initial status and change, do not invalidate the use of difference scores for comparing the amount of change across stable groups (Allison, 1990; Kenny, 1975).

Thus, based on previous research, there are numerous variables that are expected to impact the ANCOVA, difference score and residual change score methods. For example, with respect to baseline differences between groups, in some situations it makes sense to control for pretest differences (e.g., random assignment), whereas in others this does not make sense (e.g., baseline scores are related to the group variable). Also, floor/ceiling effects can be especially problematic if baseline levels are not controlled for. The current study is interested in clarifying not only the individual effects of the variables discussed above, but also how these variables interact to affect the bias, Type I error rates and power of the difference score, ANCOVA, and residualized change score methods. For example, if naturally occurring groups differ at baseline (a problem for ANCOVA), but there are floor or ceiling effects (a problem for difference scores), which procedure will perform better?

7. Method

The primary aim of this study was to assess the performance of the difference score, ANCOVA, and residual change score methods under realistic data conditions. Simulations were used to compare the three different approaches to measuring pre-post change by manipulating different population variables, including sample size, ceiling and floor effects, pre-post correlation, reliability of the measure, and effect size. Type I error rates, power, and bias were recorded for each approach, under each condition, with acceptable bounds for Type I error ranging from .025 to .075 (with $\alpha = .05$) for the Type I error rates (see Bradley, 1978). Ten thousand simulations were conducted for each condition using the open-source statistical software *R* (*R* Core Team, 2013).

7.1. Selection of Manipulated Variables

Floor and ceiling effects. In the current study, three conditions were investigated: one in which there was no floor or ceiling effect, as well as when there was either a floor or ceiling effect. For the ceiling effect all standardized scores greater than 1 were set equal to 1, whereas for the floor effect all standardized scores less than -1 were set equal to -1.

Sample size. As sample size increases, power typically increases. Although this relationship is well documented, different sample sizes are used to examine the trade-offs of the different approaches when manipulating other factors, such as floor and ceiling effects. In the current study, group sample sizes of 20, 50, and 100 were chosen to be comparable to the sample sizes utilized in the behavioural sciences.

Group Allocation. This study investigated three different situations in which groups could be related to pre-test ability. In the first condition, individuals were assigned randomly to groups; therefore, they were assumed to have no baseline imbalances. The second method involved assigning individuals to the treatment group based on some ability. Therefore, those in the treatment group scored higher on some ability than those in the control group, resulting in significant baseline imbalances. This type of group assignment is common in educational research (Wright, 2006). Lastly, ability was correlated with, but did not determine, group allocation; therefore, baseline scores are associated with group allocation. Point biserial correlations between group and ability were set at approximately .2 for a mild relationship, or .4 for a moderate relationship.

Treatment effect. The relationship between the treatment variable (control/treatment group) and change (i.e., treatment effect) was set at -.5, -.25, 0, .25, or .5. More specifically, the treatment group was set to change .5 less, .25 less, the same, .25 more, or .5 more than the control group.

Reliability. Reliabilities for pre and post were set to be equal, and they were varied to simulate low (.6) and high (.9) reliability.

Stability. The correlation between pre-test and post-test is related to individual differences in change. Large individual differences in the amount of change from pre-test to post-test result

in small stability coefficients, and small differences are associated with large stability coefficients (Petscher, 2009). In other words, as the reliability changes, stability also changes. Therefore, stability was indirectly manipulated through the manipulation of the reliability. When the reliability was set to .6 and .9, the correlation between the pre-test and post-test scores was also .6 and .9, respectively.

Correlation between initial status and change. The correlation between initial status and change can impact the performance of the different approaches studied. For example, a negative correlation has been found to reduce the power of the difference score. In the current study, the correlation between the observed pre-test scores and change was approximately -.45 and -.25 for a reliability of .6 and .9, respectively. It is important to note that the correlation between the true pre-test scores and change from pre-test to post-test is always zero regardless of reliability.

8. Results

The Type I error, power, and bias results are summarized below across the 360 conditions investigated (4 grouping conditions, 2 reliability conditions, 5 effect size conditions, 3 floor/ceiling effect conditions, and 3 sample size conditions). Relative bias [(sample estimate – population parameter)/population parameter)] was recorded for each effect size condition, except for when the effect size was set to zero. In this situation, the raw bias was recorded because division by zero prevented the calculation of relative bias.

There was no significant effect of reliability on bias or Type I error rates (except where noted), and for power the rates were higher when reliability increased. Thus, only the results for pre-test and post-test reliability of .9 are presented since the pattern of results was similar for both .6 and .9 reliability. The pattern of results for the moderate and extreme effect sizes, and the small, medium and large sample sizes were also similar, and therefore the results for the largest effect size and smallest sample size are discussed (except where otherwise noted). Finally, the pattern of results was similar for the mild and moderate pretest group difference conditions and thus the results are only discussed for the moderate condition.

8.1. Random Assignment to Groups\

Bias results for the condition in which subjects were randomly assigned to groups are presented in Table 1. Type I error and power results for this condition are presented in Table 2.

Bias. For all conditions in which the effect size was zero, the raw bias was approximately zero. When there was no floor or ceiling effect, the difference score and ANCOVA had almost no bias, however the residual change score method had a small bias of approximately .05 for both large and small effect sizes. When a floor or ceiling effect was present, bias increased for all methods with the residual change score having slightly higher bias than the difference score and ANCOVA. For the negative effect size, a floor effect produced greater bias and when the effect size was positive, a ceiling effect produced greater bias. This is expected because if one group starts lower at pre-test and there is a negative effect size, a floor effect will restrict the

amount that group can change, however if there is a positive effect the group will not approach the floor.

Type I Error. When subjects were randomly assigned to groups, the Type I error rates of the difference score, ANCOVA, and residual change score methods were all exactly equal to the nominal level (i.e., .05).

Power. Power was very similar across the three procedures for most conditions. However, power was reduced for all procedures when one group's post-test score decreased and there was a floor effect, or one group's post-test score increased and there was a ceiling effect. This effect was slightly more pronounced for the difference score method than for the ANCOVA or residual change score methods.

8.2. Grouping Based on Ability

Bias results for the condition in which group assignment is based on ability are presented in Table 3. Type I error and power results for this condition are presented in Table 4.

Bias. Raw bias results for the ANCOVA and residual change score were roughly equal to zero, with the ANCOVA having slight bias for both the ceiling and floor effect conditions. The difference score had slight bias for each condition when the effect size was zero. For each of the non-zero effect sizes, as expected, the ANCOVA had the lowest amount of bias for every condition. However, the ANCOVA had an inflated bias of .15 when the effect size was -.5 and a floor effect was present, when compared to the other conditions in which the ANCOVA had an average bias close to zero (including an effect size of .5 with a ceiling effect present). This is due to the fact that (for this condition) when groups are split by ability, the group affected by the treatment is the group with the lower score.

Regardless of the effect size, the residual change score method had consistently higher relative bias than the other two approaches. The difference score method had approximately .30 relative bias for the large effect sizes when there was no ceiling or floor effect, and when there was a ceiling effect. When there was a floor effect, the difference score had a relative bias of .75 when the effect size was -.5, and a bias of .09 when the effect size was .5. This effect is caused by the same process that affected the ANCOVA above; that is, the group affected by the treatment is the group with the lower score. In comparison to the ANCOVA and residual change score methods, the bias of the difference score was much higher when the reliability was .6.

Type I Error. When group assignment was based on ability, Type I error rates for the ANCOVA were approximately .05 when there was no ceiling or floor effect, and roughly .07 for both a ceiling and a floor effect; both are within the acceptable bounds. Type I error rates for the residual change score method were all zero, regardless of the condition. The difference score method had the highest Type I error rates; when there was no ceiling or floor effect the rate was approximately .11, and when a ceiling or floor effect was present the Type I error rate was approximately .13. The Type I error rates were consistent for the ANCOVA and residual change score methods regardless of sample size; however, as sample size increased, Type I error rates increased for the difference score method. It is also important to highlight that Type I error rates

were much higher for the difference score method when the reliability was low, ranging from approximately .35 for n = 20 to almost 1 for n = 100.

Power. Since the Type I error rates of the difference score were extremely liberal and the Type I error rates of the residual change score method were extremely conservative, it is not possible to make meaningful power comparisons with the ANCOVA approach. Power for the ANCOVA was similar for both the positive and negative effect sizes, with only slight gains or losses in power depending on whether or not a ceiling or floor effect was present; the average power was roughly .32 across each condition for the large effect sizes.

8.3. Moderate Pre-Test Group Differences

Bias results for the condition in which there were moderate pre-test group differences in ability (point biserial correlation between group and pre-test is approximately equal to .4) are presented in Table 5. Type I error and power results for this condition are presented in Table 6.

Bias. The difference score method had zero raw bias when the effect size was zero, and zero relative bias when there were no floor or ceiling effects. When the effect size was zero, the residual change score method had less bias than the ANCOVA (approximately .07 and .09, respectively). The difference score had the least amount of bias for every condition except one; when the effect size was positive and there was a ceiling effect. For example, the difference score had a relative bias of approximately .35 with an effect of .5, reliability of .9 and a ceiling effect, which was greater than that of the ANCOVA (.07) and the residual change score (.24) methods. When the effect size was -.5, the residual change score method had much higher bias for each condition in comparison to the ANCOVA and difference score methods. Conversely, when the effect size was .5, the ANCOVA had the highest amount of relative bias (.18) when there was no ceiling or floor effect. The residual change score and ANCOVA had equal bias of approximately .10 when there was a floor effect, however when there was a ceiling effect the ANCOVA had the lowest bias of roughly .07. Bias results for this grouping condition were significantly higher when the reliability was lower.

Type I Error. When subjects' ability was moderately correlated with group assignment, the Type I error rates of the difference score and residual change score methods were all approximately equal to the nominal level (i.e., .05). Floor and ceiling effects had no effect on the Type I error rates. There was only a slight variation in the Type I error rates for the ANCOVA approach; when there was no floor or ceiling effect the Type I error rate was approximately .07, and when there was a floor or ceiling effect the Type I error rate was roughly equal to .08. Type I error rates were higher when the reliability was low, and as the sample size increased the Type I error rates also increased substantially for the ANCOVA and residual change score methods. For example, when the reliability was .6 and the sample size was 100, the ANCOVA and residual change score methods had Type I error rates of approximately .39 and .35, respectively. There was no change in Type I error rates for the difference score in any of the conditions.

Power. Due to the inflated Type I error rates for the ANCOVA and residual change score methods, power comparisons between these two methods are not possible. Although the residual change score method had nominal levels when the sample size is 50 and the reliability is high, for all other conditions the Type I error rates were outside the acceptable bounds and thus power cannot be discussed. The difference score had similar power rates when there are was no ceiling or floor effect, and when there was a floor effect, regardless of the effect size magnitude. The only noteworthy reduction in power occurred when the difference score had a ceiling effect. Otherwise, power for the difference score, as expected, increased with sample size and reliability.

9. Discussion

Deciding which statistical approach to use is one of the most important decisions a researcher must make when conducting a pre-post group design. Several published articles have argued for the dismissal of the difference score, and state that researchers should use the ANCOVA to analyze their data instead (Cronbach & Furby, 1970; Linn & Slinde, 1977; Overall & Woodward, 1975). Although the primary reason for this argument is the supposed lack of reliability of the difference score, many of these premises have been found to have little effect when realistic data conditions are present. More recently, published articles have acknowledged the usefulness of the difference score (Chiou & Spreng, 1996; Petscher, 2009; Zimmerman & Williams, 1998). In fact, the difference score is often reliable, and even when it is not, the validity of its conclusions is unaffected. Despite these findings, many still advocate for the utility of the ANCOVA without proper consideration of the problems associated with this method. Regardless of the debates in the literature, Petscher and Schatschneider (2011) have noted that researchers are just as likely to use the difference score as they are the ANCOVA, and they usually offer no explanation as to why a particular approach was used. Thus, many researchers do not understand the circumstances in which applying a particular statistical method can be either detrimental or beneficial to their analysis.

Although Lord (1967) created his data to illustrate the difference score/ANCOVA paradox, the situational circumstances that created it can also occur in real data. The key issue is to gain a better understanding of the conditions that affect the difference score, ANCOVA, and residual change score methods when measuring change across two time points. Previous research has argued for additional simulation studies to better understand the wide variety of conditions that can affect these three statistical approaches (Petscher, 2009; Rogosa, 1995). Results of this study sought to extend the findings from Petscher and Schatschneider (2011) and Kisbu-Sakarya et al. (2013) by evaluating different conditions that are commonly found in psychological data. For example, different methods of group assignment other than randomization were used (e.g. non-equivalent group designs), because random allocation to conditions is often impractical (Wright, 2006). However, this study also examined the effect of floor and ceiling effects, reliability, sample size, and effect size, and how these factors directly

and jointly affect the Type I error rates, power, or bias of the difference score, ANCOVA, and residual change score methods. Thus, this study was conducted to provide guidelines for the choice of data analysis when such conditions exist.

Findings from the simulations indicated that when random assignment occurred, the difference score and ANCOVA had similar bias results for all conditions, and the residual change score had only slightly higher bias in comparison. These results are similar to those found by Wright (2006). In addition, regardless of which method was used for group assignment, the combination of a negative effect size and a floor effect, or a positive effect size and a ceiling effect caused greater bias results. When grouping was based on ability, as expected, the ANCOVA had the least amount of bias for every condition and the residual change score method had the highest bias for each condition. It is important to note that both the ANCOVA and difference score had much higher bias when there was a negative effect size and a floor effect. This was not observed when there was a positive effect and a ceiling effect because the group with the lower score is the one that is changing in this method of group assignment (note that we could just have easily have simulated the data so that the group with the highest pretest score changed due to the treatment effect, which would have left us with reverse findings; i.e., the higher bias would have been found when there was a positive effect and a ceiling effects). If there were moderate pre-test group differences, the difference score had the least amount of bias across all conditions except when there was a positive effect size in combination with a ceiling effect; when this occurred the ANCOVA had the lowest bias results. The residual change score method had the greatest bias when compared to the other two methods when the effect size was negative, and the ANCOVA had the highest bias when the effect size was positive (with the exception of a ceiling effect). When the reliability was low, the ANCOVA and residual change score methods had significantly higher bias results, especially when sample size increased. These results are similar to those presented by Wright (2006) when there are mild or moderate pre-test group differences.

If assignment to groups is random, all three approaches had excellent Type I error control. When assignment was based on ability (i.e., all participants in one group scored higher at pre-test than all the participants in the second group), the ANCOVA was the only approach within acceptable bounds, with only a slightly higher Type I error rate when either a floor or ceiling effect was present. Regardless of reliability and sample size, the difference score and residual change score methods were always outside the nominal bounds. When moderate pre-test group differences were present, the difference score method had good Type I error control across all conditions. Both the ANCOVA and residual change score methods experienced Type I error rates that deviated from the nominal bounds, and these errors were more predominate when sample size increased, and significantly more notable when the reliability was low.

Power was affected by all conditions in this study with the most notable changes occurring depending on participant group assignment. When randomization occurred, each approach had similar power except when there was a negative effect size and a floor effect, or a positive effect size and a ceiling effect. If these conditions were met, the difference score had slightly lower power than the ANCOVA and residual change score methods. When grouping was based on ability, power discussions were not possible for the difference score and residual change score

approaches due to the inflated and deflated Type I error rates. However, the ANCOVA had similar power for each condition, with only slight changes depending on whether or not there was a floor or ceiling effect. When moderate pre-test group differences occurred, the Type I error rates for the ANCOVA and residual change score methods were inflated. Therefore, power comparisons were not possible for this grouping condition. The difference score had similar power results for each condition depending on the magnitude of the effect size. However, the difference score had reduced power results when a ceiling effect was present.

9.1. Limitations and Future Directions

Due to the use of simulations, conclusions are limited to the factors investigated in this study, even though they were chosen to reflect conditions that often occur in psychological research, such as sample size, and floor or ceiling effects. Further analysis of more specific interactions and the expansion to include more conditions may be necessary to better understand the performance of the difference score, ANCOVA, and residual change score methods. Due to the scope of this study, the effects of non-normality could not be included; this is an interesting condition to explore in the future research because it is very common with psychological variables and its effects on the different statistical methods may interact with those investigated in this study (Petscher & Schatschneider, 2011). Furthermore, additional methods for simulating floor or ceiling effects could have been investigated to better understand the outcomes they have on each approach (see Cribbie & Jamieson, 2004).

Due to the method in which the data was simulated, the correlation between pre-test and post-test, as well as the correlation between initial status and change, was a direct function of the reliability. Although this is believed to be a realistic condition, other approaches for simulating these relationships could have been explored (Kisbu-Sakarya et al., 2013; Petscher & Schatschneider, 2011). In addition, continuous predictors could have been used instead of a grouping variable.

9.2. Guidelines for Researchers

These results highlight how complicated the decision making process is when deciding among the available statistical approaches for pre-post group designs. General recommendations are provided below to help researchers choose the best statistical method for their data, however the recommended approach is to consult the tables in this paper in order to try to match your sample data conditions (e.g.., floor effect, sample size) to the conditions investigated in this paper. When it is not possible to match your sample data conditions to those investigated in this paper, it is recommended that you simulate data conditions that match your data conditions. If this is not possible, then hopefully the recommendations below will be helpful.

When individuals are randomly assigned to groups, researchers can safely use the difference score, ANCOVA, and residual change score methods because each has similar Type I error rates, power, and bias. In this situation, it is important for the researcher to revisit the specific

hypotheses assessed by each procedure and ensure that the hypothesis tested by the procedure matches the specific research hypothesis being addressed. The only noteworthy difference is the slightly higher bias results of the residual change score method. Reliability, sample size, ceiling, and floor effects had small and similar effects on the Type I error rates, power, and bias of all procedures when randomization occurred. Therefore, the best advice is to use random assignment when possible because it requires fewer assumptions when making inferences, and all approaches produce good estimates. However, this is not always possible and recommendations for non-equivalent groups are also needed.

If grouping is based on ability, which often occurs when treatments are designed for a specific subset of the population, such as those designed in educational research, then the ANCOVA is the only viable option for measuring change. Due to the inflated and deflated Type I error rates of the difference score and residual change score methods, power comparisons were not possible. Wright (2006) noted that the difference score will often show that the treatment was effective or detrimental when it was not, depending on whether the treatment was given to the group that had lower or higher scores initially. Bias results for the difference score and residual change score methods were significantly higher for each condition when grouping was based on ability. Thus, for the least biased results it is important to use the ANCOVA when grouping is based on ability.

When group assignment is mildly or moderately related to ability at pre-test, the inflated Type I error rates for the ANCOVA and residual change score methods make power comparisons impossible; especially when the sample size increases, and the reliability is low. The bias results for the ANCOVA and residual change score methods were very high in comparison to the difference score. The difference score did have higher bias when a group that is influenced by a treatment is limited in the amount they can change due to a floor or ceiling effect. However, the difference score approach is still the most effective method for measuring change when there are preexisting group differences, regardless of whether these differences are mild or moderate. Although floor and ceiling effects had an impact on the difference scores, and researchers must closely investigate potential floor or ceiling effects, rarely was this impact as influential as the assignment being related to ability.

9.3. Conclusion

To summarize, the results of this study hopefully aid in the understanding of the complex decision making process required to select a statistical approach when comparing groups in pre-post designs. General recommendations and tables printed within the paper are provided to help researchers select an appropriate approach when analyzing their data. However, further simulation studies should be conducted to increase the understanding of the effects that different factors have on the Type I error rates, power, and bias of the difference score, ANCOVA, and residual change score methods.

References

[1] Allison, P. D. (1990). Change scores as dependent variables in regression analysis. *Sociological Methodology,* **20**, 93-114.

[2] Bradley, J.V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*, 144-152. DOI: 10.1111/j.2044-8317.1978.tb00581.x

[3] Chiou, J., & Spreng, R. A. (1996). The reliability of difference scores: A re-examination. Journal of Consumer Satisfaction, Dissatisfaction and Complaining Behaviour, 9, 158-167.

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

[5] Cronbach, L. J., & Furby, L. (1970). How we should measure "change": Or should we? *Psychological Bulletin,* **74**, 68-80.

[6] Furby, L. (1973). Interpreting regression toward the mean in developmental research. *Developmental Psychology,* 8, 172-179.

[7] Jamieson, J. (1995). Measurement of change and the law of initial values: A computer simulation study. *Educational and Psychological Measurement,* **55**, 38-46.

[8] Jamieson, J. (1999). Dealing with baseline differences: Two principles and two dilemmas. *International Journal of Psychophysiology,* 31, 155-161.

[9] Kenny, D. A. (1975). A quasi-experimental approach to assessing treatment effects in the nonequivalent control group design. *Psychological Bulletin*, **82**, 345-362.

[10] Kisbu-Sakarya, Y., MacKinnon, D. P., & Aiken, L. S. (2013). A monte carlo comparison study of the power of the analysis of covariance, simple difference, and residual change scores in testing two-wave data. *Educational and Psychological Measurement*, **73**, 47-62.

[11] Koch, S. C., Morlinghaus, K., & Fuchs, T. (2007). The joy dance: Specific effects of a single dance intervention on psychiatric patients with depression. *The Arts in Psychotherapy*, **34**, 340-349.

[12] Linn, R. L., & Slinde, J. A. (1977). The determination of the significance of change between pre- and posttesting periods. *Review of Educational Research,* **47**, 121-150.

[13] Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological bulletin,* **68**, 304-305.

[14] Mellenbergh, G. J. (1999). A note on simple gain precision. *Applied Psychological Measurement,* 23, 87-89.

[15] Nochajski, T. H., Stasiewicz, P. R., & Patterson, D. A. (2013). Depression, readiness for change, and treatment among court-mandated DUI offenders. *Journal of Dual Diagnosis,* **9**, 139-148.

[16] Overall, J. E., & Woodward, J. A. (1975). Unreliability of difference scores: A paradox for measurement of change. *Psychological Bulletin,* **82**, 85-86.

[17] Petscher, Y. (2009). *A simulation study on the performance of the simple difference and covariance adjusted scores in randomized experimental designs*. (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3374031)

[18] Petscher, Y., & Schatschneider, C. (2011). A simulation study on the performance of the simple difference and covariance-adjusted scores in randomized experimental designs. *Journal of Educational Measurement,* **48**, 31-43.

[19] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

[20] Rogosa, D. (1995). Myths and methods: Myths about longitudinal research plus supplemental questions. In J. M. Gottman (Ed.), *The analysis of change* (pp. 4-66). Mahwah, NJ: Lawrence Erlbaum.

[21] Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin,* 92, 726-748.

[22] Smith, P., & Beaton, D. (2008). Measuring change in psychosocial working conditions: methodological issues to consider when data are collected at baseline and one follow-up time point. *Occupational and Environmental Medicine*, 65, 288-296.

[23] Thomas, D. R., & Zumbo, B. D. (2012). 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,* 72, 37-43.

[24] Wang, L., Zhang, Z., McArdle, J. J., & Salthouse, T. A. (2008). Investigating ceiling effects in longitudinal data analysis. *Multivariate behavioral research*, **43**, 476-496.

[25] Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). Reliability coefficients and generalizability theory. In C. R. Rao & S. Sinharay, *Handbook of Statistics, Vol. 26*. New York: Elsevier.

[26] 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, 633-675.

[27] Zhang, S., Paul, J., Nantha-Aree, M., Buckley, N., Shahzad, U., Cheng, J., et al. (2014). Empirical comparison of four baseline covariate adjustment methods in analysis of continuous outcomes in randomized controlled trials. *Clinical Epidemiology*, 6, 227-235.

[28] Zimmerman, D. W., & Williams, R. H. (1982a). A note on the correlation of gains and initial status. *Journal of General Psychology*, 107, 203-207.

[29] Zimmerman, D. W., & Williams, R. H. (1982b). Gain scores in research can be highly reliable. *Journal of Educational Measurement*, 19, 149-154.

[30] Zimmerman, D. W., & Williams, R. H. (1998). Reliability of gain scores under realistic assumptions about properties of pre-test and post-test scores. *British Journal of Mathematical and Statistical Psychology*, 51, 343-351.

[31] Zimmerman, D. W., Williams, R. H., & Zumbo, B. D. (1993). Reliability of measurement and power of significance tests based on differences. *Applied Psychological Measurement*, 17, 1-9.

Received March 15, 2015; accepted November 10, 2015.

Robert Cribbie
Quantitative Methods Program
Department of Psychology
York University
Toronto, Ontario, Canada, M3J 1P3,
cribbie@yorku.ca

Table 1: Relative bias when group assignment is random.

				Difference Score			ANCOVA			Residual Change		
Rel	**ES**	**Ceil**	**Flr**	**n=20**	**n=50**	**n=100**	**n=20**	**n=50**	**n=100**	**n=20**	**n=50**	**n=100**
.6	-.5	0	0	.00	.01	.00	.00	.01	.00	.05	.01	.00
		0	1.0	.23	.23	.23	.23	.23	.23	.27	.24	.24
		1.0	0	.12	.10	.11	.11	.11	.11	.16	.13	.12
	-.25	0	0	.00	.01	.00	.03	.01	.00	.07	.01	.00
		0	1.0	.21	.20	.19	.20	.19	.20	.24	.21	.21
		1.0	0	.14	.14	.13	.14	.13	.13	.18	.15	.14
	0	0	0	.00	.00	.00	.00	.00	.00	.00	.00	.00
		0	1.0	.00	.00	.00	.00	.00	.00	.00	.00	.00
		1.0	0	.00	.00	.00	.00	.00	.00	.00	.00	.00
	.25	0	0	.00	.02	.01	.00	.02	.00	.05	.00	.00
		0	1.0	.14	.12	.13	.14	.12	.13	.19	.14	.14
		1.0	0	.19	.19	.19	.19	.19	.19	.23	.21	.20
	.5	0	0	.00	.00	.00	.00	.00	.00	.05	.02	.00
		0	1.0	.11	.11	.11	.11	.11	.11	.15	.13	.12
		1.0	0	.24	.23	.23	.24	.23	.23	.28	.25	.24
.9	-.5	0	0	.00	.00	.00	.00	.00	.00	.05	.00	.01
		0	1.0	.23	.23	.23	.23	.23	.23	.27	.24	.24
		1.0	0	.10	.11	.11	.10	.10	.11	.15	.12	.12
	-.25	0	0	.00	.01	.00	.00	.00	.00	.06	.00	.01
		0	1.0	.19	.20	.19	.19	.20	.19	.23	.21	.20
		1.0	0	.13	.13	.13	.13	.13	.13	.17	.14	.14
	0	0	0	.00	.00	.00	.00	.00	.00	.00	.00	.00
		0	1.0	.00	.00	.00	.00	.00	.00	.00	.00	.00
		1.0	0	.00	.00	.00	.00	.00	.00	.00	.00	.00
	.25	0	0	.00	.00	.01	.00	.00	.01	.05	.00	.00
		0	1.0	.14	.13	.13	.14	.12	.13	.18	.14	.14
		1.0	0	.19	.19	.20	.19	.19	.19	.23	.21	.20
	.5	0	0	.00	.00	.00	.00	.00	.00	.05	.00	.01
		0	1.0	.11	.11	.11	.11	.11	.11	.15	.13	.12
		1.0	0	.23	.23	.23	.23	.23	.23	.27	.25	.24

Table 2: Type I error and power results when group assignment is random.

				Difference Score			ANCOVA			Residual Change		
Rel	**ES**	**Ceil**	**Flr**	**n=20**	**n=50**	**n=100**	**n=20**	**n=50**	**n=100**	**n=20**	**n=50**	**n=100**
.6	-.5	0	0	.21	.49	.77	.23	.57	.86	.23	.57	.86
		0	1.0	.18	.40	.69	.22	.51	.81	.22	.51	.81
		1.0	0	.20	.48	.78	.23	.54	.84	.23	.54	.84
	-.25	0	0	.09	.16	.28	.09	.18	.34	.09	.18	.34
		0	1.0	.08	.14	.25	.09	.17	.30	.09	.17	.30
		1.0	0	.08	.15	.27	.09	.17	.31	.10	.17	.31
	0	0	0	.05	.05	.05	.05	.05	.05	.05	.05	.05
		0	1.0	.05	.05	.05	.05	.05	.05	.05	.05	.05
		1.0	0	.05	.05	.05	.05	.05	.05	.05	.05	.05
	.25	0	0	.09	.17	.29	.09	.19	.34	.09	.19	.34
		0	1.0	.08	.16	.27	.09	.18	.31	.09	.18	.32
		1.0	0	.08	.14	.24	.09	.17	.30	.09	.17	.30
	.5	0	0	.21	.48	.79	.24	.57	.87	.24	.57	.87
		0	1.0	.20	.47	.78	.23	.54	.85	.23	.54	.85
		1.0	0	.17	.40	.69	.21	.50	.82	.21	.50	.82
.9	-.5	0	0	.63	.97	1.00	.63	.97	1.00	.63	.97	1.00
		0	1.0	.49	.91	1.00	.53	.94	1.00	.53	.94	1.00
		1.0	0	.60	.96	1.00	.58	.96	1.00	.58	.96	1.00
	-.25	0	0	.20	.49	.79	.20	.50	.80	.20	.50	.80
		0	1.0	.18	.40	.69	.18	.43	.73	.19	.43	.73
		1.0	0	.19	.44	.74	.19	.44	.75	.19	.44	.75
	0	0	0	.05	.05	.05	.05	.05	.05	.05	.05	.05
		0	1.0	.05	.05	.05	.05	.05	.05	.05	.05	.05
		1.0	0	.05	.05	.05	.05	.05	.05	.05	.05	.05
	.25	0	0	.22	.48	.80	.21	.49	.81	.21	.49	.81
		0	1.0	.19	.45	.75	.19	.45	.76	.19	.46	.76
		1.0	0	.17	.41	.69	.18	.43	.73	.18	.43	.73
	.5	0	0	.64	.97	1.00	.63	.97	1.00	.63	.97	1.00
		0	1.0	.59	.96	1.00	.58	.95	1.00	.57	.96	1.00
		1.0	0	.49	.90	1.00	.53	.94	1.00	.53	.93	1.00

Table 3: Relative bias results when group assignment is based on ability.

				Difference Score			ANCOVA			Residual Change		
Rel	**ES**	**Ceil**	**Flr**	**n=20**	**n=50**	**n=100**	**n=20**	**n=50**	**n=100**	**n=20**	**n=50**	**n=100**
.6	-.5	0	0	1.24	1.26	1.26	.02	.01	.00	.65	.64	.64
		0	1.0	1.53	1.56	1.57	.27	.28	.29	.76	.77	.77
		1.0	0	1.18	1.22	1.23	.10	.13	.13	.71	.72	.72
	-.25	0	0	2.45	2.51	2.54	.04	.01	.02	.54	.63	.64
		0	1.0	2.63	2.70	2.74	.31	.30	.33	.78	.77	.78
		1.0	0	2.35	2.42	2.44	.18	.18	.19	.73	.73	.74
	0	0	0	.62	.63	.63	.00	.01	.00	.00	.00	.00
		0	1.0	.58	.59	.60	.02	.03	.03	.01	.01	.01
		1.0	0	.58	.59	.60	.03	.03	.03	.01	.01	.01
	.25	0	0	2.46	2.52	2.53	.03	.00	.01	.65	.64	.64
		0	1.0	2.11	2.17	2.17	.04	.03	.01	.67	.67	.68
		1.0	0	2.27	2.32	2.35	.02	.03	.00	.68	.69	.68
	.5	0	0	1.23	1.25	1.26	.01	.01	.00	.64	.64	.64
		0	1.0	.98	1.00	1.01	.01	.02	.03	.68	.69	.69
		1.0	0	1.07	1.11	1.12	.12	.11	.10	.71	.71	.71
.9	-.5	0	0	.31	.32	.32	.00	.00	.00	.64	.64	.64
		0	1.0	.75	.76	.76	.15	.17	.17	.73	.73	.73
		1.0	0	.30	.31	.31	.03	.03	.03	.68	.69	.69
	-.25	0	0	.62	.63	.64	.02	.00	.00	.64	.64	.64
		0	1.0	.99	1.00	1.00	.13	.13	.15	.72	.72	.73
		1.0	0	.61	.63	.63	.05	.07	.07	.69	.70	.70
	0	0	0	.15	.16	.16	.00	.00	.00	.00	.00	.00
		0	1.0	.16	.16	.16	.01	.02	.01	.00	.01	.00
		1.0	0	.15	.15	.16	.01	.01	.01	.00	.00	.00
	.25	0	0	.60	.62	.63	.01	.01	.00	.64	.64	.64
		0	1.0	.35	.37	.37	.03	.04	.04	.67	.67	.67
		1.0	0	.59	.62	.63	.01	.05	.04	.67	.66	.67
	.5	0	0	.31	.31	.32	.01	.00	.00	.63	.64	.64
		0	1.0	.09	.10	.10	.01	.01	.02	.67	.67	.67
		1.0	0	.29	.30	.31	.02	.01	.00	.68	.68	.68

Table 4: Type I error and power results when group assignment is based on ability.

				Difference Score			ANCOVA			Residual Change		
Rel	**ES**	**Ceil**	**Flr**	**n=20**	**n=50**	**n=100**	**n=20**	**n=50**	**n=100**	**n=20**	**n=50**	**n=100**
.6	-.5	0	0	.06	.08	.12	.12	.25	.46	.00	.02	.08
		0	1.0	.13	.30	.55	.14	.22	.35	.00	.02	.04
		1.0	0	.05	.08	.12	.14	.24	.40	.01	.01	.05
	-.25	0	0	.15	.35	.62	.07	.10	.15	.00	.00	.01
		0	1.0	.23	.54	.87	.08	.10	.13	.00	.00	.00
		1.0	0	.15	.38	.68	.08	.10	.14	.00	.00	.01
	0	0	0	.35	.74	.96	.05	.05	.05	.00	.00	.00
		0	1.0	.39	.81	.98	.06	.07	.07	.00	.00	.00
		1.0	0	.38	.81	.99	.06	.06	.06	.00	.00	.00
	.25	0	0	.59	.95	1.00	.07	.10	.15	.00	.00	.01
		0	1.0	.58	.95	1.00	.08	.12	.17	.00	.00	.01
		1.0	0	.68	.98	1.00	.08	.12	.19	.00	.00	.01
	.5	0	0	.80	1.00	1.00	.12	.25	.46	.00	.02	.08
		0	1.0	.78	.99	1.00	.13	.25	.44	.00	.01	.04
		1.0	0	.88	1.00	1.00	.13	.27	.46	.00	.01	.06
.9	-.5	0	0	.38	.77	.97	.30	.66	.93	.02	.16	.52
		0	1.0	.11	.20	.35	.32	.60	.85	.02	.12	.37
		1.0	0	.47	.84	.99	.33	.66	.92	.02	.14	.46
	-.25	0	0	.07	.11	.17	.11	.22	.40	.00	.02	.06
		0	1.0	.05	.05	.05	.15	.24	.38	.01	.02	.05
		1.0	0	.08	.13	.22	.14	.24	.40	.01	.02	.05
	0	0	0	.11	.23	.43	.05	.05	.05	.00	.00	.00
		0	1.0	.13	.27	.51	.07	.07	.07	.00	.00	.00
		1.0	0	.13	.27	.51	.07	.08	.07	.00	.00	.00
	.25	0	0	.48	.89	1.00	.11	.23	.40	.00	.02	.06
		0	1.0	.41	.83	.98	.13	.24	.43	.00	.01	.05
		1.0	0	.57	.95	1.00	.15	.28	.47	.00	.02	.06
	.5	0	0	.89	1.00	1.00	.31	.66	.93	.02	.16	.52
		0	1.0	.77	.99	1.00	.28	.63	.90	.01	.10	.39
		1.0	0	.94	1.00	1.00	.35	.70	.93	.02	.16	.50

Table 5: Relative bias results when group assignment is moderately related to ability.

				Difference Score			ANCOVA			Residual Change		
Rel	**ES**	**Ceil**	**Flr**	**n=20**	**n=50**	**n=100**	**n=20**	**n=50**	**n=100**	**n=20**	**n=50**	**n=100**
.6	-.5	0	0	.01	.00	.01	.57	.57	.57	.63	.62	.61
		0	1.0	.13	.13	.13	.67	.67	.67	.72	.71	.71
		1.0	0	.16	.18	.17	.62	.64	.63	.68	.68	.67
	-.25	0	0	.01	.01	.01	1.14	1.14	1.14	1.12	1.13	1.12
		0	1.0	.11	.11	.11	1.16	1.15	1.14	1.14	1.13	1.13
		1.0	0	.20	.21	.21	1.16	1.17	1.17	1.14	1.15	1.15
	0	0	0	.00	.00	.00	.28	.29	.28	.24	.25	.25
		0	1.0	.00	.00	.00	.25	.25	.25	.21	.23	.23
		1.0	0	.00	.00	.00	.25	.25	.25	.22	.22	.23
	.25	0	0	.01	.01	.01	1.16	1.15	1.14	.85	.90	.90
		0	1.0	.07	.06	.06	.91	.93	.92	.64	.71	.72
		1.0	0	.31	.29	.27	.75	.76	.78	.51	.56	.59
	.5	0	0	.00	.00	.00	.57	.57	.57	.35	.39	.40
		0	1.0	.06	.05	.06	.44	.43	.43	.24	.27	.28
		1.0	0	.34	.33	.33	.23	.23	.23	.05	.09	.10
.9	-.5	0	0	.00	.00	.00	.19	.19	.19	.34	.32	.32
		0	1.0	.11	.11	.11	.34	.35	.34	.47	.46	.45
		1.0	0	.18	.18	.18	.28	.28	.28	.42	.40	.39
	-.25	0	0	.01	.00	.00	.37	.36	.37	.49	.47	.47
		0	1.0	.07	.09	.09	.46	.48	.48	.57	.57	.56
		1.0	0	.22	.21	.21	.48	.47	.47	.58	.56	.55
	0	0	0	.00	.00	.00	.09	.09	.09	.07	.08	.08
		0	1.0	.00	.00	.00	.09	.08	.09	.07	.07	.07
		1.0	0	.00	.00	.00	.09	.08	.09	.07	.07	.07
	.25	0	0	.01	.01	.00	.37	.38	.37	.10	.15	.15
		0	1.0	.04	.04	.05	.26	.26	.26	.03	.06	.06
		1.0	0	.29	.30	.30	.16	.14	.14	.06	.04	.04
	.5	0	0	.00	.00	.00	.18	.19	.18	.04	.01	.01
		0	1.0	.04	.04	.04	.10	.11	.11	.10	.08	.07
		1.0	0	.35	.35	.35	.07	.06	.07	.24	.22	.21

Table 6: Type I error and power results when group assignment is moderately related to ability

				Difference Score			ANCOVA			Residual Change		
Rel	**ES**	**Ceil**	**Flr**	**n=20**	**n=50**	**n=100**	**n=20**	**n=50**	**n=100**	**n=20**	**n=50**	**n=100**
.6	-.5	0	0	.21	.48	.79	.08	.15	.25	.06	.12	.21
		0	1.0	.21	.49	.79	.07	.12	.21	.06	.10	.18
		1.0	0	.19	.41	.71	.08	.13	.21	.06	.10	.18
	-.25	0	0	.09	.16	.29	.05	.05	.06	.04	.04	.05
		0	1.0	.09	.16	.29	.05	.05	.06	.04	.04	.05
		1.0	0	.08	.13	.23	.06	.06	.06	.04	.05	.05
	0	0	0	.05	.05	.05	.11	.21	.39	.09	.19	.35
		0	1.0	.05	.05	.05	.11	.22	.40	.09	.20	.36
		1.0	0	.05	.05	.05	.11	.22	.40	.09	.19	.36
	.25	0	0	.09	.16	.28	.26	.59	.88	.22	.54	.85
		0	1.0	.09	.17	.31	.26	.59	.88	.22	.54	.85
		1.0	0	.07	.12	.21	.25	.56	.86	.22	.52	.84
	.5	0	0	.21	.48	.78	.48	.89	1.00	.42	.86	.99
		0	1.0	.23	.53	.83	.48	.89	.99	.43	.86	.99
		1.0	0	.13	.32	.58	.44	.86	.99	.40	.83	.99
.9	-.5	0	0	.63	.97	1.00	.41	.83	.99	.33	.77	.98
		0	1.0	.62	.97	1.00	.35	.77	.98	.28	.70	.96
		1.0	0	.51	.92	1.00	.35	.77	.97	.28	.70	.96
	-.25	0	0	.21	.48	.79	.11	.21	.37	.08	.17	.31
		0	1.0	.22	.49	.80	.10	.19	.34	.07	.15	.28
		1.0	0	.16	.37	.66	.09	.17	.31	.06	.13	.25
	0	0	0	.05	.05	.05	.07	.10	.16	.05	.07	.12
		0	1.0	.05	.05	.05	.08	.11	.17	.05	.08	.13
		1.0	0	.05	.05	.05	.08	.11	.18	.05	.08	.14
	.25	0	0	.21	.48	.79	.30	.69	.94	.24	.62	.92
		0	1.0	.23	.52	.82	.32	.70	.94	.26	.63	.91
		1.0	0	.14	.31	.58	.30	.65	.91	.24	.58	.89
	.5	0	0	.63	.97	1.00	.71	.99	1.00	.61	.98	1.00
		0	1.0	.68	.98	1.00	.71	.99	1.00	.63	.97	1.00
		1.0	0	.36	.78	.98	.63	.97	1.00	.54	.95	1.00