

Alternative Scales for Measuring Service Quality: A Comparative Assessment Based on Psychometric and Diagnostic Criteria

A. PARASURAMAN

University of Miami


VALARIE A. ZEITHAML

Principal, Partners for Service Excellence

LEONARD L. BERRY

Texas A&M University

View metadata, citation and similar papers at core.ac.uk

brought to you by  CORE

provided by Carolina Digital Repository

Service quality measurement is an area of growing interest to researchers and managers. It is also an area characterized by debate concerning the need for measuring customer expectations and how they should be measured. Building on a synthesis of the extant literature on customer expectations and service quality measurement, this article identifies unresolved issues and develops three alternative questionnaire formats to address them. It then discusses an empirical study that evaluated the three formats in four different sectors. The article concludes with practical implications and directions for further research stemming from the study's findings.

The important role played by expectations in customers' evaluations of services has been acknowledged in the service quality literature (e.g., Bolton and Drew 1991a,b; Boulding, Kalra, Staelin and Zeithaml 1993; Brown and Swartz 1989; Parasuraman, Zeithaml, and Berry 1985, 1988; Zeithaml, Berry and Parasuraman 1993) and in the customer satisfaction/dissatisfaction (CS/D) literature (e.g., Cadotte, Woodruff and Jenkins 1987; Oliver 1980; Tse and Wilton 1988; Woodruff, Clemons, Schumann, Gardial and Burns 1991; Yi 1990). Researchers generally agree that expectations serve as reference points in customers' assessment of service performance. However, there is debate about how best to incorporate expectations into service quality measurement (Babakus and Boller 1992; Brown, Churchill and Peter 1993; Carman 1990) and the empirical usefulness of expectations in terms of their explanatory power (Cronin and Taylor 1992).

In this paper, we first provide a brief synthesis of the extant literature on key conceptual and methodological issues pertaining to expectations in measuring service quality. We then discuss a multi-sector study designed to address the unresolved issues and add to our knowledge of service quality measurement. We conclude the paper with a discussion of managerial implications and directions for future research.

EXPECTATIONS AS COMPARISON STANDARDS

Service quality researchers have generally viewed expectations as *normative* standards—i.e., customers' beliefs about what a service provider *should* offer. The dominant view among CS/D researchers is that expectations are *predictive* standards—i.e., what customers feel a service provider *will* offer. CS/D researchers have also proposed and defended other comparison standards (for a comprehensive review, see Woodruff et al. 1991). In an attempt to better understand and synthesize the various comparison standards, Zeithaml, Berry and Parasuraman (1993) combined insights from past conceptualizations with findings from a multi-sector study to develop an integrative model of customers' service expectations. This model posits service expectations as existing at two different levels that customers use as comparison standards in assessing service quality:

Desired Service: The level of service representing a blend of what customers believe “can be” and “should be” provided.

Adequate Service: The minimum level of service customers are willing to accept.

Separating these two levels is a *zone of tolerance* that represents the range of service performance a customer would consider satisfactory. This expanded conceptualization of expectations served as the foundation for the empirical study to be discussed later.

Treatment of Expectations in Past Efforts to Measure Service Quality

Parasuraman et al. (1985), based on exploratory research to understand the construct of service quality and its determinants, defined service quality as the degree of discrepancy between customers' normative expectations for the service and their perceptions of the service performance. The researchers also uncovered a comprehensive set of service attributes that customers might use as criteria in assessing service performance. Subsequent empirical research based on the exploratory work produced SERVQUAL, a 22-item scale for measuring service quality along five dimensions: *reliability*, *responsiveness*, *assurance*, *empathy*, and *tangibles* (for details about SERVQUAL's structure and definitions of its dimensions, see Parasuraman et al. 1988). SERVQUAL operationalizes service quality by subtracting customers' expectation scores from their perception scores on the 22 items. While the original SERVQUAL instrument has been revised and refined, its basic content, structure, and length have remained intact (Parasuraman, Berry, and Zeithaml 1991).

The SERVQUAL instrument has been productively used for measuring service quality in many proprietary studies. It has also served as the basis for measurement approaches used in published studies examining service quality in a variety of contexts—e.g., real estate brokers (Johnson, Dotson, and Dunlop 1988); physicians in private practice (Brown and Swartz 1989); public recreation programs (Crompton and Mackay 1989); a dental school patient clinic, a business school placement center, and a tire store (Carman 1990); motor carrier companies (Brensinger and Lambert 1990); an accounting firm (Bojanic 1991); discount and department stores (Finn and Lamb 1991; Teas 1993); a gas and electric utility company (Babakus and Boller 1992); hospitals (Babakus and Mangold 1992; Carman 1990); banking, pest control, dry cleaning, and fast food (Cronin and Taylor 1992); higher education (Boulding, Kalra, Staelin and Zeithaml 1993; Ford, Joseph and Joseph 1993).

While the SERVQUAL instrument has generated considerable interest in service quality measurement, it has also raised questions about the need to measure expectations (e.g., Babakus and Mangold 1992; Cronin and Taylor 1992, 1994), the interpretation and operationalization of expectations (e.g., Teas 1993, 1994), the reliability and validity of SERVQUAL's difference-score formulation (e.g., Babakus and Boller 1992; Brown et al. 1993), and SERVQUAL's dimensionality (e.g., Carman 1990; Finn and Lamb 1991). In response to these questions, SERVQUAL's developers have presented counterarguments, clarifications, and additional evidence to reaffirm the instrument's psychometric soundness and practical value (Parasuraman et al. 1991, 1993; Parasuraman, Zeithaml and Berry 1994a). Major unresolved issues emerging from this ongoing debate include the empirical vs. diagnostic value of expectations in service quality measurement, the relative merits and demerits of SERVQUAL (i.e., difference-score) vs. direct (i.e., non-difference score) formulations of the perception-expectation gap, and the dimensionality of the instrument's items.

PURPOSE OF THE STUDY

The primary objective of the study reported here was to compare alternative service-quality measurement scales on psychometric and diagnostic criteria to address the unresolved methodological issues. The study also had a secondary objective: to incorporate the expanded conceptualization of expectations into the alternative scale formats. Specific study-related issues pertaining to the two objectives are outlined next.

Alternative Measurement Formats

Operationalizing any construct as a difference between two other constructs has been questioned for psychometric reasons (for a recent review of the concerns raised, see Peter, Churchill, and Brown 1993). SERVQUAL's difference-score formulation has also been questioned for the same reasons (Babakus and Boller 1992; Brown et al. 1993). Critics of difference scores have suggested that direct (i.e., non-difference score) measures of the

perception-expectation gap will be psychometrically superior (e.g., Carman 1990; Peter et al. 1993). Scales directly measuring perceived performance relative to expectations have also been found to be less biased and more useful than scales merely measuring performance (Devlin, Dong and Brown 1993). But the available empirical evidence comparing SERVQUAL and direct measures of service quality has not conclusively established that the alleged psychometric problems are present in SERVQUAL or that direct measures are superior (Parasuraman et al. 1993; Parasuraman et al. 1994a). Thus, there is a need for a more comprehensive examination of SERVQUAL and direct measures on psychometric as well as practical criteria. As discussed later, we developed and tested three alternative questionnaire formats to facilitate such a comparative evaluation.

Expanded Conceptualization of Expectations

The SERVQUAL instrument's expectations statements relate to the service level that customers believe they *should* get from the service provider. As such, SERVQUAL's expectations component reflects the *desired service* construct defined earlier. However, to incorporate the recently revised conceptualization of expectations (Zeithaml et al. 1993) we modified SERVQUAL's structure in the present study to capture not only the discrepancy between perceived service and desired service—labeled as *measure of service superiority* (or MSS)—but also the discrepancy between perceived service and adequate service—labeled as *measure of service adequacy* (or MSA).¹

RESEARCH METHOD

In designing the study we first sought advice from a panel of five leading academics with expertise in measurement/scale-development. We assembled the panel members for a group discussion of measurement and research-design issues pertaining to the study's general objectives. We presented to the panel a research proposal that included several scaling formats for operationalizing service quality, a sampling plan for gathering data using the different formats, and evaluative criteria for a comparative assessment of the formats. The panel critically examined the proposal and offered suggestions for strengthening it. We finetuned our proposed research design based on the panel's suggestions.

The refined research design included three alternative questionnaire formats, one incorporating the difference-score formulation and the other two incorporating direct measures of service quality. Each of these formats also incorporated the expanded conceptualization of expectations to obtain scores for the measures of service superiority (MSS) and service adequacy (MSA) defined earlier. The three alternative service quality measurement formats, illustrated in Appendix 1 and included in the pretest questionnaires, were:

1. *Three-Column Format.* This format generates separate ratings of desired, adequate, and perceived service with three identical, side-by-side scales. It requires computing

the perceived-desired and the perceived-adequate differences to quantify MSS and MSA, respectively. Thus, its operationalization of service quality is similar to that of SERVQUAL although it does not repeat the battery of items.

2. *Two-Column Format.* In contrast to SERVQUAL, this format generates *direct* ratings of the service-superiority and service-adequacy gaps (i.e., MSS and MSA scores) with two identical, side-by-side scales.
3. *One-Column Format.* This format also generates direct ratings of the service-superiority and service-adequacy gaps. However, the questionnaire is split into two parts, with Part I containing one set of scales for MSS and Part II containing the same set of scales for MSA. Thus, this format involves repeating the battery of items as in SERVQUAL.

All three formats contained the 22 attributes in the most recent version of SERVQUAL (Parasuraman et al. 1991). However, several minor modifications were made to the scale items. First, three of the 22 attribute statements were revised to eliminate redundancies and improve clarity (for additional details see Parasuraman, Zeithaml and Berry 1994b). Second, to accommodate the expanded conceptualization of expectations, and to achieve consistency in item wording across the three formats, the attribute statements were abbreviated as illustrated by the sample item in Appendix 1. Third, the response scale was changed from a 7-point to a 9-point scale to offer respondents a wider range of rating choices in view of the need to capture two different expectation levels.

Questionnaire Pretests and Refinements

Three versions of the pretest questionnaire were prepared. Each version began with a service quality section containing one of the three measurement formats described earlier. Two questions were then inserted to assess respondent ease and confidence in answering this section. Subsequent sections contained questions to measure overall service quality and value perceptions and respondents' demographics.

The questionnaires were refined through two stages of pretesting. The first stage involved obtaining respondent reactions through two focus group interviews. Banking services, familiar to all focus group participants, was chosen as the context for discussing the questionnaires. Based on the focus group feedback the directions were shortened, the sequence of the columns/parts pertaining to the adequate and desired service levels were reversed, and "no opinion" response options were added (further discussion of these changes is available in Parasuraman et al. 1994b).

In the second stage of pretesting, each of the three modified questionnaires was field tested with a separate sample of 300 customers of a retail chain, one of four nationally-known companies that sponsored and participated in the study (the other three companies were a computer manufacturer, an auto insurer, and a life insurer). The same mail survey procedures planned for the main study were used in the field test.

The field-test response rates were 13%, 16%, and 12% for the one-column, two-column, and three-column formats, respectively. These low response rates might be attributable to

the field test's being conducted during the late November and December holiday season. However, questionnaire length and appearance might also have contributed to the low response rates. Aided by suggestions of researchers in the sponsor companies, we simplified and condensed the questionnaire (without changing the meaning of the constructs being measured) and improved its layout.

The average ratings on 9-point scales used in the field test to measure respondents' ease in completing the questionnaires (1 = very difficult, 9 = very easy) and their confidence in the meaningfulness of their answers (1 = not at all confident, 9 = very confident) were as follows:

	Ease	Confidence
One-column format questionnaire	6.9	7.4
Two-column format questionnaire	5.4	6.2
Three-column format questionnaire	6.6	7.4

As these results reveal, the one-column and three-column formats fared considerably better than the two-column format.

Furthermore, for a number of SERVQUAL items in the two-column format the mean MSS (i.e., perceived service relative to desired service) score was *higher* than the mean MSA (i.e., perceived service relative to adequate service) score—a logical inconsistency since service performance relative to desired service (a higher standard) cannot exceed the same performance level relative to adequate service (a lower standard). Similar inconsistencies also surfaced in the one-column format questionnaire. In contrast, mean scores obtained from the three-column format questionnaire (which does not involve direct assessments of MSS and MSA) did not reveal any inconsistencies—i.e., as expected, for each SERVQUAL item the mean desired service rating exceeded the mean adequate service rating. Thus, when respondents were asked to directly assess MSS and MSA, the apparent complexity of the task seemed to have clouded the distinction between adequate and desired service.

The confusion evidenced by the empirical results in the field test was consistent with the difficulty expressed by some focus group participants in distinguishing between “performance compared to my desired service level” and “performance compared to my adequate service level.” Therefore, in the main study, we substituted “*minimum service*” for “*adequate service*” in the directions and the scale labels to sharpen the distinction between the two comparison standards.

We also decided to drop the set of 22 MSA items (i.e., Part II) from the one-column format questionnaire for several reasons. First, focus group participants and company executives had expressed concern about repeating the same battery of items. Second, the response rate for this version was lower than for the two-column format questionnaire that obtained the same information (i.e., direct measures of MSA and MSS). Third, by retaining just the battery of 22 MSS items, this version would be the direct-measure equivalent of the current SERVQUAL that uses a difference-score operationalization.

Directions and scale labels for the three questionnaire versions used in the main study are illustrated in Appendix 2. The full battery of revised SERVQUAL items is shown in Table 1.

TABLE 1**SERVQUAL Battery**

Reliability

1. Providing services as promised.
2. Dependability in handling customers' service problems.
3. Performing services right the first time.
4. Providing services at the promised time.
5. Maintaining error-free records.

Responsiveness

6. Keeping customers informed about when services will be performed.
7. Prompt service to customers.
8. Willingness to help customers
9. Readiness to respond to customers' requests.

Assurance

10. Employees who instill confidence in customers.
11. Making customers feel safe in their transactions.
12. Employees who are consistently courteous.
13. Employees who have the knowledge to answer customer questions.

Empathy

14. Giving customers individual attention.
15. Employees who deal with customers in a caring fashion.
16. Having the customer's best interest at heart.
17. Employees who understand the needs of their customers.
18. Convenient business hours.

Tangibles

19. Modern equipment.
 20. Visually appealing facilities.
 21. Employees who have a neat, professional appearance.
 22. Visually appealing materials associated with the service.
-

Sample Design and Mail Survey

Based on the field test results, the mail-out sample size for the main study was set at 800 customers per questionnaire version per company. This sample size was expected to yield a sufficient number of responses for anticipated multivariate analyses even if response rates did not improve over the field test results.²

The sample for the main study included business customers (of the computer manufacturer) as well as end customers (of the retail chain, auto insurer, and life insurer). The four sponsor companies generated mailing lists from their current customer bases. The retail chain, auto insurer, and life insurer provided samples of 2,400 customers each. The computer manufacturer provided a larger sample of 5,270 customers because it wanted to conduct on

TABLE 2

Sample Sizes and Response Rates by Company and Questionnaire Format

<i>Company</i>	<i>No. Sent</i>	<i>No. Received</i>	<i>% Received</i>
Computer Manufacturer			
One-Column Format	1,756	580	33
Two-Column Format	1,757	488	28
Three-Column Format	<u>1,757</u>	<u>498</u>	<u>28</u>
Total	5,270	1,566	30
Retail Chain			
One-Column Format	800	180	23
Two-Column Format	800	154	19
Three-Column Format	<u>800</u>	<u>188</u>	<u>24</u>
Total	2,400	522	22
Auto Insurer			
One-Column Format	800	205	26
Two-Column Format	800	172	22
Three-Column Format	<u>800</u>	<u>191</u>	<u>24</u>
Total	2,400	568	24
Life Insurer			
One-Column Format	800	170	21
Two-Column Format	800	111	14
Three-Column Format	<u>800</u>	<u>132</u>	<u>17</u>
Total	2,400	413	17
Total			
One-Column Format	4,156	1,135	27
Two-Column Format	4,157	925	22
Three-Column Format	<u>4,157</u>	<u>1,009</u>	<u>24</u>
Total	12,470	3,069	25

its own a detailed, segment-by-segment analysis of expectations after the completion of the main study. Each company's sample was randomly divided into three equal subsamples, one for each questionnaire version. A cover letter and postage-paid return envelope accompanied the questionnaires. The cover letter was on company letterhead and was signed by a senior company official. A reminder post card was sent two weeks after mailing the questionnaires. Respondents returned the completed questionnaires to a marketing research firm hired to assist with data collection and coding.

The content, layout, and appearance of each questionnaire version were the same across the four companies except for minor wording changes called for by contextual differences. For example, "policyholders" was used instead of "customers" in the questionnaires for the two insurance companies. Likewise, since the computer manufacturer and retail chain primarily sell goods rather than services, appropriate statements or phrases suggested by company executives were added to the directions to clarify the meaning of "quality of service"—e.g., for the computer manufacturer: "When the questionnaire mentions quality of service or [company name] service, it means how [company name] 'serves' its customers overall. Do not limit your opinions to maintenance service alone;" for the retail chain: "Please think about the quality of service [company name] offers in its stores . . ."

Response Rates and Composition

Table 2 contains a breakdown of the responses by company and questionnaire type. The questionnaire changes made after the field test apparently contributed to the improved overall response rate of 25%, considerably better than the field test response rates of 12% to 16%. Notably, for the retail chain (which participated in both the field test and the main study), the response rates for all three formats was higher in the main study than in the field test.

Demographic profiles of the respondent samples were reviewed by managers in the respective companies and considered to be representative of their customer bases. Formal statistical testing of the samples' representativeness was not possible because demographic information on the entire customer bases was not readily available. Moreover, because the primary purpose of this study was a comparative evaluation of the three formats, we felt it was more important to formally examine the compositional similarity of the three subsamples within each company. We conducted such an examination, the details of which are discussed in Section 1 of Appendix 2 (immediately following the three questionnaire formats). The results of this examination confirmed that the subsamples were similar. Thus, one can rule out sample differences as a plausible explanation for the across-format psychometric and substantive differences to be discussed in the following sections.

COMPARISON OF ALTERNATIVE FORMATS

A variety of criteria were used to assess the performance of the three questionnaire formats. These criteria pertained to the service quality scales' factor structure, reliability, validity, and diagnostic value.

Factor Structure and Reliability

To verify the dimensionality and grouping of the 22 modified SERVQUAL items, we factor analyzed the following sets of scores separately for each of the four companies:

1. Direct measures of MSS (i.e., perceptions relative to desired service) obtained from the one-column format questionnaire.
2. Direct measures of MSS and MSA (i.e., perceptions relative to adequate service) obtained from the two-column format questionnaire.
3. Difference-score measures of MSS and MSA obtained from the three-column format questionnaire.

Because SERVQUAL was hypothesized to have five distinct but correlated dimensions (Parasuraman et al. 1988; Parasuraman et al. 1991), in each factor analysis a five-factor solution was obtained and subjected to oblique rotation. Moreover, to assess the internal consistency of the *a priori* grouping of the 22 items into the five dimensions (Table 1),

TABLE 3

Reliability Coefficients (Alphas) for Service Quality Dimensions

Company	No. of Items	Questionnaire Format		
		One-Column	Two-Column	Three-Column ^a
Computer Manufacturer				
Reliability	5	.91	.91	.87 (.83)
Responsiveness	3	.87	.89	.84 (.81)
Assurance	4	.86	.88	.81 (.71)
Empathy	4	.90	.91	.85 (.82)
Tangibles	5	.83	.88	.75 (.65)
Retail Chain				
Reliability	5	.92	.96	.92 (.90)
Responsiveness	3	.83	.95	.84 (.84)
Assurance	4	.87	.91	.89 (.85)
Empathy	4	.91	.95	.93 (.91)
Tangibles	5	.90	.91	.88 (.82)
Auto Insurer				
Reliability	5	.95	.96	.95 (.92)
Responsiveness	3	.91	.92	.91 (.86)
Assurance	4	.94	.95	.87 (.82)
Empathy	4	.94	.97	.90 (.84)
Tangibles	5	.91	.94	.85 (.81)
Life Insurer				
Reliability	5	.95	.90	.94 (.91)
Responsiveness	3	.89	.83	.88 (.84)
Assurance	4	.91	.94	.90 (.87)
Empathy	4	.94	.92	.92 (.91)
Tangibles	5	.91	.97	.76 (.84)

Note a. Coefficients in parentheses were computed using the formula for the reliability of a difference score (formula shown in footnote 3 in the text).

reliability coefficients (alphas) were computed for the dimensions for each set of MSS and MSA scores.

The factor loading matrices and the reliability-analysis results revealed consistent patterns that suggested eliminating one SERVQUAL item ("maintaining error-free records") and reassigning two others ("keeping customers informed about when services will be performed" from responsiveness to reliability, and "convenient business hours" from empathy to tangibles). Section 2 in Appendix 2 discusses the rationale for these changes.

Table 3 presents the reliability coefficients by company and by questionnaire format for the reconfigured SERVQUAL dimensions. Only the coefficients based on MSS (i.e., perceptions relative to desired service) scores are reported because MSS is the only common measure across all three questionnaire formats. The coefficients based on MSA (i.e., perceptions relative to adequate service) scores obtained from the two-column and three-column formats were of the same order of magnitude as that of the reported coefficients. For

the three-column format, Table 3 reports a second set of coefficients computed by using a formula recommended specifically for assessing the reliability of a difference score (Peter et al. 1993).³

The coefficient alpha values in Table 3 are consistently high across companies and questionnaire formats, with few exceptions. Notably, except in the life insurance company, the reliability coefficients are highest for the two-column format. The second set of coefficients for the three-column format are generally lower than the corresponding coefficient alphas; however, except for two values in one company, they exceed .80. These findings by and large indicate high internal consistency among items within each SERVQUAL dimension under all three questionnaire formats.

For each of the questionnaire formats, we again factor analyzed the 21 SERVQUAL items for each company as well as for the combined sample. Table 4 reports the rotated factor loading matrices based on analyses of MSS scores for the combined sample (the factor patterns for MSA scores, and for the individual company samples, were similar to those in Table 4).

As the pattern of loadings in Table 4 reveals, the reliability items form a distinct factor (F1) in all three questionnaire formats. The responsiveness, assurance and empathy items primarily load on the same factor (F2) in the two- and three-column formats, but seem to split into two factors (F2 and F3) in the one-column format. The tangibles items, though distinct from the other dimensions, are split among two or three of the remaining factors. The splitting of tangibles into several factors has been observed in past studies (Parasuraman et al. 1991), and may be an artifact of extracting five factors (i.e., because the items for the other four dimensions are captured by just two or three factors, the tangibles items may have split up to represent the remaining factors).

To further evaluate the distinctiveness of the SERVQUAL dimensions, we conducted confirmatory factor analyses using LISREL's unweighted least squares procedure to assess the tenability of two alternative measurement models.⁴ Model 1 was a five-construct model in which the 21 indicator items loaded on the five SERVQUAL dimensions according to the groupings shown in Table 4. Model 2 was a three-construct model in which the five reliability items and five tangibles items loaded on two distinct constructs, while the 11 remaining items loaded on the third construct (acknowledging the possible unidimensionality of responsiveness, assurance and empathy suggested by Table 4). The two models were assessed separately for each company using MSS scores from each of the three questionnaire formats. Results of these analyses, discussed in Section 3 of Appendix 2, supported the tenability of both models, although the support was stronger for Model 1.

In summary, the results of the reliability, factor, and LISREL analyses suggest two broad conclusions that hold consistently across all three questionnaire formats. First, the service quality scores exhibit good internal consistency as reflected by the high reliability coefficients in Table 3. Second, although the results show evidence of discriminant validity among SERVQUAL's five dimensions, they also support the possibility of a three-dimensional structure wherein responsiveness, assurance and empathy meld into a single factor.

TABLE 4

**Factor Loading Matrices Following Oblique
Rotation of Five Factor Solutions for MSS Scores^a**

Items ^b	One-Column Format					Two-Column Format					Three-Column Format				
	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5
Reliability															
Q1	88	-	-	-	-	93	-	-	-	-	84	-	-	-	-
Q2	42	-	49	-	-	71	-	-	-	-	77	-	-	-	-
Q3	51	-	42	-	-	78	-	-	-	-	74	-	-	-	-
Q4	67	-	-	-	-	82	-	-	-	-	84	-	-	-	-
Q5	92	-	-	-	-	86	-	-	-	-	85	-	-	-	-
Responsiveness															
Q6	33	34	31	-	-	57	33	-	-	-	38	53	-	-	-
Q7	-	42	-	-	-	-	66	-	-	-	-	81	-	-	-
Q8	-	-	52	-	-	58	31	-	-	-	45	40	-	-	-
Assurance															
Q9	33	-	35	-	-	-	69	-	-	-	-	60	-	-	-
Q10	31	75	32	-	-	-	49	-	-	42	-	48	48	-	-
Q11	-	-	-	-	-	-	31	53	-	34	-	47	-	-	55
Q12	-	-	70	-	-	-	61	37	-	-	-	52	-	-	-
Empathy															
Q13	-	-	53	-	-	-	46	-	-	-	-	77	-	-	-
Q14	-	64	-	-	-	-	61	-	-	-	-	70	-	-	-
Q15	-	-	36	54	-	-	93	-	-	-	-	91	-	-	31
Q16	-	-	50	36	-	-	80	-	-	-	-	74	-	-	-
Tangibles															
Q17	-	-	-	-	85	-	-	-	-	86	-	-	56	-	-
Q18	-	-	-	-	81	-	-	-	-	76	-	-	-	-	79
Q19	-	-	-	-	80	-	-	-	-	90	-	-	-	-	83
Q20	-	-	-	89	-	-	37	-	67	-	-	-	-	87	-
Q21	-	-	-	-	69	-	-	-	-	78	-	-	73	-	-

Notes: a Numbers shown are loadings multiplied by 100. Loadings less than 0.30 have been omitted. The total variance extracted by the five factors is 80%, 83%, and 76% for the one-, two-, and three-column formats, respectively. The average interfactor correlation is 0.46, 0.30, and 0.34 for the one-, two-, and three-column formats, respectively.

b The item labels (i.e., Q's) correspond to the numbered items in Table 1 as follows: Q1-Q4 correspond to items 1-4; Q5-Q8 correspond to items 6-9 [item 5 was eliminated], Q9-Q16 correspond to items 10-17, Q17-Q20 correspond to items 19-22, and Q21 corresponds to item 18.

Validity

To assess the convergent, predictive, and discriminant validity of the different scale formats, we performed several types of analyses. First, we regressed the overall service quality ratings (measured on a 9-point scale with anchors 1 = "extremely poor" and 9 =

“extremely good”) and the overall value ratings (measured on a 9-point scale with anchors 1 = “poor value” and 9 = “excellent value”) on the scores for the five dimensions obtained from the three questionnaire formats. Specifically, we ran six separate regressions for overall service quality and for overall value with the following scores as independent variables: MSS scores from the one-, two-, and three-column formats, MSA scores from the two- and three-column formats, and perceptions-only scores from the three-column format. The R² values for these regressions are reported in Table 5.

The R² values for the quality regressions are generally high across companies and questionnaire formats, attesting to the convergent and predictive validity of all the service quality scales. The differences in R² values across formats within each company provide additional insight into possible differences in the degree of validity of the various scales.

Except in the retail chain, the perceptions-only scale consistently produced higher R²-values than did the other scales. This superior predictive validity of the perceptions-only scale is similar to findings from some previous studies (e.g., Cronin and Taylor 1992; Parasuraman et al. 1991).⁵ The direct measures of MSS (from the one- and two-column formats) and MSA (from the two-column format) have higher predictive power than the corresponding difference-score measures derived from the three-column format. The sole exception to this pattern is the higher R² value for the difference-score measure of MSS in

TABLE 5

Proportion of Variance in Overall Quality and Value Ratings Explained by Scores on the Service Quality Dimensions^a

<i>Company</i>	<i>Quality</i>	<i>Value</i>	<i>Company</i>	<i>Quality</i>	<i>Value</i>
Computer Manufacturer			Auto Insurer		
One-Column Format:			One-Column Format:		
MSS	.60	.48	MSS	.63	.46
Two-Column Format:			Two-Column Format:		
MSS	.57	.49	MSS	.64	.57
MSA	.66	.50	MSA	.66	.49
Three-Column Format:			Three-Column Format:		
P only	.74	.55	P only	.72	.52
MSS	.51	.40	MSS	.54	.31
MSA	.30	.22	MSA	.24	.10
Retail Chain			Life Insurer		
One-Column Format:			One-Column Format:		
MSS	.64	.43	MSS	.58	.43
Two-Column Format:			Two-Column Format:		
MSS	.74	.50	MSS	.45	.34
MSA	.76	.55	MSA	.55	.44
Three-Column Format:			Three-Column Format:		
P only	.73	.55	P only	.86	.69
MSS	.55	.47	MSS	.60	.58
MSA	.37	.23	MSA	.41	.43

Note: a Numbers reported are adjusted R² values; all values are significant at p < .01.

the life insurance company. Thus, the direct measures by and large have higher predictive validity than do the difference-score measures of service quality.

The results for the value regressions in Table 5 attest to the ability of all the service quality measures to explain a significant portion of the variance in perceived value, as theory predicts they should (e.g., Zeithaml 1988). Moreover, comparing these results with the quality regression results reveals that, with just one exception, the variance explained is higher for quality, the construct the scales purport to measure, than for value, a different construct (the sole exception pertains to MSA under the three-column format in the life insurance company). This consistent pattern supports the scales' discriminant validity.

We further assessed the validity of the service quality scales by comparing the mean MSS and MSA scores across subsamples of customers formed on the basis of: (1) whether they had experienced a recent service problem with the company; and, if so (2) whether the problem was resolved to their satisfaction. For all three questionnaire formats the scores were consistently higher for customers who had not experienced service problems than for those who had; and, within the latter group, the scores were consistently higher for customers who had received satisfactory resolution than for those who had not. The concurrence of these results with what one might hypothesize based on conceptual grounds further supports the scales' validity.

Finally, we examined response error, a potential threat to the scales' validity, by examining the logical consistency of the MSA and MSS ratings from the two-column format, and of the adequate-service and desired-service ratings from the three-column format. [A similar consistency check was not possible for the one-column format because this format just provided MSS ratings.] An instance of response error occurs when the MSS rating on an item exceeds the MSA rating, or when the adequate-service rating exceeds the desired-service rating. The percentages of respondents who committed one or more such errors are summarized below:

	Two-Column Format	Three-Column Format
Computer Manufacturer	8.6%	6%
Retail Chain	18.2%	1.8%
Auto Insurer	12.2%	1.6%
Life Insurer	9.9%	2.7%

The threat to the scales' validity due to response error is quite small for the three-column format but substantially higher for the two-column format. [Item responses subject to this error were deleted from all other analyses discussed in this paper.]

In summary, the measures in all three questionnaire formats possess convergent and predictive validity as evidenced by their ability to explain a significant proportion of the variance in the overall service quality measure. However, in terms of predictive power alone, the perceptions-only measure is superior to the disconfirmation measures and, within the latter, the direct measure is superior to the difference-score measure. The behavior of all measures is also consistent with theoretical predictions of their relationships with overall value, incidence of service problems, and problem-resolution experience. Finally, the three-column (difference-score) format is much less susceptible to response error than is the two-column (direct-measure) format.

Diagnostic Value

Of the three questionnaire formats investigated in this study, only the three-column format is capable of specifically indicating the position of the zone of tolerance and the perceived service level relative to the zone. The one-column format provides no information about the zone of tolerance. The two-column format scores can indicate whether the perceived service level is above the tolerance zone (MSS score greater than 5), below the tolerance zone (MSA score less than 5), or within the tolerance zone (MSS score less than or equal to 5 and MSA score greater than or equal to 5). However, the two-column format scores cannot identify the tolerance zone's position on a continuum of expectation levels. This limitation, coupled with the two-column format's higher response error, suggests that its scores may be less useful and trustworthy than the scores from the three-column format.

Figure 1 depicts the tolerance zones and perceived service levels across the dimensions for the four companies as derived from the three-column format. To gain more detailed insights, one can also construct a similar set of charts for the attributes within each dimension. The charts in Figure 1, by providing precise information about the perceived service levels across dimensions relative to the adequate and desired service levels, offer insight into the emphasis a company should place on different dimensions in initiating quality-improvement efforts.

The charts in Figure 1 also highlight the suboptimal allocation of service-improvement resources that can result from using a perceptions-only measure to assess service quality, an approach advocated by some researchers. For instance, if the computer manufacturer examined just the perceptions scores it might decide to place the same emphasis on improving its performance on tangibles as on reliability. The imprudence of such a decision is evident from examining the company's performance on tangibles and reliability *relative to the customers' tolerance zones* for these dimensions. The data suggest that the company should place significantly greater emphasis on reliability than on tangibles.

Table 6, which summarizes the mean values for the measures obtained from the three questionnaire formats, reveals several common patterns across companies and dimensions. The direct measures of service superiority (MSS scores) from the one- and two-column formats are by and large similar for each dimension, confirming that both formats are measuring the same construct.⁶ However, with just two exceptions, the mean values for the direct measures of service superiority are greater than 5, the scale point at which the desired and perceived service levels are the same (the exceptions are the mean values of 4.9 and 5.0 under the two-column format for the retail chain's responsiveness and empathy, respectively). This consistent pattern implies that perceived service performance is *above* the desired service level for virtually all dimensions in each company. In contrast, except for tangibles in the computer-manufacturer context, the difference-score values of MSS (obtained from the three-column format) are negative, implying that perceived service performance is *below* the desired service level.

The consistent pattern of discrepancies between the direct and difference-score operationalizations of MSS, and the conflicting inferences stemming from them, raise the issue of which operationalization is more trustworthy. Given that the desired service level represents a form of "ideal" standard, perceived performance falling below that level (on at least several

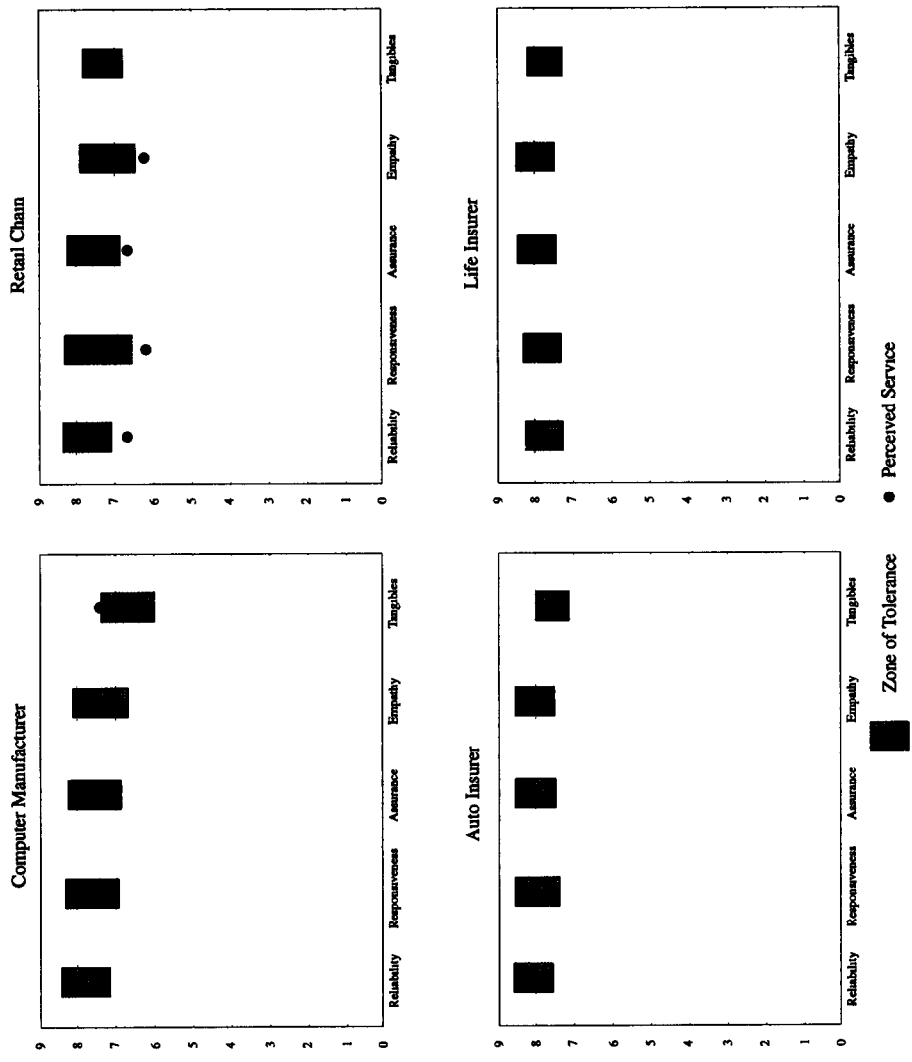


Figure 1. Service Quality Perceptions Relative to Zones of Tolerance by Dimension

TABLE 6

Mean Values of Measures Obtained from Three Formats

	Questionnaire Format					
	One-Column	Two-Column		Three-Column		
	MSS	MSS	MSA	P only	MSS	MSA
Computer Manufacturer						
Reliability	5.8	5.8	6.8	7.5	-1.0	.3
Responsiveness	5.8	5.6	6.8	7.2	-1.1	.4
Assurance	5.9	5.8	6.8	7.4	-.9	.6
Empathy	5.5	5.4	6.5	6.9	-1.2	.3
Tangibles	6.3	6.3	7.1	7.5	.1	1.5
Retail Chain						
Reliability	5.6	5.7	6.3	6.6	-1.6	-.5
Responsiveness	5.3	4.9	5.9	6.2	-1.9	-.4
Assurance	5.5	5.3	6.2	6.7	-1.5	-.2
Empathy	5.3	5.0	5.8	6.2	-1.8	-.4
Tangibles	6.3	6.2	6.8	7.2	-.5	.6
Auto Insurer						
Reliability	6.3	6.8	7.4	7.9	-.6	.3
Responsiveness	6.4	6.6	7.3	8.0	-.5	.5
Assurance	6.4	6.6	7.3	8.0	-.5	.5
Empathy	6.2	6.5	7.2	7.9	-.5	.4
Tangibles	6.3	6.8	7.3	7.8	-.3	.8
Life Insurer						
Reliability	6.2	6.1	6.8	7.5	-.8	.2
Responsiveness	6.2	6.2	6.9	7.5	-.7	.2
Assurance	6.4	6.4	7.1	7.6	-.7	.2
Empathy	6.2	6.2	6.9	7.5	-.7	.2
Tangibles	6.6	6.8	7.3	7.7	-.4	.5

dimensions) seems a more plausible and “face valid” finding than a consistent pattern of perceptions exceeding the desired service level. Thus, the direct measures may be inflating the ratings. Although it may be meaningful to interpret these ratings in a strictly *relative* sense (e.g., to compare current ratings with past ratings or with competitors’ ratings), interpreting them in isolation may cause a company to infer that its service quality is better than it actually is.

In summary, interpreting perceptions ratings in conjunction with adequate- and desired-service expectation ratings is helpful in accurately diagnosing service deficiencies and initiating appropriate improvement efforts. The three-column format provides detailed—and likely more accurate—data for these purposes than the other two formats.

The overall pattern of findings suggests that there are psychometric and practical tradeoffs in choosing the most appropriate scaling approach for measuring service quality. Table 7, summarizing the relative strengths and weaknesses of the three scaling formats examined in this study, highlights the tradeoffs, and serves as a basis for discussing the practical and research implications of the findings.

Practical Implications

While both the three-column and two-column formats provide measures of service superiority and adequacy, the three-column format calls for three separate ratings and may be more time-consuming for respondents. However, the time expended in providing additional ratings for the three-column format may be mitigated by higher respondent ease in completing this questionnaire. The apparent complexity of the two-column format questionnaire, evidenced by respondents' greater difficulty and lower confidence with it, may add to the time required to complete it despite its requiring just two sets of ratings. The similar response rates achieved by all three formats suggests that no format is likely to have a significant advantage over the others in this regard.

Regarding the diagnostic value of the information obtained through service quality measurement, the three-column format is superior to the two-column format, which, in turn, is superior to the one-column format. However, the soundness of the diagnostics provided by the two-column format questionnaire is debatable. Of particular concern is the possibility of respondent errors triggered by the apparent complexity of having to distinguish between and provide direct ratings of two different comparisons (MSA and MSS). Thus, managers preferring to directly measure perceptions relative to expectations might want to measure just MSS through the one-column format questionnaire, the direct-measure counterpart to the current two-part SERVQUAL instrument.

An important issue examined in this study is the psychometric soundness of difference-score measures of service quality relative to that of direct measures (including perceptions-only measures). The results indicate that the difference-score measures perform as well as the direct measures on all psychometric criteria except predictive power (i.e., ability to explain the variance in overall perceptions of service quality). If maximizing predictive power is the principal objective, the perceptions-only scale is the best as it outperforms all other measures on this criterion. However, if identifying critical service shortfalls is the principal objective, the three-column format questionnaire seems most useful; and, this format also provides separate perceptions ratings for those concerned with maximizing predictive power.

In summary, companies should consider adopting a service quality measurement system that produces separate measures of adequate-service and desired-service expectations, and perceptions. We recognize that radically altering current measurement systems may not be easy, especially in companies with well-entrenched systems, or with systems linked to employee compensation. However, the study's findings have implications for such compa-

TABLE 7

Comparative Summary of Alternative SERVQUAL Scales

Criteria	One-Column Format	Two-Column Format	Three-Column Format
General Scale Characteristics			
Types of Measures	Direct measure of MSS	Direct measures of MSA and MSS	Difference-score measures of MSA and MSS; perceptions ratings
Number of Ratings to Be Provided by Respondents	21	42	63
Respondent Ease	High	Medium	High
Respondent Confidence	High	Medium	High
Response Rate	27%	22%	24%
Reliability and Factor Structure			
Reliability Coefficients	High	High	High
Dimensionality	Reliability and tangibles distinct; other three dimensions overlap	Reliability and tangibles distinct, other three dimensions overlap	Reliability and tangibles distinct; other three dimensions overlap
Validity			
Predictive and Convergent Validity	High	High	High
Relative Predictive Power of Various Measures	Predictive power of MSS comparable to that of MSS in two-column format and higher than that of difference-score measure of MSS	Predictive power of MSS higher than that of difference-score measure of MSS, but <i>lower</i> than that of MSA	Predictive power of perceptions-only measure is higher than that of other measures in all formats; predictive power of MSS higher than that of MSA
Validity As Reflected by Relationships Consistent with Theoretical Predictions for:			
Value	High	High	High
Incidence of Service Problems	High	High	High
Problem-Resolution Experience	High	High	High
Response Error	N/A	High	Low
Diagnostic Value			
Ability to Determine Whether Perceptions Fall Below, Within, or Above Zone of Tolerance	No	Yes	Yes
Ability to <i>Pinpoint</i> Position of Zone of Tolerance and Perceptions Relative to the Zone	No	No	Yes
Potential for Inflated Ratings and Consequent Erroneous Inferences	High	High	Low

nies as well. Companies with perceptions-only measurement systems should consider augmenting their current system with at least a desired-service measure to be able to identify service shortfalls more accurately. If adding a new measure is not possible, then consideration should be given to converting the perceptions-only measure to a direct measure of the discrepancy between perceptions and desired-service expectations. However, in interpreting the direct-measure ratings managers should beware of possible inflation of the ratings. This problem can be neutralized if managers track the ratings over time and interpret just the *change* in ratings to determine whether performance on each attribute has improved, deteriorated, or remained the same from period to period.

Research Implications

The study's findings, while contributing to the extant knowledge about service quality and its measurement, also raise additional issues for further research. First, despite the three-column format questionnaire's superior diagnostic value, administering it in its entirety may pose practical difficulties, particularly in telephone surveys or when the list of 21 generic items is supplemented with more context-specific items as suggested by Parasuraman et al. (1991). Therefore, a fruitful avenue for additional research is to explore the soundness of administering logical subsections of the questionnaire to comparable subsamples of customers while still achieving its full diagnostic value.

For instance, the information needed to construct zone-of-tolerance charts such as those in Figure 1 can be generated through any of the following "partial" approaches: (1) Obtaining adequate-service and perceptions ratings from one half of the sample, and desired-service and perceptions ratings from the other half; (2) Obtaining adequate-service, desired-service, and perceptions ratings separately from three comparable subsamples; and (3) Dividing the total sample into five comparable subsamples and obtaining from each all three types of ratings for just one of the five SERVQUAL dimensions. Research is needed to assess the reliability and validity of these approaches, and the statistical equivalence of their results, relative to administering the full questionnaire to the entire sample.

Second, contrary to criticisms of difference-score measures on psychometric grounds, the MSA and MSS constructs operationalized as difference scores are by and large as sound as their direct-measure counterparts except in terms of predictive power. Research directed at exploring reasons for the discrepancy between the alleged deficiencies and the actual results could provide greater understanding of the pros and cons of using difference scores in service-quality measurement. Such research is especially appropriate because similar discrepancies in other studies have surfaced in recent debates about the appropriateness of using difference scores to operationalize service quality (cf. Brown et al. 1993; Parasuraman et al. 1993).

Third, the findings highlight the importance of considering practical usefulness in assessing alternative service quality scales. Therefore, consistent with recent calls issued by Parasuraman et al. (1994a) and Perreault (1992), there is a need for explicitly incorporating practical criteria such as diagnostic value into the traditional scale-assessment paradigm that is dominated by psychometric criteria.

Fourth, the apparent upward bias in the direct-measure ratings of MSS also exists in results reported in other studies that have used both direct and difference-score measures of service quality (Brown et al. 1993; Liljander and Strandvik 1992). Furthermore, based on a review of customer satisfaction studies using direct measures, Peterson and Wilson (1992) have documented a consistent pattern of potentially inflated ratings. Thus, direct measures seem to have a persistent tendency to overstate customers' assessments. Research aimed at understanding the causes of this phenomenon and estimating the extent of upward bias it produces would be helpful in reducing the bias, or at least correcting for it in interpreting direct-measure ratings.

Finally, this study's findings warrant additional research on the dimensionality of the SERVQUAL items, an issue that has produced mixed results in previous studies and has already generated debate (see, for example, Brown et al. 1993; Cronin and Taylor 1992; Parasuraman et al. 1991; Parasuraman et al. 1994a). The overall findings reveal considerable interdimensional overlap, especially among responsiveness, assurance, and empathy. Parasuraman et al. (1991) have speculated about possible reasons for similar overlaps observed in earlier studies, and have proffered directions for future research on this issue. The present study's findings reiterate the need to understand the underlying causes and managerial implications of the empirical correlations among the dimensions.

Acknowledgment: The authors gratefully acknowledge the financial assistance and cooperation provided by the Marketing Science Institute and four of its corporate sponsors. The authors also thank William O. Bearden, Gilbert A. Churchill, Jr., George S. Day, Claes Fornell, and Richard Staelin for serving on the expert panel that offered advice on research-design issues.

APPENDIX 1

Three Alternative Service Quality Measurement Formats Included in the Pretest Questionnaires

Note: Final versions of the three formats are in Appendix 2. Only one illustrative item is shown for each format.

THREE-COLUMN FORMAT

DIRECTIONS: We would like to get your impressions about how well XYZ Bank performs relative to your expectations. Please think about two different levels of expectations:

Desired Service Level—the level of service performance you believe an excellent bank can and should deliver; and

Adequate Service Level—the minimum level of service performance you would consider acceptable.

For each of the following attributes, please indicate: (a) your *desired service level* on that feature by circling one of the nine numbers in the *first* column; (b) your *adequate service level* by circling one of the nine numbers in the *second* column; and (c) your *perception of XYZ Bank's performance* by circling one of the nine numbers in the *third* column. There are no right or wrong answers—all we are interested in are three ratings on each attribute that best represent your *desired service level*, *adequate service level*, and *perception of XYZ Bank's performance*.

Note: Your desired service level is the level of performance you believe an excellent bank can and should deliver.

Your adequate service level is the minimum level of service performance you would consider acceptable.

	My Desired Service Level		My Adequate Service Level		My Perception of XYZ's Performance	
	Low	High	Low	High	Low	High
1 Modern-looking equipment	1 2 3 4 5 6 7 8 9		1 2 3 4 5 6 7 8 9		1 2 3 4 5 6 7 8 9	

TWO-COLUMN FORMAT

DIRECTIONS: Based on your experiences with XYZ Bank, think about the quality of service XYZ Bank offers compared to two different levels of service:

Desired Service Level—the level of service performance you believe an excellent bank *can and should* deliver; and

Adequate Service Level—the minimum level of service performance you would consider acceptable

For each of the following attributes, please indicate: (a) how XYZ Bank's performance compares with your *desired service level* by circling one of the nine numbers in the *first* column; and (b) how XYZ Bank's performance compares with your *adequate service level* by circling one of the nine numbers in the *second* column. There are no right or wrong answers—all we are interested in are two ratings on each feature that best represent your perception of XYZ Bank's performance compared to your *desired service level* and your *adequate service level*.

Note: Your desired service level is the level of performance you believe an excellent bank can and should deliver.

Your adequate service level is the minimum level of service performance you would consider acceptable.

	XYZ Bank's Performance Compared to My Desired Service Level Is:						XYZ Bank's Performance Compared to my Adequate Service Level Is:											
	Much Lower		The Same		Much Higher		Much Lower		The Same		Much Higher							
	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
1. Modern-looking equipment																		

One-Column Format

Part I:

DIRECTIONS: Based on your experiences with XYZ Bank, think about the quality of service XYZ Bank offers compared to the level of service you desire. Please compare your perception of XYZ Bank's performance on each of the following attributes against the performance level you believe an excellent bank *can and should* deliver (i.e., your *desired service level*). There are no right or wrong answers—just circle the number that you feel reflects your perception of how XYZ Bank's performance compares with your desired service level.

Note: Your desired service level is the level of service performance you believe an excellent bank can and should deliver.

	XYZ Bank's Performance:								
	Falls Way Short of My Desired Service Level			Meets My Desired Service Level			Far Exceeds My Desired Service Level		
	1	2	3	4	5	6	7	8	9
1 Modern-looking equipment									

Part II:

DIRECTIONS: Based on your experiences with XYZ Bank, think about the quality of service XYZ Bank offers compared to the level of service you would consider adequate. Please compare your perception of XYZ Bank's performance on each of the following attributes against the *minimum* level of performance you would consider acceptable (i.e.,

your *adequate service level*). There are no right or wrong answers—just circle the number that you feel reflects your perception of how XYZ Bank’s performance compares with your adequate service level.

Note: Your adequate service level is the minimum level of service performance you would consider acceptable.

	XYZ Bank's Performance								
	Falls Way Short of My Adequate Service Level			Meets My Adequate Service Level			Far Exceeds My Adequate Service Level		
	1	2	3	4	5	6	7	8	9
1 Modern-looking equipment									

APPENDIX 2

Final Versions of the Three Alternative Service Quality Measurement Formats

Note All formats are for the auto insurer and only one illustrative item is shown in each.

Three-Column Format

We would like your impressions about _____’s service performance relative to your expectations. Please think about the two different levels of expectations defined below:

MINIMUM SERVICE LEVEL – the *minimum* level of service performance you consider adequate.

DESIRED SERVICE LEVEL – the level of service performance you desire.

For each of the following statements, please indicate: (a) your *minimum service level* by circling one of the numbers in the *first* column; (b) your *desired service level* by circling one of the numbers in the *second* column; and (c) your perception of _____’s service by circling one of the numbers in the *third* column.

	My <i>Minimum</i> Service Level Is	My <i>Desired</i> Service Level Is	My Perception of _____'s Service Performance Is
	Low High	Low High	Low High No Opinion
When it comes to . . .	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9 N
1 Prompt service to policyholders	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9 N

Section 1: Analysis to Verify Compositional Similarity of Subsamples

The retail-chain, auto-insurer, and life-insurer questionnaires contained categorical questions about the following respondent characteristics: (1) whether the respondents had experienced a recent problem with the company; and, if so (2) whether the problem was resolved to their satisfaction; (3) frequency of contact with company employees; (4) length of association with the company; (5) gender; (6) marital status; (7) age; (8) family income; (9) education. The respondent-characteristic questions in the computer-manufacturer questionnaires included the first three characteristics listed above. The remaining personal-demographic characteristics were inappropriate because the respondents were business customers. The company's executives suggested that the following questions be substituted: (1) whether respondents were part of their company's MIS/Data Processing department; (2) their position/title; (3) their involvement in MIS-related purchasing decisions; (4) types of computer products used; (5) number of employees. Respondent profiles on the various characteristics were compared for each sponsor company across the three questionnaire formats using a series of chi-square tests.

With just one exception, the chi-square tests revealed no significant difference (at the conventional .05 level) in profiles across the three formats. The sole exception pertained to respondents' gender in the life-insurer survey—the group that responded to the two-column format questionnaire had more females (64%) than did the groups that responded to the one-column (51%) and three-column (47%) format questionnaires. Thus, the three subsamples can be considered to be compositionally similar.

Section 2: Rationale for Changes to SERVQUAL Based on Main-Study Results

Consistent patterns of findings from the factor and reliability analyses resulted in three changes to the SERVQUAL battery. First, *maintaining error-free records*, a reliability item, invariably loaded by itself, or had much weaker loadings than did the four other reliability items on the factor representing reliability. Furthermore, results from the internal-consistency analyses showed that coefficient alpha would improve by deleting this item from the reliability dimension. Discarding this item seemed to be warranted on conceptual grounds as well. Because customers generally have limited or no access to a company's records, they may experience difficulty in assessing company performance on this item. For these reasons, *maintaining error-free records* was deleted from all subsequent analyses.

Second, *keeping customers informed about when services will be performed*, a responsiveness item, consistently loaded with the reliability items. Grouping this item with the four reliability items improved coefficient alpha for the reliability dimension but left unchanged, or lowered marginally, coefficient alpha for the responsiveness dimension. Conceptually, this item relates to *providing service at the promised time*, a reliability item, and implies *dependability*, a core facet of reliability. Therefore, *keeping customers informed about when services will be performed* was reassigned to the reliability dimension.

Third, *convenient business hours*, an empathy item, had much stronger loadings on the tangibles factor than on the empathy factor. Moreover, coefficient alphas for both empathy

and tangibles improved when this item was removed from empathy and assigned to tangibles. Conceptually, relative to the other empathy items this item is more of a *search attribute* (one that customers can evaluate prior to purchasing a service) than an *experience attribute* (one that customers can evaluate only during service purchase and consumption). As such, this attribute could serve as a “tangible” clue about a company’s service orientation. Consequently, *convenient business hours* was reassigned to the tangibles dimension.

Section 3: Additional Analyses to Assess SERVQUAL’s Dimensionality

The LISREL confirmatory factor analyses showed both Model 1 (five-factor model) and Model 2 (three-factor model) to be tenable on the basis of the traditional criteria of GFI (goodness-of-fit index), AGFI (adjusted GFI), and RMSR (root-mean-squared residual). The GFI and AGFI values were consistently high for both models (ranging from .98 to .99 across companies and questionnaire formats) and the RMSR values were low (ranging from .04 to .08). In contrast, the chi-square values associated with the models were consistently high and statistically significant (at $p < .05$), indicating potential lack of model fit. However, the chi-square statistic is sensitive to sample size and is not a meaningful indicator of model fit when sample sizes are large. A more appropriate use of the chi-square statistic is in assessing the *relative* fit of alternative “nested” models such as Models 1 and 2.

To assess relative fit (i.e., improvement or deterioration in fit by using one model instead of the other), the statistical significance of the *difference* in chi-square values of two nested models is evaluated. This evaluation is done by comparing the chi-square difference with the critical chi-square value (at a specified significance level) with degrees of freedom equal to the difference in the number of paths estimated between the alternative models. If the chi-square difference is greater than the critical value, the model with the lower absolute chi-square value has a significantly better fit.

The above procedure was used to assess the fit of Model 1 relative the fit of Model 2. Of twelve such comparisons between Models 1 and 2 (across four companies and three questionnaire formats), nine revealed a significant difference (at $p = .05$) in favor of Model 1, one revealed a significant difference in favor of Model 2, and two revealed no significant difference in fit between Models 1 and 2.

Thus, although the findings collectively suggest that both models are tenable, they provide somewhat stronger support for the five-dimensional model. The distinctiveness of the five dimensions was also supported to a large extent by additional analyses suggested by Bagozzi and Phillips (1982) to examine discriminant validity among related latent constructs.

In the procedure recommended by Bagozzi and Phillips (1982), a measurement model wherein all inter-construct correlations are set to be free is first estimated. The fit of this null model is compared with that of a series of alternative models in each of which the correlation between one pair of constructs is fixed to be 1, implying no discrimination between the two constructs (all other inter-construct correlations are set to be free). Comparisons of model fit are made by using the chi-square difference test described previously. Two constructs are considered to be distinct if the model in which the correlation between them is fixed to be 1 has a significantly higher chi-square value (i.e., poorer fit) than that of the null model.

In the present study, the null model was compared with 10 alternative models corresponding to the 10 possible pairwise correlations among the five SERVQUAL dimensions. The model estimations and comparisons were made separately for the three different questionnaire formats using MSS scores from the combined sample. Of the 30 comparisons across the all three formats, 24 resulted in a significant chi-square difference in favor of the null models, providing general support for the discriminant validity among the SERVQUAL dimensions. The six comparisons not resulting in a favorable chi-square difference involved alternative models in which the correlations between the following pairs of dimensions were fixed to be 1: responsiveness/tangibles in the one-column format; responsiveness/empathy and assurance/empathy in the two-column format; and responsiveness/tangibles, responsiveness/empathy, and assurance/empathy in the three-column format.

NOTES

1. It should be noted that the MSS and MSA scores can be “positive” (when perceptions exceed expectations) or “negative” (when perceptions fall short of expectations).
2. Since the response rates in the field test ranged from 12% to 16%, a mail-out sample of 800 was expected to generate at least 96 completed questionnaires of each type for each company. This expected final sample was deemed large enough for empirically examining the factor structure of the 22 SERVQUAL items because a sample of four to five times the number of variables is considered adequate for factor analysis (Hair et al 1992).
3. The formula for the reliability (r_D) of a construct operationalized as a difference score is:

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

where r_{11} and r_{22} are the reliabilities of the first and second component scores, σ_1^2 and σ_2^2 are the variances of these component scores, and r_{12} is the correlation between the component scores

4. The unweighted-least-squares estimation procedure rather than the traditional maximum-likelihood estimation procedure was used because the latter is based on the assumption that the observed variables have a multinormal distribution, an assumption not required by the former (Dillon 1986) and not met in the present study
5. A plausible explanation for the consistently superior predictive power of the perceptions-only ratings is that in the regressions involving these ratings *both* the dependent and the independent variables are *perceptions-only* ratings, in contrast to the other regressions wherein the independent variables are *disconfirmation* ratings. In other words, the higher R^2 values could be an artifact of common method variance
6. This point is noteworthy because a less elaborate definition of the desired service level was used in the two-column format questionnaire (to keep it as concise as possible since two different expectation levels had to be defined). Specifically, as Appendix 2 shows, although the definition of the desired service level in the boxed section of both the one- and two-column format questionnaires is the same, the instructions section of the one-column format questionnaire elaborates on the construct’s meaning by including the term “can and should deliver.” The virtually identical MSS scores produced by both questionnaires suggest that any confounding effect due to this variation is negligible.

REFERENCES

- Babakus, Emin and Gregory W. Boller (1992) "An Empirical Assessment of the SERVQUAL Scale." *Journal of Business Research*, **24**: 253-268.
- Babakus, Emin and W. Glynn Mangold (1992). "Adapting the SERVQUAL Scale to Hospital Services: An Empirical Investigation." *Health Services Research*, **26**(6): 767-786.
- Bagozzi, Richard P. and Lynn W. Phillips (1982). "Representing and Testing Organizational Theories: A Holistic Construal." *Administrative Science Quarterly*, **27**(September): 459-489.
- Bojanic, David C. (1991) "Quality Measurement in Professional Services Firms." *Journal of Professional Services Marketing*, **7**(2): 27-36.
- Bolton, Ruth N. and James H. Drew (1991a). "A Longitudinal Analysis of the Impact of Service Changes on Customer Attitudes." *Journal of Marketing*, **55**(January): 1-9.
- . (1991b). "A Multistage Model of Customer's Assessments of Service Quality and Value." *Journal of Consumer Research*, **17**(March): 375-384.
- Boulding, William, Ajay Kalra, Richard Staelin and Valarie Zeithaml (1993). "A Dynamic Process Model of Service Quality: From Expectations to Behavioral Intentions." *Journal of Marketing Research*, **30**(February): 7-27
- Brensinger, Ronald P. and Douglas M. Lambert (1990). "Can The SERVQUAL Scale be Generalized to Business-to-Business Services?" P. 289 in *Knowledge Development in Marketing*. 1990 AMA's Summer Educators' Conference Proceedings.
- Brown, Stephen W. and Teresa A. Swartz (1989). "A Gap Analysis of Professional Service Quality." *Journal of Marketing*, **53**(April): 92-98.
- Brown, Tom J., Gilbert A. Churchill, Jr. and J. Paul Peter (1993). "Improving the Measurement of Service Quality." *Journal of Retailing*, **69**(Spring): 127-139
- Cadotte, Earnest R., Robert B. Woodruff and Roger L. Jenkins (1987). "Expectations and Norms in Models of Consumer Satisfaction" *Journal of Marketing Research*, **24**(August): 305-314.
- Carman, James M. (1990). "Consumer Perceptions of Service Quality. An Assessment of the SERVQUAL Dimensions." *Journal of Retailing*, **66**(Spring): 33-55.
- Crompton, John L. and Kelly J. Mackay (1989). "Users' Perceptions of the Relative Importance of Service Quality Dimensions in Selected Public Recreation Programs" *Leisure Sciences*, **11**: 367-375
- Cronin, J. Joseph, Jr. and Stephen A. Taylor (1992). "Measuring Service Quality: A Reexamination and Extension," *Journal of Marketing*, **56**(July): 55-68.
- . (1994). "SERVPERF Versus SERVQUAL: Reconciling Performance-Based and Perceptions-Minus-Expectations Measurement of Service Quality." *Journal of Marketing*, **58**(January): 125-131.
- Devlin, Susan J, H. K. Dong and Marbue Brown (1993). "Selecting A Scale for Measuring Quality" *Marketing Research A Magazine of Management and Applications*, **5**(3): 12-17.
- Dillon, William R (1986). "Building Consumer Behavior Models With LISREL: Issues in Applications" Pp 107-154 in *Perspectives on Methodology in Consumer Research*, David Brnberg and Richard J Lutz (eds). New York: Springer-Verlag.
- Finn, David W. and Charles W Lamb, Jr. (1991) "An Evaluation of the SERVQUAL Scales in a Retail Setting" P 18 in *Advances in Consumer Research*, Rebecca H Holman and Michael R Solomon (eds.) Provo, UT. Association for Consumer Research.
- Ford, John B , Mathew Joseph and Beatriz Joseph (1993). "Service Quality in Higher Education: A Comparison of Universities in the United States and New Zealand Using SERVQUAL." Pp 75-81 in *Enhancing Knowledge Development in Marketing*, 1993 AMA Educators' Proceedings.

- Hair, Joseph F. Jr., Rolph Anderson, Ronald L. Tatham and William C. Black (1992). *Multivariate Data Analysis with Readings*, 3rd ed. New York: Macmillan.
- Johnson, Linda L., Michael J. Dotson and B. J. Dunlop (1988). "Service Quality Determinants and Effectiveness in the Real Estate Brokerage Industry." *The Journal of Real Estate Research*, 3: 21-36.
- Liljander, Veronica and Tore Strandvik (1992). "The Relationship between Service Quality, Satisfaction and Intentions." Pp. 77-98 in *Proceedings of the 2nd Workshop on Quality Management in Services*, Jos Lemmink and Paul Kunst (eds.). Brussels, Belgium: European Institute for Advanced Studies in Management.
- Oliver, Richard L. (1980). "A Cognitive Model of the Antecedents and Consequences of Satisfaction Decisions." *Journal of Marketing Research*, 17(November): 460-469.
- Parasuraman, A., Leonard L. Berry and Valarie A. Zeithaml (1991). "Refinement and Reassessment of the SERVQUAL Scale." *Journal of Retailing*, 67(Winter): 420-450.
- . (1993). "More on Improving Service Quality Measurement." *Journal of Retailing*, 69(Spring): 40-147.
- Parasuraman, A., Valarie A. Zeithaml and Leonard L. Berry (1985). "A Conceptual Model of Service Quality and Its Implications for Future Research." *Journal of Marketing*, 49(Fall): 41-50.
- . (1988). "SERVQUAL: A Multiple-Item Scale for Measuring Consumer Perceptions of Service Quality." *Journal of Retailing*, 64(Spring): 12-40.
- . (1994a) "Reassessment of Expectations as a Comparison Standard in Measuring Service Quality: Implications for Further Research." *Journal of Marketing*, 58(January): 111-124.
- . (1994b). "Moving Forward in Service Quality Research: Measuring Different Customer-Expectation Levels, Comparing Alternative Scales, and Examining the Performance-Behavioral Intentions Link." Marketing Science Institute monograph, Report Number 94-114.
- Perreault, William D. Jr. (1992) "The Shifting Paradigm in Marketing Research," *Journal of the Academy of Marketing Science*, 20(Fall): 367-375.
- Peter, J. Paul, Gilbert A. Churchill Jr. and Tom J. Brown (1993). "Caution in the Use of Difference Scores in Consumer Research." *Journal of Consumer Research*, 19(March): 655-662.
- Peterson, Robert A. and William R. Wilson (1992). "Measuring Customer Satisfaction: Fact and Artifact." *Journal of the Academy of Marketing Science*, 20(1). 61-71.
- Teas, R. Kenneth (1993). "Expectations, Performance Evaluation and Consumer's Perceptions of Quality." *Journal of Marketing*, 57(October): 18-34.
- . (1994). "Expectations as a Comparison Standard in Measuring Service Quality An Assessment of a Reassessment." *Journal of Marketing*, 58(January): 132-139.
- Tse, David K. and Peter C. Wilton (1988). "Models of Consumer Satisfaction Formation: An Extension." *Journal of Marketing Research*, 25(May): 204-212.
- Woodruff, Robert B., D. Scott Clemons, David W. Schumann, Sarah F. Gardial and Mary Jane Burns (1991) "The Standards Issue in CS/D Research: A Historical Perspective." *Journal of Consumer Satisfaction, Dissatisfaction and Complaining Behavior*, 4: 103-109.
- Yi, Youjiae (1990) "A Critical Review of Consumer Satisfaction." Pp. 68-123 in *Review of Marketing*, Valarie Zeithaml (ed.). Chicago, IL: American Marketing Association.
- Zeithaml, Valarie A. (1988). "Consumer Perceptions of Price, Quality, and Value: A Conceptual Model and Synthesis of Research." *Journal of Marketing*, 52(July): 2-22.
- Zeithaml, Valarie A., Leonard L. Berry and A. Parasuraman (1993). "The Nature and Determinants of Customer Expectations of Service." *Journal of the Academy of Marketing Science*, 21(1): 1-12.