Academy of Management Review 1980, Vol. 5, No. 4, 561-566

The Measurement and Control of Beta Change[1]

ARTHUR G. BEDEIAN
ACHILLES A. ARMENAKIS
ROBERT W. GIBSON
Auburn University

Applied researchers have long been confronted with the complex statistical problems associated with the measurement of change. A method frequently used in such analyses is the pretest-posttest research design. This approach incorporates the often overlooked assumption that the pretest and posttest scores are comparable. However, for the scores to be comparable, a common metric must exist between them. Distinguishing between three types of change in test scores (alpha, beta, and gamma), we present an original statistical procedure for measuring ***and*** *controlling the confounding influence of beta change — i.e., the problem of scale recalibration in the minds of respondents.*

Within the constraints imposed by the limitations of scientific method, the primary intent of applied research is to identify causes and to evaluate their effects [Mahoney, 1978]. One method frequently employed in such efforts is to compare dependent variable values across time to determine if a behavioral change has taken place as a result of some form of planned treatment or intervention. Typically, a researcher then proceeds to select and perform those statistical tests judged appropriate for assessing the degree and extent of change achieved, if any. This approach is based on the assumption that the pre-intervention and post-intervention test scores are comparable. However, for the scores to be comparable, a common metric must exist between them, and the possibility of this requirement being violated is particularly likely in the use of self-report measures. In using self-report measures, as Howard and Dailey note, "researchers assume that a subject's standard for measurement of the dimension being assessed will not change from one testing to the next (pretest to posttest)." They go on to warn that:

> If the standard of measurement were to change, the posttest ratings would reflect the shift in addition to actual changes in the subject's level of functioning. Consequently, comparison of pretest with posttest ratings would be confounded by this distortion of the internalized scale, yielding an invalid interpretation of the effectiveness of the intervention [1979, p.144].

Our purpose here is to present a statistical technique for the measurement and control of such confounding. An obvious and immediate advantage of such a technique is that it will not only be of value in differentiating and identifying the types of change observed, but will also allow researchers to rely more on the validity of their findings.

Types of Change

Of particular pertinence to our discussion is Golembiewski, Billingsley, and Yeager's [1976]

[1]We are grateful for the helpful comments of William B. Hudson, James F. Cox, and Amitava Mitra.

identification of three types of change: alpha, beta, and gamma. *Gamma change* refers to the reconceptualization or redefinition of a referent variable. It occurs when subjects change their basic understanding, from one testing period to another, of the criterion being measured. Thus "peer leadership" may mean something quite different at Time 1 as compared to Time 2, especially if a planned treatment or intervention was directed at enhancing subjects' understanding of this or other related concepts. Such a redefinition, of course, would make a comparison of pretest and posttest responses virtually meaningless, *unless* the accomplishment of gamma change was, in fact, the intended purpose of an intervention. For example, if as part of an organization's performance-evaluation program, managers are required to evaluate subordinates on their "peer leadership," it would be important for this construct to be interpreted similarly by all raters. To this end, an intervention might be planned to induce a common understanding among all managers concerned.

Beta change occurs when the standard of measurement used by a subject to assess an item changes from one testing period to another. Such change indicates a recalibration of a subject's internalized scale of measurement. It is this change in the standard of measurement that is our focus in this paper. Beta change is defined as having occurred when, discounting for the occurrence of gamma change, a subject rates a certain behavior as a 2 (on Likert-type scale) at Time 1 and the identical behavior as a 3 at Time 2.

Finally, *alpha change* is defined as a rating change for which both gamma and beta change have been ruled out. That is to say, neither the subjects' understanding of the criterion being measured nor the measurement scale has changed. As an example, assume that a group of managers is exposed to a planned intervention and that as a result their behavior is changed. Assume further that ratings of their behavior by subordinates have been collected according to a standard pretest-posttest research design. If, after determining that neither gamma change nor beta change has occurred, the researcher observes a difference in subject responses from Time 1 to Time 2, alpha, or real change, can be said to have occurred.

Previous research on the identification and detection of the different types of change has been concerned with change as a *group* occurrence. Golembiewski, Billingsley, and Yeager discussed types of change as they related to an entire subject sample. Building on this foundation, Zmud and Armenakis [1978] developed a procedure for detecting as well as distinguishing beta from alpha change as a group phenomenon. However, the ability to gauge the confounding influences of beta and gamma change requires the analysis of *individual* subject responses. To meet this requirement, we are presenting an original statistical procedure for measuring *and* controlling the confounding influence of beta change, or what might be more descriptively referred to as the problem of scale recalibration by respondents. Although developed independently, our procedure is conceptually similar to Thompson's [1963] approach to calibrating observer bias in time-study pace estimation.

Actual versus Ideal Scores

The statistical procedure to be described has two basic requirements: (1) the collection of subject responses on at least two occasions, and (2) the use of a research instrument incorporating two measurement subsets: one to gauge respondent perceptions of *actual* conditions and one to gauge respondent perceptions of *ideal* conditions. Figure 1 shows how these requirements can be met, using a sample question taken from the widely used Survey of Organizations (SOO) questionnaire [Taylor & Bowers, 1972]. Our method utilizes both sets of scores, to obtain evidence of scale recalibration. We have also developed a recalibration function to be used when scale recalibration has occurred, whereby pretest and posttest scores can be transformed so as to be comparable. The intent is to convert (where necessary) scales for Time 1 and Time 2 to the same metric, thus making it feasible, once gamma change has been ruled out, to obtain a measurement of alpha (real) change.

Basic differences in the underlying nature of the conditions that actual and ideal scores are intended to measure should be noted, particularly in their use across time. Over time, respondent ratings of *actual conditions* may reflect not only alpha

To what extent do persons in your work group offer each other new ideas for solving job-related problems?[a]

TIME 1						TIME 2				
This is how it is *now.*										
Actual									Actual	
1	2	3	4	5		1	2	3	4	5
This is how I'd *like* it to be.										
Ideal									Ideal	
1	2	3	4	5		1	2	3	4	5

Figure 1
Identifying Beta Change

[a]Source: Taylor, J.; & Bowers, D. G. *Survey of organizations.* Ann Arbor: CRUSK, Institute for Social Research, University of Michigan, 1972.

change, but also confounding influences of beta and gamma change. In marked contrast, however, respondent ratings of *ideal conditions,* although susceptible (over time) to beta and gamma change, by definition cannot exhibit alpha change. Any alteration in a respondent's idealized notion of a referent variable would represent a reconceptualization of the particular construct in question and thus (by definition) be reflected as gamma change.

A Numerical Adjustment

For purposes of brevity, the following illustration will be developed with reference to a simple pretest-posttest research design. Once understood, the basic procedure to be presented can be readily generalized to two or more treatments and to two or more independent variables. The procedure is explicitly based on the assumption that although beta change can be expected to occur over time (i.e., from one testing to another), it is not expected to occur between responses collected at the same time (i.e., during one testing). That is, we are assuming that a monotonic relationship exists between actual and ideal ratings collected at the same time. This assumption builds on the more fundamental assumption that individuals have an internalized standard for judging the psychological distances between scale intervals. This internalized standard can be expected to differ from one testing to the next (pretest to posttest), but it is not expected to differ between responses collected on one occasion [Howard, Schmeck, & Bray, 1979].

Step One

Building on a simple linear regression model, where $Y' = a + bX$, Step 1 in the development of the proposed recalibration function is to array the Time 1 and Time 2 ideal responses for *each subject* according to the following format, where X_i represents the Time 1 and Y_i the Time 2 raw scores for each questionnaire item.

Respondent No. 1

Item	X_i Time 1 ideal (raw) scores	Y_i Time 2 ideal (raw) scores
1	x_1	y_1
2	x_2	y_2
3	x_3	y_3
•	•	•
•	•	•
n	x_n	y_n

Step Two

Step 2 calls for the fitting of a linear regression equation to the above data such that $Y'_i = a + bX_i$, where Y'_i = *Time 2 ideal (raw) scores,* a = the intercept constant, b = regression coefficient or weight, and X_i = Time 1 ideal (raw) scores. This procedure results in an estimate of the a and b coefficients using the least squares deviation criterion. The answer to the question of how good an estimate these terms provide is found by computing the resulting equation's standard error of estimate, the formulation and explanation of which can be found in any text dealing with the basics of linear regression. Of importance here is the fact that, as an index of variation of dispersion around the line of regression, the standard error of estimate enables one to state with a specific degree of certainty how good an estimate of the regression equation has been developed.

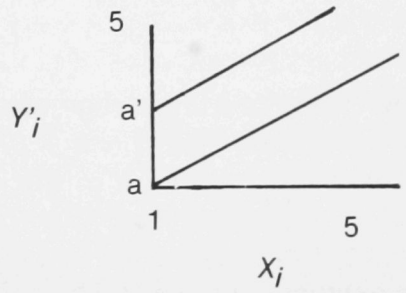
Step Three

Having now computed the regression line or the line of best fit describing the relationship between the Time 1 and Time 2 ideal scores for each sub-

ject, we are in a position to decide *(Step 3)* whether significant beta change has occurred, and if so, the nature of recalibration necessary. This determination can be divided into several parts and centers on inspection of the linear equation of regression derived above. Thus, in reference to this equation:

1. If *b* is not significantly different from 1 and *a* is not significantly different from 0, such that $Y' = 0 + 1X$ (i.e., Time 2 ideal scores = Time 1 ideal scores), no beta change has occurred. (Some computerized statistical packages provide the necessary statistics to determine if *a* is significantly different from 0 and if *b* is significantly different from 1 [e.g., Barr, Goodnight, Sall, & Helwig, 1976].) Respondent scale calibration has thus been constant.

2. If *b* is not significantly different from 1 but *a* is significantly different from 0, such that $Y' = a + 1X$, simple scale displacement has occurred. This would be a case of scale displacement of a constant magnitude (i.e., equal to the value of *a'*). Thus, between Time 1 and Time 2, all 1's may have become 2's, 2's have become 3's (or vice versa), and so forth. Graphically, this would be shown as a shift (upward or downward) in the intercept constant:

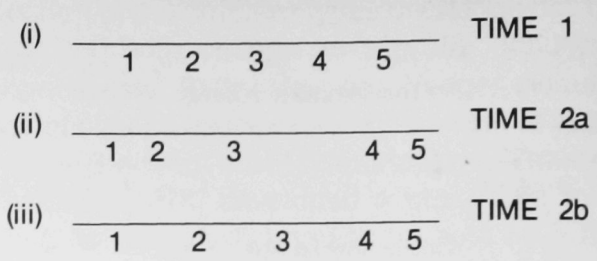


   Note that in this case (hereafter referred to as *Type I beta change),* the slope *(b)* of the regression line (representing the change in *Y* per unit change in *X),* remains the same.

3. If $b \neq 1$, regardless if $a = 0$, respondent calibration has not been constant and rescaling is necessary so that Time 2 scores can be accurately compared to Time 1 scores. Situations in which $b \neq 1$ will hereafter be referred to as *Type II beta change.*

4. Type II beta change may take at least two forms: *scale interval stretching* and *scale interval sliding.* The former occurs when respondents lengthen the psychological distance between scale intervals, so that their width is not constant. Thus, as opposed to having five equal intervals as shown for Time 1 on line (i) below, at Time 2 the respondent's judgment scale may be recalibrated in any number of ways, as in lines (ii) and (iii).

Ideal Scale

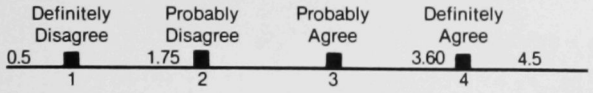


   As a result of scale interval stretching, a behavior judged a 4 at Time 1 may receive a value of 3 at Time 2. The argument advanced by Golembiewski, Billingsley, and Yeager [1976] to explain this occurrence is that the "elasticity of distance" is subject to expansion or contraction as the personal standards of respondents change.

   To determine if scale interval widths have remained fairly constant, it is necessary to compute whether there is a significant difference between the Time 1 ideal and Time 2 ideal rating frequencies. Among the tests available for this purpose, the Kolomogorov-Smirnov goodness-of-fit test (K-S) is particularly appropriate [Siegel, 1956]. Concerned with the degree of agreement between two distributions, the K-S treats individual observations separately and, unlike the chi-square test, need not lose information through the combining of categories. Additionally, the K-S is applicable to very small samples. If the Time 1 and Time 2 ideal rating frequencies are not significantly different, there is support for the interpretation that no *overall* interval stretching (or contracting) has occurred from Time 1 to Time 2. If, on the contrary, they are found to be significantly different, there is then reason to believe that respondent scale interval widths have stretched (or contracted).

5. While scale displacement is a simple displacement of equal magnitude across *all* items, scale interval sliding is the shifting of some, but not all, responses to a higher or lower interval category. A primary cause for such sliding or shifting in responses is categorization or discrete variable representation — that is, the representation of ordinal data as point or interval data. As Bohrnstedt and Carter [1971, p. 130] have correctly observed, researchers may assign the interval values 1, 2, 3, and 4 to the ordinal categories *definitely disagree, probably disagree, probably agree,* and *definitely agree,* but in doing so, they are assuming that the interval values being employed are monotonically related to the original underlying true (ordinal) scale. This, of course, may not be the

case. In particular, such scale transformations imply that all respondent selections are of an equal choice (i.e., all 1's are of equal value, all 2's are of equal value, and so on), whereas in reality, some choices may be closer than others to the next lower or higher interval category. This is depicted on the scale below. Although it is generally assumed that, for example, a 3 is simply a 3, respondent choices actually fall within a range of 3.00 to 3.99.



A slight response ambivalence could easily result in a sliding or shifting of adjacent answers (from 3.99 to a 4.00, for example) and a whole category value change.

To determine if *overall* scale interval sliding has occurred, it is recommended that the means of Time 1 and Time 2 ideal scores be compared. If they are not significantly different, there is support for the interpretation that no *overall* consistent sliding of calibration has occurred from Time 1 to Time 2; and vice versa.

6. To more completely verify the occurrence of either of the two forms of beta change, it is necessary to perform a final calculation. This is because, even though the means of the Time 1 and Time 2 ideal scores may not be significantly different, it is possible for item-to-item shifts or sliding to average out to produce no change in overall mean. For instance, at Time 1 a respondent may score one item a 2 and another a 4. At Time 2, the same respondent may mark the first item a 4 and the second a 2. As a consequence, the values offset one another (i.e., average out). Moreover, slight shifts to adjacent choices, for instance, from a three to a four, may not show up in overall mean calculations.

   To determine if either of the above has occurred, the correlation between Time 1 and Time 2 ideal scores should be computed as a measure of item response consistency. If the resulting correlation is weak, several explanations might be offered: (1) the questionnaire was improperly administered at either or both Time 1 or Time 2; (2) the questionnaire was mis-scored; or (3) the questionnaire items were interpreted differently at Time 1 and Time 2, suggesting the occurrence of gamma change.

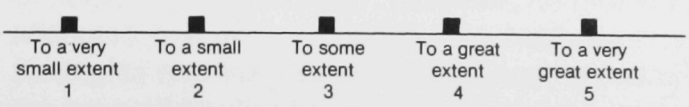
At this point, we have simply determined if there is a need to recalibrate as a result of beta change. If the Time 1 and Time 2 ideal scores are acceptably correlated, and if their rating frequencies and means *are not significantly different,* there is evidence to suggest that beta change has not occurred. Conversely, if the correlation between the Time 1 and Time 2 ideal scores is acceptable, and their rating frequencies and means *are significantly different,* beta change has most likely occurred. When it has, the questionnaire results can yet be salvaged by the transformation to be explained in Step 4.

Finally, if the Time 1 and Time 2 ideal scores *are not acceptably correlated,* regardless of whether or not their rating frequencies and means are significantly different, the value of the questionnaire for detecting alpha change is at best questionable, for the previously mentioned reasons.

Step Four

Presuming the detection of beta change, Step 4 involves the transformation of data so that the Time 1 and Time 2 actual scores will be comparable. Returning to the linear equation of regression derived earlier, we now simply insert, as the independent variable *X*, the Time 1 actual scale scores into the equation itself to compute an adjusted or recalibrated Time 1 actual score *(Y')*. Then it is a simple matter to test the adjusted Time 1 actual score against the Time 2 actual score to determine the degree of real or alpha change.

Two final methodological points deserve mention. First, in the process described above, we have chosen to adjust the Time 1 actual scores to be comparable to their Time 2 counterpart scores. Given the theoretical base, the opposite could have just as easily been performed. Second, in instances, such as that below, where scale descriptors are not stated as absolutes, it is entirely possible to derive an adjusted score of less than 1 or more than 5.



That is to say, since the scale above does not run from 0 = "absolutely none" to 5 + ∞ = "ultimately extremely all," we are consequently dealing with unbounded data. Adjusted actual scores that fall beyond five and below one are thus conceivable.

Conclusion

Our purpose has been to present a statistical technique for the detection and measurement of beta change. To the extent that the procedure corrects for such confounding, certain immediate and obvious advantages can be expected. As suggested earlier, the technique developed should be of value in differentiating and identifying the types of planned change observed in organizational settings. Furthermore, it should allow researchers to rely much more on the validity of their findings. To ignore the question of the accurate measurement of change is potentially too costly, for both organizational clients and the behavioral sciences in general.

REFERENCES

Barr, A.J.; Goodnight, J.H.; Sall, J.; & Helwig, J.T. *A user's guide to SAS 76.* Raleigh, N.C.: SAS Institute, 1976.

Bohrnstedt, G.W.; & Carter, T.M. Robustness in regression analysis. In H.L. Costner (Ed.), *Sociological methodology 1971.* San Francisco: Jossey-Bass, 1971.

Golembiewski, R.T.; Billingsley, K.; & Yeager, S. Measuring change and persistence in human affairs: Types of change generated by OD designs. *Journal of Applied Behavioral Science,* 1976, *12,* 133-157.

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

Howard, G.S.; Schmeck, R.R.; & Bray, J.H. Internal invalidity in studies employing self-report instruments: A suggested remedy. *Journal of Educational Measurement,* 1979, *16,* 129-135.

Mahoney, M.J. Experimental methods and outcome evaluation. *Journal of Consulting and Clinical Psychology,* 1978, *46,* 660-672.

Siegel, S. *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill, 1956.

Taylor, J.; & Bowers, D.G. *Survey of organizations.* Ann Arbor: CRUSK, Institute for Social Research, University of Michigan, 1972.

Thompson, D.A. Calibrating observer bias in time study pace estimation. *Journal of Industrial Engineering,* 1963, *14,* 345-346.

Zmud, R.W.; & Armenakis, A.A. Understanding the measurement of change. *Academy of Management Review,* 1978, *3,* 661-669.

Arthur G. Bedeian is Associate Professor of Management at Auburn University, Auburn, Alabama.

Achilles A. Armenakis is Associate Professor of Management and Director of the Auburn Technical Assistance Center, Auburn University, Auburn, Alabama.

Robert W. Gibson is Adjunct Associate Professor of Civil Engineering at Auburn University, Auburn, Alabama.

Received 10/9/79