Paper 909

Using The SAS® System To Examine The Reliability Of The Retrospective Measure Of Change

Jim Penny, Center for Creative Leadership, Greensboro, NC
Robert L. Johnson, University of South Carolina, Columbia, SC
Jennifer W. Martineau, Center for Creative Leadership, Greensboro, NC

ABSTRACT

Using 2 data sets for exploratory and confirmatory purposes, the psychometric characteristics of the retrospective measure of change are examined using a theoretical framework. This framework was derived from prior studies of change scores that indicate the reliability of such scores is a function of (1) the reliability of the pre- and post-measures, (2) the variance of the pre- and post-measures, and (3) the correlation of the pre- and post-measures. In the exploratory study, the retrospective change score is found to be a reliable measure, though the change score is negatively correlated with the pre-measure. The reliability of this change score is most strongly associated with the reliability of the pre- and post-measures. Moreover, evidence is presented that suggests that the raters are not estimating the pre- and post-measures as expected. Instead, it appears from these data that the raters are estimating the post-measure and the gain score. The results of the confirmatory study support the results of the exploratory study.

INTRODUCTION

The reliability of the gain score, a subject that is singularly interesting from a historical perspective, has been the focus of some study and no little debate for more than 5 decades, beginning with Gulliksen (1950), continuing with Lord & Novick (1968), Cronbach & Furby (1970), O' Connor (1972), Overall & Woodard (1975), Linn & Slinde (1977), Zimmerman & Williams (1982) and Rogosa & Willett (1983), and more recently with Willett (1988), Humphreys (1991, 1993), Zimmerman, Williams, & Zumbo (1993), Zimmerman (1994), and Feldt (1995). Using Classical True Score theory (Crocker & Algina, 1986), these studies have shown that the reliability of gain scores can vary substantially, with situations where the difference score is sufficiently reliable for the purposes of educational assessment, situations where the reliability is sufficiently low as to compromise any interpretation, and middle-ground situations where the reliability is questionable. The factors that tend to increase the reliability of gain scores are (1) increasing the reliability of both the pre- and post-tests, (2) increasing the reliability of the post-test to exceed that of the pre-test, (3) increasing the variance of the post-test to exceed that of the pre-test, and (4) reducing the correlation between the pre- and post-test.

The Definition of the Gain Score

For the purpose of this discussion, it seems reasonable to define what we mean by a gain score. Generally, we have an initial measure of some kind, then a treatment that is intended to change the attribute that was measured, and finally a follow-up measure. The difference

$$G = T_2 - T_1 \tag{1}$$

between the two measures can be seen as an indicator of change.

Types of Change

Researchers have defined 3 distinct types of change that are called "alpha," "beta," and "gamma" (Bedian & Armenakis, 1989; Golembiewski, 1989; Millsap & Hartog, 1988; Schmitt, Pulakos, & Lieblien, 1984; Tennis, 1989; Terborg, Howard, & Maxwell, 1980; Vande Vliert, Huismans, & Stok, 1985). Alpha change denotes the kind of change that one expects to find after an intervention, such as the aforementioned seminar, and is sometimes referred to as "true" change. The other two types of change, however, pose threats to the accuracy of inferences about the effectiveness of an intervention.

One form of change is the result of a heightened awareness of what is being measured. This type of change, often referred to as beta change, occurs when the anchors of a Likert-type response scale shift from pre-test to post-test in the mind of the rater. This type of change is not the kind that one generally considers when planning to evaluate the effectiveness of an intervention, though it's existence is arguably a component of effective training (Sprangers, 1988). Unfortunately, beta change generally produces observed change that is negative when it is measured with a traditional pre- and post-test design, and even though one might argue that the existence of beta change is indicative of effective and profound training, is it also generally difficult to convince consumers of such programs that the occurrence of negative gain is, in reality, good.

This shift in the interpretation of the anchors to the response scale which earlier was referred to as "beta change" has also been referred to as "response shift bias" (Bereiter, 1963; Howard & Dailey, 1979; Lord, 1958; Rogosa, Brandt, & Zimowski, 1982; Terborg, Howard, & Maxwell, 1980).

The third type of change posited by current change score theory is called "gamma" change and involves the change that is the result of an intervention that produces a change in the conceptualization of the construct that underlies the intervention.

The Retrospective Measure of Change

The retrospective measure of change (Campbell & Stanley, 1963; Hoogstraten, 1982, 1985) is a bit different than the previously defined gain score because of the manner in which the pre- and post-measures are obtained. With retrospective methodology, the retrospective pre-test and the post-test are given simultaneously at some point after the treatment is given, then the difference is computed. Such a methodology is often useful when the interpretations of the textual descriptors that anchor a metric are likely to be altered by the treatment (Collins, Graham, Hansen, & Johnson, 1985; Howard, Dailey, & Gulanick, 1979; Manthei, 1997; Martineau, 1998).

However, despite the manner in which the pre- and post-measures are obtained, the retrospective measure is still constructed as a difference in two other measures. Hence, one can argue that the retrospective measure should carry all the technical characteristics of any other gain score. A relevant question then is: *Where does the retrospective measure fit in the larger rubric of gain score methodology?* To address this question, we need to review in more detail some of the technical characteristics of the difference score.

A Few Recognized Problems with Gain Scores

Potentially Low Reliability

Gulliksen (1950, p. 353) gave the reliability of the gain score, $\rho_{DD'}$, as

$$\rho_{DD'} = \frac{\rho_{xx'}\sigma_x^2 + \rho_{yy'}\sigma_y^2 - 2\rho_{xy}\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 - 2\rho_{xy}\sigma_x\sigma_y} \tag{2}$$

where $\rho_{xx'}$ is the pretest reliability, $\rho_{yy'}$ is the posttest reliability, σ_x^2 is the pretest variance, σ_y^2 is the posttest variance, and ρ_{xy} is the correlation between the pretest and the posttest. Much has been written about this equation (Lord, 1963; Cronbach & Furby, 1970; Overall & Woodard, 1975, 1976; Fleiss, 1976; Linn & Slinde, 1977; Zimmerman, 1994; Feldt, 1995; Penny, 1996a, 1996b).

From this equation, it can be readily seen that the reliability of the difference score is far more complex than casual reflection would suggest. Not only are there five separate characteristics of the pre- and post-tests ($\rho_{xx'}$, $\rho_{yy'}$, σ_x^2, σ_y^2, and ρ_{xy}) that influence the gain score reliability, but these five quantities are frequently interrelated. To change one is likely to introduce unintended and unanticipated change in another (Feldt, 1995), making it difficult to state precisely how $\rho_{DD'}$ will change when one makes changes to either the pre-test, to the post-test, or to both.

Improving the Reliability of the Gain Score

INCREASING THE RELIABILITY OF BOTH THE PRE- AND POST-TESTS

If one assumes that the variance of the pre-test is equivalent to that of the posttest (Feldt, 1995), then Eq. (2) reduces to

$$\rho_{DD'} = \frac{\frac{1}{2}\left(\rho_{xx'} + \rho_{yy'}\right) - \rho_{xy}}{1 - \rho_{xy}} \tag{3}$$

where it is easily seen that any increase in either the reliability of the pre-test or the post-test will result in an increase in the reliability of the difference.

INCREASING THE VARIANCE OF THE POST-TEST OVER THAT OF THE PRE-TEST

Rather than attempting to increase the reliability of the pre- test and the post-test, one might choose to augment the reliability of only the post-test. For this example, we will assume that reliability can be increased without a substantial change in variance of the test. Such might be the instance of choosing a similar, though more reliable, instrument for the

post-test. A more likely approach for many researchers, however, would be the simple lengthening of the post-test. Unfortunately, lengthening a test generally increases the variance relative to that of the pre-test.

Suppose the reliability of the post-test is a factor of *k* greater than that of the pre-test. We can express this relation as

$$\rho_{yy'} = k\rho_{xx'} \tag{4}$$

which can be substituted into Eq. (2) to produce

$$\rho_{DD'} = \frac{\rho\sigma_x^2 + k\rho\sigma_y^2 - 2\rho_{xy}\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 - 2\rho_{xy}\sigma_x\sigma_y} \tag{5}$$

where the subscripts on $\rho_{xx'}$ have been dropped. If we further assume that the $\sigma_x^2 = \sigma_y^2$ (Feldt, 1995), then Eq. (5) becomes

$$\rho_{DD'} = \frac{\rho(1+k) - 2\rho_{xy}}{2(1-\rho_{xy})}. \tag{6}$$

Thus, an increase of *k* in the numerator will be associated with an increase in the reliability of difference scores.

INCREASING THE VARIANCE OF THE POST-TEST OVER THAT OF THE PRE-TEST

Although it seems unlikely that one might change the variance of the post-test without also changing the reliability, it is instructive to examine the influence that variance has on the reliability of the gain score. If we assume that the post-test variance is increased by an amount *k* over that of the pre-test,

$$\sigma_y^2 = k\sigma_x^2 \tag{7}$$

and that the reliability of the pre-test is equivalent to that of the post-test, or

$$\rho_{xx'} = \rho_{yy'} = \rho \ , \tag{8}$$

then Eq. (2) can be written as

$$\rho_{DD'} = \frac{\rho(1+k) - \sqrt{k}\rho_{xy}}{1 - \sqrt{k}\rho_{xy}} \tag{9}$$

which is very similar to Eq. (6) in form. However, the association between the gain score reliability and factor of increase in post-test variance is decidedly non-linear. Moreover, it is possible to have values of *k* for which $\rho_{DD'}$ is meaningless, and this appears to be a likely result of treating changes in variance independently of changes in reliability.

REDUCING THE CORRELATION BETWEEN THE PRE- AND POST-TESTS

It is often the case that the pre-test and the post-test are substantially correlated with each other, and, given that both tests should be measuring the same domain, some degree of correlation is to be expected. Consider the following True Score model of error and unique variance in which a pre-test and a post-test of comparable reliability are subtracted to form a gain score.

```
Pre-test:                E1E1UUUUUUUU
Post-test:           UUUUUUUUE2E2
Gain:            E1E1UUUUE2E2
```

If one considers the instance where the pre- and post-tests are parallel forms of a single instrument, then the expected shared variance of the pre- and post-test is removed by the subtraction, leaving a residual amount of unique variance with the combined error variance of both the pre- and post-tests. Substituting the conditions for equal reliability and equal variance for both the pre- and the post-tests

$$\rho_{xx'} = \rho_{yy'} = \rho$$

$$\sigma_x^2 = \sigma_y^2 = \sigma^2$$

as may happen when the pre- and post-tests are parallel forms of standardized instruments, gives

$$\rho_{DD'} = \frac{\rho - \rho_{xy}}{1 - \rho_{xy}} \tag{10}$$

In this instance, as the parallel-form reliability decreases, the reliability of difference scores increases. However, general practice is to develop parallel forms with high levels of reliability.

Negative Correlation with the Pre-test

Another frequently cited problem with the gain score is that it is likely to be negatively correlated with the pre-test score (Bereiter, 1963; Lord, 1963; Linn & Slinde, 1977; Thorndike, 1966). This observation is a consequence of the fact that those people who are likely to show the greatest gains between the pre- and post-tests are those people with the lowest pre-test scores. That is, the people with the lowest initial scores have the most potential for improvement, while those with higher scores have less potential for improvement. Such an association between a measure of change and the pre-test score seems counter to what one would prefer from an unbiased and reliable measure of change.

This relationship between the gain score and the pre-test score can be shown mathematically with

$$\rho_{xD} = \frac{\rho_{xy}\sigma_y - \sigma_x}{\sqrt{\left(\sigma_x^2 + \sigma_y^2 - 2\rho_{xy}\sigma_x\sigma_y\right)}} \tag{11}$$

where ρ_{xy} represents the correlation of the pre-test and the post-test, σ_x and σ_y represent the standard deviation of the pre- and post-tests, and ρ_{xD} represents the correlation between the gain score and the pre-test. (Linn & Slinde, 1977). The numerator of this equation, hence ρ_{xD}, will be negative when $\rho_{xy}\sigma_y$ is smaller than σ_x, or when

$$\rho_{xy} \le \frac{\sigma_x}{\sigma_y}. \tag{12}$$

Changes in Metric

Perhaps the most problematic consideration in the measurement of change is the fact that instruction is likely to alter the interpretation of the textual descriptors that anchor Likert-type items, introducing a response shift (Bereiter, 1963; Howard & Dailey, 1979; Lord, 1958; Rogosa, Brandt, & Zimowski, 1982) or beta change (Bedian & Armenakis, 1989; Golembiewski, 1989; Millsap & Hartog, 1988; Schmitt, Pulakos, & Lieblien, 1984; Sprangers, 1989; Tennis, 1989; Terborg, Howard, & Maxwell, 1980; Vande Vliert, Huismans, & Stok, 1985). This phenomenon is particularly noisome in training programs where program participants are likely to gain increased awareness of the construct. Algebraic methods posited on Classical Test Theory have yet to yield a correction for response shift, though current research using Item Response Theory is beginning to address this problem and will likely produce plausible methods by which one may identify the existence, magnitude, and direction of the response shift (Craig, 2000; Craig, Palus, & Rogolsky, 1999). Moreover, some success with the analytical separation of alpha and beta change has been demonstrated using heirarchical modeling (Hoogstraten & Koele, 1988) and structural equation modeling (Millsap & Hartog, 1988).

Research Questions

These previous studies of gain scores led to three broad questions that drove this research. The first question involves the relationship between the retrospective measure of change and the pre-measure, and arises from the criticism that gain scores tend to be negatively correlated with the pre-test score. Is this phenomenon seen with the retrospective measure, or does the peculiar nature by which both pre- and post-measures are simultaneously obtained tend to produce gain scores that are, at least, not negatively correlated with the pre-test?

The second research question involves the reliability of the retrospective measure of change. If we estimate the gain score reliability using Cronbach's alpha, do we see values that suggest that reliability changes as expected with (1) changes in the reliability of the pre- and post measures, (2) changes in the variance of the pre- and post measures, and (3) changes in the correlation between the pre- and post-measures?

Our third research question involves the process by which the retrospective measure is obtained. One can argue that the use of which the retrospective measure is posited is on the expectation that a rater can simultaneously and independently respond to both the pre- and post-measure. Of course, the concept of simultaneity, as well as that of independence, is moderated by the fact that a single person is providing the rating, and, as such, one can anticipate some residual degree of correlated error. In such an event, owing first to the erstwhile independence of the pre-and post-measures and second to the linear dependence of the gain score on both the pre- and post-measures (Gain = T_2 - T_1), it seems reasonable to suggest that the correlation between the pre- and post-

measures would be lower than the correlation between either the pre-measure and the gain score,

$$\rho_{Pre,Post} < \rho_{Pre,Gain},$$

or the post-measure and the gain score,

$$\rho_{Pre,Post} < \rho_{Post,Gain}.$$

However, one may also argue that a rater could use a different process to produce the pre- and post-ratings. That is, the rater chooses simultaneously and independently, or at least as simultaneously and independently as is possible with one rater, both the post-measure and the gain score. From these two decisions, the rater then derives the pre-measure, $T_1 = T_2$ - Gain. If such a process is in use, owing first to the degree of independence of the post-measure and the gain score and second to the linear dependence of the pre-measure on both the post-measure and also the gain score, then the correlation between the post-measure and the gain would be lower than the correlation between either the pre-measure and the gain

$$\rho_{Post,Gain} < \rho_{Pre,Gain},$$

or the pre-measure and the post-measure

$$\rho_{Post,Gain} < \rho_{Pre,Post}.$$

DESCRIPTION OF THE TEST DATA

Two studies were conducted to investigate the reliability of the retrospective measure of change within the theoretical framework of gain scores. In both studies participants in training programs completed surveys that examine program effectiveness through the use of the retrospective measure of change. The designs of the two studies are described below.

Exploratory Study

The data for the exploratory portion of this study were obtained from the pilot test of a recently developed instrument known as *REFLECTIONS* with participants in a leadership development training program known as Foundations of Leadership (FOL) that is offered by the Center for Creative Leadership. *REFLECTIONS-FOL* is a multi-rater survey designed to assess the behavioral changes that are seen in FOL participants approximately 3 to 4 months after participation in the training program. The instrument, as initially tested, was comprised of 50 items across 10 scales. The scales and items were chosen from prior impact studies conducted at the Center and from consultation with those trainers who teach the program.

The *REFLECTIONS* survey uses a retrospective measure of change to assess developmental growth in the participant. The instrument is completed by self-raters, who were participants in the training programs, and other-raters, who are colleagues of the self-raters. Our instructions to the self-raters in this pilot were:

> On the following pages, please respond to the statements regarding your leadership. Each statement needs two responses which should be written in the spaces to the right of it. The first response should represent the way that you behaved before attending FOL, and the second should represent the way that you behave now after attending FOL. We will use the difference in these two numbers to measure your developmental change since you attended FOL.

The raters were given three example statements. The stem of the example statements was identical to the stem of the 50 survey statements. The text of the stem was "Record the rating that best describes the extent to which you have been able to," and the text of the first example statement was "Become more aware of how others perceive you." The raters use a 9-point Likert-type scale to give ratings of their behaviors "Before the Program" and "Now." The odd-points on the scale were anchored with 1="Not at all," 3="To a small extent," 5="To a moderate extent," 7="To a large extent," and 9="To a very great extent." The even points of the scale did not receive textual anchors, though in the production version of the survey the even points do receive textual anchors. For the other-raters, appropriate changes in the pronouns and in the verbs of the stem and statements were made to the surveys.

The subjects for the exploratory part of this study were 39 volunteers who participated in one of the FOL programs offered by the Center for Creative Leadership over a 4 month period in 1999. Each volunteer was sent a "family" of surveys, consisting of 1 self-rater survey and 11 other-rater surveys. Resulting from these 39 families of surveys were 39 (or 14%) self surveys, 24 (9%) boss surveys, 19 (7%) other superior surveys, 67 (24%) surveys from direct reports, and 105 (38%) surveys from peers.

The feedback reports from multi-rater surveys, or 360-degree surveys, frequently contain information from each rater group. As well, there is often a composite group that is called "all other raters" which contains information from all of the other-raters, but not the self-rater. For the purpose of this study, we chose to examine the change scores for the combined group of all other-raters (N=237).

Confirmatory Study

The data for the confirmatory portion of this study were obtained from the pilot test of a recently developed version of *REFLECTIONS* for use with participants in a leadership development training program known as the Looking Glass Experience (LGE) that is also offered by the Center for Creative Leadership. As with *REFLECTIONS-FOL*, *REFLECTIONS-LGE* is a multi-rater survey that was designed to assess the behavioral changes that are seen in LGE participants approximately 3 to 4 months after participation in the training program. The instrument was comprised of 96 items across 20 scales. The scales and items were chosen from prior impact studies conducted at the Center and from consultation with those trainers who teach the program.

The subjects for the confirmatory part of this study were 55 volunteers who participated in one of the LGE programs offered by the Center for Creative Leadership over a 4 month period in 1999. Each volunteer was sent a family of surveys with instructions identical to the FOL volunteers. In these 55 families of surveys, there were 408 surveys returned. Of these 408 surveys, there were 55 (or 14%) self surveys, 41 (10%) boss surveys, 29 (7%) other superior surveys, 131 (32%) surveys from direct reports, and 152 (37%) surveys from peers. For the purpose of this study, we chose to examine the change scores for the combined group of all other-raters (N=353).

RESULTS OF EXPLORATORY STUDY

The Association between the Retrospective Pre-test and the Gain Score

The first research question involves the common criticism that gain scores are often negatively correlated with the pre-test. The next to the last column of Table 1 shows the correlation between the pre-measure and the gain score for each scale of the REFLECTIONS instrument for FOL. In every instance, the correlation is negative, with a mean correlation of -.49 and a standard deviation of .07. The pre-test, on average, explains about 24 percent of the variance in the gain score. Moreover, the smaller pre-measures are more likely to be associated with the large gain scores.

In addition, there appears to be a relation between a reduction in the standard deviation between the pre and post test and the magnitude of the correlation $r_{Before,Gain}$. More specifically, the largest negative correlations between gain scores and pre-tests are associated with scales in which the standard deviation decreased from pre- to post-measures. For example, in Scale 2 the standard deviation went from 1.5 on the pre-measure to 1.2 on the post-measure and the correlation between $r_{Before,Gain}$ is -.60. Conversely, in the case of Scale 6 the standard deviation remained the same (SD=1.4) and the magnitude of the correlation is -.41.

From these data, it would appear that there are circumstances under which the retrospective methodology can produce gain scores that are negatively correlated with the pre-measure. These results conform to the expectations presented in Linn & Slinde (1977).

The Internal Consistency of the Retrospective Measure of Gain

We used Cronbach's alpha to estimate the reliability of the gain score for each scale. The mean reliability of the gain scores across the 10 scales is .90 with a standard deviation of .02. These values are given in Table 1. We find them surprisingly high given the common notions that seem to surround the reliability of gain scores. Moreover, these values compare quite well with the reliability of the pre- and post-measures, both of which have a mean reliability and standard deviation of .93 and .02, respectively.

From Table 1, it is apparent that when the reliability of both the pre- and post-measure increased, then the reliability of the gain score generally increased ($r_{r\_Pre\_r+r\_Post,r\_Gain}$=.89, p=.0005). However, changes in the variance of the pre- and post measures and changes in the correlation between the pre- and post-measure did not exhibit statistically significant correlations with the values of Cronbach's alpha that were computed for the gain scores.

What Process Are the Raters Using?

Our third research question involved the process by which raters arrived at the retrospective pre-measure. Do they choose the pre- and post-measures independently, or do they choose a post-measure and a gain and then compute the pre-measure? The mean correlation between the pre- and post-measures is .70 with a standard deviation of .06. The mean correlation between the post-measure and the gain score is .29 with a standard deviation of .11. This evidence appears to suggest greater independence between the post-measure and the gain score, supporting the hypothesis that raters are estimating the gain score and post-measure instead of the pre-measure and the post-measure.

Discussion of the Exploratory Study

The Negative Correlation of the Change Score and the Retrospective Pre-test

Popham (1978) suggests that a negative correlation between a set of gain scores and a set of pre-test scores can be attributed to the test ceiling. In the exploratory study, accompanying a reduction of the standard deviation between pre and post measures was a concomitant rise in the magnitude of the correlation between the gain scores and pre-test scores. Smaller deviations in the post-test, then, may occur due to a ceiling or floor

effect and appear to support Popham's contention. Zimmerman and Williams (1982) additionally suggest that even in the case of a negative correlation between gain scores and pre-test scores that gain scores can still be highly reliable.

Factors that Influence the Reliability of the Retrospective Change Score

The influence of pre- and post-test reliability on the observed reliability of the gain scores appears in keeping with the expectations that have been formed from prior studies. However, we were a bit dismayed when the post-test variance and the correlation of the pre- and post-test exhibited little association with the reliability of the gain scores. We expect that this occurrence is an artifact, first, of the sample size of 10 scales, and, second, of the bivariate analysis which ignores inter-relations between the factors (Feldt, 1995).

How the Raters Estimate the Ratings

We find the evidence that is given by the correlational analysis to be the most interesting. There is indeed evidence to suggest that the other-raters were not arriving at their pre- and post-measures by the process that was expected. Indeed, one can easily argue that the raters are computing the pre-measure from their perspective of both the current standing on the item and their perspective of the change that they have witnessed in the past 3 to 4 months.

Expectations of the Confirmatory Study

In general, we anticipated that our confirmatory study would support the finding from the exploratory study. This expectation was posited on the observation that the training programs were similar, the participants in those programs come from similar environments, the two surveys were of similar content and style, and the retrospective measure was presented similarly in both studies. However, there were some differences that could influence the results of the two studies. The LGE program required 4.5 days of participation whereas the FOL program required only 3 days. The LGE program was centered around an intensive simulation in which the participants role-play a day in the life of a corporation that is under-going significant change. Moreover, the LGE survey was longer by nearly a factor of two, owing primarily to the increased areas of impact that the LGE program has over the FOL program.

Results and Discussion of the Confirmatory Study

The results of the confirmatory study were highly congruent with the results of the exploratory study. Again, the reliability of the gain score was higher than one might expect with a mean of reliability .88 and a standard deviation of .05 across the 20 scales of the instrument. As well, the reliability of the gain score was comparable to both the reliability of the pre-measure where the mean was .91 and the standard deviation was .05 and also the reliability of the post-measure where the mean was .90 with a standard deviation of .06.

The correlation of the retrospective measure of change and the pre-score continued to be negative without exception for each of the 20 scales on the instrument, with a mean value of -.45 and a standard deviation of .07, which are comparable to the values seen with the exploratory instrument. On this instrument, the pre-test accounted for 20 percent of the variability in the gain score, 4 percent less than in the exploratory study.

With 20 scales in the confirmatory study instead of the 10 scales in the exploratory study, we had hoped for sufficient power to be able to demonstrate more effectively the association between the gain score and the various factors that tend to affect the gain scores. Unfortunately, the extra scales did not produce results that were different from those of the exploratory study. The correlation between gain score and the combined reliability of the pre- and post-measure was .94 and statistically significant. The correlations between the gain score and (1) the ratio of post-measure reliability to pre-measure reliability, (2) the ratio of the variance of the post-measure to that variance of the pre-measure, and (3) the correlation between the pre-measure and the post measure all failed to achieve statistical significance. As with the exploratory study, it is our belief that these bivariate relationships likely require a larger sample size for dependable detection.

Our third expectation of correlational evidence to suggest that the raters were estimating the post-measure and the gain score in lieu of the pre- and post measures was supported in each of the 20 sub-scales except the first one, where we had a small difference in the opposite direction. The average correlation between the pre- and post-measures was .70 with a standard deviation of .07; between the post-measure and the gain score, .31 and .12, respectively. As with the exploratory study, this evidence suggests greater independence between the post-measure and the gain score than between the pre- and post-measures.

From this evidence, one could argue, then, that the raters are not implementing the retrospective measure as suggested in the survey instructions, where they are explicitly asked to give pre- and post-measures for the individual who is being rated. Instead, these data suggest that the raters are deciding on the post-measure and the gain score, then recording the difference as the pre-measure. A study that included think-aloud protocols and post-rating interviews could likely uncover just how the raters are thinking when they complete a retrospective survey, and we would like to see just such a study undertaken.

Concluding Remarks

Although this paper represents an empirical study of the psychometric properties of gain scores using retrospective methodology, and by no means represents a study of appropriate measures of gain, it seems reasonable at this point to reflect upon the use of the gain score as a measure of change in general and of learning in specific. Argument for the importance of the assessment of change abound (Linn, 1981; Ewell, 1984; Nuttall, 1986; and Astin, 1987) and arguments for the usefulness of gain scores as measures of educational improvement are plentiful (Astin & Ewell, 1985; and McMillan, 1988). Indeed, if one assumes that the test scores are unbiased estimators of ability or knowledge, then the difference between the two scores should be an equally unbiased estimator of change, regardless of the issues embedded within reliability; however, the problematic nature of gain score reliability is well-documented (Cronbach & Furby, 1970; Overall & Woodard, 1975; Warren, 1984; and Pike, 1992), though perhaps overstated in Cronbach & Furby (1970, p. 78) who write: "There appears to be no need to use measures of change as dependent variables and no virtue in using them."

Perhaps the issue should turn from the reliability of gain scores to the validity of gain scores (Kane, 1996). One can argue that a change score can potentially represent a tangled mess of three types of change, one that is expected, and two others that may tag along for the ride and render the interpretation of the gain score problematic. Moreover, we wonder how one might interpret any composite, of which the gain score is but one, when the constructs that underlie the components of the composite are not equivalent, and we would suggest that the validity of the interpretation would be compromised.

However, the retrospective measure of change, regardless of how raters actually implement it, is designed to avoid the problem that the existence of beta and gamma change can present to the measurement of alpha change; and, if you can accept that raters are able to provide valid reflective pre-measures and post-measures, then the retrospective measure of change can indeed provide a reliable measure of alpha change, as these data suggest.

REFERENCES

Astin, A. W. (1987). *Achieving Educational Excellence*. San Francisco: Jossey-Bass.

Astin, A. W., & Ewell, P. T. (1985). The value added debate. . . continued. *AAHE Bulletin*, 37, 11-13.

Baird, L. L. (1988). Value-added: Using student gains as yardsticks of learning. In C. Adelman (Ed.), *Performance and Judgement: Essays on Principles and Practice in the Assessment of College Student Learning* (pp. 205-216). Washington, DC: U.S. Government Printing Office.

Bedian, A. G., & Arkmenkais, A. A. (1989). Promise and prospects: The case of the alpha, beta, gamma change typology. *Group and Organizational Studies, 14,* 155-160.

Bereiter, C. (1963). Some persisting problems in the measurement of change. In C. W. Harris (Ed.), *Problems in measuring change*. Madison: University of Wisconsin Press, pp. 3-20.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research. Chicago*: Rand McNally.

Collins, L. M., Graham, J. W., Hansen, W. B., Johnson, C. A. (1985). Agreement between retrospective accounts of substance use and earlier reported substance use. *Applied Psychological Measurement, 9(3)*, 301-309.

Craig, S. B. (2000). Differentrial Item Functioning in the Measurement of Training Effectiveness: An Examination Using Item Response Theory. Unpublished doctoral dissertation.

Craig, S. B., Palus, C. J., & Rogolski, S. (1999). Using item response theory to correct for response shift bias in measurements of training impact. In S. Craig (Chair), New Strategies for Old Problems in Evaluation: Coping with Missing Data, Scale Compression, and Response Shift Bias. Symposium presented at the annual conference of the American Evaluation Association, Orlando, FL.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, and Winston, Inc.

Cronbach, L. J., & Furby, L. (1970). How should we measure "change"-or should we? *Psychological Bulletin*, 74, 68-70.

Ewell, P. T. (1984). *The Self-Regarding Institution: Information for Excellence*. Boulder, CO: National Center for Higher Education Management Systems.

Feldt, L. S. (1995). Estimation of the reliability of differences under revised reliabilities of component scores. *Journal of Educational Measurement*, *32*, 295-301.

Golembiewski, R. T. (1989). The alpha, beta, gamma change typology: Perspectives on acceptance as well as resistance. *Group and Organizational Studies, 14*, 150-154.

Golembiewski, R. T., Billingsley, K., & Yeager, S. (1976). Measuring change and persistence in human affairs: Types of change

generated by OD designs. *Journal of Applied Behavioral Science*, *12*, 133-157.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Hoogstraten, J. (1982). The retrospective pretest in an educational training context. *Journal of Experimental Education*, *50*, 200-204.

Hoogstraten, J. (1985). Influence of objective measures on self-reports in a retrospective pretest-posttest design. *Journal of Experimental Education*, *53*, 207-210.

Hoogstraten, J., & Koele, Pieter. (1988). A method for analyzing retrospective pretest/posttest designs. *Bulletin of the Psychonomic Society, 26(2)*, 124-125.

Howard, G. S., Dailey, P. R., & Gulanick, N. A. (1979). The feasibility of informed pretests in attenuating response-shift bias, *Applied Psychological Measurement, 3(4)*, 481-494.

Howard, G. S., Ralph, K. M., Gulanick, N. A., Maxwell, S. E., Nance, D. W., & Gerber, S. K. (1979). Internal validity in pretest-posttest self-report evaluations and a reevaluation of retrospective pretests. *Applied Psychological Measurement*, *3*, 1-23.

Howard, G. S., & Dailey, P. R. (1979). Response shift bias: A source of contamination in self-report measures. *Journal of Applied Psychology*, *64*, 144-150.

Humphreys, L. G. (1991). The relation of power of statistical tests to range of talent: A correction and amplification. *Applied Psychological Measurement*, 15, 267.

Humphreys, L. G. (1993). Further comments on reliability and power of significance tests. *Applied Psychological Measurement*, 17, 11-14.

Kane, M. (1996). The precision of measurements. *Applied Measurement in Education*, *9*, 355-379.

Kennedy, J. J., & Bush, A. J. (1985). *An Introduction to the Design and Analysis of Experiments in Behavioral Research*. New York: University Press of America.

Linn, R. L. (1981). Measuring pretest-posttest performance changes. In R. A. Berk (Ed.), *Educational Evaluation Methodology: The State of the Art* (pp. 94-109). Baltimore: Johns Hopkins University Press.

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

Lord, F. M. (1958). Further problems in the measurement of growth. *Educational and Psychological Measurement*, *18*, 437-454

Lord, F. M. (1963). Elementary models for measuring change. In C. W. Harris (Ed.), *Problems in measuring change*. Madison: University of Wisconsin Press, pp. 21-38.

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Erlbaum.

Martineau, J. W. (1998). Using 360-degree surveys to assess change. In Tornow, W. W., & London, M. (Eds.) *Maximizing the Value of 360-Degree Feedback*. San Francisco: Josey-Bass Publishers.

Manthei, R. J. (1997). The response-shift bias in a counselor education programme. *British Journal of Guidance & Counseling. 25(2)*, 229-237.

McMillan, J. H. (1988). Beyond value-added education: Improvement alone is not enough. *Journal of Higher Education*, 59, 564-579.

Millsap, R. E., & Hartog, S. B. (1988). Alpha, beta, and gamma change in evaluation research: A structural equation approach. *Journal of Applied Psychology, 73*, 574-584.

Nuttall, D. L. (1986). Problems in the measurement of change. In D. L. Nuttall (Ed.), *Assessing Educational Achievement* (pp. 153-167). London: Falmer Press.

O' Connor, D. F., Jr. (1972). Extending classical test theory to the measurement of change. *Review of Educational Research*, 42, 73-97.

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

Penny, J. (1996). Using the SAS® System to assess the reliability of gain scores. *SESUG '96 Users Group Proceedings*. Cary, NC: SAS Institute, Inc., 428-435.

Penny, J. (1996). *The case for the use of gain scores in educational assessment*. Paper presented at the annual meeting of the North Carolina Association for Research in Education, Chapel Hill, NC.

Pike, G. R. (1992). Lies, damn lies, and statistics revisited: A comparison of three methods of representing change. *Research in Higher Education*, 33, 71-84.

Popham, W. J. (1978). *Criterion-referenced measurement*. Englewood Cliffs, NJ: Prentice-Hall.

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

Rogosa, D. R., & Willett, J. B. (1983). Demonstrating the reliability of the difference score in the measurement of change. *Journal of Educational Measurement*, 20, 335-343.

Roskam, E. E. (1976). Multivariate analysis of change and growth: Critical review and perspectives. In D. N. M. De Gruijter and L. J. T. van der Kamp (Eds.), *Advances in Psychological and Educational Measurement* (pp. 111-133). New York: John Wiley.

Schmitt, N., Pulakos, E. D., & Lieblein, A. (1984). Comparison of three techniques to assess group-level beta and gamma change. *Applied Psychological Measurement, 8*, 249-260.

Sprangers, M. (1988). *Response Shift and The Retrospective Pretest: On The Usefulness of Retrospective pretest-post-test designs in detecting training related Response Shifts.* Den Haag, Netherlands: Het Institut Voor Onderzoek Van Het Onderwijs.

Sprangers, M. (1989). Subject bias and the retrospective pretest. *Bulletin of the Psychonomic Society, 27(1)*, 11-14.

Tennis, C. N. (1989). Responses to the alpha, beta, gamma change typology. *Group and Organizational Studies, 14*, 134-149.

Terborg, J. R., Howard, G. S., & Maxwell, S. E. (1980). Evaluating planned organizational change: A method for assessing alpha, beta, and gamma change. *The Academy of Management Review, 5*, 109-121.

Thorndike, R. L. (1966). *The concepts of over- and underachievement*. New York: Columbia University, Teachers College, Bureau of Publications.

Van de Vliert, E., Huismans, S. E., & Stok, J. J. (1985). The criterion approach to unravelling beta and alpha change. *Academy of Management Review, 10*, 269-274.

Willett, J. B. (1988). Questions and answers in the measurement of change. In E. Z. Rothkopf (Ed.), *Review of research in education* (Vol. 15, pp. 345-422). Washington, DC: American Educational Research Association.

Zimmerman, D. W. (1994). A note on the interpretation of formulas for the reliability of differences. *Journal of Educational Measurement*, 31, 143-147.

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

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

CONTACT INFORMATION

Jim Penny
Center for Creative Leadership
One Leadership Place
Greensboro, NC 27438-6300
Work: 336-286-4442
Fax: 336-286-4434
Email: pennyj@leaders.ccl.org
pennyjim@aol.com

Table 1.

Before, Now, and Gain scores for exploratory FOL data.

	Before			Now			Gain			Correlations		
Scale	mean	sd	α	mean	sd	α	mean	sd	α	$r_{Before,Now}$	$r_{Befor,Gain}$	$r_{Now,Gain}$
1	4.4	1.4	.89	6.2	1.5	.91	1.9	1.4	.88	.55	-.45	.52
2	6.0	1.5	.91	7.1	1.2	.89	1.5	1.1	.86	.66	-.60	.21
3	5.3	1.6	.94	6.7	1.4	.92	1.5	1.1	.89	.73	-.50	.24
4	5.2	1.6	.95	6.6	1.5	.93	1.5	1.2	.91	.68	-.44	.35
5	5.2	2.0	.92	6.7	1.6	.94	1.5	1.4	.91	.71	-.59	.16
6	5.1	1.4	.93	6.5	1.4	.92	1.4	1.1	.88	.72	-.41	.34
7	5.1	1.5	.94	6.6	1.4	.94	1.6	1.1	.91	.69	-.42	.37
8	5.1	1.6	.97	6.4	1.5	.97	1.4	1.1	.94	.75	-.46	.24
9	5.3	1.5	.94	6.6	1.3	.95	1.3	1.1	.92	.71	-.53	.24
10	5.6	1.5	.90	6.8	1.3	.90	1.2	1.0	.89	.75	-.49	.22

Note: N=237