STRUCTURAL EQUATION MODELS AND THE REGRESSION BIAS FOR MEASURING CORRELATES OF CHANGE

ROBERT A. CRIBBIE
University of Manitoba

JOHN JAMIESON
Lakehead University

ANCOVA and regression both exhibit a directional bias when measuring correlates of change. This bias confounds the comparison of changes between naturally occurring groups with large pretest differences (ANCOVA), or for identifying predictors of change when the predictor is correlated with pretest (regression). This bias is described in some detail. A computer simulation study is presented, which shows that properly identified structural equation models are not susceptible to this bias. Neither gain scores (posttest minus pretest) nor structural equation models exhibit the "regression bias." Other factors, such as skewness, that may confound measurement of change are also discussed.

The issue of measuring change and predictors of change is fundamental to many areas of research in education and psychology. Although much attention has been directed at the importance of measurements at multiple time periods and the value of growth curves (e.g., Lawrence & Hancock, 1998; Rogosa & Willett, 1985; Willett, 1997), there has been renewed interest in the two phase, pretest-posttest design (e.g., Maris, 1998; Williams & Zimmerman, 1996). Although this simple design does not permit the detailed analysis of individual growth patterns, it has many useful features and is widely used. This design has two main advantages over a posttest-only design: (a) it permits error variance due to consistent individual differences to be removed, thereby increasing power; and (b) it permits the groups to be

Robert A. Cribbie is now at York University.

Educational and Psychological Measurement, Vol. 60 No. 6, December 2000 893-907

equated for baseline differences thereby increasing the internal validity of the design.

Two statistical methods are generally used to analyze data from this design. One method uses "gain scores," $d = y - x$, where y is the posttest score and x is the pretest score. Gain scores are sometimes called *change scores* or *difference scores*. The independent t test comparing mean gain scores for two groups yields the identical conclusion as the interaction term in a 2 (groups) by 2 (pretest/posttest) mixed-ANOVA and answers the question of whether the two groups changed differently. The second method for analyzing this design is ANCOVA, in which the pretest is used as the covariate and the posttest is used as the dependent (criterion) variable. ANCOVA is the discrete case of a partial correlation or R-squared change from multiple regression and answers the question of whether group membership predicts posttest scores after pretest differences between groups are "controlled for" or "removed" (i.e., Did the groups change differently?).

These two methods each achieve the dual goals of controlling for pretest (sometimes called *baseline*) differences and removing error variance. Gain scores achieve these goals through subtraction: The consistent individual differences are subtracted out, which leaves less error variance and removes individual differences from between-group differences. ANCOVA achieves these goals through use of regression lines. The regression line relating pretest and posttest scores within each group is used to calculate a residual score for each participant that contains the amount of posttest not predictable from pretest, which leaves less error variance. The regression lines are also used to adjust the posttest mean to what the mean would have been expected to be if the two groups had started at the same baseline. Because this latter correction is a focus of attention later in this article, it will be described in a bit more detail now.

ANCOVA involves an adjusted treatment mean, $\overline{Y}'_j$, that estimates what performance in the jth group would have been if the group mean on the covariate $\overline{X}_j$ had been equal to the grand mean for the covariate $\overline{X}$. This is calculated from

$$\overline{Y}'_j = \overline{Y}_j - b_w\,(\overline{X}_j - \overline{X}), \tag{1}$$

where b_w is the pooled slope from the within-group regression and $\overline{Y}_j$ is the observed mean on the posttest, which will contain the treatment effect. To understand the nature of this adjustment, it is necessary to focus on what is predicted under the null hypothesis. For example, in a two-group problem, the expected values of $\overline{Y}'_1$ and $\overline{Y}'_2$ are equal for H_o: $\mu_1 = \mu_2$. Hence,

$$\overline{Y}_1 - b_w(\overline{X}_1 - \overline{X}) = \overline{Y}_2 - b_w\,(\overline{X}_2 - \overline{X}). \tag{2}$$

This reduces to

$$\overline{Y}_1 - \overline{Y}_2 = b_w\,(\overline{X}_1 - \overline{X}_2). \tag{3}$$

Equation 3 can be used to illustrate how the expected differences in the posttest ($\overline{Y}_j$) will change as a function of differences in the pretest ($\overline{X}_j$) and the slope (b_w). An example of this effect is presented in Table 1.

From Table 1 it is clear that when b_w is equal to 1, the expected difference in the posttest means equals the difference in the pretest means. However, when b_w is less than 1, the posttest means are expected to come closer together. For illustration, with a pretest difference of +20 and b_w = .5, the expected difference in the posttest means is +10. Thus, ANCOVA expects the posttest means to come closer together than they were at pretest, to regress toward zero difference. The expected difference in posttest means is greater when either the pretest means are farther apart or the slope of the within-group regression (b_w) is lower. The value of b_w will generally be less than 1.0 because

$$b_w = r_{xy}\,(s_y/s_x) \tag{4}$$

and r_{xy} will be less than 1.0 due to measurement error. If the data are approximately normally distributed, the variability should be fairly constant from pretest to posttest. Problems associated with skewed data are addressed in the discussion.

It has been known for some time that the gain score and ANCOVA methods can produce different conclusions. The most dramatic demonstration of this difference is known as Lord's Paradox. Lord (1967) presented a hypothetical example of a group of male and a group of female adolescents each weighed on 2 consecutive years. Even though both groups gained the identical number of pounds (the posttest difference = pretest difference), ANCOVA resulted in the conclusion that males increased significantly more than did females, whereas a *t* test on gain scores yielded a *t* of zero. The resolution of Lord's paradox is quite simple. Assuming b_w is less than 1.0, the observed posttest difference will be greater than expected. The observed difference in posttest means did not regress as expected by ANCOVA under H_o, so the difference in posttest means was interpreted by ANCOVA as being significantly higher than expected. The group with the higher pretest mean (males) was found by ANCOVA to gain significantly more weight than females because they did not regress at posttest as expected.

It is clear that ANCOVA involves a quite different method for measuring change than analysis of gain scores. With gain scores, the expected posttest difference under H_o will be equal to the pretest difference in means. With ANCOVA, the expected posttest difference will be closer to zero than the pre-

Table 1
Expected Difference in Posttest Means ($\overline{Y}_1 - \overline{Y}_2$)
From ANCOVA as a Function of Differences in Pretest Means and Slope

	b_w		
$\overline{X}_1 - \overline{X}_2$	1.0	0.8	0.5
20	20	16	10
10	10	8	5
0	0	0	0
–10	–10	–8	–5
–20	–20	–16	–10

test difference. Thus, ANCOVA expects the pretest mean difference to decrease at posttest.

The important issue is when each of gain scores and ANCOVA is the correct model. When should each be used? The advice most often given is that ANCOVA should only be used with randomized experiments and should be avoided with naturally occurring groups (Huitema, 1980; Rogosa, 1988; Schafer, 1992). The rationale for this advice is that when groups are assigned at random, the group with the higher mean will have received more positive random errors, whereas the group with the lower mean will have received more negative random errors, both of which will tend toward zero on posttest. So, regression toward the mean is expected when groups are randomly assigned. On the other hand, if groups are naturally occurring, there is no reason to expect the mean differences to dissipate. On the contrary, it is more reasonable to assume the mean differences will be maintained on posttest (the gain score assumption).

Although the advice to avoid ANCOVA with naturally occurring groups is widely offered, this advice is conspicuously absent from two articles that were published in influential journals (Maris, 1998; Wainer, 1991). These articles based their recommendations on a causal model designed by Rubin that was specifically applied to resolve Lord's Paradox (Holland & Rubin, 1983). Rubin's model focuses on untested control groups—basically, the question of how the individuals would have changed had there been no treatment. Rather than simply focusing on the question, Did the groups change differently? Rubin's model asks, Did the groups change differently, relative to how they would have changed had they received the control condition? Unfortunately, Rubin's model appears to have distracted attention from the main issue determining when ANCOVA should be used, namely the reason for the baseline differences. Furthermore, there are many possible control groups that might be considered because there really is no single "control condition" implied by a question of differential response to a treatment.

Moreover, Rubin's model led Wainer (1991) to make an incorrect statement that ANCOVA might be preferred to gain scores when there is an underlying baseline drift (scores are increasing even without treatment). That statement is quite incorrect because ANCOVA will still be the wrong method in this situation unless the groups are assigned at random, and gain scores are not affected by the problem of baseline drift.

By failing to provide clear guidelines about when ANCOVA should be used, these two articles (as well as others) have left this issue open, so researchers have the option of choosing either ANCOVA or gain scores. Jamieson (1999) recently pointed out that this option creates an ethical dilemma for researchers because the two methods are too differentially powerful to detect differences in changes. Simply by looking at the direction of pretest differences and knowing which group is expected to change more, it is possible to identify which of ANCOVA or gain scores will be more powerful to detect this hypothesized difference (see Figure 4 in Jamieson, 1999, for an illustration of the patterns of differences that each of ANCOVA or gain scores are better able to detect). Jamieson (1999) pointed out that ANCOVA has a directional bias that is a function of the direction of pretest differences and the direction of differences in change: "ANCOVA is biased to find significance in means which stay apart (parallel lines, as in Lord's paradox) or which diverge. Conversely it is biased against detecting significant effects when the lines converge" (p. 159). Because the difference in power between gain scores and ANCOVA is always in a predictable direction, it is essential to have clear guidelines so that each method will be used only when it is the correct model, to avoid the ethical dilemma and associated Type 1 errors that result from capitalizing on chance differences.

The directional bias also extends beyond the ANCOVA case to the continuous case, where a continuous "third variable" replaces group membership. Thus, instead of asking which group changed more, the question becomes whether a third variable is correlated with amount of change. This question arises in psychophysiology as identifying personality or lifestyle predictors of physiological reactivity to stress, and in clinical or educational contexts as identifying what types of individuals will benefit most from a program. What variables predict who will change the most? Jamieson (1994) showed that the "ANCOVA bias" also appears in the case of the continuous third variable. For generality, the bias will now be referred to as the *regression bias*.

Using computer simulations, Jamieson (1994) compared the partial correlation between a third variable and posttest, controlling for pretest, with the correlation between the third variable and the gain score ($d = y - x$). The partial correlation is the generalization of ANCOVA to a continuous variable, whereas the correlation with the gain score is the generalization of the gain score approach. Simulations were examined that manipulated two conditions: (a) the correlation of the third variable with pretest and (b) the correla-

tion of the third variable with the gain score. When the third variable was uncorrelated with pretest, both gain score correlations and the partial correlations yielded comparable values. However, when the third variable was correlated with pretest, the partial correlations and gain score correlations differed. The pattern of differences showed that partial correlations were biased to detect statistically significant relationships with change, which were in the same direction as the correlation with baseline, and were biased against detecting statistical significance when the correlations were in opposite direction. When the correlation between the third variable and change was zero, large correlations with baseline resulted in many significant partial correlations (Type 1 errors). Thus, the baseline correlation could result in the finding of statistically significant correlation of the third variable with change, even when the real correlation was zero. This finding is the generalization of Lord's Paradox to a continuous variable.

Jamieson (1994) also showed when there was a very large correlation between the third variable and pretest, the regression method sometimes detected statistically significant correlations with change that were in the opposite direction from the real correlation. In every case where there was a large correlation of the third variable with baseline, regression was biased to detect correlations with change in only one direction. These findings are extensions of the ANCOVA bias: When the third variable is highly correlated with pretest (which is comparable to large baseline differences in ANCOVA), regression measures are biased to detect a correlation of change with the third variable that is in the same direction as the correlation with baseline (which is comparable to detecting mean differences that diverge from pretest to posttest in ANCOVA). The ethical dilemma described in Jamieson (1999) also applies to the case of a continuous variable. For example, suppose one predicts that physically fit individuals will show lower heart rate increases to stress than physically unfit. Pretest heart rate shows a large difference because fit individuals have lower heart rates (heart rate is negatively correlated with fitness level). If the study was done with two groups (fit, unfit), then ANCOVA will be biased to find the group with the higher pretest heart rate (unfit) to show a larger increase to stress. On the other hand, if, instead of dichotomizing fitness, a fitness score was used as a predictor of heart rate increase, the regression approach will be biased to detect a negative partial correlation of fitness and posttest heart rate (unfit increasing more). In both cases, the effect could simply be an artifact of the statistical method, not a real finding.

The reason for the regression bias is well understood. Basically, the problem with regression is that the covariate (pretest) is assumed to be measured without error (see, e.g., Darlington, 1990, p. 203). The greater the amount of error variance in the pretest, the less effective regression will be in eliminating pretest variation from the measure of change. The findings of Jamieson

(1994) show that one is in precisely the same situation when a third variable is strongly correlated with pretest, as when there are large baseline differences with naturally occurring groups. In both cases, the use of regression-based methods can result in erroneous conclusions.

The present study was designed to determine whether the regression bias also appears when a structural equation model is used to measure change. Raykov (1992, 1993, 1994) and others have presented structural models that measure change using two congeneric variables (both measures of the same construct), each measured at pretest and posttest. The models include a latent variable for the construct of "change" and allow examination of the relationship between change and a third variable construct, also measured by two congeneric variables. For example, heart rate and systolic blood pressure might both be measured before and after exposure to a stressor. The latent variable change is then a measure of increased cardiovascular activity. A third variable construct, perhaps Type A behavior measured by both a questionnaire and a structured interview, could then be examined as a predictor of the cardiovascular response to stress.

Because structural equation models (SEM) involve explicit modeling of error variance for all observed and latent measures, it should not be susceptible to the regression bias. The purpose of the following study is to evaluate whether this is in fact true. Simulations were created, similar to those of Jamieson (1994) except that two measures are included instead of one. Different conditions are then created to manipulate the correlation of the third variable with pretest, and the correlation of the third variable with change. Three summary statistics are then examined to determine the effect of these manipulations: (a) the correlation between the gain scores and the third variable; (b) the partial correlation between the third variable and posttest, controlling for pretest; and (c) the path from the structural model linking the latent third variable to the latent change measure. The first two statistics are included to replicate the regression bias demonstrated by Jamieson (1994). The third statistic will show whether SEM are also affected by the regression bias.

Method

Pseudorandom normal variates were generated by the SAS generator RANNOR (SAS Institute, 1990a). The data generation was similar to methods used in Jamieson and Howk (1992) and Jamieson (1994). Three normally distributed variables were generated with means of zero and standard deviations of two to represent (a) a pretest (baseline) for two congeneric measures (X), (b) a measure of amount of change (C) in the two measures from pretest to posttest, and (c) a construct underlying a third variable or correlate of change (Z). In addition, six normally distributed variables were generated

with means of zero and standard deviations of one to account for measurement error in the dependent variables.

Pretest measures of the repeatedly assessed variables were computed from the construct (X) and error variables (e_1, e_2):

$$\text{PRE}_1 = X + e_1, \text{PRE}_2 = X + e_2.$$

Posttest measures were computed by adding a measure of change (C) to the pretest construct (X) and individual error variables (e_3, e_4):

$$\text{POST}_1 = X + C + e_3, \text{POST}_2 = X + C + e_4.$$

The advantage of using computer simulated data is that specific relationships can be built into the data and tested using various statistical methods. In this study, the data were generated such that there was either a positive, negative, or null relationship between the two congeneric measures of the third variable ($Z1$, $Z2$) and pretest, as well as a positive or negative relationship between the two measures of the third variable and change. The positive and negative relationships were identical in magnitude. Therefore, six conditions were created by manipulating the relationships between the third variable measures and baseline (0, +, –), and the third variable measures and change (+, –). For example, to generate congeneric third variable measures that correlate positively with baseline and negatively with change,

$$Z1 = Z + .75X - .75C + e_5, Z2 = Z + .75X - .75C + e_6. \quad (5)$$

The variability of the latent pretest (X) and change (C) constructs were reduced (i.e., premultiplied by .75) to generate empirically realistic correlations between the third variable measures and the pretest/posttest measures. In addition, to generate third variable measures that are not correlated with pretest scores, the pretest construct (X) is removed from the equation. For example, to generate third variable measures that are correlated positively with change, but *not* correlated with baseline,

$$Z1 = Z + .75C + e_5, Z2 = Z + .75C + e_6. \quad (6)$$

Gain Scores

Gain scores (G1, G2) for the two repeatedly assessed variables were computed by taking the difference between pretest and posttest scores:

$$\text{G1} = \text{POST}_1 - \text{PRE}_1, \text{G2} = \text{POST}_2 - \text{PRE}_2.$$

Correlations were computed between each of the gain scores and each of the congeneric measures of the third variable. A mean correlation was computed by averaging the four correlations between the gain scores and third variables.

Regression-Based Method (partial correlations)

Partial correlations were computed between posttest scores and the congeneric measures of the third variable after controlling for the corresponding pretest scores. An average partial correlation was computed by averaging the four partial correlations between the posttest scores and the third variable measures.

Structural Equation Modeling

The structural model used in this study (see Figure 1) was derived from research by Raykov (1993, 1994, 1997); MacCallum, Kim, Malarkey, and Kiecolt-Glaser (1997); Duncan et al. (1997); Steyer, Eid, and Schwenkmezger (1997); and others for measuring change and identifying correlates of change with SEM. The observed variables in this model are the two pretest measures, the two posttest measures, and the two congeneric measures of the third variable. Latent measures in this model represent the baseline (X), the true change between pretest and posttest (C), and the third variable construct (Z). The model definition equations for this model are

$$\mathrm{PRE}_1 = X + e_1,\ \mathrm{PRE}_2 = X + e_2,$$

$$\mathrm{POST}_1 = X + (a)C + e_3,\ \mathrm{POST}_2 = X + (b)C + e_4,$$

$$Z1 = (c)Z + e_5,\ Z2 = Z + e_6,$$

where a, b, and c are unknown path coefficients to be estimated from the data, and $e_1 - e_6$ are the corresponding errors of measurement, also estimated from the data, with the constraint that $e_1 = e_3$ and $e_2 = e_4$. The variances of the latent baseline and third variable constructs were estimated from the data, and to achieve identification of the model, the variance of the latent change variable was set equal to one. The model contains 12 degrees of freedom, with 9 unknowns being estimated from 15 covariances and 6 variances.

To test the fit of the model to the data, the Goodness of Fit Index (GFI) (Jöreskog & Sörbom, 1989), Adjusted Goodness of Fit Index (AGFI) (Jöreskog & Sörbom, 1989), and Root Mean Square Error of Approximation (RMSEA) (Browne & Cudeck, 1993) were recorded for each analysis. Cutoff values of .95 were used to establish a good fit for the GFI and AGFI (Shevlin & Miles, 1998), and .05 was used to establish a good fit for the RMSEA (Browne & Cudeck, 1993; Hu & Bentler, 1999). Fan, Thompson, and Wang (1999) found that, relative to other fit indices, the GFI, AGFI, and RMSEA were sensitive to model misspecifications, the GFI and AGFI were

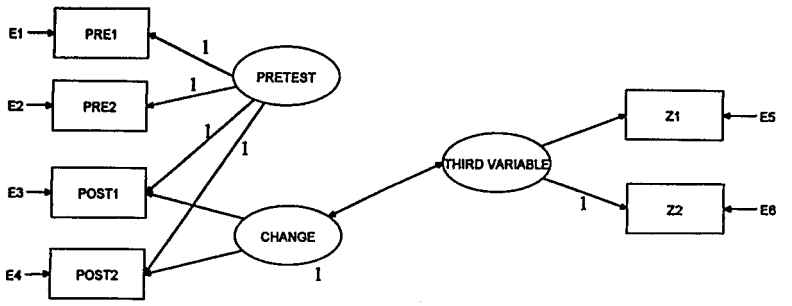

Figure 1. The structural model used to identify correlates of change.
Note. All parameters not fixed at a set value (1) are estimated from the data, with the only constraint that E1 = E3 and E2 = E4.

insensitive to estimation method, and the RMSEA was insensitive to sample size variation. Based on these results, Fan et al. (1999) recommended the use of these fit indices in research. As always, it is important to examine multiple fit indices to get a good overall "picture" of the fit of the model to the data.

Change score correlations, partial correlations, and structural model analyses were computed over 1,000 simulations of 200 cases each, for the six conditions. All analyses were performed using SAS (SAS Institute, 1990a). The structural models were tested against the data using SAS PROC CALIS (SAS Institute, 1990b) and maximum likelihood estimation.

Results

Gain Scores

The mean correlations between the gain scores and the third variables for the six conditions can be found in Table 2. The results were equivalent to those of Jamieson (1994), with the correlations between the gain scores and the third variables consistent across each of the conditions. In other words, the correlations between the gain scores and the third variable scores were unaffected by the relationship of the third variable with pretest. As expected, the correlations between the gain scores and the third variable were larger when there was no correlation between the pretest scores and the third variables as a result of the increased proportion of variability accounted for by the correlation between change and the third variables (see Equations 5 and 6).

Regression (partial correlation)

The mean partial correlations between the posttest scores and each of the third variables after controlling for pretest scores can be found in Table 3. The

Table 2
Mean Correlations Between the Gain Scores and the Third Variables

	Baseline and Third Variable Relationship		
Change and Third Variable Relationship	Negative	Zero	Positive
Negative	–.397	–.455	–.399
Positive	.398	.453	.396

Table 3
Mean Partial Correlations Between Posttest Scores and Third Variable Scores After Controlling for Pretest Scores

	Baseline and Third Variable Relationship		
Change and Third Variable Relationship	Negative	Zero	Positive
Negative	–.539	–.460	–.360
Positive	.360	.463	.538

results were again consistent with those found by Jamieson (1994). When there was no relationship between pretest scores and third variable scores, the average partial correlations were quite similar to the gain score correlations, although when a relationship existed between baseline scores and third variable scores, the partial correlations were biased. Specifically, when the relationship between the baseline scores and the third variable scores was congruent with the relationship between change and the third variable scores (i.e., both relationships positive or both relationships negative), the partial correlations were inflated. In contrast, when the relationship between the baseline scores and the third variable scores was incongruent with the relationship between change and the third variable scores (i.e., one relationship positive and one relationship negative), the partial correlations were deflated. Thus, partial correlations of a third variable with change are very much affected by the correlation of the third variable with baseline.

Structural Modeling

To verify that the proposed structural model provided a good fit to the simulated data, the mean GFI, AGFI, and RMSEA were computed for each condition. The mean GFI, AGFI, and RMSEA were similar across conditions and indicate a good overall and parsimonious fit of the model (.983 to .984, .965 to .966, and .018 to .019, respectively).

Table 4
Mean Correlations Between the Latent Change and the Latent Third Variable

	Baseline and Third Variable Relationship		
Change and Third Variable Relationship	Negative	Zero	Positive
Negative	–.513	–.597	–.512
Positive	.514	.597	.512

The mean correlations between the latent third variable and the latent change variable for the six conditions can be found in Table 4. The results for the proposed model were similar to the results found for the change score correlations; namely, the relationship between the third variables and change was unaffected by the relationship between the third variables and baseline. Again, as expected, the correlations between the change scores and the third variable were larger when there was no correlation between the pretest scores and the third variables as a result of the increased proportion of variability accounted for by the correlation between change and the third variables (see Equations 5 and 6).

Discussion

These simulations replicate the regression bias demonstrated by Jamieson (1994) and show that SEM are not affected by this bias. Although the correlation of a third variable with pretest confounded regression measures of change, SEM and gain scores were unaffected. This finding, although not surprising, is nevertheless reassuring and confirms that SEM are an unbiased method for identifying correlates of change. SEM are becoming more widely used and have great potential for testing unambiguous, predetermined, and theoretically important models. However, in situations where SEM are not appropriate, either because of small sample size or because two congeneric measures are not available, gain scores remain a good alternative for analyzing change because they are not affected by correlations of a third variable with pretest.

However, gain scores are not perfect. Although earlier concerns about poor reliability of gain scores are no longer seen as a serious problem (Llabre, Spitzer, Saab, Ironson, & Schneiderman, 1991; Williams & Zimmerman, 1996), the question remains of whether gain scores might also be biased measures of change under some circumstances. Wainer (1991) referred to "the very difficult problems associated with restriction of range" (p. 150). Specifically, when data are skewed, perhaps because of floor or ceiling effects, are gain scores unbiased measures of change? The answer is *no*! Jamieson

(1995) used computer simulations to compare gain scores and ANCOVA when variance decreased from pretest to posttest. The decrease in variance was created through simulating a floor/ceiling effect, through skewness, and through just manipulating variance. In all three cases, gain scores lost power relative to ANCOVA when the variance decreased from pretest to posttest.

Jamieson (1999) described a principle that gain scores are confounded with pretest whenever data are skewed. As an illustration, he offers the following example:

> Another way to illustrate this principle is to consider a measurement model in which the true dimension is normally distributed, but the observed measure is positively skewed. Suppose a square root transformation will change the observed measure to a normal distribution, eliminating the skewness and thereby reflecting the true underlying process. Consider a numerical illustration, in which observed scores of 25 and 100 correspond to true scores of 5 and 10 (the square roots). If a task produced an increase of 1 in each true score, from 5 to 6 and from 10 to 11, this would correspond to changes in the observed scores from 25 to 36 and from 100 to 121. The changes for the observed scores are 36 – 25 = 9 and 121 – 100 = 21. Thus the observed scores show a difference (21 – 9 = 15) in the amount of change, even though the true scores both have the identical increase of 1 unit. The positive skewness caused the higher score to change more. (pp. 156-157)

Thus, there is a real concern about the confound of comparing changes in groups that start from different baselines when the data are skewed. At present, the best solution appears to be to examine the shape of the distribution and apply an appropriate transformation, if possible. Williams and Zimmerman (1996) recently dealt in some depth with the issue of changing shape of distributions. They pointed out that in many educational contexts, data may start from a floor (positively skewed distribution) and increase in variance in response to a treatment to show a more symmetrical distribution. If the data are only skewed at pretest, a transformation applied to both pretest and posttest would not be appropriate. It is not clear how to obtain an unbiased measure of change in such circumstances. Although there are still unanswered questions about how to measure change in the 2-phase design, the two guidelines suggested by Jamieson (1999) are a good starting point: Avoid ANCOVA except for randomized experiments, and avoid using gain scores with skewed data.

The potential of SEM for measuring correlates of change is an important topic for further research. The present study showed that SEM are not affected by the regression bias and are able to accurately estimate the relationship between a third variable and change, independent of whether the third variable is also correlated with pretest. One question of particular importance is whether SEM will show the same confounds with skewed data that are evident with gain scores. Previous research has reported that maximum likelihood estimation in SEM is relatively robust to moderate violations

of normality (Ip & Willson, 1998) and that asymptotically distribution-free test statistics may provide accurate parameter estimates and standard errors for nonnormal data. Therefore, it is possible that SEM, unlike gain scores, may provide accurate estimates (and statistical tests) of the relationship between change and correlates of change when the variables are skewed. We are currently exploring this question.

References

Browne, M. W. & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Darlington, R. B. (1990). *Regression and linear models*. New York: McGraw-Hill.

Duncan, T. E., Duncan, S. C., Alpert, A., Hops, H., Stoolmiller, M., & Muthen, B. (1997). Latent variable modeling of longitudinal and multilevel substance abuse data. *Multivariate Behavioral Research*, *32*, 275-318.

Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling*, *6*, 56-83.

Holland, P. W., & Rubin, D. B. (1983). On Lord's Paradox. In H. Wainer & S. Messick (Eds.), *Principles of modern psychological measurement* (pp. 3-35). Hillsdale, NJ: Lawrence Erlbaum.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1-55.

Huitema, B. E. (1980). *The analysis of covariance and alternatives*. New York: Wiley.

Ip, C., & Willson, V. L. (1998, April). *Parameter estimates and fit indices in covariance structure analysis with nonnormal data*. Paper presented at the 1998 Annual Meeting of the American Educational Research Association, San Diego, CA.

Jamieson, J. (1993). The law of initial values: Five factors or two? *International Journal of Psychophysiology*, *14*, 233-239.

Jamieson, J. (1994). Correlates of reactivity: Problems with regression based methods. *International Journal of Psychophysiology*, *17*, 73-78.

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

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

Jamieson, J., & Howk, S. (1992). The law of initial values: A four factor theory. *International Journal of Psychophysiology*, *12*, 53-61.

Jöreskog, K., & Sörbom, D. (1989). *LISREL 7: A guide to the program and applications* (2nd ed.). Chicago: SPSS.

Lawrence, F. R., & Hancock, G. R. (1998). Assessing change over time using latent growth modeling. *Measurement and Evaluation in Counseling and Development*, *30*, 211-224.

Llabre, M. M., Spitzer, S. S., Saab, P. G., Ironson, G. H., & Schneiderman, N. (1991). The reliability and specificity of delta versus residualized change as measures of cardiovascular reactivity to behavioral challenges. *Psychophysiology*, *28*, 701-711.

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

MacCallum, R. C., Kim, C., Malarkey, W. B., & Kiecolt-Glaser, J. K. (1997). Studying multivariate change using multilevel models and latent curve models. *Multivariate Behavioral Research*, *32*, 215-253.

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

Raykov, T. (1992). Structural models for studying correlates and predictors of change. *Australian Journal of Psychology*, *44*, 101-112.

Raykov, T. (1993). A structural equation model for measuring residualized change and discerning patterns of growth or decline. *Applied Psychological Measurement*, *17*, 53-71.

Raykov, T. (1994). Studying correlates and predictors of longitudinal change using structural equation modeling. *Applied Psychological Measurement*, *18*, 63-77.

Raykov, T. (1997). Simultaneous study of individual and group patterns of latent longitudinal change using structural equation modeling. *Structural Equation Modeling*, *4*, 212-236.

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

Rogosa, D. R., & Willett, J. B. (1985). Understanding correlates of change by modeling individual differences in growth. *Psychometrika*, *50*, 203-228.

SAS Institute. (1990a). *SAS procedures guide* (Version 6, 3rd ed.). Cary, NC: Author.

SAS Institute. (1990b). *SAS/STAT user's guide* (Version 6, 4th ed.). Cary, NC: Author.

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

Shevlin, M., & Miles, J.N.V. (1998). Effects of sample size, model specification and factor loadings on the GFI in confirmatory factor analysis. *Personality and Individual Differences*, *25*, 85-90.

Steyer, R., Eid, M., & Schwenkmezger, P. (1997). Modeling true intraindividual change: True change as a latent variable. *Methods of Psychological Research* [Online]. Available: www.hsp.de/MPR

Wainer, H. (1991). Adjusting for differential base rates: Lord's paradox again. *Psychological Bulletin*, *109*, 147-151.

Willett, J. B. (1997). Measuring change: What individual growth modeling buys you. In E. Amsel & K. A. Reninger (Eds.), *Change and development* (pp. 213-243). Mahwah, NJ: Lawrence Erlbaum.

Williams, R. H., & Zimmerman, D. W. (1996). Are simple gain scores obsolete? *Applied Psychological Measurement*, *20*, 59-69.