Reassessment of Expectations as a Comparison Standard in Measuring Service Quality: Implications for Further Research

The authors respond to concerns raised by Cronin and Taylor (1992) and Teas (1993) about the SERVQUAL instrument and the perceptions-minus-expectations specification invoked by it to operationalize service quality. After demonstrating that the validity and alleged severity of many of those concerns are questionable, they offer a set of research directions for addressing unresolved issues and adding to the understanding of service quality assessment.

Two recent articles in this journal—Cronin and Taylor (1992) and Teas (1993)—have raised concerns about our specification of service quality as the gap between customers’ expectations and perceptions (Parasuraman, Zeithaml, and Berry 1985) and about SERVQUAL, a two-part instrument we developed for measuring service quality (Parasuraman, Zeithaml, and Berry 1988) and later refined (Parasuraman, Berry, and Zeithaml 1991). In this article we respond to those concerns by reexamining the underlying arguments and evidence, introducing additional perspectives that we argue are necessary for a balanced assessment of the specification and measurement of service quality, and isolating the concerns that seem appropriate from those that are questionable. Building on the insights emerging from this exchange, we propose an agenda for further research.

Because many of the issues raised in the two articles and in our response are still unsettled, we hope that the research agenda will encourage all interested researchers to address those issues and add to the service quality literature in a constructive manner.

Though the concerns raised by both C&T and Teas relate to service quality assessment, by and large they focus on different issues. For this reason, we discuss C&T’s concerns first, then address the issues raised by Teas. We conclude with a discussion of further research for addressing unresolved issues, including the relationship between service quality and customer satisfaction.

Response to Cronin and Taylor (1992)

C&T surveyed customers in four sectors (banking, pest control, dry cleaning, and fast food) using a questionnaire that contained (1) the battery of expectations and perceptions questions (Parasuraman, Zeithaml, and Berry 1988) and later refined (Parasuraman, Berry, and Zeithaml 1991). In this article we respond to those concerns by reexamining the underlying arguments and evidence, introducing additional perspectives that we argue are necessary for a balanced assessment of the specification and measurement of service quality, and isolating the concerns that seem appropriate from those that are questionable. Building on the insights emerging from this exchange, we propose an agenda for further research.

C&T conclude that it is unnecessary to measure customer expectations in service quality research. Measuring perceptions is sufficient, they contend. They also conclude that service quality fails to affect purchase intentions. We do not believe that these conclusions are warranted at this stage of research on service quality, as we attempt to demonstrate in this section. Our comments on C&T’s study focus on (1) conceptual issues, (2) methodological and analytical issues, and (3) practical issues.

Conceptual Issues

The perceptions-expectations gap conceptualization of SQ. C&T contend (p. 56) that, “little if any theoretical or empirical evidence supports the relevance of the expectations-performance gap as the basis for measuring service quality.” They imply here and elsewhere that PZB’s extensive focus group research (PZB 1985; ZPB 1990) suggest only the attributes of SQ and not its expectations-performance gap formulation. We wish to emphasize that our research provides strong support for defining SQ as the discrepancy between customers’ expectations and perceptions, a point we make in PZB (1985) and ZPB (1990). C&T’s assertion...
also seems to discount prior conceptual work in the SQ literature (Gronroos 1982; Lehtinen and Lehtinen 1982; Sasser, Olsen, and Wyckoff 1978) as well as more recent research (Bolton and Drew 1991a, b; ZPB 1991) that supports the disconfirmation of expectations conceptualization of SQ. Bolton and Drew (1991b, p. 383), in an empirical study cited by C&T, conclude:

Consistent with prior exploratory research concerning service quality, a key determinant of overall service quality is the gap between performance and expectations (i.e., disconfirmation). For residential customers, perceived telephone service quality depended on the disconfirmation triggered by perceived changes in existing service or changes in service providers. . . It is interesting to note that disconfirmation explains a larger proportion of the variance in service quality than performance.

Therefore, C&T’s use of the same Bolton and Drew article as support for their claim (p. 56) that “the marketing literature appears to offer considerable support for the superiority of simple performance-based measures of service quality” is surprising and questionable. Moreover, a second citation that C&T offer to support this contention—Mazis, Ahtola, and Klippel (1975)—is an article that neither dealt with service quality nor tested performance-based measures against measures incorporating expectations.

There is strong theoretical support for the general notion that customer assessments of stimuli invariably occur relative to some norm. As far back as Nelson (1964), adaptation-level theory held that individuals perceive stimuli only in relation to an adapted standard. More recently, Kahneman and Miller (1986) provide additional theoretical and empirical support for norms as standards against which performance is compared and can be measured.

**Attitude formation versus attitude measurement.** The empirical, longitudinal studies (e.g., Bolton and Drew 1991a, b; Churchill and Surprenant 1982; Oliver 1980) that C&T cite to support their arguments focus on the roles of expectations, performance, and disconfirmation in the formation of attitudes. However, the relevance of those arguments for raising concerns about SERVQUAL is moot because the instrument is designed merely to measure perceived SQ—an attitude level—at a given point in time, regardless of the process by which it was formed. SERVQUAL is a tool to obtain a reading of the attitude level, not a statement about how the level was developed.

**Relationship between CS and SQ.** An important issue raised by C&T (as well as Teas) is the nature of the link between CS and SQ. This is a complex issue characterized by confusion about the distinction between the two constructs as well as the causal direction of their relationship. SQ researchers (e.g., Carman 1990; PZB 1988) in the past have distinguished between the two according to the level at which they are measured: CS is a transaction-specific assessment (consistent with the CS/D literature) whereas SQ is a global assessment (consistent with the SQ literature). On the basis of this distinction, SQ researchers have posited that an accumulation of transaction-specific assessments leads to a global assessment (i.e., the direction of causality is from CS to SQ). However, on careful reflection, we now believe that this distinction may need to be revised, particularly because some recent studies (Reidenbach and Sandifer-Smallwood 1990; Woodside, Frey, and Daly 1989) have modeled SQ as an antecedent of CS.

In the research implications section, we propose and discuss an integrative framework that attempts to reconcile the differing viewpoints about the distinction, and the direction of causality, between CS and SQ. Here we wish to point out an apparent misinterpretation by C&T of our published work (PZB 1985, 1988). Specifically, in PZB (1985) we do not discuss the relationship between SQ and CS, and in PZB (1988) we argue in favor of the “CS leads to SQ” hypothesis. Therefore, we are surprised by C&T’s attributing to us the opposite view by stating (p. 56) “PZB (1985, 1988) proposed that higher levels of perceived SQ result in increased CS” and by referring to (p. 62) “the effects hypothesized by PZB (1985, 1988) [that] SQ is an antecedent of CS.” Moreover, as we discuss in the next section, C&T’s empirical findings do not seem to warrant their conclusion (p. 64) that “perceived service quality in fact leads to satisfaction as proposed by PZB (1985, 1988).”

**Types of comparison norms in assessing CS and SQ.** In critiquing the perceptions-minus-expectations conceptualization of SERVQUAL, C&T raise the issue of appropriate comparison standards against which perceptions are to be compared (p. 56):

[PZB] state that in measuring perceived service quality the level of comparison is what a consumer should expect, whereas in measures of satisfaction the appropriate comparison is what a consumer would expect. However, such a differentiation appears to be inconsistent with Woodruff, Cadotte, and Jenkins’ (1983) suggestion that expectations should be based on experience norms—what consumers should expect from a given service provider given their experience with that specific type of service organization.

Though the first sentence in this quote is consistent with the distinction we have made between comparison norms in assessing SQ and CS, the next sentence seems to imply that the debate over norms has been resolved, the consensus being that the “experience-based norms” of Woodruff, Cadotte, and Jenkins (1983) are the appropriate frame of reference in CS assessment. We believe that such an inference is unwarranted because conceptualizations more recent than Woodruff, Cadotte, and Jenkins (1983) suggest that CS assessment could involve more than one comparison norm (Forbes, Tse, and Taylor 1986; Oliver 1985; Tse and Wilton 1988; Wilton and Nicosia 1986). Cadotte, Woodruff, and Jenkins (1987) themselves have stated (p. 313) that “additional work is needed to refine and expand the conceptualization of norms as standards.” Our own recent research on customers’ service expectations (ZPB 1991, a publication that C&T cite) identifies two different comparison norms for SQ assessment: desired service (the level of service a customer believes can and should be delivered) and adequate service (the level of service the customer considers acceptable).

Our point is that the issue of comparison norms and their interpretation has not yet been resolved fully. Though
recent conceptual work (e.g., Woodruff et al. 1991) and empirical work (e.g., Boulding et al. 1993) continue to add to our understanding of comparison norms and how they influence customers’ assessments, more research is needed.

**Methodological and Analytical Issues**

C&T’s empirical study does not justify their claim (p. 64) that “marketing’s current conceptualization and measurement of service quality are based on a flawed paradigm” and that a performance-based measure is superior to the SERVQUAL measure. A number of serious questions about their methodology and their interpretation of the findings challenge the strong inferences they make.

**Dimensionality of service quality.** One piece of evidence C&T use to argue against the five-component structure of SERVQUAL is that their LISREL-based confirmatory analysis does not support the model shown in their Figure 1. Unfortunately, C&T’s Figure 1 is not a totally accurate depiction of our prior findings pertaining to SERVQUAL (PZB 1988) because it does not allow for possible intercorrelations among the five latent constructs. Though we have said in our previous work that SERVQUAL consists of five distinct dimensions, we also have pointed out that the factors representing those dimensions are intercorrelated and hence overlap to some degree. We report average interfactor correlations of .23 to .35 in PZB (1988). In a more recent study in which we refined SERVQUAL, reassessed its dimensionality, and compared our findings with four other replication studies, we acknowledge and discuss potential overlap among the dimensions and arrive at the following conclusion (PZB 1991, p. 442):

> Though the SERVQUAL dimensions represent five conceptually distinct facets of service quality, they are also interrelated, as evidenced by the need for oblique rotations of factor solutions in the various studies to obtain the most interpretable factor patterns. One fruitful area for future research is to explore the nature and causes of these interrelationships.

Therefore, we would argue that the fit of C&T’s SERVQUAL data to their model in Figure 1 might have been better than that implied by their Table 1 results if the model had allowed the five latent constructs to intercorrelate.

C&T use results from their oblique factor analyses to reiterate their inference that the 22 SERVQUAL items are unidimensional. They do so by merely referring to their Table 2 and indicating (p. 61) that “all of the items loaded predictably on a single factor with the exception of item 19.” But a careful examination of C&T’s Table 2 raises questions about the soundness of their inference.

First, except for the factor representing SERVQUAL’s perceptions component (i.e., SERVPERF) in the pest control context, the percentage of variance in the 22 items captured by the one factor for which results are reported is less than 50%, the cut-off value recommended by Bagozzi and Yi (1988) for indicators of latent constructs to be considered adequate. Moreover, across the four contexts, the variance captured for the SERVQUAL items is much lower than for the SERVPERF items (mean of 32.4% for SERVQUAL versus 42.6% for SERVPERF). These results strongly suggest that a unidimensional factor is not sufficient to represent fully the information generated by the 22 items, especially in the case of SERVQUAL, for which over two-thirds of the variance in the items is unexplained if just one factor is used. The difference in the variance explained for SERVQUAL and SERVPERF also suggests that the former could be a richer construct that more accurately represents the multifaceted nature of SQ posited in the conceptual literature on the subject (Gronroos 1982; Lehtinen and Lehtinen 1982; PZB 1985; Sasser, Olsen, and Wyckoff 1978).

Second, though C&T say that they conducted oblique factor analysis, they report loadings for just one factor in each setting (Table 2). It is not clear from their article whether the reported loadings are from the rotated or unrotated factor-loading matrix. If they are prerotation loadings, interpreting them to infer unidimensionality is questionable. If they are indeed postrotation loadings, the variance explained uniquely by each factor might be even lower than the percentages shown in Table 2. Because oblique rotation allows factors to correlate with one another, part of the variance reported as being explained by one factor might be shared with the other factors included in the rotation. The possibility that the unique variance represented by the single factors might be lower than the already low percentages in Table 2 further weakens C&T’s claim that the SQ construct is unidimensional.

Finally, C&T’s inference that SERVQUAL and SERVPERF can be treated as unidimensional on the basis of the high alpha values reported in Table 2 is erroneous because it is at odds with the extensive discussion in the literature about what unidimensionality means and what coefficient alpha does and does not represent (Anderson and Gerbing 1982; Gerbing and Anderson 1988; Green, Lissitz, and Mulaik 1977; Hattie 1985; Howell 1987). As Howell (1987, p. 121) points out, “Coefficient alpha, as an estimate of reliability, is neither necessary nor sufficient for unidimensionality.” And, according to Gerbing and Anderson (1988, p. 190), “Coefficient alpha ... sometimes has been misinterpreted as an index of unidimensionality rather than reliability. Unidimensionality and reliability are distinct concepts.”

In short, every argument that C&T make on the basis of their empirical findings to maintain that the SERVQUAL items form an unidimensional scale is questionable. Therefore, summing or averaging the scores across all items to create a single measure of service quality, as C&T have done in evaluating their structural models (Figure 2), is questionable as well.

**Validity.** Citing evidence from the correlations in their Table 3, C&T conclude that SERVPERF has better validity than SERVQUAL (p. 61): “We suggest that the proposed performance-based measures provide a more construct-valid explication of service quality because of their content validity ... and the evidence of their discriminant validity.” This suggestion is unwarranted because, as we demonstrate subsequently, SERVQUAL performs just as well as SERVPERF on each validity criterion that C&T use.
C&T infer convergent validity for SERVPERF by stating (p. 61), "A high correlation between the items SERVPERF, importance-weighted SERVPERF, and service quality indicates some degree of convergent validity." However, they fail to examine SERVQUAL in similar fashion. According to Table 3, the average pairwise correlation among SERVPERF, importance-weighted SERVPERF, and overall service quality is .6892 (average of .9787, .5430, and .5572). The corresponding average correlation for SERVQUAL is .6870 (average of .9787, .5430, and .5394). Clearly, the virtually identical average correlations for SERVPERF and SERVQUAL do not warrant the conclusion that the former has higher convergent validity than the latter.

C&T claim discriminant validity for SERVPERF by stating (p. 61), "An examination of the correlation matrix in Table 3 indicates discriminant validity ... as the three service quality scales all correlate more highly with each other than they do with other research variables (i.e., satisfaction and purchase intentions)." Though this statement is true, it is equally true for SERVQUAL, a finding that C&T fail to acknowledge. In fact, though the average pairwise correlation of SERVPERF with satisfaction and purchase intentions is .4812 (average of .5978 and .3647), the corresponding value for SERVQUAL is only .4569 (average of .5605 and .3534). Because the average within-construct intercorrelations are almost identical for the two scales, as demonstrated previously, these results actually imply somewhat stronger discriminant validity for SERVQUAL.

Regression analyses. C&T assess the predictive ability of the various multiple-item service quality scales by conducting regression analyses wherein their single-item overall SQ measure is the dependent variable. On the basis of the \( R^2 \) values reported in Table 4, they correctly conclude that SERVPERF outperforms the other three scales. Nevertheless, one should note that the average improvement in variance explained across the four contexts by using SERVPERF instead of SERVQUAL is 6%. (The mean \( R^2 \) values, rounded to two decimal places, are .39 for SERVQUAL and .45 for SERVPERF.) Whether this difference is large enough to claim superiority for SERVPERF is arguable. The dependent variable itself is a performance-based (rather than disconfirmation-based) measure and, as such, is more similar to the SERVPERF than the SERVQUAL formulation. Therefore, at least some of the improvement in explanatory power achieved by using SERVPERF instead of SERVQUAL could be merely an artifact of the "shared method variance" between the dependent and independent variables. As Bagozzi and Yi (1991, p. 426) point out, "When the same method is used to measure different constructs, shared method variance always inflates the observed between-measure correlation."

The overall pattern of significant regression coefficients in C&T's Table 4 offers some insight about the dimensionality of the 22 items as well. Specifically, it can be partitioned into the following five horizontal segments corresponding to the a priori classification of the items into the SERVQUAL dimensions (PZB 1988):

<table>
<thead>
<tr>
<th>Segment</th>
<th>Tangibles</th>
<th>Variables V1–V4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Segment</td>
<td>Reliability</td>
<td>Variables V5–V9</td>
</tr>
<tr>
<td>Segment</td>
<td>Responsiveness</td>
<td>Variables V10–V13</td>
</tr>
<tr>
<td>Segment</td>
<td>Assurance</td>
<td>Variables V14–V17</td>
</tr>
<tr>
<td>Segment</td>
<td>Empathy</td>
<td>Variables V18–V22</td>
</tr>
</tbody>
</table>

The total number of significant regression coefficients in the various segments of Table 4 as follows:

<table>
<thead>
<tr>
<th>Segment</th>
<th>Tangibles</th>
<th>Reliability</th>
<th>Responsiveness</th>
<th>Assurance</th>
<th>Empathy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Segment</td>
<td>6</td>
<td>32</td>
<td>9</td>
<td>14</td>
<td>11</td>
</tr>
</tbody>
</table>

As this summary implies, the items under the five dimensions do not all contribute in like fashion to explaining the variance in overall service quality. The reliability items are the most critical drivers, and the tangibles items are the least critical drivers. Interestingly, the relative importance of the five dimensions implied by this summary—that is, reliability being most important, tangibles being least important, and the other three being of intermediate importance—almost exactly mirrors our findings obtained through both direct (i.e., customer-expressed) and indirect (i.e., regression-based) measures of importance (PBZ 1991; PZB 1988; ZPB 1990). The consistency of these findings across studies, researchers, and contexts offers strong, albeit indirect, support for the multidimensional nature of service quality. They also suggest that techniques other than, or in addition to, factor analyses can and should be used to understand accurately the dimensionality of a complex construct such as SQ. As we have argued in another article comparing several SERVQUAL-replication studies that yielded differing factor patterns (PBZ 1991, p. 442–3):

Respondents' rating a specific company similarly on SERVQUAL items pertaining to different dimensions—rather than the respondents' inability to distinguish conceptually between the dimensions in general—is a plausible explanation for the diffused factor patterns. To illustrate, consider the following SERVQUAL items:

- Employees at XYZ give you prompt service (a responsiveness item)
- Employees of XYZ have the knowledge to answer your questions (an assurance item)
- XYZ has operating hours convenient to all its customers (an empathy item)

If customers happen to rate XYZ the same or even similarly on these items, it does not necessarily mean that customers consider the items to be part of the same dimension. Yet, because of high inter-correlations among the three sets of ratings, the items are likely to load on the same factor when the ratings are factor analyzed. Therefore, whether an unclear factor pattern obtained through analyzing company-specific ratings necessarily implies poor discriminant validity for the general SERVQUAL dimensions is debatable.

In discussing the rationale for their first proposition, which they later test by using regression analysis, C&T state (p. 59), "The evaluation [of] \( P_i \) calls for an assessment of whether the addition of the importance weights suggested by ZPB (1990) improves the ability of the
SERVQUAL and SERVPERF scales to measure service quality." This statement is misleading because in ZPB (1990), we recommend the use of importance weights merely to compute a weighted average SERVQUAL score (across the five dimensions) as an indicator of a company’s overall SQ gap. Moreover, the importance weights that we use in ZPB (1990) are weights for the dimensions (derived from customer responses to a 100-point allocation question), not for the individual SERVQUAL items. Neither in ZPB (1990) nor in our other work (PBZ 1991; PZB 1988) have we suggested using survey questions to measure the importance of individual items, let alone using weighted individual-item scores in regression analyses as C&T have done. In fact, we would argue that using weighted item scores as independent variables in regression analysis is not meaningful because a primary purpose of regression analysis is to derive the importance weights indirectly (in the form of beta coefficients) by using unweighted or “raw” scores as independent variables. Therefore, using weighted scores as independent variables is a form of “double counting.”

Relationships among SQ, CS, and PI. Figure 2 in C&T’s article shows the structural model they use to examine the interrelationships among SQ, CS, and PI. Their operationalization and testing of this model suffer from several serious problems. C&T’s use of single-item scales to measure SQ, CS, and PI fails to do justice to the richness of these constructs. As already discussed, SQ is a multifaceted construct, even though there is no clear consensus yet on the number of dimensions and their interrelationships. Though a single-item overall SQ measure may be appropriate for examining the convergent validity and predictive power of alternative SQ measures such as SERVQUAL, it is not as appropriate for testing models positing structural relationships between SQ and other constructs such as PI, especially when a multiple-item scale is available. Therefore, in Figure 2, using SERVQUAL or SERVPERF directly to operationalize overall SQ (i.e., \( \eta_2 \)) would have been far better than using a single-item SQ measure as C&T have done. In other words, because C&T’s \( P_2, P_3, \) and \( P_4 \) relate to just the bottom part of their Figure 2, the procedure they used to operationalize \( \xi_1 \) should have been used to operationalize \( \eta_2 \) instead.

C&T’s failure to use multiple-item scales to measure CS and PI, particularly in view of the centrality of these constructs to their discussion and claims, is also questionable. Multiple-item CS scales (e.g., Westbrook and Oliver 1981) are available in our literature and have been used in several customer-satisfaction studies (e.g., Oliver and Swan 1989; Swan and Oliver 1989). Multiple-item scales have also been used to operationalize purchase or behavioral intentions (e.g., Bagozzi 1982; Dodds, Monroe, and Grewal 1991).

Another serious consequence of C&T’s use of single-item measures as indicators for their model’s four latent constructs is the lack of degrees of freedom for a robust test of their model. The four measures yield only six intermeasure correlations (i.e., \( 4 \times 3 / 2 = 6 \)) to estimate the model’s parameters. Because five parameters are being estimated as per Figure 2, only one degree of freedom is available to test the model, a fact that C&T do not report in their LISREL results summarized in Table 5 or their discussion of the results. Insufficient degrees of freedom to estimate the parameters of a structural model will tend to inflate the fit of the data to the model. For example, a model with zero degrees of freedom (i.e., the number of intermeasure correlations equal to the number of model parameters) will fit the observed data perfectly.

Therefore, the meaningfulness of C&T’s observation (p. 63) that “Model 1 (SERVQUAL) had a good fit in two of the four industries ... whereas Model 2 (SERVPERF) had an excellent fit in all four industries,” should be assessed carefully, keeping in mind two important points: (1) The fit values for both Model 1 and Model 2 could be inflated due to insufficient degrees of freedom and (2) The somewhat better fit values for Model 2 could be an artifact of the “shared method variance” between the SERVPERF and the “Overall Service Quality” measures in the model. Because of the questionable meaningfulness of the structural model tests, C&T’s interpreting the test results (p. 63) “as additional support for the superiority of the SERVPERF approach” is debatable as well.

A comparison of the results in C&T’s Tables 3 and 5 reveals several serious inconsistencies and interpretational problems that reiterate the inadequacies of their measures and structural model test. For example, the correlation between the single-item measures of SQ and CS is .8175 (Table 3), implying that the measures are highly collinear and exhibit little or no discriminant validity. Yet C&T seem to ignore this problem and rely solely on the Table 5 results (which themselves are questionable because of the problems described previously) to make strong claims pertaining to the direction of causality between the two constructs (i.e., SQ leads to CS and not vice versa) and the relative influence of the two constructs on PI (i.e., CS has a significant effect in all four contexts and SQ has no significant effect in any context).

Regarding the latter claim, it is instructive to note from Table 3 that CS and SQ have virtually identical correlations with PI (.5272 for SQ and .5334 for CS). The most logical inference on the basis of this fact is that SQ and CS have a similar effect on PI, particularly because SQ and CS themselves are highly correlated. Such an inference is much more plausible and justified than C&T’s conclusion that only CS has a significant effect on PI. That the findings from C&T’s structural-model test led them to this questionable inference is perhaps a manifestation of the problems with their measures (single-item measure, multicollinearity) and their model (questionable causal path, lack of degrees of freedom).

Practical Issues

Arguing in favor of performance-based measures of SQ, C&T state (p. 58), “Practitioners often measure the determinants of overall satisfaction/perceived quality by having customers simply assess the performance of the company’s business processes.” Though the practice of measuring only perceptions is widespread, such a practice does not necessarily...
mean performance-based measures are superior to disconfirmation-based measures. As demonstrated in PBZ (1990), SQ measurements that incorporate customer expectations provide richer information than those that focus on perceptions only. Moreover, executives in companies that have switched to a disconfirmation-based measurement approach tell us that the information generated by this approach has greater diagnostic value.

In a recent study we administered the SERVQUAL scale to independent samples of customers of five nationally known companies (details of the study are in PBZ 1991). Table 1 summarizes perceptions-only and SERVQUAL scores by dimension for each of the five companies. As Table 1 shows, the SERVQUAL scores for each company consistently exhibit greater variation across dimensions than the perceptions-only scores. This pattern of findings suggests that the SERVQUAL scores could be superior in terms of pinpointing areas of deficiency within a company. Moreover, examining only performance ratings can lead to different actions than examining those ratings relative to customer expectations. For example, Ins. Co. 1 might focus more attention on tangibles than on assurance if it relied solely on the perceptions-only scores. This would be a mistake, because the company's SERVQUAL scores show a serious shortfall for assurance and no shortfall for tangibles.

An important question to ask in assessing the practical value of SERVQUAL vis-à-vis SERVPERF is, Are managers who use SQ measurements more interested in accurately identifying service shortfalls or explaining variance in an overall measure of perceived service? (Explained variance is the only criterion on which SERVPERF performs better than SERVQUAL, and, as discussed previously, this could be due to shared method variance.) We believe that managers would be more interested in an accurate diagnosis of SQ problems. From a practical standpoint, SERVQUAL is preferable to SERVPERF in our judgment. The superior diagnostic value of SERVQUAL more than offsets the loss in predictive power.

Response to Teas (1993)

The issues raised by Teas (1993) fall under three main topics: (1) interpretation of the expectations standard, (2) operationalization of this standard, and (3) evaluation of alternative models specifying the SQ construct.

Interpretation of Expectations

A key issue addressed by Teas is the impact of the interpretation of the expectations measure (E) on the meaningfulness of the P-E specification invoked by the SERVQUAL framework. Specifically, on the basis of a series of conceptual and mathematical arguments, Teas concludes that increasing P-E scores may not necessarily reflect continuously increasing levels of perceived quality, as the SERVQUAL framework implies. Though his conclusion has merit, several assumptions underlying his arguments must be reexamined to assess accurately the severity of the problem.

The P-E specification is problematic only for certain types of attributes under certain conditions. As Teas's discussion suggests, this specification is meaningful if the service feature being assessed is a vector attribute—that is, one on which a customer's ideal point is at an infinite level. With vector attributes, higher performance is always better (e.g., checkout speed at a grocery store). Thus, as portrayed under Case A in Figure 1, for vector attributes there is a positive monotonic relationship between P and SQ for a given level of the expectation norm E, regardless of how E is interpreted. This relationship is consistent with the SERVQUAL formulation of SQ. Moreover, customers are likely to consider most of the 22 items in the SERVQUAL instrument to be vector attributes—a point we discuss subsequently and one that Teas acknowledges in footnote 16 of his article.

The P–E specification could be problematic when a service attribute is a classic ideal point attribute—that is, one on which a customer's ideal point is at a finite level and, therefore, performance beyond which will displease the customer (e.g., friendliness of a salesperson in a retail store). However, the severity of the potential problem depends on how the expectations norm E is interpreted. Teas offers two interpretations of E that are helpful in assessing the meaningfulness of the P–E specification: a "classic attitudinal model ideal point" interpretation and a "feasible ideal point" interpretation. If E is interpreted as the classic ideal point (i.e., E = I) and the performance level P exceeds E, the relationship between P and SQ for a given level of E becomes negative monotonic, in contrast to the positive monotonic relationship implied by the SERVQUAL formulation. As Case B in Figure 1 shows, under the classic ideal point interpretation of E the P–E specification is meaningful as long as P is less than or equal to E, but becomes a

### Table 1

<table>
<thead>
<tr>
<th>Service Quality Dimensions</th>
<th>Tel. Co. 1</th>
<th>Ins. Co. 2</th>
<th>Ins. Co. 1</th>
<th>Bank 1</th>
<th>Bank 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>P SQ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tangibles</td>
<td>5.6</td>
<td>5.3</td>
<td>5.6</td>
<td>5.4</td>
<td>5.8</td>
</tr>
<tr>
<td>Reliability</td>
<td>5.1</td>
<td>4.8</td>
<td>5.4</td>
<td>5.1</td>
<td>5.4</td>
</tr>
<tr>
<td>Responsiveness</td>
<td>5.1</td>
<td>5.1</td>
<td>5.6</td>
<td>4.8</td>
<td>5.4</td>
</tr>
<tr>
<td>Assurance</td>
<td>5.4</td>
<td>5.4</td>
<td>5.8</td>
<td>5.2</td>
<td>5.7</td>
</tr>
<tr>
<td>Empathy</td>
<td>5.2</td>
<td>5.1</td>
<td>5.6</td>
<td>4.7</td>
<td>5.2</td>
</tr>
</tbody>
</table>

*The numbers under "P" and "SQ" represent mean perceptions-only and SERVQUAL scores respectively.*
FIGURE 1
Functional Relationship Between Perceived Performance (Pj) and Service Quality (SQ)

Case A: Attribute j is a vector attribute and, therefore, the general form of the function below holds regardless of how the expectations standard (Ej) is defined.

\[ SQ = \begin{cases} 0 & \text{if } P \leq E \text{ and } E < I \text{ is the feasible ideal point,} \text{ or if } P > E < I \text{ is the ideal point,} \\
 0 - (P - E) & \text{if } P > E > I \end{cases} \]

Case B: Attribute j is a classic ideal-point attribute and Ej is interpreted as the classic ideal point (i.e., Ej = Ij).

\[ MQ = -I |P - I| - (E - I) \]

Case C: Attribute j is a classic ideal-point attribute and Ej is interpreted as the feasible ideal point (i.e., Ej < Ij).

\[ MQ = (P - I) - (E - I) \]

Reassessment of Expectations as a Comparison Standard / 117
Teas raises several concerns about the operationalization of E, including a concern we ourselves have raised (PZB 1990) about the use of the word "should" in our original expectations measure (PZB 1988). Because of this concern, we developed a revised expectations measure (E*) representing the extent to which customers believe a particular attribute is "essential" for an excellent service company. Teas wonders if E* is really an improvement over E, our original measure (p. 21-22):

Defining the revised SERVQUAL E* in this way, in conjunction with the P–E measurement specification, suggests high performance scores on essential attributes (high E* scores) reflect lower quality than high performance on attributes that are less essential (low E* scores). It is difficult to envision a theoretical argument that supports this measurement specification.

Teas's argument is somewhat misleading because high performance on an "essential" attribute may not be high enough (from the customer's standpoint) and, therefore, logically can reflect lower quality on that attribute (a key phrase missing from Teas's argument) than equally high performance on a less essential attribute. In fact, as we have argued in our response to C&T, this possibility is an important reason why measuring only perceived service performance can lead to inaccurate assessment of perceived service quality.

The operationalization issue addressed and discussed most extensively by Teas is the congruence (or lack thereof) between the conceptual and operational definitions of the expectation measures. The systematic, comprehensive, and unique approach Teas has used to explore this issue is commendable. His approach involves obtaining ratings on three different comparison norms (E, E*, and I), conducting follow-up interviews of respondents providing non-extreme ratings (i.e., less than 7 on the 7-point scales used), using multiple coders to classify the open-ended responses, and computing indexes of congruence on the basis of relative frequencies of responses in the coded categories.

Though the results from Teas's content analysis of the open-ended responses (summarized in his Appendix B and Tables 4 and 5) are sound and insightful, his interpretation of them is open to question. Specifically, "under the assumption that the feasibility concept is congruent with the excellence norm concept" (p. 25), he concludes that the index of congruence is only 39.4% for E and 35.0% for E*, though he later raises these indexes to 52.4% and 49.1%, respectively, by including several more response categories. On the basis of these values, Teas further concludes (p. 25) that "a considerable portion of the variance in the SERVQUAL E and E* measures is the result of measurement error induced by respondents misinterpreting the scales." As discussed subsequently, both conclusions are problematic.

The purpose of the follow-up interviews in Teas's study was to explore why respondents had lower than the maximum expectation levels—as Teas acknowledges (p. 24), "respondents answered qualitative follow-up questions concerning their reasons for non-extreme (non-7) responses." The issue of how respondents interpreted the expectations questions was not the focus of the follow-up interviews. Therefore, inferring incongruence between intended and interpreted meanings of expectations (i.e., measurement error) on the basis of the open-ended responses is questionable.

Pertinent to the present discussion is a study that we conducted to understand the nature and determinants of customers' service expectations (ZBP 1993, a revised version of the monograph ZBP 1991). In this study, we developed a conceptual model explicating the general antecedents of expectations and how they are likely to influence expectation levels. In the first column of Table 2, we list five general antecedents from our study that are especially likely to lower expectations. Shown in the second column are illustrative open-ended responses from Teas's study that relate to the general antecedents. The correspondence between the two columns of Table 2 is additional evidence that the items in Appendix B of Teas's article reflect reasons for expectation levels rather than interpretations of the expectation measures.

A key assumption underlying Teas's indexes of congruence is that the only open-ended responses congruent with the excellence norm standard are the ones included under the "not feasible," and perhaps "sufficient," categories. The rationale for this critical assumption is unclear. Therefore, what the index of congruence really represents is unclear as well. Given that the open-ended responses are reasons for customers' nonextreme expectations, no response listed in Appendix B of Teas's article is necessarily incom-
TABLE 2
Correspondence Between Antecedent Constructs in ZBP's Expectations Model and Follow-up Responses Obtained by Teas

<table>
<thead>
<tr>
<th>Antecedent Constructs from ZBP (1993)</th>
<th>Illustrative Follow-up Responses from Teas (1993) Reflecting Each Construct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personal Needs: States or conditions essential to the physical or psychological well-being of the customer.</td>
<td>Not necessary to know exact need.</td>
</tr>
<tr>
<td>Perceived Service Alternatives: Customers' perceptions of the degree to which they can obtain better service through providers other than the focal company.</td>
<td>All of them have pretty much good hours.</td>
</tr>
<tr>
<td>Self-Perceived Service Role: Customers' perceptions of the degree to which they themselves influence the level of service they receive.</td>
<td>I've gone into the store not knowing exactly myself what I want.</td>
</tr>
<tr>
<td>Situational Factors: Service-performance contingencies that customers perceive as beyond the control of the service provider.</td>
<td>Can't wait on everybody at once.</td>
</tr>
<tr>
<td>Past Experience: Customers' previous exposure to the service that is relevant to the focal service.</td>
<td>You don't get perfection in these kind of stores.</td>
</tr>
</tbody>
</table>

The response-category frequencies in Teas's Tables 4 and 5 can be used to compute a different type of "index of congruence" that reflects the extent to which reasons for nonextreme expectations ratings are compatible with a vector-attribute assumption. This index is given by

\[ 100 - \Sigma \text{Percentages of responses in the "Ideal" categories} \]

From Table 4, this index is 97.6% for E and 96.6% for E*, implying strong support for the vector-attribute assumption. From Table 5, which shows the distribution of respondents rather than responses, this index is 91.6% for E and 90.8% for E*. These are conservative estimates because Table 5 allows for multiple responses and, therefore, some respondents could be included in more than one category percentage in the summation term of the previous expression, thereby inflating the term. Therefore, for a vast majority of the respondents, the vector-attribute assumption is tenable for all service attributes included in Teas's study.

**Evaluation of Alternative Service Quality Models**

Using the empirical results from his study, Teas examines the criterion and construct validity of eight different SQ models: unweighted and weighted versions of the SERVQUAL (P-E), revised SERVQUAL (P-E*), normed quality (NQ), and evaluated performance (EP) formulations. On the basis of correlation coefficients reflecting criterion and construct validity, he correctly concludes that his EP formulation outperforms the other formulations. However, apart from the statistical significance of the correlation differences reported in his Tables 7 and 8, several additional issues are relevant for assessing the superiority of the EP formulation.

In introducing these issues we limit our discussion to the unweighted versions of just the P-E, P-E*, and EP formulations for simplicity and also because (1) for all formulations, the unweighted version outperformed its weighted counterpart on both criterion and construct validity and (2) the NQ formulation, in addition to having the lowest criterion and construct validity among the four formulations, is not one that we have proposed. From Teas's equation 6 and his discussion of it, the unweighted EP formulation for any service is as follows (the service subscript "i" is suppressed for notational simplicity):

\[ \text{EP} = -1 \left( \sum_{j=1}^{m} | P_j - I_j | \right) \]

where "m" is the number of service attributes. This expression assumes that the attributes are classic ideal point attributes as evidenced by the fact that the theoretical maximum value for EP is zero, which occurs when \( P_j = I_j \) for all attributes j. As discussed in the preceding section, the evidence from Teas's follow-up responses strongly suggests that the service attributes included in the study are much more likely to be vector attributes than classic ideal point attributes. Therefore, the conceptual soundness of the EP formulation is open to question.

Teas's findings pertaining to nonextreme ratings on his \( I_j \) scale raise additional questions about the soundness of
the classic ideal point attribute assumption invoked by the EP formulation. As his Table 4 reveals, only 151 non-7 ratings (13% of the total 1200 ratings) were obtained for the ideal point (I) scale, prompting him to observe (p. 26), "It is interesting to note, however, that the tendency for nonextreme responses was much smaller for the ideal point measure than it was for the expectations measures." Ironically, this observation, though accurate, appears to go against the classic ideal point attribute assumption because the fact that only 13% of the ratings were less than 7 seems incompatible with that assumption. If the service features in question were indeed classic ideal point attributes, one would expect the ideal-point levels to be dominated by nonextreme ratings.

Even in instances in which nonextreme ratings were obtained for the ideal point (I) measure, a majority of the reasons for those ratings (as reflected by the open-ended responses to the follow-up questions) do not imply the presence of finite ideal-point levels for the attributes investigated. From Teas's Table 4, only 40.5% of the reasons are classified under "Ideal" (including double classifications with "Ideal" as one of the two categories). Teas does acknowledge the presence of this problem, though he labels it as a lack of discriminant validity for the ideal-point measure.

Finally, it is worth noting that a semantic-differential scale was used for the "P" and "I" measures in the EP formulation, whereas a 7-point "strongly disagree"—"strongly agree" scale was used for the three measures (P, E, and E*) in the SERVQUAL and revised SERVQUAL formulations. Therefore, the observed criterion and construct validity advantages of the EP formulation could be an artifact of the scale format differences, especially because the correlation differences, though statistically significant, are modest in magnitude: For the unweighted formulations, the mean criterion validity correlation difference between EP and the two SERVQUAL formulations is .077 ([(.081 + .073)/2]; the corresponding mean construct validity correlation difference is .132 ([(.149 + .114)/2]). However, because these correlation differences are consistently in favor of EP, the semantic-differential scale format could be worth exploring in the future for measuring the SERVQUAL perceptions and expectations components.

Implications and Directions for Further Research

C&T and Teas have raised important, but essentially different, concerns about the SERVQUAL approach for measuring SQ. C&T's primary concerns are that SERVQUAL's expectations component is unnecessary and the instrument's dimensionality is problematic. In contrast, Teas questions the meaningfulness of SERVQUAL's P-E specification and wonders whether an alternate specification—the "evaluated performance" (EP) specification—is superior. We have attempted to demonstrate that the validity and alleged severity of many of the issues raised are debatable. Nevertheless, several key unresolved issues emerging from this exchange offer a challenging agenda for further research on measuring SQ.

Measurement of Expectations

Of the psychometric concerns raised by C&T about SERVQUAL, the only one that has consistent support from their empirical findings is the somewhat lower predictive power of the P–E measure relative to the P-only measure. Our own findings (PBZ 1991) as well as those of other researchers (e.g., Babakus and Boller 1992) also support the superiority of the perceptions-only measure from a purely predictive-validity standpoint. However, as argued previously, the superior predictive power of the P-only measure must be balanced against its inferior diagnostic value. Therefore, formally assessing the practical usefulness of measuring expectations and the trade-offs involved in not doing so is a fruitful avenue for additional research. More generally, a need and an opportunity exist for explicitly incorporating practical criteria (e.g., potential diagnostic value) into the traditional scale-assessment paradigm that is dominated by psychometric criteria. Perreault (1992, p. 371), in a recent commentary on the status of research in our discipline, issues a similar call for a broadened perspective in assessing measures:

Current academic thinking tends to define measures as acceptable or not primarily based on properties of the measure. However, research that focuses on defining what is "acceptable" in terms of a measure's likely impact on interpretation of substantive relationships, or the conclusions that might be drawn from those relationships, might provide more practical guidelines—not only for practitioners but also for academics.... There is probably a reasonable 'middle ground' between the 'all or nothing' contrast that seems to have evolved.

The most appropriate way to incorporate expectations into SQ measurement is another area for additional research. Though the SERVQUAL formulation as well as the various SQ models evaluated by Teas use a difference-score formulation, psychometric concerns have been raised about the use of difference scores in multivariate analyses (for a recent review of these concerns, see Peter, Churchill, and Brown 1993). Therefore, there is a need for studies comparing the paired-item, difference-score formulation of SERVQUAL with direct formulations that use single items to measure customers' perceptions relative to their expectations.

Brown, Churchill, and Peter (1993) conducted such a comparative study and concluded that a nondifference score measure had better psychometric properties than SERVQUAL. However, in a comment on their study, we have raised questions about their interpretations and demonstrated that the claimed psychometric superiority of the nondifference score measure is questionable (PBZ 1993). Likewise, as our response to C&T's empirical study shows, the psychometric properties of SERVQUAL's difference score formulation are by and large just as strong as those of its P-only component. Because the cumulative empirical evidence about difference score and nondifference score measures has not established conclusively the superiority of one measure over the other, additional research in this area is
warranted. Such research also should explore alternatives to the Likert-scale format ("strongly agree"—"strongly disagree") used most frequently in past SERVQUAL studies. The semantic-differential scale format used by Teas to operationalize his EP formulation is an especially attractive alternative—as we pointed out previously, this scale format is a plausible explanation for the consistently superior performance of the EP formulation.

**Dimensionality of Service Quality**

As already discussed, C&T’s conclusion that the 22 SERVQUAL items form a unidimensional scale is unwarranted. However, replication studies have shown significant correlations among the five dimensions originally derived for SERVQUAL. Therefore, exploring why and how the five service quality dimensions are interrelated (e.g., could some of the dimensions be antecedents of the others?) is a fertile area for additional research. Pursuing such a research avenue would be more appropriate for advancing our understanding of SQ than hastily discarding the multidimensional nature of the construct.

The potential drawback pointed out previously of using factor analysis as the sole approach for assessing the dimensionality of service quality implies a need for developing other approaches. As we have suggested elsewhere (PBZ 1991), an intriguing approach is to give customers definitions of the five dimensions and ask them to sort the SERVQUAL items into the dimensions only on the basis of each item’s content. The proportions of customers “correctly” sorting the items into the five dimensions would reflect the degree to which the dimensions are distinct. The pattern of correct and incorrect classifications also could reveal potentially confusing items and the consequent need to reword the items and/or dimensional definitions.

**Relationship Between Service Quality and Customer Satisfaction**

The direction of causality between SQ and CS is an important unresolved issue that C&T’s article addresses empirically and Teas’s article addresses conceptually. Though C&T conclude that SQ leads to CS, and not vice versa, the empirical basis for their conclusion is questionable because of the previously discussed problems with their measures and analysis. Nevertheless, there is a lack of consensus in the literature and among researchers about the causal link between the two constructs. Specifically, the view held by many service quality researchers that CS leads to SQ is at odds with the causal direction implied in models specified by CS researchers. As Teas’s discussion suggests, these conflicting perspectives could be due to the global or overall attitude focus in most SQ research in contrast to the transaction-specific focus in most CS research. Adding to the complexity of this issue, practitioners and the popular press often use the terms service quality and customer satisfaction interchangeably. An integrative framework that reflects and reconciles these differing perspectives is sorely needed to divert our attention from merely debating the causal direction between SQ and CS to enhancing our knowledge of how they interrelate.

---

**FIGURE 3**

Components of Transaction-Specific Evaluations

| Evaluation of Service Quality (SQ) | Transaction Satisfaction (TSAT) |
| Evaluation of Product Quality (SQ) | |
| Evaluation of Price (P) | |

Teas offers a thoughtful suggestion that could be the foundation for such a framework (p. 30):

One way to integrate these two causal perspectives is to specify two perceived quality concepts—transaction-specific quality and relationship quality—and to specify perceived transaction-specific quality as the transaction-specific performance component of contemporary consumer satisfaction models. This implies that transaction-specific satisfaction is a function of perceived transaction-specific performance quality. Furthermore, transaction-specific satisfaction could be argued to be a predictor of perceived long-term relationship quality.

A key notion embedded in Teas’s suggestion is that both SQ and CS can be examined meaningfully from both transaction-specific as well as global perspectives, a viewpoint embraced by other researchers as well (e.g., Dabholkar 1993). Building on this notion and incorporating two other potential antecedents of CS—product quality and price—we propose (1) a transaction-specific conceptualization of the constructs’ interrelationships and (2) a global framework reflecting an aggregation of customers’ evaluations of multiple transactions.

Figure 3 portrays the proposed transaction-specific conceptual model. This model posits a customer’s overall satisfaction with a transaction to be a function of his or her assessment of service quality, product quality, and price. This conceptualization is consistent with the "quality leads to satisfaction" school of thought that satisfaction researchers often espouse (e.g., Reidenbach and Sandifer-Smith 1990; Woodside, Frey, and Daly 1989). The two separate quality-evaluation antecedents capture the fact that virtually all market offerings possess a mix of service and product features and fall along a continuum anchored by "tangible-product or good quality.

---

2 Though the term "product" theoretically can refer to a good (i.e., tangible product) or a service (i.e., intangible product), here we use it only in the former sense. Thus, "product quality" in Figure 3 and elsewhere refers to tangible-product or good quality.
FIGURE 4 Components of Global Evaluations

<table>
<thead>
<tr>
<th>Transaction 1</th>
<th>Transaction n</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQ₁</td>
<td>SQₙ</td>
</tr>
<tr>
<td>PQ₁</td>
<td>PQₙ</td>
</tr>
<tr>
<td>P₁</td>
<td>Pₙ</td>
</tr>
<tr>
<td>TSAT₁</td>
<td>TSATₙ</td>
</tr>
</tbody>
</table>

Global Impressions about Firm

*Satisfaction  
*Service Quality  
*Product Quality  
*Price

In Figure 4, we present our proposed global framework. It depicts customers’ global impressions about a firm stemming from an aggregation of transaction experiences. We posit global impressions to be multifaceted, consisting of customers’ overall satisfaction with the firm as well as their overall perceptions of the firm’s service quality, product quality, and price. This framework, in addition to capturing the notion that the SQ and CS constructs can be examined at both transaction-specific and global levels, is consistent with the “satisfaction (with specific transactions) leads to overall quality perceptions” school of thought embraced by service quality researchers (e.g., Bitner 1990; Bolton and Drew 1991a, b; Carman 1990). The term “transaction” in this framework can be used to represent an entire service episode (e.g., a visit to a fitness center or barber shop) or discrete components of a lengthy interaction between a customer and firm (e.g., the multiple encounters that a hotel guest could have with hotel staff, facilities, and services).

The SERVQUAL instrument, in its present form, is intended to ascertain customers’ global perceptions of a firm’s service quality. In light of the proposed framework, modifying SERVQUAL to assess transaction-specific service quality is a useful direction for further research. The proposed framework also raises other issues worth exploring. For example, how do customers integrate transaction-specific evaluations in forming overall impressions? Are some transactions weighted more heavily than others because of “primacy” and “recency” type effects or satisfaction/dissatisfaction levels exceeding critical thresholds? Do transaction-specific service quality evaluations have any direct influence on global service quality perceptions, in addition to the posited indirect influence (mediated through transaction-specific satisfaction)? How do the four facets of global impressions relate to one another? For example, are there any halo effects? Research addressing these and related issues is likely to add significantly to our knowledge in this area.

**Specification of Service Quality**

A primary contribution of Teas’s article is that it highlights the potential problem with the P-E specification of SQ when the features on which a service is evaluated are not vector attributes. Though we believe that the collective evidence from Teas’s study strongly supports the vector-attribute assumption, the concern he raises has implications for correctly specifying SQ and for further research on service attributes. Teas’s EP specification, in spite of its apparent empirical superiority over the other models he tested, suffers from several deficiencies, including its being based on the questionable assumption that all service features are classic ideal point attributes. A mixed-model specification that assumes some features to be vector attributes and others to be classic ideal point attributes would be conceptually more appropriate (cf. Green and Srinivasan 1978). The flowchart in Figure 2 implies the following mixed-model specification that takes into account service-attribute type as well as interpretation of the comparison standard E:

\[
SQ = \sum_{j=1}^{n} [(P_j - E_j)D_{ij} - (P_j - E_j)D_{3j}] + [(E_j - E) - (P_j - I_j)]D_{4j}
\]

where

\[
D_{ij} = 1 \text{ if } j \text{ is a vector attribute, or if it is a classic ideal point attribute and } P_j \leq I_j;
\]

\[
= 0 \text{ otherwise.}
\]

\[
D_{3j} = 1 \text{ if } j \text{ is a classic ideal point attribute and } E_j \text{ is interpreted as the classic ideal point (i.e., } E_j = I_j \text{) and } P_j > I_j;
\]

\[
= 0 \text{ otherwise.}
\]

\[
D_{4j} = 1 \text{ if } j \text{ is a classic ideal point attribute and } E_j \text{ is interpreted as the feasible ideal point (i.e., } E_j < I_j \text{) and } P_j > I_j;
\]

\[
= 0 \text{ otherwise.}
\]

Operationalizing this specification has implications for data collection and analysis in SQ research. Specifically, in addition to obtaining data on customers’ perceptions (P_j) and comparison standards (I_j and/or E_j), one should ascer-
and comparison standards (I_j and/or E_j), one should ascertain whether customers view each service feature as a vector or classic ideal point attribute. Including a separate battery of questions is one option for obtaining the additional information, though this option will lengthen the SQ survey. An alternate approach is to conduct a separate study (e.g., focus group interviews in which customers discuss whether "more is always better" on each service feature) to preclassify the features into the two attribute categories.

The expression for SQ is implicitly an individual-level specification because a given attribute j can be viewed differently by different customers. In other words, for the same attribute j, the values of D_j, D_2j, and D_3j can vary across customers. This possibility could pose a challenge in conducting aggregate-level analyses across customers similar to that faced by researchers using conjoint analysis (Green and Srinivasan 1990). An interesting direction for further research is to investigate whether distinct customer segments with homogeneous views about the nature of service attributes exist. Identification of such segments, in addition to offering managerial insights for market segmentation, will reduce the aforementioned analytical challenge in that the expression for SQ can be treated as a segment level, rather than an individual-level specification.

**Conclusion**

The articles by C&T and Teas raise important issues about the specification and measurement of SQ. In this article, we attempt to reexamine and clarify the key issues raised. Though the current approach for assessing SQ can and should be refined, abandoning it altogether in favor of the alternate approaches proffered by C&T and Teas does not seem warranted. The collective conceptual and empirical evidence casts doubt on the alleged severity of the concerns about the current approach and on the claimed superiority of the alternate approaches. Critical issues remain unresolved and we hope that the research agenda we have articulated will be helpful in advancing our knowledge.

---

**REFERENCES**


