

Monte Carlo Analysis of Change Scores

David J. Mangen, Mangen Research Associates, Inc.

Introduction

Issues in understanding processes of change are central to almost all fields of scientific inquiry. Methodological developments in the design of studies that address the issues associated with change include new approaches to cross-sequential designs and other methods for examining time-based trends. Appropriate designs and statistical models that decompose change effects into age, period, and cohort effects are now recognized as methodologically necessary conditions for an accurate understanding of change in populations.

Related to this trend, researchers now recognize that early longitudinal research was plagued by a reliance on simple difference change scores (Bohrstedt, 1969; Cronbach and Furby, 1970). Simple change scores are negatively correlated with time-one (T_1) scores due to regression to the mean. The near consensus - with some notable disagreements - is that analysis of simple change scores *per se* is not appropriate. More commonly, analysis of change is now accomplished by analyzing time-two (T_2) outcomes after controlling for T_1 origin status.

These developments mean that current research using change models generally present accurate results, although it is clear that unusual data will necessarily influence any analysis. An unfortunate consequence of the conclusion that simple difference scores are biased is that it is not unusual to see much of early, ground-breaking research that used simple change scores as criterion measures discounted solely for that reason. Since science is cumulative, this may be inefficient.

The purpose of this paper is threefold: (1) To what extent has previous research using simple change scores distorted the true underlying data? This question goes to the heart of the tendency to discount prior change studies. (2) To what extent are currently accepted models for analyzing change biased by "typical" data problems inherent in longitudinal data? (3) Can the SAS® system be used to design a Monte Carlo simulation study to address these issues? More particularly, can a simulation be created that empirically examines the biasing effects of typical data problems on both the analysis of longitudinal data?

Model Presentation

For purposes of this analysis, we consider a simple four variable model including two exogenous characteristics (X_1 and X_2) and one endogenous

characteristic measured at two points (T_1 and T_2). Using regression equations, early research examining simple difference change scores would commonly have looked at an equation such as the following:

$$(T_2 - T_1) = a + \beta_1 X_1 + \beta_2 X_2 + e \quad [\text{Eq. 1}]$$

The currently recommended strategy for analyzing data such as these, assuming that there is no problem with the left or right-censoring of the data, would usually frame the analysis as:

$$T_2 = a + \beta_3 T_1 + \beta_4 X_1 + \beta_5 X_2 + e \quad [\text{Eq. 2}]$$

In essence, this model implies that we are not analyzing change *per se* but rather that we are analyzing T_2 outcomes after controlling for T_1 origin status. It is readily demonstrated, however, that if the researcher wants to analyze simple difference change scores, it is quite legitimate to do so if T_1 origin status is properly controlled. That is:

$$(T_2 - T_1) = a + \beta_6 T_1 + \beta_7 X_1 + \beta_8 X_2 + e \quad [\text{Eq. 3}]$$

The model in Eq. 3 is functionally equivalent to the model in Eq. 2. The coefficients of Eq. 2 and Eq. 3 are identical, or vary by a constant factor of 1.0, i.e., $\beta_4 = \beta_7$, $\beta_5 = \beta_8$, and $\beta_3 - 1.0 = \beta_6$ (Carlton-Ford, Simmons, and Blyth, 1983). Thus, if researchers are interested in analyzing change scores they can in fact do so provided that adjustment for origin status is incorporated into the model.

The results yielded by Eq. 2 and Eq. 3 are sample estimates subject to normal sampling fluctuations, the population parameter correlations among the four variables, measurement reliability, and known and unknown selection factors intervening between T_1 and T_2 (Reichardt, 1979).

For purposes of this paper, selection factors that influence the coefficients in the equations are ignored. This analysis focuses on: 1) defining populations with known parameters among the four variables in the model; 2) sampling from that population and introducing varying degrees of measurement error, and varying degrees of correlated measurement error; and 3) since the population parameters among the four variables are known, comparing the population coefficients to the sample-specific coefficients obtained in Eq. 1 and Eq. 3.

With this design, we can examine the degree to which studies using simple difference change

scores yield biased results under varying conditions. Because we know the true scores and the true parameters, we can test the biasing factor. Furthermore, we can extend the analysis to examine the degree to which measurement error factors bias the observed coefficients in the covariate adjustment model (Eq. 3).

Method

The DATA step capabilities of the SAS® system were used to program a Monte Carlo probability simulation that is used to analyze these issues. Five test factors that influence population regression coefficients and represent typical research data are systematically varied, and repeated samples from each theoretical population are drawn. For each sample, we analyze both the simple difference and covariate-adjustment models of change, and compare the obtained results with the known population parameters.

In order to assist in presentation of the characteristics of the design, and to simplify subscripts of coefficients, throughout the remainder of this presentation the variable T_1 in Eq. 1-3 is henceforth referenced as X_3 , and T_2 is referenced as X_4 . The five factors included in the simulation design are:

1. Measurement error levels analogous to α reliability for X_3 and X_4 of 0.55, 0.70, and 0.85 are systematically varied. An α of 0.85 is generally considered quite good for research purposes, but α levels of 0.55 are seen in the published literature. These are referenced as test Factor I.
2. The degree of over-time correlation in the measurement error is also systematically varied. In longitudinal research, failures in measurement get repeated across the periods of measurement, and since the same subjects are studied, the failures get repeated - usually systematically. Thus, three levels of increasingly severe correlated measurement error (0.10, 0.40, and 0.70) constitute Factor J.
3. The time-one to time-two correlation (r_{34}) constitutes Factor K. It is usual to see at least *some* relationship between origin and outcome status; we use correlations between X_3 and X_4 of 0.20, 0.40, and 0.60 in this analysis.
4. The correlations of the two exogenous variables with the time-one measure

(r_{13} and r_{23}) are Factor L. These correlations are varied as a block, with the first level specifying $r_{13} = 0.30$ and $r_{23} = 0.40$, and the second level specifying $r_{13} = 0.40$ and $r_{23} = 0.50$.

5. Factor M is the correlation of the two exogenous variables with the time-two measure (r_{14} and r_{24}). These correlations are also varied as a block, with the first level specifying $r_{14} = 0.30$ and $r_{24} = 0.20$, and the second level specifying $r_{14} = 0.40$ and $r_{24} = 0.30$.

Note that r_{12} never varies and is fixed at 0.25; this is done to simplify the model. Further, perfect reliability of X_1 and X_2 is assumed. The complete cross-classification of these factors yields a $3 \times 3 \times 3 \times 2 \times 2$ matrix of 108 cells. While the levels of these test factors are arbitrarily chosen values, we feel that they represent a range of values typical in social science research.

For each of the 108 cells in the design matrix, a SAS® computer program generates a theoretically infinite population of four true scores with the specified correlations among the variables. Each true score is specified to have a mean of 0.0 and a standard deviation of 1.0. From this population, 25 random samples of $n = 300$ are drawn, and error vectors corresponding to the specified level of reliability and correlated measurement error are added to the true scores of X_3 and X_4 . Thus, a total of 2,700 (25×108) random samples of $n = 300$ are drawn for this analysis.¹

For each sample, three regression equations (PROC REG) are computed. First, as a check on the accuracy of the program that generates the data, Eq. 3 above is tested using the *true score* data for all variables. This was done merely to verify the accuracy of the simulation programming. Second, a raw change score model (Eq. 1) is estimated. Finally, the covariate adjustment model (Eq. 3) is estimated on the *observed scores* incorporating correlated measurement error. We never estimate Eq. 2, since we know that the coefficients are identical to those obtained in Eq. 3.

One of the primary benefits to programming this simulation in SAS® lies in the fact that we have a composite tool that first generates the data with known characteristics plus performs all of the statistical tests. The flexibility of the SAS® system in creating summary datasets containing model statistics greatly facilitates this analysis. We output a dataset of the regression statistics for each random sample of 300, and append these files for each of the 25 simulations in that cell of the design to build the analysis file. These data are supplemented with the population parameters. This allows us to create

the dependent variables for this analysis. Specifically, by simply subtracting the 25 obtained sample coefficients from the known population parameters, we obtain measures of *the inaccuracy in the regression coefficients that have been introduced by the test factors in the experiment*. If there is no systematic bias, then these variables will show a mean of 0.0, with only sampling variation in the values. A *positive* difference implies that the observed regression coefficient is smaller than the population parameter, i.e., the coefficient is attenuated, while a *negative* difference implies that the observed coefficient is greater than the population parameter.

The criterion variables include the following measures: 1) B42DIFF, which is the inaccuracy in β_{42} , 2) B41DIFF, which is the inaccuracy in β_{41} , and 3) B43DIFF, computed only in the models controlling for origin status, which is the inaccuracy in β_{43} .

Data Analysis

Since the resulting data matrix is a fully-crossed design with equal sample sizes for each cell, analysis of these data used standard analysis of variance techniques. For each dependent variable, we analyzed an hierarchical set of models testing first the main effects of the five test factors and higher-order interactions.

With this approach, it became apparent that for each model and each dependent variable, the main effects model was clearly the most important. The main effects model explains at worst 50% of the variance in the dependent variables. Each set of higher-order interactions significantly contributes to explanation, although the three through five-way interactions never add more than 2% to the explained variance. Analysis of the total explained variation for each of the coefficients in the two models shows greater predictability of errors in β_{42} than β_{41} , with the greatest predictability for errors in the stability coefficient β_{43} .

Since the two-way interactions yield significant improvements in explanation of the errors in the regression coefficients while the higher-order effects add comparatively little, the remainder of this analysis is restricted to the two-way model, with greater emphasis placed on the main effects. Since the sample size for this analysis is quite large, the environment highly controlled, and a relatively large number of effects tested, we discuss those effects significant at $p < 0.0001$.

Effects on the Simple Difference Model: Four of the five test factors exhibit strong main effects. For both B42DIFF and B41DIFF, measurement reliability, the time-one to time-two correlation of the endogenous variables, and the correlations of the independent

variables with both the time-one and time-two measures strongly influence the accuracy of the coefficients. The degree of correlated measurement error influences the accuracy of the coefficients, but this effect is comparatively weak.

Of the interaction effects, the two-way interaction K^*L , which tests the joint effect of varying the time-one to time-two correlation and the correlations of the exogenous variables with the time-one endogenous measure, is a strong predictor of both B41DIFF and B42DIFF. The interactions of measurement reliability with all of the other test factors are determinants of both B41DIFF and B42DIFF, although in only one case is a strong effect exhibited. For B41DIFF, the interaction of correlated measurement error with time-one to time-two correlation is a strong determinant (J^*K), while for B42DIFF the interaction of correlations of the exogenous measures with the time-one and time-two endogenous measures (L^*M) is significant.

Effects on the Covariate Adjustment Model:

Significant main effects are noted for four of the five test factors. As expected, measurement reliability, correlated measurement error, the magnitude of the time-one to time-two correlation, and the correlation of exogenous measures with the time-two endogenous measure all significantly influence errors. The magnitude of the correlation between exogenous variables and the time-one endogenous measure significantly influences errors in the stability coefficient (β_{43}), but does not strongly influence errors in β_{41} or β_{42} .

Measurement reliability interacts with correlated measurement error (I^*J), with the magnitude of the time-one to time-two correlation (I^*K), and with the correlation of exogenous measures with the time-two endogenous measure (I^*M). The magnitude of the correlated measurement error (J^*L), the size of the time-one to time-two correlation (K^*L), and the strength of the correlation between exogenous measures and the time-two endogenous measure (L^*M) all interact with the strength of the correlation between exogenous measures and time-one endogenous data to influence the accuracy of regression coefficients.

It is apparent from these results that the test factors used to structure the Monte Carlo analysis strongly influence the accuracy of regression coefficients obtained in any study analyzing change. This is not unexpected; the degree of measurement reliability has long been recognized to influence the accuracy of sample estimates of parameters, and the other test factors included in this analysis may be seen as proxies for the degree of collinearity present in the data.

In Table 1, the mean main effects for test factors on the simple difference and covariate adjustment models are presented. Positive means imply coefficients that are attenuated, while negative means imply coefficients that are inflated.

As expected, the inaccuracy of the regression coefficients is much greater for coefficients estimated in the simple difference than those estimated in the covariate adjustment model. In all cases, the coefficients estimated in the simple difference model are attenuated. This addresses the initial goal of this paper. Early research that used simple difference scores should not be discounted merely for that reason. Indeed, any significant results that were obtained in early longitudinal studies are, in all likelihood, even more significant than previously thought. Thus, discounting the results of these earlier studies for this reason alone is clearly not warranted.

The main effect of reliability on errors in the regression coefficients is not as clear as expected. We had anticipated that increasing reliability should yield increased accuracy of coefficients. This is true of all coefficients estimated in the covariate adjustment factor, and is also true of β_{41} in the simple difference model. In contrast, however, the accuracy of β_{42} in the simple difference model is actually greater under conditions of poorer reliability ($B42DIFF = 0.17$) vs higher reliability ($B42DIFF = 0.24$).

The main effect for correlated measurement error is noteworthy only for the covariate adjustment model. A low degree of correlated measurement error produces an inflated estimate of the stability coefficient β_{43} , while greater degrees of correlated error lead to attenuation of this coefficient. Thus, we over-estimate stability when correlated error is substantial, and under-estimate when it is small. The accuracy of β_{41} and β_{42} is also influenced by the magnitude of correlated measurement error, with greater attenuation noted for increasing levels of error correlation. These results clearly suggest that adjustment for correlated measurement error using procedures such as LISREL (Jöreskog and Sörbom, 1979) or the recently released PROC CALIS is necessary in studies such as these.

Not surprisingly, the main effect for the magnitude of the time-one to time-two correlation is greater in the simple difference model than the covariate adjustment model. In the simple difference model, increasing time-one to time-two correlations produces substantially decreased attenuation of coefficients. In the covariate adjustment model, a similar pattern is noted for β_{41} , while for β_{42} and β_{43} the highest degree of time-one to time-two correlation produces coefficients that are slightly inflated.

The effect of the correlations of the exogenous measures on both the time-one and time-two endogenous measures is clear. In the simple difference model, increasing degrees of correlation with both the time-one (Factor L) and time-two (Factor M) measures yields increasing degrees of attenuation in the obtained coefficients. In the covariate adjustment model, varying the correlations with the time-one measure (Factor L) has no effect on β_{41} and β_{42} , and only a small albeit statistically significant effect on the accuracy of β_{43} . Varying the magnitude of correlations with the time-two measure (Factor M) leads to increasing attenuation of all coefficients. This is likely due to collinearity effects.

Discussion and Conclusion

The results of this Monte Carlo analysis have implications for both the credence given to longitudinal studies that have used simple difference change scores as well as current longitudinal research efforts. First, the results of this analysis clearly suggest that studies that examined simple difference change scores without controlling for time-one origin status in the dependent variables have in all likelihood underestimated the significance of possible correlates. Studies that found significant effects should not be completely discounted for flawed methods; rather, the importance of these findings should probably be highlighted.

Published work that yielded inconsistent or nonsignificant findings may well warrant reanalysis. Alternatively, with the results presented here, it may be possible to estimate the actual parameters present in some of this earlier research.

For new longitudinal research analyses, the results presented here suggest that covariance adjustment procedures, even when estimated by ordinary least squares techniques, are reasonably robust to problems in estimation. To the extent that problems are present in OLS models of change, these problems are most severe in estimation of stability coefficients, an important but often less substantively interesting issue.

A strong degree of correlated measurement error can also attenuate coefficients to a considerable degree, as can poor reliability. Techniques such as LISREL and PROC CALIS are available that begin to address these concerns.

Finally, the magnitude of the time-one to time-two correlation in the change variable of interest also strongly influences effect coefficients in longitudinal analysis. Researchers who are examining models of change in their longitudinal data are encouraged to carefully address these issues in subsequent analyses carried out with their data.

FOOTNOTES

1. Readers interested in the SAS® code may contact the author.

REFERENCES

- Bahrnstedt, G. "Observations on the Measurement of Change." Pages 113-133 in E.F. Borgatta (ed.), *Sociological Methodology 1969*. San Francisco: Jossey-Bass, 1969.
- Carlton-Ford, S., R. Simmons, and D. Blyth. "Analyzing Longitudinal Data From Quasi-Experimental Designs: The Analysis of Change." Paper presented to the Annual Meeting of the American Sociological Association, Detroit, MI, 1983.
- Cronbach, L. and L. Furby. "How Should We Measure 'Change'--Or Should We?" *Psychological Bulletin*, Vol. 74(1): 68-80, 1970.
- Jöreskog, K. and D. Sörbom. *Advances in Factor Analysis and Structural Equation Models*. Cambridge, Massachusetts: Abt Books, 1979.
- Reichardt, C. "The Statistical Analysis of Data from Non-Equivalent Group Designs." Pages 147-206 in T. Cook and D. Campbell (eds.), *Quasi-Experimentation*. Chicago: Rand McNally, 1979.

Author Name and Address:

David J. Mangen
Mangen Research Associates, Inc.
Butler Square, Suite 220C
100 North Sixth Street
Minneapolis, MN 55403-1504
(612) 332-2764

Table 1
Means and Standard Deviations for Main Effects

Simple Difference Model						Covariate Adjustment Model					
		B42Diff		B41Diff		B42Diff		B41Diff		B43Diff	
Factor											
I	N	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	900	0.17	0.11	0.21	0.09	0.04	0.10	0.12	0.08	0.08	0.26
2	900	0.21	0.11	0.20	0.09	0.03	0.09	0.08	0.07	0.06	0.20
3	900	0.24	0.11	0.18	0.09	0.02	0.07	0.04	0.06	0.04	0.12
J											
1	900	0.21	0.12	0.20	0.10	-0.01	0.09	0.05	0.08	-0.11	0.15
2	900	0.21	0.12	0.20	0.09	0.04	0.09	0.08	0.08	0.06	0.14
3	900	0.20	0.10	0.19	0.08	0.07	0.08	0.11	0.07	0.23	0.16
K											
1	900	0.31	0.09	0.26	0.08	0.10	0.08	0.13	0.08	0.21	0.17
2	900	0.21	0.08	0.20	0.07	0.03	0.07	0.08	0.07	0.06	0.15
3	900	0.10	0.07	0.12	0.06	-0.04	0.07	0.03	0.06	-0.09	0.16
L											
1	1350	0.19	0.10	0.18	0.08	0.03	0.08	0.08	0.07	0.06	0.19
2	1350	0.22	0.12	0.21	0.10	0.03	0.10	0.08	0.09	0.07	0.22
M											
1	1350	0.18	0.11	0.17	0.09	0.01	0.09	0.06	0.08	0.03	0.20
2	1350	0.24	0.11	0.22	0.09	0.06	0.09	0.10	0.08	0.09	0.20