Journal of Management
1994, Vol. 20, No. 3, 673-674

Point/Counterpoint

Difference Scores: Rationale, Formulation, and Interpretation

Moderators

Arthur G. Bedeian
Louisiana State University

David V. Day
Pennsylvania State University

Panelists

Jeffrey R. Edwards
University of Michigan

John Tisak and Carlla S. Smith
Bowling Green State University

Opening Statement

Difference scores have long been of interest in management research because of the need to assess agreement or congruence between constructs. For example, some form of difference, dissimilarity, or discrepancy score is typically at the core of investigations contrasting subordinate and supervisor work values (e.g., Posner, Kouzes & Schmidt, 1985), perceived and desired job characteristics or attributes (e.g., Dawis & Lofquist, 1984), job demands and worker abilities (French, Caplan & Harrison, 1982), and self and referent other rewards (e.g., Goodman, 1977). In such instances, difference scores are typically used as correlates or predictors of individual or organizational outcomes, such as changes in personal health and adjustment, and job-related attitudes and performance.

Despite decades of use, the appropriateness of difference scores for estimating discrepancies between measurement units continues to be a source of much debate. Some have advised that the statistical and psychometric properties of difference scores are so problematic that their use should be discontinued (e.g., Cronbach & Furby, 1970; Johns, 1981). Others have concluded that such criticisms are unfounded, declaring difference scores to be both relatively reliable and unbiased (e.g. Rovine, in press; Zimmerman, Brotohusodo & Williams, 1981). More recently, Smith and Tisak (1993) and Edwards (1994) have each commented—independent of the other's work—on the adequacy and appropriateness of using difference scores in management research, and arrived at opposite conclusions.

Direct all correspondence to: Arthur G. Bedeian, Louisiana State University, Department of Management, Baton Rouge, LA 70803, e-mail: MGBEDE @LSUVM. See acknowledgment.

Such contrasting positions are not unique to the management discipline. Cronbach (1958) has suggested that methodological developments in all disciplines proceed through four stages. In the initial stage, a new method is introduced with wide approbation and little criticism. The second stage sees the introduction of small improvements in the method to address recognized weaknesses. At the third stage, it is generally acknowledged that the method is imperfect with limitations. In the final stage, discussion advances to a higher level of thinking. The method's intended purpose and admitted limitations are restated more rigorously, clarifying previous confusion.

The ongoing debate concerning difference scores suggests that their methodological development may be well into Cronbach's third stage. Accordingly, our principal purpose in the present exchange was to bring together the recent proponents of these divergent perspectives in a give-and-take forum to identify primary issues of disagreement, clarify open issues, and ascertain areas for future work. Although it is possible that the ensuing discussion may prompt a return to a previous stage, we hope this intellectual interchange will advance our knowledge of difference scores toward Cronbach's fourth stage of thinking.

The following discussion is organized as a quasi-formal debate (somewhere between the formal, single perspective article or position paper, and the relatively informal and dynamic symposium), with authors having the opportunity to respond to each others' points. This type of dialectic is (in our opinion) sorely lacking in the management discipline, and we hope will provide the reader with a synthesis of apparently contradictory views.

John Tisak and Carlla S. Smith first outline their position regarding difference scores, provide background information, and address previously noted measurement concerns with difference scores. They conclude by proposing alternative approaches to conceptualizing difference scores, but that maintain the scores' essential character. Jeffrey R. Edwards then responds to points raised by Tisak and Smith, argues the merits of his regression-based procedure for testing response surfaces, and proposes that the use of difference scores should be abandoned. Tisak and Smith conclude with a final commentary on Edwards's proposal.

—Arthur G. Bedeian
Louisiana State University

—David V. Day
Pennsylvania State University

Journal of Management
1994, Vol. 20, No. 3, 675-682

Defending and Extending Difference Score Methods

John Tisak
Carlla S. Smith
Bowling Green State University

We define difference scores as the difference between distinct but conceptually linked constructs. This definition should not be confused with change scores, or the difference between a single construct measured at two or more points in time.

In the disciplines of education and human development, the attack against difference scores has stemmed from their use for assessing change on multiple measurements of some within-person characteristic (e.g., changes in abilities or skills) over time, usually in response to some type of treatment. Critics note that these change or difference scores must have some variability to function as good predictors (or outcomes), which they often do not, and that they frequently correlate with the initial level of the characteristic measured. As a consequence of these problems, several researchers (e.g., Cronbach & Furby, 1970; Lord, 1958; Werts & Linn, 1970) suggest that difference measures should be abandoned in favor of other techniques, such as residualized gain scores and regression-based estimates of change (Cronbach & Furby, 1970). Other researchers (e.g., Rogosa, Brandt & Zimowski, 1982; Rogosa & Willett, 1983; Zimmerman, Brotohusodo & Williams, 1981), however, disagree with this position, claiming that difference scores provide unique information on intraindividual change and should not be dismissed simply because they may not always be useful.

The historical arguments against difference scores that have arisen in educational and developmental research, however, often do not directly translate to management research. For example, there are notable distinctions between the difference scores criticized by psychometricians and the difference scores used by organizational researchers. Traditional psychometric arguments have mostly concerned change scores, or scores on identical variables over time. These measures are usually single pre and post scores collected from individual subjects. The difference scores collected by organizational researchers are often composite (multiple item), multiple source measures collected at a single point in time. Many of the measurement concerns about single item, single source,

Direct all correspondence to: John Tisak, Bowling Green State University, Department of Psychology, Bowling Green, OH 43403-0028.

longitudinal data are not as relevant to multiple item, multiple source, cross-sectional data. As further evidence of their utility, differences among measures are implicit in our commonly used statistical procedures, such as analysis of covariance and repeated measures analysis of variance. Therefore, difference scores in general are certainly useful and acceptable measures.

Most studies that have used some type of difference representation have operationalized difference scores by combining two or more measures into a single index. For example, the most common bivariate indices of agreement are the algebraic, absolute, and squared difference measures. The algebraic difference index is the algebraic difference between two measures $(X - Y)$; the absolute difference index is simply the absolute difference between two measures $(|X - Y|)$; and the squared difference index is the squared difference between two measures $(X - Y)^2$. Although different types of difference scores may yield different patterns of results, the selection of a specific type, as far as we can discern, has typically not been based on any identifiable, objective criteria.

Whereas not as widely used as the simpler types of difference measures, the more complex profile similarity indices are often preferred because they consider profile (i.e., dimension) level, dispersion, and shape, whereas simpler indices consider only level. Furthermore, as we will discuss later, they also ameliorate some of the traditional criticisms (e.g., reliability and model evaluation) of difference scores. The most commonly used profile similarity indices are the sum of absolute differences $(\Sigma|X_i - Y_i|)$; the sum of squared differences $(\Sigma(X_i-Y_i)^2)$; the square root of the sum of squared differences $(\Sigma(X_i-Y_i)^2)^{1/2}$; and the correlation between profiles of the component variables (r_{xy}, where the correlation is calculated between entities, e.g., respondent, rather than between measures).

Regardless of type, difference scores have been roundly criticized. For example, Cronbach (1958) argued against the use of profile similarity measures in person perception research. Johns (1981) admonished researchers for using any type of simple difference or profile similarity measure. More recently, Edwards critically examined several types of difference and profile similarity measures specifically within the theoretical framework of the person-environment fit model of stress (Edwards & Cooper, 1990) and organizational behavior research in general (Edwards, in press). All of these authors raised several issues related to the use of difference scores, although their primary concerns involved either reliability, or other measurement problems (e.g., validity, model evaluation).

Whereas a detailed account of the measurement issues is beyond the scope of this paper, we hope to stimulate critical thinking about difference measures. Further, we would suggest that researchers proceed in a manner analogous to that of Rogosa, Brandt and Zimowski (1982) in their examination of the criticisms of change measures (cf. Cronbach & Furby, 1970). Specifically, we next summarize the major concerns of difference score critics and then suggest approaches that should respond to some of those criticisms.

Measurement Concerns with Difference Scores

Reliability

The most consistent criticism against difference measures has been their presumed unreliability, particularly in relation to their component variables (i.e., X and Y; Cronbach, 1958; Johns, 1981). The reliability of a difference score is defined as the proportion of true score variance to observed score variance. This relationship can be expressed as a function of the reliability of its component variables (X and Y), the component variances (σ^2_x and σ^2_y), the correlation between the components (ρ_{xy}), and the reliability for each component ($\rho_{xx'}$, $\rho_{yy'}$).

$$\rho_{\text{diff}} = \sigma_x^2\rho_{xx'} + \sigma_y^2\rho_{yy'} - 2\rho_{xy}\sigma_x\sigma_y \;/\; \sigma_x^2 + \sigma_y^2 - 2\rho_{xy}\sigma_x\sigma_y \qquad (1)$$

From this formula, it is obvious that the reliability of a difference score may be less than the average reliability of its component variables, particularly if the component variables are positively correlated. The reliability of a difference score equals the average of its component reliabilities only if the correlation between component variables is zero. In the unlikely event that the components are negatively correlated, however, the reliability of the difference score is magnified. The presumed unreliability of difference scores has arisen because the component variables are often at least moderately positively correlated. If the components are generated by a single source (i.e., within subject), an even larger positive component intercorrelation is expected. In situations where within-subject self-report measures are used (as component variables) and the measures themselves are not very reliable, the reliability of their difference would probably be quite low. Of course, if the components are reliable or not highly positively correlated, the reliability of their difference may be quite acceptable.

As discussed previously, OB researchers frequently use composite, multiple-source measures collected at a single time. However, many of the reliability concerns about single item, single source, longitudinal data are not as relevant to multiple item, multiple source, cross-sectional data. For example, when a difference measure is determined across multiple components, the lower bound on reliability for that measure may be empirically assessed with coefficient alpha (see Smith & Tisak, 1993). Our point is that difference scores are not inherently unreliable, only that, in some circumstances (e.g., unreliable and highly intercorrelated component variables), they may prove to be unreliable. Further, by increasing the number of items, the reliability may also be increased. Unfortunately, the presumption of unreliability has often been made without empirical verification.

Validity

Other technical criticisms of (most types of) difference scores related to their validity are that they cannot be unambiguously interpreted, confound the effects of their component variables, and do not explain variance beyond that associated with their components (see Edwards, in press; Johns, 1981). First, the criticism that difference scores cannot be unambiguously interpreted refers to the fact that a difference will primarily reflect that component with the larger variance, and therefore will not represent equal but opposite contributions of each component. Second, the allegation that difference scores confound the effects of their components means that difference scores conceal the relative contributions of their components, particularly when explained variance is mostly attributable to one component. Third, critics have maintained that difference scores often do not explain variance beyond that of their components, and, in fact, are not very useful because they cannot explain more variance than both of their components jointly.

We maintain that all three criticisms are basically empirical questions that can be addressed within the context of data. For example, the effects of difference scores, as well as the relative effects of their component variables, can easily be assessed within a regression framework. We disagree with the position that difference scores are not useful because they cannot explain more variance than their components jointly. Beyond some pre-determined level of statistical significance or practical meaningfulness, the relative usefulness of difference scores is a value judgment that researchers must determine on a situational basis. These criticisms are reasons why differences may not function as expected, but *not* reasons to abandon them *a priori*!

Critics have also alleged that users of difference scores have paid inadequate attention to issues of construct validity, or the meaningfulness of difference scores. Johns (1981) maintained that complex measures, such as difference scores, are often not scrutinized as carefully as the simpler measures, and that, in fact, difference scores are usually presumed to be valid measures simply because their components are valid. He also speculated that difference scores may not be as theoretically relevant as their component variables, which, in their present or altered form, may be substitutable for the difference score. Edwards and Cooper (1990) extended these arguments in their discussion of problems associated with the person-environment fit approach to stress. They presented the different forms of fit (congruence) that have been used in p-e fit research: discrepancy or difference scores, interaction or multiplicative scores, and proportional scores. They further asserted that these different forms of fit represent different theoretical perspectives (see Edwards & Cooper, 1990, for details). This fact, according to Edwards and Cooper, has been largely ignored by p-e fit researchers, who have often applied the different forms of fit in a seemingly cavalier manner.

Although our interest here lies only in difference scores, we concur with the gist of Edwards and Cooper's arguments. In general, researchers who have used difference scores have not conceptually attended to the difference scores they selected. For example, in some situations, the signs or direction of the difference may be of great theoretical importance; in other situations, the size of the difference, regardless of direction, may be the primary concern. Researchers should address these issues with more care than has been apparent in much previous research, in which the selection of a difference measure often appeared to be predicated on the researcher's whim or the easiest analytical approach.

We hasten to add, however, that evidence of construct validity is obviously not divorced from empirical scrutiny. After a careful a priori determination of the appropriate type of difference score needed to answer a particular research question, a researcher should empirically assess the incremental validity of the difference score beyond the effects of its components within a nomological network. Unlike Johns (1981), we do not believe that difference scores are often theoretically weaker than and/or similar to their component variables. Compared to their component variables, difference scores may be more or less meaningful constructs; however, they certainly capture something conceptually different (also see, Smith & Tisak, 1993). For example, an assessment of the importance a worker attaches to various dimensions of his job is *conceptually* quite different from a comparison between the worker's assessment of the importance ratings and a supervisor's assessment of the importance of the same dimensions. The combination of a thoughtful selection of difference scores and a well-guided empirical assessment should provide evidence of a difference score's validity.

Other Approaches: Simplistic vs. Complex Response Functions

We previously mentioned some proposed statistical alternatives to traditional difference scores (e.g., residualized gain scores) that have been generally criticized because they do not capture the essence of the difference score construct (e.g., Rogosa & Willett, 1983). Recently, Edwards and his colleagues (Edwards, in press; Edwards & Cooper, 1990) added their criticisms of difference scores to those of earlier researchers and proposed a new method for the study of congruence. Edwards assumes that the relationship between difference (X − Y) and outcome (Z) measures should be considered in three dimensions (X, Y, and Z) and viewed as a three-dimensional response surface (Box & Draper, 1987) relating component measures to the outcome (in contrast to the two-dimensional function relating the difference score to the outcome). He astutely points out that most of the difference measures used in organizational (and other) research may be depicted as special cases of a more elaborate (three dimensional) response surface (Box & Draper, 1987). That is, a traditional difference score model is:

$$Z = \beta_0 + \beta_1 (X_1 - Y_1) + E, \qquad (2)$$

where E is the error or residual term and where the βs represent unstandardized population regression coefficients. When expanded, Equation 2 becomes:

$$Z = \beta_0 + \beta_1 X_1 - \beta_1 Y_1 + E, \qquad (3)$$

which is a specific case of the more general response surface model:

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 Y_1 + E, \tag{4}$$

when certain constraints are placed on the general response surface model (i.e., $\beta_1 = -\beta_2$).

Furthermore, in practice, Edwards contends that highly restricted difference score models are rarely evaluated against more general response surface models. Although we agree with these issues, we believe that it is very unclear whether only two approaches exist: the simplistic difference score model and the response surface model.

In this section, we suggest models that maintain the notion of a difference between conceptually linked constructs and those that have a multidimensional parameter space (i.e., contain more than one parameter). To make this discussion very concrete, let us consider two constructs, each of which consists of three homogeneous items. Specifically, consider the construct of Role Conflict (Rizzo, House & Lirtzman, 1970) with components or items such as "I have to do things that should be done differently" (X_1), "I work with two or more groups who operate quite differently" (X_2), and "I work on unnecessary things" (X_3); and the construct of Role Ambiguity (Rizzo, House & Lirtzman, 1970) with items such as "I have clear planned goals and objectives in my job" (V_1), "I know that I have divided my time properly" (V_2), and "I know what my responsibilities are" (V_3). Notice that we have labeled Role Conflict and Role Ambiguity as X and V, respectively, when they represent the employee's response and Y and W, respectively, when they represent the supervisor's evaluation of the employee. The outcome variate, Job Satisfaction, is labeled Z.

Further, for this illustration, let us initially assume a squared difference between the first component of Role Conflict for employees and their supervisors (i.e., between X_1 and Y_1). A common difference model would be:

$$Z = \beta_0 + \beta_1 (X_1 - Y_1)^2 + E, \tag{5}$$

where E is the error or residual term. The quadratic response surface for this example would be given by:

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 Y_1 + \beta_3 X_1 Y_1 + \beta_4 X^2{}_1 + \beta_5 Y^2{}_1 + E \tag{6}$$

Clearly, Equation 6 is the same as Equation 5 if $\beta_1 = \beta_2 = 0$ (i.e., the individual component measures, X_1, and Y_2, are not present), $\beta_4 = \beta_5$, (i.e., the coefficients of the squared components, $X^2{}_1$ and $Y^2{}_1$, are equal), and $\beta_3 = -2\beta_4$ (i.e., the weight of the interaction term, X_1Y_1, is equal to twice the coefficient of either squared components and opposite in sign). Since Equation 5 is nested in Equation 6, Edwards (in press) recommends statistically testing the hypothesis $H_0 = \beta_1 = \beta_2 = 0$, $\beta_4 = \beta_5$; and $\beta_3 = -2\beta_4$, as a way of evaluating the meaningfulness of the difference model against the more general model.

We feel that such a procedure would be inherently unfair to difference measures because, as currently formulated, they contain only one parameter (i.e., β_1 in Equation 5). Instead, before discarding difference models of the form Equation 5 in favor of the more general models in Equation 6, we suggest that researchers first consider some generalizations of the difference model. Specifically, Equation 5 may be easily generalized to:

$$Z = \beta_0 + \beta_1 (X_1 - Y_1) + \beta_2 (X_1 - Y_1)^2 + E \qquad (7)$$

Notice that unlike Eqation 6, Equation 7 still maintains the idea of a difference between the component of the Role Conflict between employees and supervisors. However, (7) permits the individual components, X and Y , to be represented. Furthermore, it also allows for a signed or directional difference. Clearly, the models are nested, with Equation 5 within 7 and 7 within 6, so that relative evaluations or statistical tests of these models are possible.

Continuing our illustration, if we have available multiple components, X_1, X_2, X_3, and Y_1, Y_2, Y_3, respectively, of Role Conflict for employees and supervisors (i.e., we are considering profile differences), then Equation 7 may be generalized to include the information, or:

$$Z = \beta_0 + \beta_1 \Sigma(X_i - Y_i) + \beta_2 \Sigma(X_i - Y_i)^2 + E \qquad (8)$$

Although Equation 8 has not increased the number of parameters over Equation 7, as discussed earlier, it has allowed the researcher to improve the reliability of each of the difference measures, $(X_i - Y_i)$ and $(X_i - Y_i)^2$. Notice that this increase in reliability is analogous to the increase in reliability when moving from component to composite reliability in scale construction.

Finally, if one has the two constructs, Role Conflict and Role Ambiguity, each with the multiple components indicated at the beginning of this example, then (8) may be generalized once again to:

$$Z = \beta_0 + \beta_1 \Sigma(X_i - Y_i) + \beta_2 \Sigma(X_i - Y_i)^2 + \beta_3 \Sigma(V_i - W_i) + \beta_4 \Sigma(V_i - W_i)^2 + E \qquad (9)$$

Again, notice that Equation 9 is more general than Equations 5, 7, or 8, but not nearly as general as the corresponding response surface model. Also, Equation 9 permits the use of heterogeneous sets of homogeneous items. For example, the component measures are homogeneous with Role Conflict and Role Ambiguity, but these two constructs are distinct (heterogeneous). Further, Equation 9 generalizes in an obvious manner to additional constructs, with each containing multiple components. Note that there are similar developments for other difference functions (e.g., the absolute difference measure).

In summary, whereas we recognize the contribution of Edwards's work for testing difference models versus response surfaces, we suggest that researchers first evaluate some generalizations of the difference model before discarding it. Although we have provided specific examples, we suggest that researchers also consider other functions, such as generalizations of the absolute difference, which may be more appropriate for their needs. Clearly, one parameter difference functions are extremely restrictive. However, before we discard this (potentially) theoretically rich concept, more complex difference score functions should be investigated. Finally, we issue the caveat that the generalized difference models place additional burdens on the researcher in that he or she must now consider and justify these more complex difference models (i.e., be concerned about the additional terms and the nature of the difference).

Journal of Management
1994, Vol. 20, No. 3, 683-689

Regression Analysis as an Alternative to Difference Scores

Jeffrey R. Edwards
University of Michigan

For nearly 40 years, it has been asserted (see, e.g. Cronbach, 1958, 1992; Cronbach & Furby, 1970; Cronbach & Gleser, 1953; Edwards, 1994; Edwards & Cooper, 1990; Johns, 1981; Wall & Payne, 1973; Werts & Linn, 1970) that difference scores suffer from various methodological problems (about which more anon). In their position statement, Tisak and Smith argue that some of these problems have been overstated and suggest alternative procedures (e.g., the expanded difference equation) intended to overcome certain problems while maintaining the use of difference scores. Although certain points made by Tisak and Smith have merit, they minimize or overlook several important problems with difference scores, and their recommended procedures fail to overcome these problems. I will elaborate my position according to the two primary issues addressed by Tisak and Smith, the reliability and validity of difference scores. I will then note shortcomings with the Tisak and Smith procedure and contend that the regression procedure described by Edwards (1994) mitigates or avoids arguable problems with difference scores, but permits comprehensive tests of conceptual models that difference scores are intended to represent.

Reliability

In defense of the reliability of difference scores, Tisak and Smith argue that difference scores are not inherently unreliable, but may prove unreliable when the component measures comprising the difference are unreliable and positively correlated. Tisak and Smith also point out that, unlike bivariate difference scores, profile similarity indices are often based on composite (multi-item) multiple source measures. Because of this, profile similarity indices are likely to yield higher reliability estimates than bivariate difference scores.

As pointed out by Tisak and Smith, it is undeniable that the reliability of any measure is ultimately an empirical question that should be addressed on a study-by-study basis. However, the primary message of Johns (1981) and others is that the conditions under which difference scores are unreliable (i.e., positively correlated component measures with modest reliabilities) are common in empirical research. This is not surprising, given that difference score

Direct all correspondence to: Jeffrey R. Edwards, University of Michigan, School of Business Administration, Ann Arbor, MI 4809-1234.

components are usually measured with the same instrument and often represent constructs that should be positively correlated on conceptual grounds. For example, Schneider (1987) argues that people gravitate toward work settings that are similar to themselves, thereby generating a positive correlation between measures of the person and job. Because of this, it is reasonable to assert *a priori* that difference scores may well exhibit poor reliabilities. Furthermore, the reliability of a difference score should be evaluated not only in an absolute sense, but also in relation to viable alternatives, such as using both component measures jointly in multiple regression analysis (Edwards, 1994; Edwards & Cooper, 1990). If a difference score exhibits adequate reliability, then it is almost certain that its component measures will exhibit superior reliabilities, indicating that the latter should be used in place of the former (Edwards, 1994).

Unlike bivariate difference scores, profile similarity indices (e.g., D^2) will often exhibit reliabilities that are substantially larger than their component measures. This is due in part to the number of dimensions involved in the calculation of the index, which has a dramatic impact on its estimated reliability (Nunnally, 1978). For example, if 10 squared differences exhibiting reliabilities of .50 and intercorrelations of .10 were standardized and summed to form D^2, the reliability of the resulting index would be .74. Studies using profile similarity indices (e.g., Caldwell & O'Reilly, 1990; Chatman, 1991; Dougherty & Pritchard, 1985; Rounds, Dawis & Lofquist, 1987) often incorporate a much larger number of dimensions, virtually guaranteeing that the index will demonstrate high reliability.

Although profile similarity indices may yield high reliability estimates, the interpretation of these estimates can be problematic. Reliability is typically defined as the proportion of true score variance in a measure, or the squared correlation between a measure and its associated underlying construct (Lord & Novick, 1968; Nunnally, 1978). Unless the items comprising a measure share a common meaning, it is difficult to define the construct underlying the measure, and the interpretation of the reliability of the measure therefore becomes suspect (Gerbing & Anderson, 1988; Hattie, 1985; Wolins, 1982). In my experience, the items typically comprising profile similarity indices represent conceptually distinct dimensions and, hence, do not share a common meaning. For example, dimensions measured by Chatman (1991) included aggressiveness, risk taking, precision, and social responsibility, those measured by Dougherty and Pritchard (1985) included making presentations, keeping records, and providing written advice to clients, and those measured by Smith and Tisak (1993) included data entry, obtaining information from clients, and interpreting company policies and procedures. In these cases, it seems difficult to define a construct that encompasses such diverse dimensions. Although it may be argued that indices that combine diverse dimensions represent similarity in a global sense, Cronbach and Gleser (1953) and Lykken (1956) have forcefully argued that similarity is meaningful only in terms of specific dimensions, not as a general quality. Without a clear definition of the construct underlying a profile similarity index, the concept of a "true score" is meaningless, and the reliability of the index becomes moot.

Validity

Tisak and Smith acknowledge several problems pertaining to the validity of difference scores, such as ambiguous interpretation, confounding the effects of their component measures, and failure to explain variance beyond their component measures. Nonetheless, they assert that these problems do not provide sufficient justification to abandon difference scores *a priori,* arguing that the severity of each problem should be assessed empirically within the context of the data. Tisak and Smith further argue that, even when evidence for these problems is found (e.g., a difference score explains less variance than its component measures), the utility of difference scores remains a value judgment for the researcher.

Tisak and Smith are correct in pointing out that the severity of problems regarding the validity of difference scores can be assessed empirically. For example, the degree to which an algebraic difference explains less variance than its components can be assessed by comparing the R^2 from Equation 4 to that obtained from Equation 2, using a conventional F-test (Edwards, 1994). If Equation 4 explains significantly more variance than Equation 2, then the functional form associated with the algebraic difference (i.e., equal but opposite effects for the two component measures) is rejected, and the form indicated by Equation 4 should be preferred. If Equation 4 does not explain significantly more variance than Equation 2, then the functional form for the algebraic difference may be considered tenable (in both cases, it is also necessary to ensure that the overall R^2 is significant and no significant higher-order terms are found, thereby establishing that a linear equation adequately represents the functional form relating the component measures to the outcome; see Edwards, 1994). In neither case is it necessary or desirable to resort to Equation 2 once Equation 4 has been estimated. Moreover, the F-test comparing the R^2 values from Equations 2 and 4 can be replaced by a direct test of whether β_1 and β_2 in Equation 4 are equal in magnitude but opposite in sign (Cohen & Cohen, 1983, pp. 479-480), which makes Equation 2 superfluous (for analogous tests pertaining to absolute and squared difference scores, see Edwards, 1994).

The use of Equation 4 also avoids other problems regarding the validity of algebraic difference scores. For example, the interpretational ambiguity created by combining the component measures into a single composite is eliminated, given that the component measures are used as separate predictors. In addition, the effects of the component measures are no longer confounded, because separate coefficients are obtained for each measure. Of course, these advantages also pertain when Equation 6 is used in place of Equation 5, or when the piecewise linear equation described by Edwards (1994) is used instead of an absolute difference.

Tisak and Smith also attempt to bolster the validity of difference scores by arguing that they capture something distinct from their component measures. However, because difference scores are simply composites of their component measures, they cannot contain information beyond that available when these measures are considered *jointly* (Johns, 1981). Furthermore, as shown by comparing regression equations using difference scores (e.g., Equations 2 and 5) to their unconstrained counterparts (Equations 4 and 6, respectively), the former equations are simply special cases of the latter. Because of this, it is logically impossible for equations using difference scores as predictors to capture anything beyond that represented by equations using difference score components. Moreover, equations using difference score components can capture theoretically meaningful effects that cannot be detected when equations relying on difference scores are used (for examples, see Edwards, 1994; Edwards & Harrison, 1993).

The Viability of the Tisak and Smith Procedure

Tisak and Smith contend that tests comparing constrained regression equations using difference scores (e.g., Equation 5) to their unconstrained counterparts (e.g., Equation 6) are "inherently unfair," given that difference score equations contain only one parameter. As an alternative, they propose a generalized difference equation, Equation 7, that uses an algebraic and a squared difference as predictors.

There are two fundamental problems with the generalized difference equation proposed by Tisak and Smith. First, beyond the argument that it "maintains the idea of a difference between the components," there is no apparent conceptual justification for Equation 7. The central issue in testing the effects of congruence (i.e., fit, similarity, or agreement) is not whether a difference score is used in the equation, but whether the functional form relating the component measures to the outcome is consistent with that represented by the difference score. This cannot be determined by merely inserting a difference score into the equation, because a significant coefficient on a difference score can be generated by a substantial variety of functional forms, only one of which is consistent with that represented by the difference score itself (for examples of this, see Edwards, 1994; Edwards & Harrison, 1993). Further inspection of Equation 7 reveals that it is conceptually similar to Equation 5, but can depict minima at locations other than the point where X and Y are equal (specifically, if β_1 in Equation 7 is positive, the minimum is shifted to the region where $X < Y$, whereas if β_1 is negative, the minimum is shifted to the region where $X > Y$).

Second, when compared to Equation 5, Equation 7 simply replaces one set of constraints on Equation 6 with another (for the ensuing discussion, it is assumed that all coefficients in Equation 6 are estimated simultaneously). In particular, Equations 5 and 7 both impose the constraints $\beta_4 = \beta_5$ and $\beta_3 = -2\beta_4$. However, whereas Equation 5 constrains $\beta_1 = \beta_2 = 0$, Equation 7 constrains $\beta_1 = -\beta_2$. To test the constraints imposed by Equation 7, it is necessary to estimate Equation 6 and test the increment in R^2 yielded by Equation 6 over Equation 7 or, equivalently, directly test whether the coefficients from Equation 6 follow the pattern corresponding to Equation 7 (Dwyer, 1983). If the constraints imposed by Equation 7 are rejected and the set of cubic terms composed of X_1 and Y_1 is not significant (Edwards, 1994), then interpretation should focus on Equation 6, using procedures described by

Edwards and Parry (1993). If the constraints are not rejected, then the functional form corresponding to Equation 7 may be considered tenable. This, however, does not mean that Equation 7 should then be estimated, because the functional form relating the component measures to the outcome can be obtained directly from Equation 6. Furthermore, additional information that could be found by estimating Equation 7, such as its R^2 and coefficient estimates, can be calculated from the results of Equation 6, provided the constraints imposed on Equation 6 to yield Equation 7 are known (e.g., Johnston, 1984). The primary utility of Equation 7 is that it allows a researcher to construct hypotheses regarding the pattern of coefficients from Equation 6 that would yield support for the functional form corresponding to it. However, once Equation 6 has been estimated, it is unnecessary and redundant to then estimate Equation 7.

Tisak and Smith propose two generalizations of Equation 7, one using the sum of algebraic and squared differences across multiple dimensions (i.e., Equation 8), and another adding a second set of analogous summed difference measures (i.e., Equation 9). Unfortunately, Equations 8 and 9 simply compound the problems associated with Equation 7. This can be seen by considering the following equation, which is an expanded version of Equation 8:

$$\begin{aligned} Z = \beta_0 &+ \beta_1(X_1 - Y_1) + \beta_2(X_1 - Y_1)^2 \\ &+ \beta_1(X_2 - Y_2) + \beta_2(X_2 - X_2)^2 \\ &+ \beta_1(X_2 - Y_3) + \beta_2(X_3 - Y_3)^2 + e \end{aligned} \tag{10}$$

As Equation 10 shows, Equation 8 imposes the same constraints as Equation 7 on the algebraic and squared differences corresponding to each dimension. Moreover, Equation 8 constrains coefficients across dimensions, such that the coefficients on each algebraic difference are the same, and the coefficients on each squared difference are the same. Conceptually, this implies that the functional form relating each paired X_i and Y_i to the outcome is the same, regardless of the substantive distinctions among the dimensions. Obviously, such an elaborate set of constraints should be tested empirically, not simply imposed on the data. This can be accomplished using the following equation, which is a generalization of Equation 6:

$$\begin{aligned} Z = \beta_0 &+ \beta_1 X_1 + \beta_2 Y_1 + \beta_3 X_1 Y_1 + \beta_4 X_1^2 + \beta_5 Y_1^2 \\ &+ \beta_6 X_2 + \beta_7 Y_2 + \beta_8 X_2 Y_2 + \beta_9 X_2^2 + \beta_{10} Y_2^2 \\ &+ \beta_{11} X_3 + \beta_{12} Y_3 + \beta_{13} X_3 Y_3 + \beta_{14} X_3^2 + \beta_{15} Y_3^2 + e \end{aligned} \tag{11}$$

The constraints imposed by Equation 8 can be evaluated by testing the increment in R^2 yielded by Equation 11 or by directly testing whether the coefficients obtained from Equation 11 conform to the pattern associated with Equation 8. As before, once Equation 11 has been estimated, it is unnecessary to estimate Equation 8, regardless of whether the constraints imposed by Equation 8 are supported. An analogous unconstrained equation corresponding to Equation 9 can be derived and tested in a similar manner.

Estimating equations such as Equation 11 carries the obvious disadvantage of requiring large samples, particularly when the number of dimensions is large. However, the additional degrees of freedom provided by Equation 8 over 11 are obtained only by imposing constraints that are highly restrictive and, based on prior work with similar equations (Edwards, 1993), are unlikely to receive empirical support. Fortunately, this disadvantage is ameliorated when the dimensions are conceptually homogeneous, in which case the X_i and Y_i should be summed prior to analysis to form composite X and Y scales. For example, if the Role Conflict items described by Tisak and Smith represent a single underlying construct and satisfy the requirements for unidimensional measurement (Gerbing & Anderson, 1988; Hattie, 1985), then scales representing the employee's and supervisor's responses should be constructed by summing the corresponding items, and these scales should be used in Equation 6. When a larger number of dimensions is involved, as in studies using profile similarity indices (Caldwell & O'Reilly, 1990; Chatman, 1991; Dougherty & Pritchard, 1985; Rounds et al., 1987), it is likely that the dimensions can be distilled into a more parsimonious set (O'Reilly, Chatman & Caldwell, 1991) which, provided sample sizes were adequate, would permit the use of an equation such as Equation 11.

Applications of the Edwards Procedure

The preceding discussion has contended that the aforementioned methodological problems with difference scores can be mitigated or avoided by applying the regression procedure described by Edwards (1994). The merits of this procedure over difference scores is not simply a matter of intellectual debate, but has also been demonstrated empirically. For example, Edwards (1994) found that, on average, when the constraints imposed by the algebraic, absolute, and squared differences between actual and desired job attributes were relaxed, the variance explained in job satisfaction nearly tripled. Similarly, Edwards and Harrison (1993) reanalyzed data from the classic P-E fit study conducted by French, Caplan and Harrison (1982) and found that, when the constraints imposed by the difference scores used by French et al. (1982) were relaxed, the variance explained in strain more than doubled. In both studies, the unconstrained regression equations indicated three-dimensional surfaces that were theoretically meaningful but notably more complex than the simplistic two-dimensional functions corresponding to bivariate difference scores. Furthermore, results from Edwards and Harrison (1993) required modifying or abandoning many of the substantive conclusions drawn by French et al. (1982), thereby altering the theoretical implications of the study.

Is Anything Lost by Abandoning Difference Scores?

Despite the apparent advantages of the regression procedure, Tisak and Smith maintain that it is premature to abandon difference scores, arguing that "before we discard this (potentially) theoretically rich concept, more complex difference score functions should be investigated." This apparently reflects the assumption that, by abandoning difference scores, we are unable to examine theoretical questions of congruence. This assumption is mistaken. As the preceding discussion has shown, the constrained regression equations represented by difference scores are special cases of the unconstrained equations described by Edwards (1994), and any theoretically meaningful functional form depicted by the former can be fully represented by the latter. Furthermore, the unconstrained equations can depict an extensive variety of theoretically meaningful functional forms that difference scores simply cannot represent. Thus, rather than discarding the concept of congruence, the regression procedure permits more rigorous and comprehensive tests of congruence hypotheses while avoiding various problems with difference scores that have plagued this area of investigation for decades.

Journal of Management
1994, Vol. 20, No. 3, 691-694

Rejoinder to Edwards's Comments

John Tisak
Carlla S. Smith
Bowling Green State University

In his position paper, Edwards critiqued several of our comments concerning the reliability and validity of difference scores. We believe our differences of opinion occur not only because Edwards has endorsed historical arguments against difference scores, but also because he conceptualizes certain issues quite differently than we do. We address his major points of criticism and then reiterate (and perhaps clarify) our position.

Edwards assumes that it is reasonable to assert a priori that difference scores will often exhibit poor reliabilities because the conditions under which poor reliabilities can occur (i.e., unreliable and highly positively intercorrelated component measures) are very common in empirical research. Although these circumstances may be common, they should not be sufficient to condemn the use of difference scores a priori because reliability may be empirically investigated and because, as we suggested, reliabilities can be improved.

We take exception to Edwards's statements, "... the reliability of a difference score should be evaluated not only in an absolute sense, but also in relation to viable alternatives, such as using both component measures jointly in multiple regression analysis .... If a difference score exhibits adequate reliability, then it is almost certain that its components will exhibit superior reliabilities, indicating that the latter should be used in place of the former." To us, this presumes that the difference and component measures in question are conceptually interchangeable, a blanket assumption we are unwilling to make. For example, the concept of role conflict obtained from the differences between subordinate and supervisor job ratings is not the same as conceptualizations of the components of subordinate and supervisor job ratings. Also, we do not agree, given adequate difference score reliabilities, that difference scores should be discarded because their component measures show higher reliabilities. What about the theory being tested or research goals? Finally, notice that we and Edwards (1994) agree, that response surfaces do not eliminate reliability problems.

We disagree with Edwards's suggestion that the reliabilities of profile similarity measures can be "problematic" because dimensions are often formed by large numbers of heterogeneous items. Our position was never that

Direct all correspondence to: John Tisak, Bowling Green State University, Department of Psychology, Bowling Green, OH 43403-0228.

dimensions should be formed from conceptually unrelated items. In addition, we suggested (and subsequently illustrated) that researchers may sometimes want to use heterogeneous items to achieve a general dimension. (Of course, dimensionally can be assessed statistically.) Our arguments are analogous to standard scale construction procedures.

One of Edwards's criticisms concerning the validity of difference scores is that difference scores can never explain more variance than their component measures because they are simply composites of their components. Whereas we agree with Edwards's statistical assessment, we steadfastly maintain that difference scores often represent something conceptually quite distinct from their components. Whether the contribution of a particular difference measure is statistically or practically meaningful is best left to the determination of individual researchers.

Edwards advances two criticisms of our proposal to consider generalized difference functions prior to more complex functions. Specifically, he maintains that, beyond the general notion of construct differences, there is no conceptual justification for Equation 7 and that it replaces one set of constraints with another. To begin, we hold that the general notion of construct differences is quite simply the point in question; given 40 years of research that has developed a theoretical basis for difference scores, we view this point as extremely important!

Further, Edwards maintains that a significant coefficient on a difference score can be generated by a variety of functional forms. We concur; however, we do not advocate accepting unconstrained equations (i.e., full response surface models) merely because they are consistent with the functional forms found in the raw data of a specific sample, even when cross-validation procedures have been applied (e.g., Edwards & Harrison, 1993). In other words, if both unconstrained and constrained models fit the data reasonably well, we do not advocate accepting the unconstrained model simply because its fit is somewhat (or even significantly) better. Our primary concern is whether the data fit a predetermined theory rather than whether the data fit an empirical model.

As we indicated, Equation 5 is nested within Equation 7, which, in turn, is nested within Equation 6; of course, Equations 5 and 7 must be variations of constraints on Equation 6. The point here is that the set of constraints placed on Equation 6 to obtain Equation 5 is more restrictive than those constraints placed on Equation 6 to obtain Equation 7. We attempted to suggest generalizations of the highly restrictive "one parameter" difference models while permitting the potential falsification of these models against a general response surface model. Moreover, Edwards asserts that if Equation 7 is considered tenable by the appropriate testing procedure, then "... additional information that could be found by estimating Equation 7 ... can be calculated from the results of Equation 6 ..." and "... once Equation 6 has been estimated, it is unnecessary and redundant to then estimate Equation 7." If this means that there is a *simple* linear transformation that would allow one to obtain the regression coefficients and/or their associated covariance estimates of Equation 7 from those of Equation 6 by merely applying the hypothesized constraints, then the above assertions are mistaken. Although a detailed discussion is inappropriate here, at question is how the coefficients were initially determined (i.e., was a simultaneous or sequential estimation strategy utilized; cf. Rozeboom, 1966, or Tisak, 1994). In particular, the regression coefficients of Equation 6 are estimated by a simultaneous procedure; on the other hand, those for Equation 7 could best be considered sequential estimates. The fundamental issue, however, is whether one should develop models that depict predetermined theoretical concerns (i.e., the variables and parameters are of predetermined theoretical interest) or use general purpose models that attempt to fit data. Finally, Equation 7 was only meant to serve only as an example of the type of generalization that might be used.

Edwards goes on to argue that "once Equation 11 has been estimated, it is unnecessary to estimate Equation 8, regardless of whether the constraints imposed by Equation 8 are supported." We disagree because one might be interested in the estimation of the regression parameters of Equation 8 and the incremental increase in the squared multiple correlation coefficients from Equation 8 to Equation 11. These estimates are not readily available from Equation 11. Furthermore, Equation 8 provides a way of conceptualizing generalized difference measures that we believe is not readily apparent from Equation 11.

In general, we fail to understand the distinction Edwards seems to draw between difference measures and response surfaces. When we read Edwards's position paper, we noted comments such as, "The preceding discussion has contended that the aforementioned methodological problems with differences scores can be mitigated or avoided by applying the regression procedure described by Edwards" (1994). In our opinion, this statement gives the impression that two distinct techniques exist, difference functions and response surfaces. Because the former is nested in the latter, we find this distinction artificial. We maintain that this issue can only be resolved by empirical verification.

It remains to be determined if "... the unconstrained equations can depict an extensive variety of theoretically meaningful functional forms...." As far as we can determine, no current psychological theories exist to support the complex higher-order response surfaces endorsed by Edwards (1994). We interpret Edwards (1994) and Edwards and Harrison (1993) as advocating tests of theoretically derived submodels against full response surface models and, if a full model is superior, accepting it. (*N.B.*: Professor Edwards, pace Professors Tisak and Smith, abjures this interpretation, referring interested readers to Edwards, 1994, and Edwards and Harrison, 1993, for his position on this matter.) Finally, we believe too much emphasis has been given to the relatively simple one-parameter difference measure over some of the more complex difference measures we suggested. As previously indicated, the statistical cards are stacked in favor of the full models because they estimate more parameters than submodels.

We would also like to emphasize that the use of multiple dimensions, both in our generalized difference functions and particularly Edwards's complex functions, has the disadvantage of requiring large samples. With Edwards's procedure, this liability adds to the necessity of selecting and interpreting appropriate higher order terms (see Equation 11). For example, do cubic terms adequately address a particular research question such that the additional power (i.e., subjects) needed to include them in the regression equation is deemed worthwhile?

We summarize what we perceive to be the following points of disagreement between our and Edwards's perspectives:

1. There are two prevailing schools of thought on the viability of difference scores. Edwards has aligned his arguments with one school (e.g., Cronbach & Furby, 1970) whereas we have aligned our arguments with the other (e.g., Rogosa et al., 1982).
2. Whereas Edwards wants to contrast a simple difference model against a much more general response surface model, we suggest that intermediate models first need to be evaluated. Specifically, intermediate models should be developed from theory. Only if they prove inadequate, researchers may elect to explore the viability of full "response surface" models. We, therefore, are not necessarily advocating a different statistical approach than Edwards, but rather a different order in which the approach should be applied.

In conclusion, historical arguments against difference scores assume that they should be discarded because they are typically unreliable. Further, Edwards asserts that difference scores should be evaluated (and hence usually discarded) against more complex response surface models. Although we believe Edwards's approach has merit, we do not advocate such an extreme position. We attempted to address both of these issues by recommending an approach that can increase the reliability of measures and generalize the notion of differences, yet, compared to simple submodels, still allow them to be intermediate (or less restrictive) submodels of more general response surfaces.

Regardless of the position adopted, we beseech researchers to embrace a less myopic perspective while taking a more systematic approach to the use of difference scores. We believe blanket condemnation of difference scores over the years has stifled research on congruence. On balance, we do not advocate thoughtless application of any procedure (including the use of difference scores). Before investigators "throw out the baby with the bath water," or use difference scores with the abandonment obvious in much previous research, we ask that they carefully weigh the issues presented here.

Journal of Management
1994, Vol. 20, No. 3, 695-698

Concluding Statement

Arthur G. Bedeian
Louisiana State University

David V. Day
Pennsylvania State University

Restated, the main purpose for the preceding exchange was to provide a give-and-take forum for the on-going debate concerning difference scores. Primary issues of disagreement have been identified, certain open issues clarified, and recommendations for future research highlighted. Clearly, there can be no serious question that difference scores are an integral aspect of management research, but that their use is replete with conceptual and methodological considerations. It is evident from the present forum that researchers need to exercise caution prior to committing to a particular technique, whether it be a difference score measure or response surface. This is an especially important concern, because discrepancy constructs (e.g., similarity, congruence, fit) have been and most likely will continue to play a central role management research.

By articulating some of the basic philosophical and statistical assumptions underlying their divergent perspectives, our panelists have revealed both the advantages and disadvantages of alternative difference score representations. From our perspective, three unresolved general issues are at the root of the panelists' disagreement. First, are difference scores conceptually distinct from their components? If researchers work from the Tisak and Smith assumption that difference scores are, indeed, conceptually distinct, then additional construct validity evidence should be examined. After all, if difference scores have a meaning that is distinguishable from their individual components, then they must establish their own nomological net. Second, are intermediate submodels appropriate? Edwards maintains that his approach incorporates the submodels proposed by Tisak and Smith. The latter authors, however, argue that a "fairer" test of difference scores would be to first compare these intermediate models against more complex response surfaces. Space limitations preclude the possibility of empirically testing these various alternatives in the present forum. Moreover, it is unlikely that one test would yield generalizable conclusions. Third, do present theories fit the higher-order response surfaces recommended by Edwards? If not, then rejecting basic difference-score models in favor of more complex response surfaces may be letting the empirical tail wag the theoretical dog. This, however, places future researchers in the unusual

Direct all correspondence to: Arthur G. Bedeian, Louisiana State University, Department of Management, Baton Rouge, LA 70803, e-mail: MGBEDE@LSUVM. See acknowledgment.

position of trying to formulate more complex theories to account for the empirical relations that a particular technique has suggested. The general message here seems to be that rather than routinely applying a prototypic technique, the research hypotheses being tested and the theory on which they are based should dictate the analytic procedure to be used. The role of theory as a precondition for selecting an analytic strategy for testing research hypotheses has been recently emphasized by Schoorman, Bobko, and Rentsch (1991).

Burr and Nesselroade (1990) note that, as scientists, researchers are "in the business of drawing inferences from the data at hand to the larger, not-ever-to-be-fully-explored data box of nature" (p. 29). They further observe that the means for exploring such data lie in the continued evolution of statistical tools. We believe that this evolution can be furthered through the interaction of alternative and even conflicting beliefs. Hence, the present forum.

At the same time, we recognize there can be no final knowledge, only temporary suggestions and techniques that are subsequently overturned by more adequate, but still necessarily inconclusive suggestions and techniques. Thus, the road to further knowledge will always be open. It is hoped that this forum has taken a few steps along this road in the direction of Cronbach's (1958) fourth stage of methodological development. The fact that it has perhaps raised as many questions as it has answered is a reminder that the road to knowledge is an endless avenue of continuing exploration and high adventure. Perhaps this is why the search for knowledge is so much fun.

Acknowledgment: Correspondence concerning this manuscript and requests for reprints can be addressed to the authors as follows: Arthur G. Bedeian, Department of Management, Louisiana State University, Baton Rouge, LA 70803-6312, e-mail: MGBEDE @ LSUVM; David V. Day, Department of Psychology, Pennsylvania State University, University Park, PA 16802-3104, e-mail: DVD1 @ PSUVM; Jeffrey R. Edwards, School of Business Administration, University of Michigan, Ann Arbor, MI 48109-1234, e-mail: JEFF.EDWARDS @ UM.CC.UMICH.EDU; John Tisak, Department of Psychology, Bowling Green State University, Bowling Green, OH 43403-0228, e-mail: JTISAK @ TRAPPER; Carlla S. Smith, Department of Psychology, Bowling Green State University, Bowling Green, OH 43403-0228, e-mail: CSMITH @ TRAPPER. The names of the first three authors are listed alphabetically.

References

Box, G. E. P. & Draper, N. R. (1987). *Empirical model building and response surfaces*. New York: Wiley.

Burr, J. A. & Nesselroade, J. R. (1990). Change measurement. Pp. 3-34 in A. von Eye (Ed.), *Statistical methods in longitudinal research*, Vol. 2. New York: Academic Press.

Caldwell, D. F. & O'Reilly, C. A. (1990). Measuring person-job fit with a profile comparison process. *Journal of Applied Psychology, 75*: 648-657.

Chatman, J. A. (1991). Matching people and organizations: Selection and socialization in public accounting firms. *Administrative Science Quarterly, 36*: 459-494.

Cohen, J. & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral science*, 2nd ed. Hillsdale, NJ: Erlbaum.

Cronbach, L. J. (1958). Proposals leading to analytic treatment of social perception scores. Pp. 353-379 in R. Tagiuri & L. Petrullo (Eds.), *Person perception and interpersonal behavior*. Stanford, CA: Stanford University Press.

———. (1992). Four *Psychological Bulletin* articles in perspective. *Psychological Bulletin, 112*: 389-392.

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

Cronbach, L. J. & Gleser, G. C. (1953). Assessing the similarity between profiles. *Psychological Bulletin, 50*: 456-473.

Dawis, R. V. & Lofquist, L. H. (1984). *A psychological theory of work adjustment*. Minneapolis: University of Minnesota Press.

Dougherty, T. W. & Pritchard, R. D. (1985). The measurement of role variables: Exploratory examination of a new approach. *Organizational Behavior and Human Decision Processes, 35*: 141-155.

Dwyer, J. H. (1983). *Statistical models for the social and behavioral sciences*. New York: Oxford University

Edwards, J. R. (1993). Problems with the use of profile similarity indices in the study of congruence in organizational research. *Personnel Psychology, 46*: 641-665.

———. (1994). The study of congruence in organizational behavior research: Critique and a proposed alternative. *Organizational Behavior and Human Decision Processes, 58*: 51-100.

Edwards, J. R. & Cooper, C. L. (1990). The person-environment fit approach to stress: Recurring problems and some suggested solutions. *Journal of Organizational Behavior, 11*: 293-307.

Edwards, J. R. & Harrison, R. V. (1993). Job demands and worker health: A three-dimensional reexamination of the relationship between person-environment fit and strain. *Journal of Applied Psychology, 78*: 628-648.

Edwards, J. R. & Parry, M. E. (1993). On the use of polynomial regression equations as an alternative to difference scores in organizational research. *Academy of Management Journal, 36*: 1577-1613.

French, J. R. P., Caplan, R. D. & Harrison, R. V. (1982). *The mechanisms of job stress and strain*. Chichester: Wiley.

Gerbing, D. W. & Anderson, J. C. (1988). An updated paradigm for scale development incorporating unidimensionality and its assessment. *Journal of Marketing Research, 25*: 186-192.

Goodman, P. S. (1977). Social comparison processes in organizations. Pp. 97-132 in B. Staw & G. Salancik (Eds.), *New directions in organizational behavior*. Chicago, IL: St. Clair Press.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*: 139-164.

Johns, G. (1981). Difference score measures of organizational behavior variables: A critique. *Organizational Behavior and Human Performance, 27*: 443-463.

Johnston, J. (1984). *Econometric methods*. New York: McGraw-Hill.

Lord, F. M. (1958). Further problems in the measurement of growth. *Educational and Psychological Measurement, 18*: 437-451.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lykken, D. T. (1956). A method of actuarial pattern analysis. *Psychological Bulletin, 53*: 102-107.

Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.

O'Reilly, C. A., Chatman, J. A. & Caldwell, D. F. (1991). People and organizational culture: A profile comparison approach to assessing person-organization fit. *Academy of Management Journal, 34*: 487-516.

Posner, B. Z., Kouzes, J. M. & Schmidt, W. H. (1985). Shared values make a difference: An empirical test of corporate culture. *Human Resource Management, 24*: 293-309.

Rizzo, J., House, R. & Lirtzman, S. (1970). Role conflict and ambiguity in complex organizations. *Administrative Science Quarterly, 15*: 150-163.

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

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

Rounds, J. B., Dawis, R. W. & Lofquist, L. H. (1987). Measurement of person-environment fit and prediction of satisfaction in the theory of work adjustment. *Journal of Vocational Behavior, 31*: 297-318.

Rovine, M. J. (in press). Estimating non-shared environment using sibling discrepancy scores. In D. Reiss, E. M. Hetherington & R. Plomin (Eds.), *Nonshared environment*. Hillsdale, NJ: Earlbaum.

Rozeboom, W. W. (1966). *Foundations of the theory of prediction*. Homewood, IL: Dorsey Press.

Schneider, B. (1987). The people make the place. *Personnel Psychology, 40*: 437-453.

Schoorman, F. D., Bobko, P. & Rentsch, J. (1991). The role of theory in testing hypothesized interactions: An example from the research on escalation of commitment. *Journal of Applied Social Psychology, 21*: 1338-1355.

Smith, C. S. & Tisak, J. (1993). Discrepancy measures of role stress revisited: New perspectives on old issues. *Organizational Behavior and Human Decision Processes, 56*: 285-307.

Tisak, J. (1994). Determination of the regression coefficients and their associated standard errors in hierarchical regression analysis. *Multivariate Behavioral Research, 29*: 185-201.

Wall, T. D. & Payne, R. (1973). Are deficiency scores deficient? *Journal of Applied Psychology, 58*: 322-326.

Werts, C. E. & Linn, R. L. (1970). A general linear model for studying growth. *Psychological Bulletin, 73*: 17-22.

Wolins, L. (1982). *Research mistakes in the social and behavioral sciences.* Ames: Iowa State University Press.

Zimmerman, D. W., Brotohusodo, T. L. & Williams, R. H. (1981). The reliability of sums and differences of test scores: Some new results and anomalies. *Journal of Experimental Education, 49*: 177-186.