Change Scores as Dependent Variables in Regression Analysis
Author(s): Paul D. Allison
Source: *Sociological Methodology*, Vol. 20 (1990), pp. 93-114
Published by: American Sociological Association
Stable URL: http://www.jstor.org/stable/271083
Accessed: 09/02/2015 10:18

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at
http://www.jstor.org/page/info/about/policies/terms.jsp

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

American Sociological Association is collaborating with JSTOR to digitize, preserve and extend access to *Sociological Methodology*.

http://www.jstor.org

This content downloaded from 160.39.38.113 on Mon, 9 Feb 2015 10:18:11 AM
All use subject to JSTOR Terms and Conditions

3

CHANGE SCORES AS DEPENDENT VARIABLES IN REGRESSION ANALYSIS

*Paul D. Allison**

Change scores have been widely criticized for their purported unreliability and for their sensitivity to regression toward the mean. These objections are shown to be unfounded under a plausible regression model for the nonequivalent control group design. This model leads to inferences that are intuitively correct, as judged by changes in means over time, while the conventional model leads to inferences that are intuitively false. Moreover, the conventional model implies that regression toward the mean within groups leads to regression toward the mean between groups, an implausible result for naturally occurring groups. Nevertheless, the conventional model may be more appropriate when there is a true causal effect of the pretest on the posttest, or when cases are assigned to groups on the basis of their pretest scores.

1. INTRODUCTION

The measurement of a dependent variable at two or more points in time is widely regarded as a powerful tool for making

An earlier version of this paper was presented at the 1988 Annual Meetings of the American Sociological Association. I am indebted to Herbert Smith for helpful comments and suggestions. Michael Pertschuk generously permitted me to report results from his data.

*University of Pennsylvania

This content downloaded from 160.39.38.113 on Mon, 9 Feb 2015 10:18:11 AM
All use subject to JSTOR Terms and Conditions

causal inferences with nonexperimental data. If the aim is to show that X causes Y, there is supposedly great merit in examining the relationship between X and Y_2 while "controlling" for Y_1, where Y_1 and Y_2 are measurements of the same variable at times 1 and 2, respectively. Such a procedure, it is argued, allows one to rule out the rival hypothesis that Y causes X. It also greatly reduces the threat of spuriousness, i.e., that some other variable causes both X and Y.

Using the terminology of experimental design, I shall hereafter refer to Y_1 as the pretest and to Y_2 as the posttest. In the early 1960s, there was a great deal of confusion about the proper statistical analysis for designs with both pretest and posttest measurements (Bereiter 1963; Lord 1963). Specifically, the issue was how to control for Y_1. Of the many proposed treatments, the two most commonly considered were the *change score method* and the *regressor variable method*. In the change score method, $Y_2 - Y_1$ is regressed on X. In the regressor variable method, Y_2 is regressed on both Y_1 and X. Thus, the regressor variable method treats the pretest like any other control variable, and the change score method assigns it a special status.

Before proceeding further, it is worth pointing out that each of these methods has a computational equivalent. First, the regression of $Y_2 - Y_1$ on both Y_1 and X is equivalent to the regressor variable method (Werts and Linn 1970). Both procedures produce the same coefficient of X and the same estimated standard error. Hence, it must be understood that in the change score method, Y_1 cannot also appear as a regressor variable. Second, when X is categorical, the change score method is equivalent to a repeated-measures analysis of variance in which a test for an effect of X on Y is achieved by testing the interaction of X with the within-subject factor (Maxwell and Howard 1981).

Debate among psychometricians soon led to a consensus against the change score method and in favor of the regressor variable method (Cronbach and Furby 1970), and this opinion became conventional wisdom among sociologists (Bohrnstedt 1969). There were two major objections to the use of change scores.

1. *Unreliability*. Change scores tend to be much less reliable than the component variables (Kessler 1977). Consider the simplified case in which Y_1 and Y_2 are equally reliable and have the same

This content downloaded from 160.39.38.113 on Mon, 9 Feb 2015 10:18:11 AM
All use subject to JSTOR Terms and Conditions

variance. The reliability of $Y_2 - Y_1$ is then given by

$$\frac{\rho_Y^2 - \rho_{12}}{1 - \rho_{12}},$$

where ρ_{12} is the correlation between Y_1 and Y_2 and ρ_Y^2 is their common reliability. If this correlation is positive (as it almost always is), then the reliability of the change score must be less than ρ_Y^2, often much less. For example, if $\rho_Y^2 = 0.7$ and $\rho_{12} = 0.6$, the reliability of the change score is only 0.25.

2. *Regression effects*. Because of the almost universal phenomenon of regression toward the mean from pretest to posttest measurements, Y_1 will usually be negatively correlated with $Y_2 - Y_1$. Thus, individuals with high pretest scores will tend to move down on the posttest, while individuals with low pretest scores will tend to move up. Consequently, if X (or any other variable) is correlated with Y_1, it will tend to have a spuriously negative relationship with $Y_2 - Y_1$ (Markus 1980). For these reasons, methodologists in the social sciences have repeatedly warned against the use of change scores.

Since 1975 there has been some tempering in the attitude of psychometricians toward change scores. Zimmerman and Williams (1982) and Sharma and Gupta (1986) showed that there are common circumstances in which change scores can be highly reliable. Moreover, Overall and Woodward (1975) demonstrated the paradoxical result that change scores can yield powerful tests of causal hypotheses even when they are extremely unreliable. Their result was elaborated by Maxwell and Howard (1981), who claimed that change scores are sometimes appropriate for randomized experimental designs. Most importantly, Kenny (1975) and Kenny and Cohen (1979) argued that regression toward the mean is not a problem when the objective is to compare two or more stable groups. In such circumstances, the change score method can give results with less bias than the regressor variable method.

A major aim of this paper is to review and clarify these results. Nevertheless, I believe that psychometricians have not gone far enough. I claim that the change score method is superior to the regressor variable method whenever X is temporally subsequent to Y_1 and uncorrelated with the *transient* component of Y_1. I shall

This content downloaded from 160.39.38.113 on Mon, 9 Feb 2015 10:18:11 AM
All use subject to JSTOR Terms and Conditions

argue that this is a commonly satisfied condition. When it holds, the problem of measurement error in Y_1 disappears entirely. Moreover, when the effects of other variables on Y are invariant from pretest to posttest, those variables can be omitted from the analysis without introducing bias. In short, the use of change scores under appropriate conditions can greatly enhance our ability to make causal inferences from nonexperimental data.

Like much of the psychometric work on change scores, this paper focuses almost exclusively on the nonequivalent control group design (Campbell and Stanley 1963), largely because it is the simplest design that embodies all the relevant issues. The principal advantage of this restriction is that parameters in alternative models are simple functions of pretest and posttest means. This allows for a more direct, intuitive evaluation of the models and their implications. Many of the arguments given here also apply in more general settings, however. Liker, Augustyniak, and Duncan (1985), for example, described the benefits of change scores in general two-wave panel designs.

2. LORD'S PARADOX AND THE NONEQUIVALENT CONTROL GROUP

In the nonequivalent control group design, individuals in the sample are divided into two groups, labeled *treatment* and *control*. (Multiple treatment groups are also possible.) A pretest (Y_1) is conducted; then something happens to the treatment group that does not happen to the control group; finally, a posttest measurement (Y_2) is taken. What distinguishes this design from a true experiment is that individuals are not randomly assigned to the two groups, which may therefore differ substantially in their distributions of Y_1. Though the terminology of this design suggests that the treatment is under the control of the investigator, the design also encompasses studies in which the treatment group experiences some naturally occurring event, such as divorce, retirement, or marriage. It is essential, however, that the pretest measurements be made before the event occurs.

In analyzing such data, the aim is to compare the treatment and control groups on Y_2 while somehow controlling for Y_1. Clearly this can be accomplished by either the change score method or the

This content downloaded from 160.39.38.113 on Mon, 9 Feb 2015 10:18:11 AM
All use subject to JSTOR Terms and Conditions

regressor variable method, defining X to be 1 for the treatment group and 0 for the control group. To decide between these methods, it is helpful to imagine what would be observed if the treatment had no effect. Lord (1967) developed the following hypothetical example. Suppose that the mean and standard deviation for Y in the treatment group are exactly the same on the pretest and the posttest and that the within-group correlation between Y_1 and Y_2 is substantially less than 1. Suppose that the same conditions hold in the control group, except that the means are lower than those in the treatment group. Thus, the mean differences on the posttest reproduce those on the pretest. Intuitively, such a pattern seems to be consistent with no treatment effect.

The data in Table 1, panel A, come from a real quasi-experiment that had exactly this outcome. The treatment group consisted of 18 children who underwent plastic surgery for craniofacial abnormalities. The control group consisted of 30 normal children. The two groups had approximately the same age range. The dependent variable was a measure of the frequency of negative social encounters, based on parental reports. In the treatment group, this variable was measured shortly before the surgery and again 18 months later. In the control group, the measurements were also

TABLE 1
Means and Standard Deviations for Two Measures of Adjustment, by Time and Treatment

		Time 1	Time 2
A.	Frequency of negative social encounters		
	Treatment group	48.3	48.6
		(7.6)	(6.5)
	Control group	41.6	41.1
		(9.2)	(8.1)
B.	Trait anxiety		
	Treatment group	37.3	31.5
		(6.8)	(5.1)
	Control group	32.1	30.3
		(5.9)	(6.7)

This content downloaded from 160.39.38.113 on Mon, 9 Feb 2015 10:18:11 AM
All use subject to JSTOR Terms and Conditions

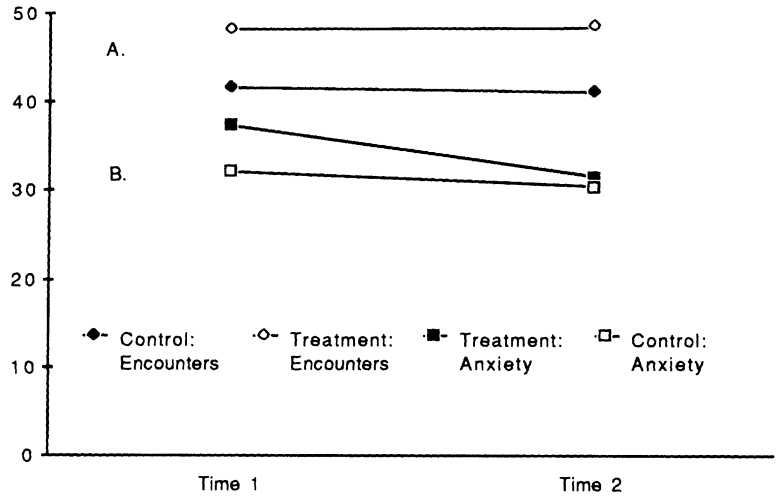

FIGURE 1 Means from Table 1.

taken 18 months apart at comparable ages. In both groups, the means and standard deviations hardly changed at all from the pretest to the posttest. On the other hand, at each occasion the control group scored about 7 points lower than the treatment group. The means are displayed graphically in Figure 1, panel A.

Two analyses were performed. The regressor variable approach, in which Y_2 was regressed on Y_1 and X (the treatment indicator), yielded a coefficient of X that was positive and significant at the 0.03 level. This seems to imply that plastic surgery had a deleterious effect on children's social experiences. But the change score method, in which $Y_2 - Y_1$ was regressed on X, yielded a coefficient for X that was near zero and far from statistical significance.

This is what has come to be known as Lord's paradox. The standard regressor variable approach seems to give the wrong answer. The change score method seems to confirm the intuitive impression that the treatment produced no change. Lord (1967, p.305) concluded that "with the data usually available for such studies, there is simply no logical or statistical procedure that can be counted on to make proper allowances for uncontrolled preexisting differences between groups." An alternative response is to question

This content downloaded from 160.39.38.113 on Mon, 9 Feb 2015 10:18:11 AM
All use subject to JSTOR Terms and Conditions

the preferential status of the regressor variable method and to reconsider the utility of the change score method.

The regressor variable method can also lead to the conclusion that there is no treatment effect when a straightforward examination of the means suggests otherwise. This phenomenon is exhibited in Table 1, panel B, which reports results from the same study of craniofacial surgery. Here the outcome variable is a measure of trait anxiety. In the control group, the mean declined slightly from 32 to 30. In the treatment group, the mean declined from 37 to 31 (see Figure 1, panel B). While it is not entirely clear what produced this decline in the treatment group, it seems desirable that the statistical analysis should be sensitive to the differential change. The change score method does, in fact, find an effect of the treatment that is significant at the 0.02 level. But the regressor variable method yields a treatment effect that is far from significant.

Why does the regressor variable approach seem to go astray in these examples? A definitive answer requires the formal analysis presented in the next section, but it is helpful first to consider some informal arguments. The basic problem is that the use of Y_1 as a regressor variable seems to underadjust for prior differences. Thus, in Figure 1, panel A, the fact that the pretest difference is the same as the posttest difference is not fully captured by including Y_1 as an independent variable. Similarly, in Figure 1, panel B, the regressor variable approach emphasizes the fact that the treatment and control groups were almost the same on the posttest but puts little weight on the fact that they were different on the pretest.

In the psychometric literature, measurement error was quickly seized upon as an explanation for such underadjustment. If Y_1 is a fallible measure of some true score, then the estimated coefficient of Y_1 will be biased toward zero. This, in turn, could bias the coefficient of X. One proposed solution was to get estimates of the reliability of Y_1 and use these to produce corrected regression estimates (Preece 1982).

Some have argued, however, that even with perfect measurement, there is usually underadjustment when Y_1 is used as a regressor variable (Reichardt 1979). There is typically inherent instability in any variable Y, leading to a less than perfect correlation between Y_1 and Y_2. When Y_2 is regressed on Y_1 and X, therefore, the regression coefficient for Y_1 is usually between zero and one.

This content downloaded from 160.39.38.113 on Mon, 9 Feb 2015 10:18:11 AM
All use subject to JSTOR Terms and Conditions

The lower the correlation between Y_1 and Y_2, the smaller the regression coefficient. If b is that coefficient, then the estimated coefficient for X (the treatment effect) is $\bar{Y}_{2T} - \bar{Y}_{2C} - b(\bar{Y}_{1T} - \bar{Y}_{1C})$, where T and C indicate treatment and control, respectively. That is, the estimated treatment effect is the posttest difference in means minus some fraction of the pretest difference in means. If the correlation between Y_1 and Y_2 is small, the adjustment accomplished by the regressor variable approach will also be small. When the change score approach is used, the estimated treatment effect is $\bar{Y}_{2T} - \bar{Y}_{2C} - (\bar{Y}_{1T} - \bar{Y}_{1C})$, a quantity that is unaffected by the correlation between Y_1 and Y_2.

The problem with these informal arguments is that they rest on a number of implicit assumptions that need more careful examination. This will be accomplished in the next section. Before proceeding further, however, it is worth emphasizing that these arguments apply whenever one includes Y_1 in a regression predicting Y_2. The only thing special about the nonequivalent control group design is that it makes it easy to compare mean differences, and thus easy to see when something is amiss.

3. MODELS FOR THE NONEQUIVALENT CONTROL GROUP DESIGN

A problem with much of the work comparing change score and regressor variable methods is that the conclusions are rarely based on an explicit model for the generation of the data. Thus, it is often difficult to evaluate the arguments on such basic statistical criteria as bias and efficiency. In this section, I consider two alternative models for the nonequivalent control group design. Model 1 is implicitly assumed in the regressor variable approach. Model 2, which leads to the change score method, is a variant of a model introduced by Kenny (1975). It also has precedents in econometric treatments of panel data (Chamberlain 1984; Heckman and Robb 1985).

3.1. *Model* 1

Model 1 can be written as

$$Y_{i2} = \alpha + \beta Y_{i1} + \delta X_i + \epsilon_i, \qquad i = 1, \ldots, n, \tag{1}$$

This content downloaded from 160.39.38.113 on Mon, 9 Feb 2015 10:18:11 AM
All use subject to JSTOR Terms and Conditions

where $X_i = 1$ for individuals in the treatment group, 0 for those in the control group. Thus, δ is the treatment effect. With the further assumption that $E(\epsilon_i \mid Y_{i1}, X_i) = 0$ for all i, ordinary least squares (OLS) regression of Y_2 on Y_1 and X yields unbiased estimates of the coefficients.

3.2. *Model 2*

Model 2 posits separate equations for Y at the two time points. To facilitate interpretation, we define G_i to equal 1 or 0 depending on whether or not an observation is in the group that eventually gets the treatment. This variable is conceptually distinct from X_i, which denotes whether or not the treatment has actually been received. However, in the present context, $G_i = X_i$ at time 2. The model is

$$Y_{i1} = \alpha + \gamma G_i + \epsilon_{i1}, \tag{2}$$

$$Y_{i2} = \alpha + \tau + \gamma G_i + \delta X_i + \epsilon_{i2}, \qquad i = 1 \ldots n. \tag{3}$$

The coefficient γ represents a group difference that is assumed to be constant over time. This assumption is essential in the present treatment, but some possible modifications will be considered in a later paper. The parameter τ is the change over time that applies to all individuals in both treatment and control groups. The treatment effect is δ. The variables ϵ_{i1} and ϵ_{i2} are random disturbances. For the moment, we assume that $E(\epsilon_{i1} \mid X_i) = E(\epsilon_{i2} \mid X_i) = 0$ for all i, although we shall shortly relax this assumption. It is not necessary to assume that ϵ_1 and ϵ_2 are uncorrelated; indeed, it would be very unusual if they were.

Equation (2) can be estimated without bias by OLS regression of Y_1 on G. But equation (3) cannot, because X is completely confounded (collinear) with G. By subtracting equation (2) from equation (3), however, we get

$$Y_{i2} - Y_{i1} = \tau + \delta X_i + \epsilon_i^*, \tag{4}$$

where $\epsilon_i^* = \epsilon_{i2} - \epsilon_{i1}$. Since $E(\epsilon_i^* \mid X_i) = E(\epsilon_{i2} \mid X_i) - E(\epsilon_{i1} \mid X_i) = 0$ for all i, it follows that (4) can be estimated without bias by OLS regression. Thus, model 2 justifies the change score method.

This content downloaded from 160.39.38.113 on Mon, 9 Feb 2015 10:18:11 AM
All use subject to JSTOR Terms and Conditions

Model 2 can be informatively extended by decomposing the disturbance terms in (2) and (3) into three components:

$$\epsilon_{it} = U_i + V_{it} + W_{it}, \qquad t = 1,2; \qquad i = 1, \ldots n, \tag{5}$$

where W_{it} is random measurement error at time t, V_{it} is period-specific variation in Y_{it}, and U_i is a component of Y that is stable across time. U_i can be thought of as including all those explanatory variables that are stable over time and that have identical effects at both times. Although V_{it} and W_{it} are formally equivalent, they are substantively distinct: V_{it} captures true variation in the phenomenon of interest, and W_{it} refers to random variation that is specific to the measurement process.

We assume that $\mathrm{E}(V_{it} \mid X_i) = \mathrm{E}(W_{it} \mid X_i) = 0$ for all i and t, but we do *not* have to assume that $\mathrm{E}(U_i \mid X_i) = 0$ for all i. In other words, we now allow X to be correlated with the stable component of Y. This causes no difficulty, because the U_i term drops out when the difference is taken in equation (4); i.e.,

$$\epsilon_i^* = V_{i2} - V_{i1} + W_{i2} - W_{i1}. \tag{6}$$

This implies that stable explanatory variables whose effects are constant over the two time points may be omitted from the equation without biasing the estimate of the treatment effect.

4. SOME IMPLICATIONS OF THE MODELS

The most obvious distinction between models 1 and 2 is that model 1 has Y_1 on the right-hand side of the equation and model 2 does not. This is not quite accurate, however, because adding Y_1 to both sides of equation (4) yields

$$Y_{i2} = \tau + Y_{i1} + \delta X_i + \epsilon_i^*. \tag{7}$$

Thus, model 2 does not actually exclude Y_1 from the right-hand side of the equation; it just forces its coefficient to be equal to 1. In fact, equation (7) can be estimated directly by using a regression routine that allows restrictions on the coefficients.[1] It is tempting,

[1] Kessler and Greenberg (1981, p. 13) erroneously claimed that such constrained estimation will yield biased estimates because Y_1 is correlated with the error term. When a coefficient is constrained to any specific value, it no longer matters whether that associated variable is uncorrelated with the error term. This is easily demonstrated by comparing the residual sum of squares (the minimization criterion) for OLS estimation of (7) and (4). The two are identical.

This content downloaded from 160.39.38.113 on Mon, 9 Feb 2015 10:18:11 AM
All use subject to JSTOR Terms and Conditions

then, to conclude that model 2 is just a special case of model 1 and that model 1 is therefore preferable because it avoids needless restrictions. But this is erroneous, because model 1 assumes that Y_1 is uncorrelated with the disturbance term in equation (1), and model 2 implies that Y_1 *must* be negatively correlated with the disturbance term in equation (7). Hence, if model 2 is correct, unrestricted OLS estimation with Y_1 on the right-hand side will yield biased estimates.

This conclusion can be stated more precisely. If we assume that model 2 is true and that $\delta = 0$ (no treatment effect), it is relatively easy to derive the population linear regression function[2] of Y_2 on Y_1 and X. The regression of Y_2 on Y_1 alone has the slope coefficient

$$\beta_{21} = \frac{\text{cov}(Y_1, Y_2)}{\text{var}(Y_1)}$$

$$= \frac{\text{var}(\gamma X + U)}{\text{var}(\gamma X + U) + \text{var}(V_1 + W_1)},$$

which is clearly less than one. If we let ρ_{1X} be the correlation between X and Y_1, the partial regression coefficient for Y_1 controlling for X is

$$\beta_{21 \cdot X} = \frac{\beta_{21} - \rho_{1X}^2}{1 - \rho_{1X}^2},$$

which is easily shown to be less than 1. Thus, model 2 accounts for the finding that empirical regressions of Y_2 on Y_1 and X nearly always have slope coefficients for Y_1 that are less than 1. More importantly, the partial coefficient for X is

$$\beta_{2X \cdot 1} = \frac{(1 - \beta_{21})\left[\gamma + \dfrac{\text{cov}(X, U)}{\text{var}(X)}\right]}{1 - \rho_{1X}^2}.$$

Hence, under model 2, the regression of Y_2 on Y_1 and X will yield a nonzero coefficient of X even when there is no real treatment

[2] Under model 2, the population linear regression function will not be equivalent to the conditional expectation function, which will be nonlinear. However, OLS will always give consistent estimates of the population linear regression function. The derivations reported in this section depend on the additional assumptions that $\text{E}(V_{it} \mid U) = \text{E}(W_{it} \mid U) = 0$ for all i and t, and that $\text{E}(W_{i2}W_{i1}) = \text{E}(V_{i2}V_{i1}) = \text{E}(W_{i2}V_{i1}) = \text{E}(V_{i2}W_{i1}) = 0$ for all i. If these assumptions are not satisfied, the results are much more complicated but have essentially the same qualitative interpretation.

This content downloaded from 160.39.38.113 on Mon, 9 Feb 2015 10:18:11 AM
All use subject to JSTOR Terms and Conditions

effect. It will be zero (a) when there is no pretest difference between the two groups (i.e., when $\gamma = 0$), *and* (b) when the assignment to the treatment group is uncorrelated with the stable component of Y (i.e., when $\text{cov}(X,U) = 0$). These two conditions are guaranteed to hold in a true experiment with random assignment to treatments, but not in general. In sum, model 2 fully accounts for the empirical results that are typically obtained when one estimates model 1.

5. THE IMPACT OF MEASUREMENT ERROR

What about the notorious unreliability of change scores? The problem with this claim is that it is typically considered in isolation from the estimation and testing of a causal model. In that context, what we need to be concerned about is not reliability but rather the error variance in (4), which is what determines the precision of the estimates of δ. The error variance can decrease even as the reliability of the change scores decreases. Here is how it works. Equations (4) and (5) imply that the reliability of the change score is given by

$$\frac{\text{var}(\delta X + V_2 - V_1)}{\text{var}(\delta X + V_2 - V_1 + W_2 - W_1)},$$

where $V_2 - V_1$ represents true change over time, and the W's are random measurement errors. If we make the further assumption that the V's are uncorrelated with the W's, the reliability becomes

$$\frac{\text{var}(\delta X) + \text{var}(V_2 - V_1)}{\text{var}(\delta x) + \text{var}(V_2 - V_1) + \text{var}(W_2 - W_1)}. \tag{8}$$

Notice that $\text{var}(V_2 - V_1) + \text{var}(W_2 - W_1)$ is the error variance in equation (4), so that any reduction in $\text{var}(V_2 - V_1)$ will reduce the error variance and thereby increase the power of statistical tests of hypotheses about the coefficients in (4). But (8) is also an increasing function of $\text{var}(V_2 - V_1)$. In short, $\text{var}(V_2 - V_1)$ is true-score variance for the purpose of estimating reliability, but it is error variance for the purpose of estimating parameters in equation (4). Consequently, any reduction in $\text{var}(V_2 - V_1)$ will simultaneously *decrease* the reliability of the change score and *increase* statistical power. Hence, the low reliability of change scores is irrelevant for the purpose of causal inference.

This content downloaded from 160.39.38.113 on Mon, 9 Feb 2015 10:18:11 AM
All use subject to JSTOR Terms and Conditions

This seemingly paradoxical result is actually quite consistent with intuition. The ideal situation for detecting a treatment effect is one in which subjects who do not receive the treatment hardly change at all from pretest to posttest while subjects who do receive the treatment all change by about the same amount. But this is just the situation that produces high correlations between pretest and posttest and, as we have seen, low reliability of change scores. The low reliability results from the fact that in calculating the change score we difference out all the stable between-subject variation, except for that due to the treatment effect.

In addition to this result, model 2 has several advantages over model 1 with respect to measurement error. Under model 1, measurement error in Y_1 produces biased estimates of the treatment effect; but this is not the case with model 2. Given the generally low reliability of measurement in the social sciences, this insensitivity to measurement error is highly desirable. Although the estimates under model 1 can be corrected for unreliability (Porter 1967), such corrections typically require data that are unavailable and assumptions that are unrealistic and untestable. Model 2 not only allows for measurement error in Y_1 but is also surprisingly robust to characteristics of the measurement error. For example, correlation over time in the measurement error (between W_1 and W_2) does not bias the treatment estimate, nor does differential reliability between treatment and control groups. (Differential reliability may cause inefficiency, but this can be remedied by weighted least squares.) Of course, the fact that model 2 is more robust to measurement error is not a sound basis for choosing it over model 1. That decision should be based on our beliefs about which model better represents reality.

We now proceed to a consideration of that question.

6. CHOOSING BETWEEN THE MODELS

We have seen that the choice between the regressor variable method and the change score method can be recast as a choice between models 1 and 2. We have also seen that the criticisms commonly leveled at the change score method do not stand up when examined in the framework of model 2. Now we are faced with a choice between these two models. What criteria are appropriate for

This content downloaded from 160.39.38.113 on Mon, 9 Feb 2015 10:18:11 AM
All use subject to JSTOR Terms and Conditions

choosing between them? Purely mathematical or statistical considerations will not suffice because the models are incommensurable. Moreover, both models do a good job of accounting for the typical empirical patterns found in data produced by the nonequivalent control group design.

The arguments I shall consider depend to some degree on the fit between the models' implications and what we intuitively expect for a given phenomenon. It is unrealistic to expect either model to be best in all situations; indeed, I shall argue that each of these models has its appropriate sphere of application. Since model 2 is the underdog, I shall concentrate on arguments that support it. Nevertheless, the choice will rarely be obvious, and there will almost always be some residual uncertainty. One should also consider the possibility that neither of these models is appropriate (more on that later).

Clearly, the empirical examples discussed above provide an argument in favor of model 2. The fact that the posttest means are identical to the pretest means for both groups strongly suggests that there was no effect of the treatment. We ordinarily expect a treatment to produce some change, and no change means no treatment effect. Since model 2 leads to that answer, we should have greater confidence in the model. Nevertheless, while the intuition is essentially correct, it rests on assumptions that will not always hold in the nonequivalent control group design. These assumptions will be examined in detail shortly.

Model 1 has a closely related problem: It implies that regression toward the mean within groups will be translated into regression toward the mean between groups. For example, if we assume that $\delta = 0$ (no treatment effect), it is easily shown that model 1 implies

$$\mu_{2T} - \mu_{2C} = \beta(\mu_{1T} - \mu_{1C}),$$

where μ_{ij} is a population mean at time i, with $j = T$ for the treatment group and $j = C$ for the control group. As suggested earlier, in most empirical applications β will be a positive number less than 1, often substantially less. In that case, the model implies that the expected mean difference on the posttest will be less than the mean difference on the pretest. In other words, the mean for each of the two groups will regress toward the grand mean.

This content downloaded from 160.39.38.113 on Mon, 9 Feb 2015 10:18:11 AM
All use subject to JSTOR Terms and Conditions

If the two groups are the result of some existing social demarcation, such an implication is quite unreasonable. Consider the following hypothetical example. Suppose that the treatment group consists of all the males in some organization and that the control group consists of all the females. These two groups are exposed to different conditions. The dependent variable is a measure of productivity, taken at two points in time. Suppose, further, that the correlation in productivity from time 1 to time 2 is 0.50 and that the variances are stable over time. Model 1 then implies that in the absence of a treatment effect, the sex difference in productivity at time 2 should be only 50 percent of the sex difference at time 1. Model 2 says that the sex difference should be the same at the two points in time when there is no treatment effect.

Despite these arguments, model 1 may be preferable when Y_1 has a true causal effect on Y_2 or when the values of X are determined, in part, by the period-specific components of Y_1.

I shall consider each of these possibilities in turn. First, let us specify what it means to say that Y_1 has a causal effect on Y_2. To avoid philosophical difficulties with the concept of causality, it is sufficient for our purpose to consider a hypothetical randomized experiment. Suppose that individuals are randomly assigned[3] to values of Y_1. If the excepted value of Y_2 varies as a function of Y_1, we shall say that Y_1 is a cause of Y_2.

When might such a functional relationship occur? To answer this question it is convenient to distinguish between "stocks" and "flows." Stocks are those quantities that have an inherent persistence over time unless altered by some specific process. Flows are quantities that must be created anew at each time point or time interval. An obvious example of the difference between stocks and flows is the difference between income and wealth. Organizational size and body weight are examples of stock variables. Most measures of behavior and attitude, on the other hand, are considered flow variables.

In most cases, stock variables satisfy our definition of a causal effect of Y_1 on Y_2. For example, if we could randomly assign people

[3] Lord (1969) pointed out that the mechanisms by which such assignment is made may have important consequences for the outcome.

This content downloaded from 160.39.38.113 on Mon, 9 Feb 2015 10:18:11 AM
All use subject to JSTOR Terms and Conditions

to different levels of wealth, we would expect them to have different levels of wealth one year later. If such were the case, then Y_1 would have to appear on the right-hand side of a linear equation for Y_2. But in most cases, we can expect the coefficient for Y_1 to have a structural value[4] of 1. That is, Y_2 will be equal to Y_1 plus some increment or decrement determined by other variables. Thus, despite the causal impact of Y_1, we would again be led to equation (7).

In the case of flow variables, there is usually much less reason to expect a causal effect of Y_1 on Y_2. For example, if we could randomly alter people's incomes by giving them additional amounts of money during a single year, we would not expect those increments to change the expected values of income in the following year. Nevertheless, there are some flow variables for which it *is* plausible to expect a true effect of Y_1 on Y_2. For the income example, if the increments were substantial, a portion could be saved for income-producing investments. Attitudes and behaviors are subject to reinforcement or habit formation, and either of these processes could be interpreted as a causal effect of Y_1 on Y_2. For instance, Allison, Long, and Krauze (1982) considered the hypothesis that each time a scientist publishes a paper, the probability of future publications is increased. If this hypothesis is represented by a contagious Poisson process (Allison 1980), it follows that the expected number of publications in year t is a linear function of the number of publications in year $t-1$.

Another type of mechanism that leads to a causal effect of a variable on itself at a later point in time is the "halo effect." If an individual is searching for a job, the prestige of his current job may be an important factor in potential employers' evaluation of him. Thus, a high-prestige origin job may lead directly to a high-prestige destination job, irrespective of any characteristics of the job holder.

If any of these processes is operative to a substantial degree, model 2 may not be an adequate representation. On the other hand, model 1 may also be too simple a formulation for such complex effects. My own view is that while such causal processes are often operative, the effects are typically so small relative to the action of other variables that it is not essential to build them into the model.

[4] Because of measurement error, the estimated coefficient may be less than 1.

This content downloaded from 160.39.38.113 on Mon, 9 Feb 2015 10:18:11 AM
All use subject to JSTOR Terms and Conditions

This should not be interpreted as a blanket admonition to disregard them, however. Each case should be evaluated on its own merits.

The other situation in which model 1 may be preferable to model 2 is when the *transient* (period-specific) components of Y_1 (V_1 and W_1 in equation (5)) are correlated with X. Such a correlation is a violation of the assumption that $E(V_{it} \mid X_i) = E(W_{it} \mid X_i) = 0$ for all i and t. How might this occur? Suppose that the treatment is participation in an SAT training program and that the aim of the study is to determine whether the program improves SAT scores. We can imagine several different ways in which high school seniors might be selected for the treatment:

1. All seniors in high school A are enrolled in the program, and the SAT is administered before and after the program. All seniors in high school B serve as controls.
2. The SAT is administered as a pretest to a group of high school seniors. Those who score below 400 are enrolled in the program, and those who score above are not.
3. Seniors self-select into the program *before* seeing the results of a pretest administration of the SAT.
4. Seniors self-select into the program *after* seeing the results of a pretest administration of the SAT.

For cases 1 and 3 we would *not* expect the treatment assignment to be correlated with the transient components of Y. In case 1 such a correlation would occur only if the mean of $V_1 + W_1$ happened to be different for the treatment and control groups. Even if this occurred by chance, the induced correlation between X and $V_1 + W_1$ would likely be small. In case 3, while we would expect a student's ability to affect self-selection, the stable rather than the transient component of ability would be the determining factor.

For cases 2 and 4, on the other hand, it should be obvious that the transient components of Y are likely to affect the treatment assignment. In case 2, the observed Y_1 completely determines treatment assignment; hence, its transient components must be correlated with X. This is just the well-known regression-discontinuity design (Campbell and Stanley 1963). In case 4, although we don't know for certain, it would be surprising if Y_1 did not have some impact on the selection into treatment. For either 2 or 4, it

This content downloaded from 160.39.38.113 on Mon, 9 Feb 2015 10:18:11 AM
All use subject to JSTOR Terms and Conditions

can be shown that the regressor variable method is more appropriate than the change score method (Goldberger 1972; Reichardt 1979), although in many cases it may be necessary to use more elaborate methods (Rubin 1977).

7. DISCUSSION AND EXTENSIONS

The objective of this paper has been to make a case for the use of change scores as dependent variables in regression models. In the first part of that case, I refuted standard objections to the use of change scores: that they are unreliable and prone to biases resulting from regression toward the mean. I showed that both objections are groundless under the specification of a plausible causal model (model 2) whose most distinctive feature is that Y_1 does not appear as an independent variable. While this refutation does not resolve the controversy, it raises it from a technical issue to a choice between alternative models and assumptions.

In the second part of the case, I considered reasons why model 2 might be more sensible than the model leading to the usual regressor variable approach (model 1). The most compelling argument against model 1 is that it leads to the conclusion that there is a treatment effect when a straightforward examination of means indicates that nothing has happened. Moreover, model 1 implies that regression to the mean within groups implies regression to the mean between groups, a conclusion that seems quite implausible for many applications.

Unfortunately, arguments about model choice are notoriously difficult to resolve. In fact, I also argued that the standard model 1 might be more appropriate for some applications, notably when Y_1 has a true causal effect on Y_2, or when X is correlated with the transient components of Y_1. The important point is that there should be no automatic preference for either model and that the only proper basis for a choice is a careful consideration of each empirical application. Even then, the choice may not be clear cut. In ambiguous cases, there may be no recourse but to do the analysis both ways and to trust only those conclusions that are consistent across methods.

All these arguments were presented for the case of the nonequivalent control group design, but they are by no means limited

This content downloaded from 160.39.38.113 on Mon, 9 Feb 2015 10:18:11 AM
All use subject to JSTOR Terms and Conditions

to that application. On the other hand, neither are they as general as one might like. I shall now sketch some possible extensions and current limitations. I already mentioned that the treatment variable X can have more than two categories, and indeed, there is nothing to prevent it from being a continuous variable. What *is* necessary is a clear temporal ordering from Y_1 to X to Y_2. This allows for most designs in which X indexes some event or intervention that occurs between time 1 and time 2, but it does not include the popular two-wave, two-variable (2W2V) panel design in which X is measured contemporaneously with Y at both points in time. While it is straightforward to generalize model 2 to allow for 2W2V designs (Liker et al. 1985), the causal interpretation is more problematic.

The generalization is accomplished by specifying a model for the regression of Y on X and other variables at each time point:

$$Y_t = \alpha_t + \delta X_t + \gamma_t Z + \beta_t V_t + \lambda U + \epsilon_t, \qquad t = 1,2. \tag{9}$$

In this equation there are four kinds of independent variables: X is a vector of variables whose values change from time 1 to time 2 but whose effects do not change; Z is a vector that is constant over time but whose effects change; V is a vector whose values and effects both change; U is a vector of time-constant variables with constant effects. When we substract the equation for time 1 from the equation for time 2, we get

$$Y_2 - Y_1 = (\alpha_2 - \alpha_1) + \delta(X_2 - X_1) + (\gamma_2 - \gamma_1)Z + \beta_2 V_2 - \beta_1 V_1 + (\epsilon_2 - \epsilon_1).$$

With appropriate assumptions for the error term, this equation can be estimated by OLS regression of the change score for Y on the change score for X and the other independent variables. But, of course, the equations in (9) can also be estimated directly by OLS in a straightforward manner. The advantage of the difference score formulation is that the U vector has dropped out of the equation, making it unnecessary to statistically control for stable variables with stable effects. The problem with this result is that we must *assume* that the independent variables are causally prior to Y_t at each t, rather than rely on the design to ensure that causal priority. In evaluating the appropriateness of this model, we must again ask, (a) Is there a true causal effect of Y_1 on Y_2? or (b) Is $(X_2 - X_1)$ correlated with any omitted Z or V variables? If the answer to

This content downloaded from 160.39.38.113 on Mon, 9 Feb 2015 10:18:11 AM
All use subject to JSTOR Terms and Conditions

either of these questions is "yes," then other models may be more appropriate.

Attention has also been restricted to linear models with normal errors, leading to OLS regression estimates. Similar results can be obtained for logit models with binomial errors and for loglinear models with Poisson errors, in both cases by using conditional maximum likelihood estimation. These results will be reported elsewhere. On the other hand, I have not been successful in generalizing model 2 to allow for interaction between the treatment variable X and the initial level of Y_1. This is easily accomplished with the standard model 1, but it leads to great complexities and possible underidentification with model 2.

Another unresolved problem (for both models) is how to distinguish treatment effects from "differential growth." Model 2 (and implicitly, model 1) presumes that the parameter γ, representing group differences in the absence of a treatment effect, is constant over time. If this parameter is allowed to vary, it is almost impossible to distinguish such differential changes from true treatment effects (Blumberg and Porter 1983; Bryk and Weisberg 1977). Solutions are available if there are two pretest measurements rather than just one, but that is a subject for another paper.

Finally, it must be stressed that the two models discussed here do not exhaust those that may be appropriate for the nonequivalent control group design. Heckman and Robb (1985), for example, considered many different models and estimation methods for data of this sort; some of these are similar to model 1, and others are closer in spirit to model 2. They also generalize the models to allow for more than two time points. Working in the observational studies tradition, Holland and Rubin (1983) took a somewhat more general approach that did not presume a particular functional form or error distribution. They concluded, as I have, that the choice of methods is essentially dictated by a choice among competing models.

REFERENCES

Allison, Paul D. 1980. "Estimation and Testing for a Markov Model of Reinforcement." *Sociological Methods and Research* 8: 434–53.

Allison, Paul D., J. Scott Long, and Tad K. Krauze. 1982. "Cumulative Advantage and Inequality in Science." *American Sociological Review* 47: 615–25.

This content downloaded from 160.39.38.113 on Mon, 9 Feb 2015 10:18:11 AM
All use subject to JSTOR Terms and Conditions

Bereiter, Carl. 1963. "Some Persisting Dilemmas in the Measurement of Change." Pp. 3–20 in *Problems in Measuring Change*, edited by Chester W. Harris. Madison: University of Wisconsin Press.

Blumberg, C. J., and A. C. Porter. 1983. "Analyzing Quasi-Experiments: Some Implications of Assuming Continuous Growth Models." *Journal of Experimental Education* 51: 150–59.

Bohrnstedt, George W. 1969. "Observations on the Measurement of Change." Pp. 113–33 in *Sociological Methodology 1969*, edited by E. Borgatta. San Francisco: Jossey-Bass.

Bryk, A. S., and H. I. Weisberg,. 1977. "Use of the Nonequivalent Control Group Design When Subjects are Growing." *Psychological Bulletin* 84: 950–62.

Campbell, Donald T., and Julian C. Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand-McNally.

Chamberlain, Gary. 1984. "Panel Data." Pp. 1248–1318. in *Handbook of Economics*, vol. 2, edited by Zvi Griliches and M. D. Intriligator. Amsterdam: North-Holland.

Cronbach, L. J., and L. Furby. 1970. "How We Should Measure Change—Or Should We?" *Psychological Bulletin* 74: 32–49.

Goldberger, Arthur S. 1972. "Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations." Discussion paper. Madison: University of Wisconsin, Institute for Research on Poverty.

Heckman, J. J., and R. Robb, Jr. 1985. "Alternative Methods for Evaluating the Impact of Interventions." Pp. 156–246 in *Longitudinal Analysis of Labor Market Data*, edited by R. W. Pearson and R. F. Boruch. New York: Springer-Verlag.

Holland, P. W., and D. B. Rubin. 1983. "On Lord's Paradox." Pp. 3–25 in *Principles of Modern Psychological Measurement: A Festschrift for Frederic M. Lord*, edited by H. Wainer and S. Messick. Hillsdale, NJ: Lawrence Erlbaum.

Kenny, D. A. 1975. "A Quasi-Experimental Approach to Assessing Treatment Effects in the Nonequivalent Control Group Design." *Psychological Bulletin* 82:345–62.

Kenny, D. A., and S. H. Cohen. 1979. "A Reexamination of Selection and Growth Processes in the Nonequivalent Control Group Design." Pp. 290–313 in *Sociological Methodology 1980*, edited by Karl Schuessler. San Francisco: Jossey-Bass.

Kessler, R. C. 1977. "Use of Change Scores in Criteria in Longitudinal Survey Research." *Quality and Quantity* 11: 43–66.

Kessler, R. C., and D. F. Greenberg. 1981. *Linear Panel Analysis: Models of Quantitative Change*. New York: Academic Press.

Liker, J. K., S. Augustyniak, and G. J. Duncan. 1985. "Panel Data and Models of Change: A Comparison of First Difference and Conventional Two-Wave Models." *Social Science Research* 14: 80–101.

Lord, Frederic M. 1963. "Elementary Models for Measuring Change." Pp. 21–38 in *Problems in Measuring Change*, edited by Chester W. Harris. Madison: University of Wisconsin Press.

This content downloaded from 160.39.38.113 on Mon, 9 Feb 2015 10:18:11 AM
All use subject to JSTOR Terms and Conditions

———. 1967. "A Paradox in the Interpretation of Group Comparisons." *Psychological Bulletin* 68: 304–305.

———. 1969. "Statistical Adjustments When Comparing Preexisting Groups." *Psychological Bulletin* 72: 336–37.

Markus, G. 1980. *Models for the Analysis of Panel Data*. Beverly Hills: Sage.

Maxwell, Scott E., and George S. Howard. 1981. "Change Scores—Necessarily Anathema?" *Educational and Psychological Measurement* 41: 747–56.

Overall, J. E., and J. A. Woodward. 1975. "Unreliability of Difference Scores: A Paradox for Measurement of Change." *Psychological Bulletin* 82: 85–86.

Porter, A. C. 1967. "The Effects of Using Fallible Variables in the Analysis of Covariance." Ph.D. diss., University of Wisconsin.

Preece, P. F. W. 1982. "The Fan-Spread Hypothesis and the Adjustment for Initial Differences Between Groups in Uncontrolled Studies." *Educational and Psychological Measurement* 42: 759–62.

Reichardt, C. S. 1979. "The Statistical Analysis of Data from the Nonequivalent Control Group Design." Pp. 147–206 in *Quasi-Experimentation: Design and Analysis Issues in Field Settings*, edited by T. D. Cook and D. T. Campbell. Chicago: Rand-McNally.

Rubin, Donald B. 1977. "Assignment to Treatment Group on the Basis of a Covariate." *Journal of Educational Statistics* 2: 1–26.

Sharma, K. K., and J. K. Gupta. 1986. "Optimum Reliability of Gain Scores." *Journal of Experimental Education* 54: 105–108.

Werts, C. E., and R. L. Linn. 1970. "A General Linear Model for Studying Growth." *Psychological Bulletin* 73: 17–22.

Zimmerman, Donald W., and Richard H. Williams. 1982. "Gain Scores in Research can be Highly Reliable." *Journal of Educational Measurment* 19: 149–54.

This content downloaded from 160.39.38.113 on Mon, 9 Feb 2015 10:18:11 AM
All use subject to JSTOR Terms and Conditions