Journal of Consulting and Clinical Psychology
1990, Vol. 58, No. 4, 502–505

0022-006X/90/$00.75

Complex Statistics Are Not Always Clearer Than Simple Statistics: A Reply to Woody and Costanzo

John M. Gottman
University of Washington

Lowell J. Krokoff
University of Wisconsin

This reply to Woody and Costanzo's (1990) critique responds on both methodological and substantive grounds. The methodological grounds are that Woody and Costanzo misrepresented the issues in deciding whether to use change scores, partial correlations, or regression analyses of residuals. Regression toward the mean is misunderstood in their comment with respect to the issue of extreme groups. Furthermore, the problems they point out with the correlation of Time 1 social behavior with change scores in marital satisfaction remain intact with the part and partial correlations *if the test–retest correlations are high. Five alternative statistics are discussed in this reply.* Each statistic may have its own problems, but they tend to be equivalent if the test–retest correlation is high. Substantively, as suggested in our paper (Gottman & Krokoff, 1989), the analyses presented are robust to employing partial correlations, controlling Time 1 marital satisfaction.

There are two issues at stake in this discussion. The first issue is about how to best study longitudinal change. Actually, the problem is more general than the study of longitudinal change. It applies to any use of a difference score. The second issue is whether the results we reported (Gottman & Krokoff, 1989) of two longitudinal studies of marital satisfaction and change in marital satisfaction are fragile or robust with respect to this dialogue.

Methodological Issues

The discussion of statistical problems in measuring change has a venerable history. In one classic volume (Lord, 1967) edited by Harris, Fredric Lord discussed many of the issues raised anew by Woody and Costanzo (1990). The first issue is whether one should use change scores and search for correlates of the change scores or use some other approach such as partial correlations or the analysis of residuals with Time 1 scores as covariate.

The choice of methods is actually highly dependent on exactly the issues that Woody and Costanzo (1990) raised, but, unfortunately, they may not have been clear in their discussion. For example, they discussed the problem of regression toward the mean. Indeed, this is an essential part of this issue. However, regression to the mean is exacerbated by a distribution more humped near the mean than at its tails. The problem of regression to the mean is reduced when the distribution is rectangular (i.e., when each part of the sampling distribution is equally likely). Because this is the case, oversampling the tails of a distribution (that is, the oversampling of extreme groups) forces the distribution to be more rectangular and reduces regression to the mean. This was the logic of our sampling procedure (Gottman & Krokoff, 1989), and our distribution is indeed nearly rectangular. Hence, rather than exacerbating the problems, as Woody and Costanzo contend, the oversampling of the tails is actually at the core of *solving* the problem of regression toward the mean. Thus, it is not the case that extreme groups may "exacerbate this contamination" (p. 500), as Woody and Costanzo suggest.

Contrary to Woody and Costanzo's (1990) comment and use of arithmetic examples, regression to the mean does not imply that the variance decreases from initial to final score. Indeed, in statistical discussions of regression toward the mean, both the mean and standard deviations are assumed to be stable from pre to post (see Cohen & Cohen, 1983, p. 46). Hence, Woody and Costanzo's discussion and numerical examples following Equation 3 reflect a misunderstanding of regression to the mean and must be considered spurious. In our own data for Study 1, the ratio of post to pre marital satisfaction standard deviations was 0.82 for the husband and 1.02 for the wife; cumulative experience with these scores suggests that the expected value of the ratio is probably unity.

As an example of a logical error, there are many ways to rewrite an equation such as Equation 1 in Woody and Costanzo's (1990) comment. It is indeed a mathematical tautology that one way negative correlations with change scores can be obtained is by a decreasing correlation of a predictor variable with the criterion from Time 1 to Time 2. It is nonetheless the case that subjects lower on the criterion increased more on the predictor than subjects higher on the criterion. This follows by virtue of the fact that this is what the correlation means. Thus, it is not a confound, or "an artificial reversal of the $r(vb)$ result, masquerading as something new" (Woody & Costanzo, p. 500), but rather it follows mathematically from the definition of the correlation coefficient as a measure of linear association. Furthermore, as we shall see, if this represents a problem, it is also a problem with the alternative statistics that Woody and Costanzo propose.

The authors wish to acknowledge consultation with Alan Edwards and Robert Levenson in connection with the statistical aspects of this paper.

Correspondence concerning this article should be addressed to John M. Gottman, University of Washington, Department of Psychology, NI-25, Seattle, Washington 98195.

Related to the problem of regression toward the mean is another problem, which is at the heart of which statistic to use. This problem arises when a variable is correlated with initial scores and also correlated with change scores. Lord (1967) gives the example of weight loss and its relation to initial weight. If initial weight discriminates between those who gain and those who lose weight, variables related to initial weight will be related to change in weight as well. He recommends using the change score but partialling out the initial score from both the final score and the predictor variable. This is conceptually equivalent to using the partial correlation, controlling Time 1 scores. The regression alternative that Woody and Costanzo (1990) suggested (residualizing Time 2 scores with Time 1 scores) is actually mathematically equivalent to using what is usually called the *part correlation*, not, as they suggest, the partial correlation (see Glass & Stanley, 1970, pp. 182 ff). The choice of statistic depends in part on the kind of statement one wishes to make. This, in turn, is a function of the choice of a variable and a statistic that are not very far removed from the phenomenon under investigation.

There are times when one can obtain different results by using partial correlations rather than by using simple correlations with the change score. In our case the residual from regression and the simple difference gave very similar results. As we noted in our article (Gottman & Krokoff, 1989), change scores and the deviation of change controlling for Time 1 scores were highly correlated in both samples: in the Champaign sample, $r(23) = .97$, and in the Bloomington sample, $r(23) = .68$. For this reason, we used only change scores as the criterion variables.

Five Alternative Statistics

We will discuss the use of five alternative statistics that could be used in this analysis and the potential problems with each. We will use Woody and Costanzo's (1990) notation, and we will assume, as they did, that $SD(a) = SD(b)$, where a represents the Time 2 score and b represents the Time 1 score.

1. Simple correlation with change score. The first statistic is the one that we used. This correlation is denoted by Woody and Costanzo as $r_{v(a-b)}$. As they showed, this statistic is equal to

$$(r_{va} - r_{vb})/(2[1 - r_{va}])^{1/2}$$

They suggested that the numerator of this equation can be negative if v has different correlations with Time 1 scores and with Time 2 scores. This is indeed the case.

2. Part correlation of v *with* a*, controlling* b. This is the first suggestion Woody and Costanzo (1990) made, namely, residualizing Time 2 scores with Time 1 scores and correlating the residual with v. Following Glass and Stanley (1970), we can write the part correlation of a variable x with a variable y controlling a third variable, z as

$$r_{xe(y.z)} = \frac{(r_{xy} - r_{xz}r_{yz})}{(1 - r_{yz}^2)^{1/2}}$$

where e represents the residual of y when z is regressed on y. The partial correlation is defined as

$$r_{xy.z} = \frac{r_{xy} - r_{xz}r_{yz}}{(1 - r_{xz}^2)^{1/2}(1 - r_{yz}^2)^{1/2}}.$$

Now let us apply these general equations to our situation. By the definition of the part correlation, for our application this is

$$r_{ve(a.b)} = \frac{r_{va} - r_{vb}r_{ab}}{(1 - r_{ab}^2)^{1/2}}.$$

Two things are apparent from this equation. First, the numerator of this expression is nearly identical with the numerator of the first statistic if test–retest scores are high. Given that our Study 1 test–retest scores were .96 for husbands and .81 for wives, we are faced with exactly the same supposed problems with the part correlation that Woody and Costanzo proposed as with the correlation of the predictor variable with just the change score, namely, changing correlations of the predictor with Time 1 and Time 2 marital satisfaction can produce a negative sign. A second problem arises if we examine the denominator. If the test–retest correlation is high, this denominator can, at times, be extremely low. In our case of the husband's marital satisfaction, this denominator equals .078. We have no way of knowing in the general case whether this will be a problem. Unfortunately, concocting numerical examples will not shed light on the question.

3. Partial correlation of v *with* a *scores, controlling* b. The equation for this statistic is identical with the equation for the part correlation, except that the denominator contains an additional term, namely, the square root of $(1 - r_{va}^2)$. This is quite similar in form to the denominator of the first statistic.

4. The part correlation of v *with (*a − b*), controlling* b. After some algebra,[1] it can be shown that this correlation is

$$r_{ve(a-b)b} = \frac{r_{va} - r_{vb} + r_{vb}(1 - r_{ab})^{1/2}(1 - r_{va})^{1/2}}{(1 - r_{va})^{1/2}(1 + r_{ab})^{1/2}}.$$

If the test–retest correlation is near 1.0, this is approximately

$$\frac{r_{va} - r_{vb}}{(2[1 - r_{va}])^{1/2}}.$$

This is simply the correlation of v with $(a - b)$.

5. Partial correlation of v *with (*a − b*), controlling* b. This statistic is approximately equal to the following expression if the test–retest correlation is near 1.0:

$$\frac{r_{va} - r_{vb}}{(2[1 - r_{va}])^{1/2}} \frac{1}{(1 - r_{vb}^2)^{1/2}}.$$

This statistic is quite similar to the third statistic.

To summarize, the suggestion made by Woody and Costanzo (1990) on the issues of statistical approaches to the study of longitudinal change is no real improvement in a statistical sense. Most of these statistics are quite similar if the test–retest correlation is high, and the supposed problems they pointed to appear to be ubiquitous. Additionally, there could be some inflation of Statistics 2 and 3 if the test–retest correlation is high.

[1] A page of algebraic equations is available from John M. Gottman on request that presents the derivation of the equations in the section on five alternative statistics.

Woody and Costanzo's (1990) final suggestion is that causal or structural equations will provide a new insight. However, their suggestions reduce to their original proposal. It is important not to be misled by the use of the term *causal.* The analysis still involves the study of the correlation matrix (or the covariance matrix), and correlation still does not imply causation. It is difficult to see what they suggest structural equations can provide in this instance, because they begin the discussion with another error. They suggest that in the change score structural equation, the weight of the predictor must be negative. This contention cannot be true. The weight of the predictor must depend on the covariance structure between X and the Ys (and assumptions about the es) and not only on the fact that the weight of the posttest score is -1 (see, for example, Hayduk, 1987, for the precise equations). Next comes their suggestion that the most appropriate approach is to weight the posttest with another parameter. This suggestion is actually once again equivalent to a suggestion that analysis of residuals with the pretest score as a covariate be used. In analysis of covariance this is precisely the equation. Hence, we are back where we started. We do not get anything conceptually new in their suggestion of using structural equations.

What Is the True Measure of Change?

Woody and Costanzo (1990) contend that "the raw change model tests the wrong substantive hypothesis" (p. 501). Once again, we can turn to Lord's (1967) paper. He wrote:

> All this has led some people to assert that deviation from the regression line is the real measure of change, and that the ordinary difference between initial and final measurements is not a measure of change. This can hardly be correct. If certain individuals gained 300 ounces, this is a definite fact, not a result of an improper definition of growth. (p. 23)

In fact, the use of change scores is quite clear and simple in the sense that it has a precise interpretable physical meaning. It is, quite simply, the amount of change. The deviation from a regression line is a more complex statistic to interpret. This is a contrast from Woody and Costanzo's (1990) notion that "the prediction of raw change may be devoid of interest" (p. 500). In fact, rather than being devoid of interest, it is precisely the question of interest in this research. A similar point was made by Cronbach and Furby (1970), who wrote:

> Residualizing removes from the posttest score, and hence from the gain, the portion that could have been predicted linearly from pretest status. One cannot argue that the residualized score is a "corrected" measure of gain, since in most studies the portion discarded includes some genuine and important change in the person. The residualized score is primarily a way of singling out individuals who changed more (or less) than expected. (p. 74)

Despite Cronbach and Furby's question about whether change scores or even the difference between any two variables should ever be used, researchers across many fields have continued to use such differences between variables. The proof of the usefulness of differencing two variables must lie in the empirical performance of the difference, as with any other transformation.

Substantive Issues

Next we need to address the substantive issue related to our findings. First, Woody and Costanzo (1990) refer to our results as a "counterintuitive portrayal of a temporally related dynamic relationship of marital conflict and marital satisfaction" (p. 499). However, the results are not at all counterintuitive. It makes some sense to suggest that if a couple confronts an existing problem, their marital satisfaction will improve over the long run. We are not saying, as their title suggests, that marital "agony" precedes marital "ecstasy." We are saying that some forms of confrontation during marital conflict precede increases in marital satisfaction and that concurrently these behaviors are correlated negatively with marital satisfaction and interaction at home. These findings are consistent with a growing body of research that suggests that conflict avoidance can have deleterious long-term consequences for a marriage (Krokoff, in press).

Woody and Costanzo (1990) wrote that reversals of sign were "disturbingly pervasive" (p. 500). It should be noted that in our original article's three tables, significant reversals occurred 9 out of 34 times. This is an interesting enough pattern to attract attention, but it can hardly be called pervasive. In fact, what was interesting about the results in both studies was that some patterns predicted both concurrent distress and deterioration, whereas some patterns predicted concurrent distress and improvement in marital satisfaction. Thus, there is not just one pattern of results in these data. This pattern of results is not likely to be merely a statistical artifact. It is hard to account for such a complex pattern across two studies.

As we noted in our article (Gottman & Krokoff, 1989), the results are quite robust to analysis with partial correlations or regression controlling Time 1 marital satisfaction. For example, husband conflict engagement correlates with wife change scores .44, whereas in the Woody and Costanzo (1990) analysis, the partial of husband conflict engagement with Time 2 wife marital satisfaction, controlling Time 1 wife marital satisfaction is .56 (df for partials = 22); wife positive verbal correlates with wife change scores $-.41$, whereas the partial is $-.48$; and so on. The conclusions are generally robust across all variables. Furthermore, although we did not publish multiple regression analyses, our effects are not colinear. For example, if we use husband conflict engagement and wife positive verbal, we account for 22% ($R = .47$, $df = 20$) of the variance of the residual of Time 2 wife marital satisfaction, regressing Time 1 scores.

In truth, the hypotheses we suggested have far-reaching implications for the study of how marriages function. Previous research on the relation between marital interaction and marital quality had assumed that concurrent and longitudinal patterns would be identical, and interventions for distressed marriages were based on that assumption. We expected to find this result, but it appears to be wrong. The data are consistent with the hypothesis that some interaction patterns actually appear to be harmful to marital satisfaction and the change in marital satisfaction. Some, however, may operate in an opposite fashion. Furthermore, the results are consistent across the two studies. Although the results are not counterintuitive, we did not expect to find these results, and that is why the paper reported *two* longitudinal studies. Replication is essential when results are new and of some import. We (Gottman & Levenson, 1990) are currently analyzing another replication study with a larger N, and the results appear quite similar. Conflict avoidance (particularly by wives) is not always positive in a longitudinal sense, nor is conflict engagement necessarily negative.

References

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Cronbach, L. J., & Furby, L. (1970). How should we measure "change" —or should we? *Psychological Bulletin, 74,* 68–80.

Glass, G. V., & Stanley, J. C. (1970). *Statistical methods in education and psychology.* Engelwood Cliffs, NJ: Prentice-Hall.

Gottman, J. M., & Krokoff, L. J. (1989). Marital interaction and satisfaction: A longitudinal view. *Journal of Consulting and Clinical Psychology, 57,* 47–52.

Gottman, J. M., & Levenson, R. W. (1990). [The social psychophysiology of marriage]. Unpublished raw data.

Hayduk, L. A. (1987). *Structural equation modeling with LISREL.* Baltimore, MD: Johns Hopkins University Press.

Krokoff, L. J. (in press). The role of hidden agendas for understanding emotional reactions and satisfaction in marriage. *Communication Research.*

Lord, F. M. (1967). Elementary models for measuring change. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 12–27). Madison: University of Wisconsin Press.

Woody, E. Z., & Costanzo, P. R. (1990). Does marital agony precede marital ecstasy? A comment on Gottman and Krokoff's "Marital interaction and satisfaction: A longitudinal view." *Journal of Consulting and Clinical Psychology, 58,* 499–501.

Received October 27, 1989
Revision received December 4, 1989
Accepted December 4, 1989 ■