Caution in the Use of Difference Scores in Consumer Research

J. PAUL PETER
GILBERT A. CHURCHILL, JR.
TOM J. BROWN

This article illustrates the common use of difference scores in consumer research and discusses a number of potential problems with using them. Difference scores often have problems in the areas of reliability, discriminant validity, spurious correlations, and variance restriction. The article concludes that difference scores should generally not be used in consumer research and offers alternatives that overcome their limitations.

DIFFERENCE SCORES ARE OFTEN USED AS MEASURES OF CONSTRUCTS IN CONSUMER RESEARCH. A DIFFERENCE SCORE INVOLVES THE SUBTRACTION OF ONE MEASURE FROM ANOTHER TO CREATE A MEASURE OF A DISTINCT CONSTRUCT, FOR EXAMPLE, SUBTRACTING EXPECTATIONS FROM PERCEPTIONS TO CREATE A MEASURE OF PERCEIVED SERVICE QUALITY. AS SHOWN IN TABLE 1, DIFFERENCE SCORES HAVE BEEN USED IN AT LEAST FIVE AREAS OF CONSUMER AND MARKETING RESEARCH, INCLUDING SATISFACTION, SERVICE QUALITY, PRICE PERCEPTIONS, DECISION MAKING, AND ORGANIZATIONAL BUYING AND SELLING.

A NUMBER OF POTENTIAL PROBLEMS WITH THE USE OF DIFFERENCE SCORES CAN CAUSE THEM TO PERFORM POORLY AS MEASURES OF CONSTRUCTS. THESE PROBLEMS ARE WELL-KNOWN IN THE PSYCHOMETRIC LITERATURE (E.G., CRONBACH AND FURBY 1970; LORD 1958; MOSIER 1951; WALL AND PAYNE 1973). HOWEVER, CONSUMER RESEARCHERS GENERAL SEEM TO BE UNAWARE OF THEM, GIVEN THE FREQUENCY WITH WHICH DIFFERENCE SCORES ARE USED AND NEW DIFFERENCE SCORE MEASURES ARE PROPOSED. THE PURPOSE OF THIS ARTICLE IS TO ILLUSTRATE AND EXPLAIN THESE PROBLEMS SO THAT CONSUMER RESEARCHERS MAY CONSIDER THEM WHEN DEVELOPING NEW MEASURES. IN ADDITION, WE DISCUSS TWO ALTERNATIVE APPROACHES THAT ELIMINATE THE NEED TO USE DIFFERENCE SCORES IN MANY SITUATIONS.

This article is divided into two sections. First, it briefly reviews the problems that can arise when using difference scores. These problems are illustrated with the examples shown in Table 1 from the consumer behavior and marketing literature. In the second section, the article proposes alternative approaches that overcome these problems and offers general recommendations regarding the use of difference scores in consumer research.

PROBLEMS WITH THE USE OF DIFFERENCE SCORES

A variety of problems can occur when difference scores are used in consumer research, although not all problems necessarily occur in all cases. The potential problems involve reliability, discriminant validity, spurious correlations, and variance restriction.

Reliability Problems

Difference scores are usually less reliable than their component variables (JOHNS 1981; LORD 1958; MOSIER 1951; PRakash and LOINSbury 1983). The reason for this unreliability can be seen by examining the formula for the reliability of a difference score ($\rho_d$):

$$r_d = \frac{\sigma_1^2 r_{11} + \sigma_2^2 r_{22} - 2r_{12} \sigma_1 \sigma_2}{\sigma_1^2 + \sigma_2^2 - 2r_{12} \sigma_1 \sigma_2},$$

where $r_{11}$ and $r_{22}$ are the reliabilities of the first and second component scores, $\sigma_1^2$ and $\sigma_2^2$ are the variances of these component scores, and $r_{12}$ is the correlation between the component scores (JOHNS 1981). The reliability of a difference score thus hinges on the corre-

1Difference scores may also be used to measure change in some variable over time. This type of difference score appears to be used less frequently in consumer research than other types. CRONBACH and FURBY (1970) provide a detailed discussion of the use of change scores and conclude that there is little to justify the use of such scores as measures of constructs.

2Note that Table 1 is a convenience sample of studies that have utilized difference scores. No attempt was made to provide a census of studies using such measures.

All rights reserved. 0093-3061/93/0012-252.00

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
## TABLE 1
EXAMPLES OF THE USE OF DIFFERENCE SCORES IN MARKETING RESEARCH

<table>
<thead>
<tr>
<th>Marketing area and authors</th>
<th>General study area</th>
<th>Construct conceptualized as a difference score</th>
<th>Operationalization of difference score variable</th>
<th>Primary use of difference score variable</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Satisfaction: Oliver and Swan (1989)</td>
<td>Equity theory</td>
<td>Equity difference</td>
<td>Absolute difference between consumer equity and consumer perceptions of salesperson/dealer equity for eight automobile purchase outcomes.</td>
<td>A predictor of fairness, preference, and satisfaction.</td>
<td>One of four hypothesized relationships involving difference score measure of equity difference was supported; two of the others were in opposite direction of hypotheses. Equity difference was dropped from further analysis.</td>
</tr>
<tr>
<td>Tse and Wilton (1988)</td>
<td>Consumer satisfaction (with a product)</td>
<td>Disconfirmation</td>
<td>Algebraic differences between perceived and, in turn, expected, ideal, and equitable product performance overall and for 14 attributes of a record player.</td>
<td>A predictor of satisfaction.</td>
<td>Difference score measure was outperformed by a nondifference score measure of disconfirmation.</td>
</tr>
<tr>
<td>Swan and Trawick (1981)</td>
<td>Consumer satisfaction (with retail service)</td>
<td>Inferred disconfirmation</td>
<td>Algebraic difference between preorder and postorder ratings of five attributes of restaurant food.</td>
<td>A predictor of satisfaction and intentions and a consequence of expectations, beliefs, and anticipated satisfaction.</td>
<td>Difference score measure of inferred disconfirmation found to be strongly related to satisfaction and intentions but weakly related to expectations and inversely related to anticipated satisfaction.</td>
</tr>
<tr>
<td>Service quality: Brown and Swartz (1989)</td>
<td>Service quality (for professional services)</td>
<td>Service quality gaps</td>
<td>Three sets of algebraic differences between various patient and doctor perceptions of medical services.</td>
<td>Predictors of overall evaluation of medical service encounters.</td>
<td>Two of three hypotheses using difference score measures of service gaps were supported; most of the correlations for the other were in opposite direction of hypothesis.</td>
</tr>
<tr>
<td>Parasuraman, Zeithaml, and Berry (1988)</td>
<td>Service quality for retail service sectors</td>
<td>Perceived service quality</td>
<td>Algebraic difference between consumer's perceptions and expectations; five dimensions measured by 22 attributes across four service categories.</td>
<td>Development of a measure of service quality.</td>
<td>Service quality measure was developed and investigated solely as a difference score.</td>
</tr>
</tbody>
</table>
TABLE 1 (Continued)

<table>
<thead>
<tr>
<th>Marketing area and authors</th>
<th>General study area</th>
<th>Construct conceptualized as a difference score</th>
<th>Operationalization of difference score variable</th>
<th>Primary use of difference score variable</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price perceptions:</td>
<td>Accuracy of price perceptions</td>
<td>Price recall error</td>
<td>Algebraic difference between price recalled and stimulus price treatment divided by stimulus price treatments for four products.</td>
<td>A criterion to be predicted by price expectation and level of involvement.</td>
<td>Limited support was found for two hypotheses involving a difference score measure of price recall error.</td>
</tr>
<tr>
<td>Helgeson and Beatty (1987)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zeithaml (1982)</td>
<td>Price recall</td>
<td>Exact-price recall error</td>
<td>Absolute difference between recalled price and correct price divided by correct price for 12 grocery products.</td>
<td>A criterion to be predicted by price information format.</td>
<td>One of three hypotheses employing difference score measure of exact-price recall error was supported; in one of the other two comparisons, the means were ordered opposite the hypothesis.</td>
</tr>
<tr>
<td>Decision making:</td>
<td>Decision making</td>
<td>Net perceived return</td>
<td>Algebraic difference between perceived return and perceived risk for six facets of automobile purchase.</td>
<td>A predictor of brand preference.</td>
<td>Difference score measure of net perceived return factored differently than nondifference score measures, but the net perceived return model explained more of the variance in brand preference than did the other models.</td>
</tr>
<tr>
<td>Peter and Tarpey (1975)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Organizational buying and</td>
<td>Franchisor-franchisee goal compatibility</td>
<td>Preference incongruity</td>
<td>Four algebraic differences between various measures of actual and perceived goal importance.</td>
<td>A correlate of perceived conflict.</td>
<td>Five of 24 hypothesized relationships involving difference score measures of preference incongruity were supported using p ≤ .10 as a criterion.</td>
</tr>
<tr>
<td>selling:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elashberg and Michie (1984)</td>
<td></td>
<td>Perceptual differences</td>
<td>Two algebraic differences between franchisor and franchisee ratings of business goals.</td>
<td>A correlate of perceived conflict.</td>
<td>Three of eight hypothesized relationships involving difference score measures of perceptual differences were supported using p ≤ .10 as a criterion.</td>
</tr>
<tr>
<td>Ross and Lusch (1982)</td>
<td>Channel conflict and cooperation</td>
<td>Domain dissensus</td>
<td>Absolute difference between food broker and food wholesaler perceptions of relative influence on five activities or functions.</td>
<td>A predictor of channel conflict and channel cooperation.</td>
<td>One of two relationships involving difference score measure of domain dissensus was supported.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Perceptual incongruity</td>
<td>Absolute difference between food broker and food wholesaler importance ratings for five types of assistance.</td>
<td>A predictor of channel conflict and channel cooperation.</td>
<td>One of two relationships involving difference score measure of perceptual incongruity was supported.</td>
</tr>
<tr>
<td>Marketing area and authors</td>
<td>General study area</td>
<td>Construct conceptualized as a difference score</td>
<td>Operationalization of difference score variable</td>
<td>Primary use of difference score variable</td>
<td>Comments</td>
</tr>
<tr>
<td>---------------------------</td>
<td>-------------------</td>
<td>-----------------------------------------------</td>
<td>-----------------------------------------------</td>
<td>------------------------------------------</td>
<td>-----------</td>
</tr>
<tr>
<td>Rosenberg and Stern (1971)</td>
<td>Conflict between channel members</td>
<td>Channel conflict</td>
<td>Absolute difference between manufacturers, distributors, and dealers on composite scores (32 items) and on four conflict dimension scores (eight items per dimension).</td>
<td>A determinant of differences in levels of channel conflict.</td>
<td>Six of 15 hypothesized differences involving difference score measure of channel conflict were supported.</td>
</tr>
<tr>
<td>Teas, Wacker, and Hughes (1979)</td>
<td>Salespeople’s role clarity</td>
<td>Need dissatisfaction</td>
<td>Eight sets of algebraic differences between desired and actual need fulfillment.</td>
<td>An alternative to actual need fulfillment to investigate convergent and discriminant validity.</td>
<td>Evidence of convergent and discriminant validity was obtained by correlating difference score measures of need dissatisfaction with one of the components used in forming the difference scores.</td>
</tr>
<tr>
<td>Ford, Walker, and Churchill (1975)</td>
<td>Salespeople’s role conflict</td>
<td>Role conflict</td>
<td>Absolute difference in salespersons’ perceptions of the expectations held by each of four role partners for 30 items.</td>
<td>Development of a measure of salesperson role conflict.</td>
<td>Difference score measure of role conflict had the lowest reliability and convergent validity correlations of the three measures investigated.</td>
</tr>
</tbody>
</table>

lation between its component scores and on their reliabilities. As the reliability of either (or both) component score decreases, the reliability of the difference score decreases. Moreover, as the correlation between the component scores becomes larger, the reliability of the difference score also decreases. Assuming equal component score variances, Figure 1 demonstrates the impact of component score reliabilities and intercorrelation on the reliability of a difference score. For example, the reliability of the difference score variable formed by subtracting scores on two measures with average reliability of .70 and a correlation of .40 is only .50. Thus, while the average reliability of the two component measures is .70 (considered acceptable for many research purposes), the correlation between the measures attenuates the reliability of the difference score measure to a level most researchers would find unacceptable.

Although several of the studies in Table 1 addressed the reliability of measures, none of the studies computed the reliability of the difference scores correctly, nor did any of them provide the necessary information to do so. Four of the studies (Ford et al. 1975; Parasuraman et al. 1988; Ross and Lusch 1982; Teas et al. 1979) attempted to assess the reliability of difference score variables by computing coefficient alphas on the differences themselves. Two additional studies (Brown and Swartz 1989; Oliver and Swan 1989) considered reliability but only for one or both of the component variables forming the difference score and not for the difference score variable itself. These approaches to assessing the reliability of a difference score are inappropriate because they do not adequately consider the correlation between the components.

The components of difference scores can often be expected to be positively correlated (Johns 1981). In many cases, both components are provided by the same respondent on identical scales, and the only operational difference in component measures is the perspective the respondent is asked to take (e.g., Brown and Swartz 1989; Ford et al. 1975; Oliver and Swan 1989; Parasuraman et al. 1988; Swan and Trawick 1981; Teas et al. 1979; Tse and Wilton 1988). For example, in the study on need satisfaction reported by Teas et al. (1979, p. 357), respondents were first asked “how much there is now” and then “how much there should be” of some attribute such as performance feedback. Their responses were recorded on identical scales and subtracted from

---

7In the event of a negative correlation between the component scores, the reliability of the difference score will increase beyond the average of the component score reliabilities.
one another. Similarly, Parasuraman et al. (1988) first asked respondents for the level of service that should be provided by firms and then for the level of service that is actually provided. A follow-up study using their scales found the correlation between component measures to be .34 (Brown, Churchill, and Peter 1992). Even when the component scores come from different respondents, it is possible that they will be correlated, since the same method of measurement is being used.

The lower reliability of difference score measures attenuates their observed correlations with other variables. Since most of the studies in Table 1 used some form of correlational analysis, it is not surprising that many of them found weak relationships, as reported in the comments.

Discriminant Validity Problems

Discriminant validity refers to the degree to which measures of theoretically distinct constructs do not correlate too highly; without evidence of discriminant validity, a measure cannot demonstrate construct validity (Churchill 1979; Peter 1981). Difference scores often have low reliability for the reasons mentioned above. Moreover, low reliability attenuates the correlations between a measure and measures of other constructs. Thus, it is possible that the resulting correlations between a difference score measure and other measures may create the illusion of meeting discriminant validity standards simply because of low reliability. This problem is, of course, magnified when one difference score measure is correlated with another. For example, it seems likely that one reason the difference score measures used by Teas et al. (1979) had low heterotrait correlations in a multitrait-multimethod matrix and thus appeared to demonstrate discriminant validity is that they had low reliability or at least lower reliability than the nondifference score method. In addition, in one heterotrait-monomethod triangle, discriminant validity
was examined by correlating two sets of difference scores, which thus amplified the problem of low reliability. Since low correlations with measures of other constructs is a criterion for discriminant validity, it is easier to give the appearance of discriminant validity using difference scores.

There is a second discriminant validity problem that is unique to measures formed as linear combinations of scale scores, as is true with difference scores. This problem applies when the linear combination is formed by summing measures of different constructs or dimensions of constructs. Conceptually, a difference score is a measure of a construct that is distinct from the constructs represented by the component variables. Operationally, however, the difference between two variables will be strongly correlated with at least one of the two variables (Johns 1981; Wall and Payne 1973). If the difference score cannot be distinguished from its components, it does not demonstrate discriminant validity. For example, a recent study comparing a difference score measure with a nondifference score measure of service quality found that the difference score measure was correlated quite highly \((r = .79)\) with one of its components (Brown et al. 1992). In such cases, it is difficult to argue that the difference score is measuring something unique from the component variable. With the exception of Teas et al. (1979), none of the studies in Table 1 report the correlations between a difference score and its components. However, it is common for such correlations to be high (John 1981; Wall and Payne 1973); these correlations in Teas et al. ranged from \(-.66\) to \(-.88\), for example.

**Spurious Correlations**

A third problem with difference scores follows directly from the fact that difference scores are not unique from their components. More specifically, because of the relationship between difference scores and their components, any correlation between difference scores and other variables is likely to be spurious (Wall and Payne 1973). A correlation between a difference score and other variables is an artifact of the relationship of the other variables with one or both of the components of the difference score. In fact, the difference between two variables provides no additional information for predicting or explaining a criterion beyond that held in the components themselves. Moreover, in some cases, one of the component variables may perform better than the difference score itself. For example, in the service quality study cited above, the perceptions component of the difference score predicted behavioral intentions better than the difference score measure itself.

The high correlations between difference scores and their components are particularly troublesome when all three variables are included in the same analysis. For example, consumer satisfaction research often includes expectations, performance, and their difference, disconfirmation, as independent variables in regression equations. The multicollinearity among these variables can produce unstable parameter estimates and misleading results.

**Variance Restriction Problem**

A fourth problem with the use of difference scores concerns a possible restriction of the variance of the difference score variable. This problem arises when one of the components used in calculating a difference score is consistently higher than the other. This phenomenon will almost always occur when one of the components used to create a difference score is one for which more is always better (Wall and Payne 1973). For example, consider the difference score measure of professional service quality used by Brown and Swartz (1989). In one part of the study, patients were asked to indicate their expectations of how much of certain qualities their physicians *should* have: they were also asked for their actual experiences with their physicians on these same issues. Service quality was assessed by subtracting experience scores from expectation scores. One issue dealt with the degree to which the doctor was up-to-date on the latest medical technology. It seems reasonable to assume that only rarely would patients’ expectations with respect to what a doctor *should* know about medical technology be exceeded by what they believe the doctor actually *does* know in this area. Could a doctor ever have too much knowledge of medical technology? Assuming, then, that expectations will usually equal or exceed actual experiences, patients whose experiences indicate that their doctor is not up-to-date in this area have a wider potential range on the difference score than do patients who indicate that their doctor is very knowledgeable in this area. This restriction in range for some patients leads to a corresponding restriction in the variance of the difference score. Moreover, this variance restriction is systematic; as the value of the difference score increases (indicating less perceived professional service quality), so too does the variance of the difference score. This can cause problems in many types of statistical analysis.

For example, in ordinary least squares (OLS) regression, constant variance of the dependent variable is assumed. When the variance is not constant, the estimators no longer have minimum variance. Moreover, the condition is usually associated with nonnormality in the dependent variables (Neter and Wasserman 1974, pp. 131–132). Such conditions can lead to erroneous conclusions about the statistical significance of regres-
sion coefficients. While this problem can be overcome through weighted least squares or a transformation, the analysis becomes more complex. In addition, the studies reported in Table 1 typically used OLS regression without transformation.

The variance restriction problem may have been manifested in several of the studies in Table 1 in which one of the component scores was likely to have been consistently higher than the other component score (e.g., Brown and Swartz 1989; Eliahsberg and Michie 1984; Ford et al. 1975; Parasuraman et al. 1988; Ross and Lusch 1982; Swan and Trawick 1981; Teas et al. 1979; Tse and Wilton 1988). For example, consider the perceptual incongruity construct proposed and operationalized as a difference score by Ross and Lusch (1982). Perceptual incongruity was defined as the difference between wholesalers' and brokers' ratings of the importance of five types of assistances offered by the broker to the wholesaler. Brokers may well rate their own services as more important because they may wish to justify their existence and value to wholesalers and other channel members.

**SUMMARY AND RECOMMENDATIONS**

This article explained a variety of potential problems with the use of difference scores as measures of constructs. Difference scores (1) are typically less reliable than other measures, (2) may appear to demonstrate discriminant validity when this conclusion is not warranted, (3) may be only sparsely correlated to other measures since they typically do not discriminate from at least one of their components, and (4) may exhibit variance restriction.

However, there may be circumstances in which, on balance, a researcher believes that the advantages of using a difference score warrant it. For example, a researcher may wish to assess the level of conflict between a seller and buyer indirectly without explicitly questioning them about it. Differences in seller and buyer responses might be hypothesized to reflect such conflict. Thus, there may be cases in which calculating a difference score has value. Yet caution must be taken when such scores are used as input into further statistical analysis. In sum, difference scores should not be used simply because they are intuitively appealing or computationally convenient.

Below we offer two alternatives to the use of difference scores in consumer research that may prove useful in some situations. These include (1) using a direct comparison operationalization of the proposed difference score construct and (2) reframing the research question to avoid using difference scores.

**Using Direct Comparison Operationalizations**

It is often possible to operationalize constructs more directly so that the subsequent calculation of difference scores is unnecessary (Johns 1981). This approach is applicable when a single respondent provides both of the measures used in computing a difference score. For example, in a study of consumer satisfaction, Tse and Wilton (1988, p. 206) included a direct comparison measure of disconfirmation in addition to a difference score measure of it. Their direct comparison measure, "Overall how close did the [product] come to your expectations?" was scored on a scale anchored "very much poorer than expected" to "very much better than expected." As noted in Table 1, this measure outperformed the difference score measure empirically. Brown et al. (1992) empirically compared the difference score approach to measuring perceived service quality with a direct comparison operationalization of the same construct. The results indicated that the direct comparison approach had better psychometric properties, was less taxing on respondents (because they responded to only half as many items), and produced favorable empirical results.

Note that the direct comparison operationalization approach does no damage to a substantive theory that involves a construct considered to be a difference between two other constructs. The direct comparison operationalization approach requires subjects to mentally consider differences rather than have the researcher calculate an arithmetic difference for them. In fact, until we have greater knowledge of cognitive algebra that supports the idea that consumers use arithmetic differences in forming their thoughts, the direct comparison approach has the advantage of allowing consumers to combine their thoughts as they wish rather than have an arbitrary combination rule forced on them.

**Reframing Research Questions**

A second alternative is to reframe research questions to avoid the use of difference scores (Johns 1981). For example, Oliver and Swan (1989) operationalized equity advantage as the difference between consumers' perceptions of their own equity in a transaction and their perceptions of dealer equity and used this difference to predict such things as the level of consumer satisfaction. To avoid the use of difference scores, the research question might be reframed as the following: Do consumers' perceptions of dealer equity add incrementally to the prediction of satisfaction beyond consumers' perceptions of their own equity? Although the idea of equity advantage is no longer researched, the fundamental research question remains quite similar. Namely, is customer satisfaction influenced by both customers' perceptions of their own equity and their perceptions of dealer's equity? Moreover, this research question can be addressed without the use of difference scores by employing hierarchical regression. With a measure of consumer satisfaction as the dependent variable, consumers' perceptions of their own equity would first be entered into the regression equation followed by their
perceptions of dealer equity. The analysis would focus on the incremental variance accounted for by the addition of dealer equity. Reframing research questions in this manner may be an especially attractive alternative in situations in which difference score data have already been collected. Doing so and using hierarchical regression are perhaps the best approaches available for dealing with extant difference score data. Alternatively, if a theory does not suggest the ordering of the components, they could be modeled simultaneously with structural equation methods.

Reframing research questions need not, but can, change a substantive theory and the hypotheses that are derived from it. However, given that an arithmetic difference score adds no additional information for predicting a criterion beyond that held in the components themselves, there is at least some question as to the value of conceptualizing constructs as differences between two other constructs. Reframing research questions in the manner suggested may also offer more creative and insightful hypotheses about consumer behavior since it requires new ways of thinking about relationships between component constructs.

In sum, we recommend that difference scores should generally not be used in consumer research, particularly where other methods of measurement can be applied. If difference scores are to be used, adequate theoretical justification should be provided. In addition, the reliability of the difference score should be examined and reported, along with the means, variances, reliabilities, and intercorrelations of the component variables. The correlations of the components with the difference score and with other variables should also be examined and reported. This information allows researchers to judge whether the use of a difference score aided or hindered our understanding of consumer behavior.

[Received April 1992. Revised September 1992.]

REFERENCES


