Measuring Service Quality with SERVPERF

Fogarty, G., Catts, R., & Forlin, C.

University of Southern Queensland

Full reference: Fogarty, G., Catts, R., & Forlin, C. (2000). Identifying shortcomings in the measurement of service quality. *Journal of Outcome Measurement*, 4(1), 425-447.

Abstract

SERVPERF, the performance component of the Service Quality scale (SERVQUAL), has been shown to measure five underlying dimensions corresponding to Tangibles, Reliability, Responsiveness, Assurance, and Empathy (Parasuraman, Zeithaml, & Berry, 1988). This paper describes a validation study, employing four different datasets, of a shortened 15-item version of the SERVPERF scale to be called SERVPERF-M. Exploratory and confirmatory factor analytic techniques were used to explore the dimensionality of the scale. Although internal consistency estimates for all scales were very satisfactory for all four datasets, the factor structure was somewhat unstable with Responsiveness, Assurance and Empathy tending to define a single factor in three of the sets and Reliability and Tangibles to define two other factors. Rasch analysis was employed to gain further insights into the behaviour of the items. These analyses suggested that the five factors can be treated as five different stages of service quality, rather than as five qualitatively different dimensions. The Rasch analysis also suggested that the items in both SERVPERF and SERVPERF-M are too easy to rate highly and that more "difficult" items need to be added to the scale. If this is done, it is likely that some of the confusion that exists over the dimensionality of this scale will disappear.

Measuring Service Quality with SERVPERF

The concept of service quality is not universally understood and is often used as an umbrella term to cover a range of impressions gathered by customers when dealing with vendors. These impressions, however, are important factors that influence buying behaviour and firms are very conscious of the need to improve this aspect of their operations, either by staff training or direct investment in facilities. If training programmes aimed at improving service quality are to be effective, there is a need to identify the specific constructs underlying this generic term. The present study reports on the construct validity of one of the main instruments used to measure this construct, the SERVQUAL scale developed by Parasuraman, Zeithaml, and Berry (1988), and a shortened form of this scale (SERVPERF-M) developed by the researchers for use in an Australian small-business setting.

The psychometric properties of the SERVQUAL scale have been the subject of considerable research in recent times. The scale was developed from an initial pool of 97 items generated through a series of focus group sessions conducted with consumers (Parasuraman et al., 1988). The initial pool of 97 items was reduced to 22 to form the SERVQUAL scale with a reported reliability above .90 (Parasuraman et al., 1988). The scale was said to tap five different underlying dimensions of customer service termed Tangibles, Reliability, Responsiveness, Assurance, and Empathy. Other researchers have questioned the validity of the five-factor structure. Partial support was reported by Gagliono and Hathcote (1994) who obtained four factors corresponding to Personal Attention, Reliability, Tangibles, and Convenience. Babakus and Mangold (cited in Brown et al., 1989) found a single factor when SERVQUAL was used in a hospital setting. The five-factor structure was also rejected factor by Cronin and Taylor (1992) who conducted studies across five retail settings. Finn and Lamb (1991) tested the five-factor structure in a retail context. They reported that while the reliabilities for each of the five factors were acceptable with estimates ranging from .59 to .83, confirmatory factor analysis indicated that the data did not fit the model.

Whilst the questions that make up the SERVPERF scale might well cover most of the broad domain of service quality, the issue of dimensionality is a fundamental concern since the utility of SERVPERF in guiding management and staff training decisions depends on the capacity to segment customer service into factors that are meaningful to staff and trainers in an Australian small business context. Whilst other studies have already explored the question of dimensionality, the present study took the additional step of not only identifying but also examining the relative difficulty of achieving high standards in the different dimensions in a training context. An associated aim of the present study was to see what improvements could be made to the scale that would lead to a better definition of the core dimensions of service quality. Three main statistical techniques were used to achieve these aims. The first was traditional item analysis focussing particularly on developing subscales that had high internal consistency. The SPSS RELIABILITY routines were used for this purpose. The second set of techniques included exploratory and confirmatory (LISREL) factor analysis. These were used to help define the main constructs underlying the SERVPERF scale and the relations among these constructs. The final technique involved the application of the principles of Rasch analysis to gain a clearer picture of the role of individual items and groups of items in the measurement of service quality.

The study is presented as two parts. The first part is based on what will be termed Dataset 1 and concerns the initial analysis of the 22-item scale and the derivation of a 15-item SERVPERF scale. The second part is based on three further datasets and describes further validation of the shortened version of SERVPERF which is to be called SERVPERF-M.

Method

Participants

The project involved case studies of four small retail businesses within provincial cities in South East Queensland. Each business employed both full-time and part-time staff. All firms were owner-managed and had operated in the area for many years. The majority of staff were long-term employees and some had family links with the company. In recent years, the firms had faced increased competition from major retail companies, especially through the development of regional shopping complexes. All four firms operated in strip retail locations.

Materials

SERVQUAL consists of 22 pairs of items: one member of each pair assessing the customer's expectations, the other assessing perceptions of service quality. Service quality is determined by calculating the difference between expectations and perceptions for each item. This aspect of the administration of SERVQUAL has been criticised on the grounds that there is a lack of evidence supporting the expectation-performance gap as a predictive measure of service quality (Cronin, Steven & Taylor, 1992). Other researchers suggested that the calculation of difference scores could result in poor reliability, expecially if the expectations scale was truncated by ceiling effects (Brown, Churchill, & Peter, 1993). This would happen if customer expectations of service are very high, as is likely for some types of retail provider. Cronin and Taylor (1992) found that the performance component out-performed SERVQUAL in terms of reliabilities, providing some evidence to support these concerns. Parasuraman, Berry and Zeithaml (1993), on the other hand, although they conceded the logic of the criticism, argued that truncation had little effect on reliabilities in practice. Given the uncertain benefits to be gained from the collection of both expectation and perception scores for each item, and the continued support for the perception-based aspect of the measure (Cronin & Taylor, 1992), it was decided to use the one-stage (SERVPERF) form of the survey in the present research.

Prior to administration, minor wording changes were made to SERVPERF to convert negatively worded items to positive items consistent with the recommendations of earlier researchers (Babakus & Boller, 1991; Parasuraman, Berry, & Zeithaml., 1991). Minor changes were also made to some questions to ensure that the text was familiar to Australian consumers. The first stage of this study involved administering the 22-item SERVPERF to 113 customers of the first firm. Following extensive psychometric analysis, a modified version of the instrument, SERVPERF-M, was developed. The new scale was basically a shortened form of the original SERVPERF. The new scale was administered to the remaining three businesses, thus forming three additional datasets. The second dataset was based on a sample of 63, the third on a sample of 75, and the fourth on a sample of 74. When pooled, the combined dataset contained 212 cases which were used for validation of SERVPERF-M.

Procedure

With the exception of the fourth sample, where data was collected by survey some months after a purchase, the SERVPERF and SERVPERF-M scales were administered at the point-of-sale by two researchers who were trained in the use of the instruments. Customers were approached only after they had made a purchase. This was to ensure that they were actual customers and not merely "browsers" and therefore familiar with the service they were being asked to evaluate. Customers at firm 1 (Dataset 1) were asked to give their perceptions of service quality for the 22-items forming the original SERVPERF scale. Customers at firms

2, 3, and 4 answered the shortened SERVPERF-M scale. All responses were recorded on a 7-point Likert sale using categories ranging from “strongly agree” (7) to “strongly disagree” (1).

Results

Given the challenges already made to the supposed structure of service quality scales (e.g., Cronin & Taylor, 1992), the first stage of data analysis involved the use of exploratory factor analytic routines from the SPSS package to check the dimensionality of the full 22-item SERVPERF scale. In accordance with the procedures followed by Parasuraman et al. (1991) in their validation study, the principal axis factoring technique was used with the solution constrained to five factors subjected to oblique (oblimin) rotation. Although a confirmatory factor analytic solution is to be reported later, it is worth showing the solution obtained with exploratory routines with just these constraints. The pattern matrix and factor intercorrelation matrix is shown in Table 1.

Table 1
Pattern Matrix for Full SERVPERF Scale Using Exploratory Factor Analysis

	Factors				
Items	I	II	III	IV	V
Q1	.10	.74	.01	-.08	.08
Q2	-.03	.67	.11	.21	-.11
Q3	-.14	.34	.15	.27	.28
Q4	.12	.65	-.00	.00	.08
Q5	.14	-.13	.07	.36	.29
Q6	-.08	.09	-.01	.89	-.05
Q7	.40	.12	.10	.56	-.07
Q8	.22	-.21	.14	.47	.41
Q9	.13	-.18	-.07	.08	.37
Q10	.64	.12	.03	.19	.19
Q11	-.03	.05	.76	.03	.03
Q12	.05	-.05	.50	.21	.34
Q13	-.01	.02	.84	.11	-.11
Q14	.56	.02	.33	.13	.11
Q15	.46	.24	.44	-.06	.00
Q16	.15	.23	.17	.39	-.06
Q17	.47	.19	-.14	.12	.28
Q18	-.19	.30	.53	-.12	.49
Q19	.63	.20	.04	-.03	-.09
Q20	.17	.11	.22	-.11	.58
Q21	.43	.03	.64	-.13	.07
Q22	.50	.15	-.11	.04	.32

Factor Intercorrelation Matrix

I	1.00				
II	.35	1.00			
III	.36	.43	1.00		
IV	.34	.23	.33	1.00	
V	.40	.27	.39	.32	1.00

The five factor solution accounted for 70% of the variance and happens to coincide with the solution that would have been offered under root one criterion. The first factor was defined by items mostly from the Assurance and Empathy scales, with item 7 (Reliability) and item 10 (Responsiveness) also showing some tendency to load here. The second factor

was defined by items 1-4 and is clearly the Tangible factor. The third factor picked up variance from items 11-13 (Responsiveness) and also from items 14-15 (Assurance) plus item 18 and 22 (Empathy). This may be the Personal Attention factor from Cronin and Taylor. The fourth factor was defined mostly by items 4-8 (Reliability). The last factor picks up variance from the Reliability and Empathy scales and, again, seems to be picking up aspects of Personal Attention. Whatever the labels, it is apparent that the five factors identified by Parasuraman et al. (1988, 1991) have not emerged clearly here. Reliability and Tangibles were readily identifiable but the other three were somewhat mixed. This is similar to the solution reported by Cronin and Taylor (1992). It is also interesting to note that the first principal component in the unrotated solution accounted for almost 45% of the variance, suggesting that there may be a strong common factor underlying the five factors obtained here. This is supported by the correlations among the factors themselves.

The solution obtained above was close enough to the theoretical structure of SERVQUAL to suggest that a traditional item analysis may show that the five scales have good internal consistency estimates. In the next stage of analysis, Cronbach's alpha was calculated for each of the five scales. These analyses indicated that all five scales had at least reasonable reliability with estimates ranging from .69 for the Reliability scale to .86 for the Empathy scale. In all cases, however, the analyses indicated that reliability could be improved if some items were deleted. In a second series of reliability analyses, three criteria were applied to help decide whether an item should be deleted. The first criterion was the impact of the item on the reliability of the scale - items that lowered the alpha estimates for each scale were considered for removal. The second criterion was the location of the item in the exploratory analysis shown in Table 1 - items that did not line up with other items in its scale were considered for deletion. The third criterion was the suitability of an item for use in an Australian context, as judged by feedback from the interviewers who administered the scale. For example, Item 7 ("They are dependable") contributed to the internal consistency of the Reliability scale but interviewers had some trouble explaining its meaning and thus it became a candidate for deletion. It is interesting to see that this item loaded on two factors in Table 1, perhaps reflecting the confusion in interpretation. All these criteria were applied jointly to make decisions about item deletions. Descriptive statistics for the revised scales are shown in Table 2.

Table 2
Means, Standard Deviations, and Reliabilities of SERVPERF-M for Dataset 1

Variables	Items	Mean	SD	Alpha
Tangibles	1,2,4	17.02	3.0	.80
Reliability	5,6,8	18.84	2.12	.74
Responsiveness	11,12,13	18.94	2.25	.83
Assurance	14,15,17	19.29	1.99	.82
Empathy	18,20,22	18.96	2.41	.82

It can be seen that this shortened form of the SERVPERF (SERVPERF-M) contains 15 items with three items per scale. Internal consistency estimates for the various scales were

generally satisfactory, although the estimate for the Reliability scale itself was only moderate. This could have been improved by choosing a slightly different mix of items but it was felt that the three items chosen were the most appropriate and that significant improvements could be made by changing the wording in future administrations.

The 22-item version of SERVPERF was developed to measure perceptions of five different dimensions of service quality. To support an argument for the use of the 15-item version in its place, it was important to demonstrate that the shorter version taps these same latent traits. Confirmatory factor analytic (CFA) procedures from the LISREL 8 (Joreskog & Sorbom, 1993) package were used for this purpose. In CFA, the researcher posits an *a priori* structure and tests the ability of a solution based on this structure to fit the data by demonstrating that: a) the solution is well defined; b) the parameter estimates are consistent with theory and *a priori* predictions; and c) the χ^2 likelihood ratio and subjective indices of fit are reasonable (McDonald & Marsh, 1990). For present purposes, the Non-Normed Fit Index (NNFI) recommended by McDonald and Marsh (1990) and the Root Mean Square Error of Approximation (RMSEA) recommended by Browne and Cudeck (1993) were considered as well as the usual χ^2 measure of goodness of fit. The NNFI varies along a 0-1 continuum in which values greater than .9 are taken to reflect an acceptable fit. Browne and Cudeck (1993) suggest that an RMSEA value of .05 indicates a close fit and that values up to .08 are still acceptable.

The first model tested was based on the full set of 22 items described in Parasuraman et al. (1988) with items 1-4 acting as indicator variables for the Tangibles factor, items 5-9 for Reliability, items 10-13 for Responsiveness, items 14-17 for Assurance, and items 18-22 for the Empathy factor. The second model used only the 15 items in SERVPERF-M with each factor tapped by the three indicator variables shown in Table 1. In both cases, the model allowed for five correlated latent traits. Factor loadings and factor intercorrelations for both solutions are shown in Table 3.

Table 3
LISREL Factor Pattern Matrices for 22-item and 15-item Versions of SERVPERF

Items	Full Version of SERVPERF Factors					Shortened Version of SERVPERF Factors				
	I	II	III	IV	V	I	II	III	IV	V
Q1	.78					.82				
Q2	.77					.74				
Q3	.53					----				
Q4	.73					.75				
Q5		.55					.56			
Q6		.64					.56			
Q7		.81					----			
Q8		.81					.92			
Q9		.47					----			
Q10			.53					----		
Q11			.73					.75		
Q12			.81					.83		
Q13			.77					.77		
Q14				.94					.94	
Q15				.87					.88	
Q16				.58					----	
Q17				.55					.55	
Q18					.80					.85
Q19					.56					----
Q20					.78					.82
Q21					.86					----
Q22					.77					.73
	Factor Intercorrelations					Factor Intercorrelations				
I	1.00					1.00				
II	.58	1.00				.34	1.00			
III	.64	.77	1.00			.57	.70	1.00		
IV	.71	.82	.83	1.00		.68	.70	.79	1.00	
V	.75	.70	.84	.87	1.00	.72	.61	.78	.82	1.00

Note:
---- indicates that this item omitted in shortened version

It can be seen that all items load on their respective factors. Despite this, the fit is not particularly good. For the full version of SERVPERF, $\chi^2 = 503.2$, df = 199; NNFI = .77, and RMSEA = .12. These indices suggest that the model approaches a fit, but that it should be rejected. In such cases, it is usual to inspect the modification indices to see where the misfit occurs. In the present case, that is hardly necessary. The high factor intercorrelations suggest that there is a great deal of overlap among these factors and, as a consequence, it is unrealistic to expect an item to load on just one factor. The modification indices confirm this: if one allows more factorial complexity - that is, some of the items loading on more than one factor, a good fit can be obtained.[1] The fit for SERVPERF-M ($\chi^2 = 172.08$, df = 80; NNFI = .88; RMSEA = .10), was better than that for the longer form (χ^2 difference = 331.12, df = 119, $p < .05$) but still just outside the boundaries of what would be considered acceptable fit. The reasons are the same as before: if items are allowed to load on more than one factor or if the error terms for some items are allowed to covary, the fit becomes acceptable. For example, if the error terms for items 2 and 6 and also for items 22 and 18 are allowed to covary, two of

[1] Given the high factor intercorrelations, it was considered advisable to also test a single-factor model at this point. The fit was only marginally poorer than that given by the five-factor solution.

the fit indices for SERVPERF-M move within acceptable ranges. This indicates that a model that contrains items to serve as indicators for single factors is overly restrictive.

Item Response Analysis of SERVPERF-M

The high factor intercorrelations for both the long and short forms of the SERVPERF scale suggested that there is a common service quality dimension underlying all the items in this scale. Other aspects of the data analysis also suggested that this is the case: a) in exploratory factor analyses, the first eigenvalue accounted for 45% of the variance, more than five times that accounted for by the next eigenvalue; b) reliability analyses with a single 15-item scale indicated that it has an internal consistency estimate of .91; c) LISREL analyses revealed that a model with all items serving as indicator variables for a single underlying dimension produced indices of fit that were not much different than those associated with a five-factor model. For this reason, it was decided to treat the scale as unidimensional and to analyse the sections of the service quality dimension sampled by SERVPERF-M.

It is difficult to do this using classical item analysis procedures because these procedures give little direction with respect to quantifying the affective value of items. In order to help decide such questions, the latent trait models can be particularly useful. One pertinent model within the latent-trait family of models is the Rasch Model (1960) which has been applied to dichotomous scales (Wright & Stone, 1979) and rating scales (Wright and Masters, 1982; Andrich, 1975, 1981, 1982). The particular model introduced by Andrich (1975) is called the multiplicative binomial and is of interest here because it provides a perspective for unifying Thurstone's (1928) procedures for item scaling and the Likert procedure for attitude measurement. A brief introduction to some aspects of this model will help to illustrate how it can be used to help provide additional information about service quality.

The main task of Rasch analysis consists of defining a latent continuum and then estimating the location of items and individuals on this continuum. When attitudes are being studied, the same latent continuum represents both the *affectivity* of the items and the *attitudes* of the persons taking the test. In a simple dichotomous test, the probability of a person selecting a particular item is a function of the affectivity of the item and the attitude of the person. Because the probability is a function of both of these, it is not a single value and is usually represented diagramatically as an item characteristic curve. Curves for a number of different items are shown in Figure 1.

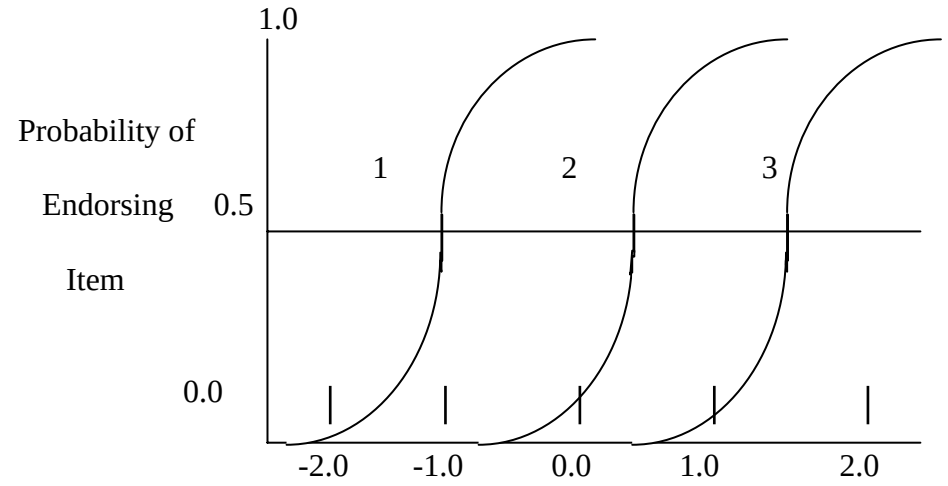

Figure 1
Item Characteristic Curves

If this were an ability test, item one would be described as an easy item because one would not need a lot of ability to get the item correct. In the case of attitude scales, the term "affectivity" is used instead of "ability" and we speak of "endorsing an item" rather than "getting an item correct". Otherwise, the item characteristic curves can be interpreted in the same way for attitude scales and ability tests. The shape of the item characteristic curve shows the increasing probability of endorsing the item as affectivity increases. Item three will only be endorsed by those with higher levels of affectivity. Someone with an affectivity level of about zero is scarcely likely to endorse the item but someone with a level of one is almost certain to endorse it. Item characteristic curves are very useful because they provide valuable information about the sections of the latent trait that are sampled by the scale items. This can lead to decisions to include more items to cover areas not presently covered or to delete some items in areas that are well-covered.

The ASCORE program (Andrich, Sheridan, & Lyne, 1991) contains an implementation of the multiplicative binomial that is capable of handling data generated by the response format used in SERVPERF. Among other things, it calculates threshold estimates for the multiple categories of each item. For the purposes of these analyses, SERVPERF was treated as a unidimensional scale. As indicated above, and also in Cronin and Taylor (1992), there is some support for this interpretation of the factorial structure of SERVPERF and the internal consistency (alpha) estimate for the shortened (15-item) version was .91 in the present study. Although the probability of an item being endorsed varies according to the person and is properly described by the item characteristic curve, affectivity estimates for each item are shown as location estimates. The estimates represent the point on the curve where the probability of endorsing the item is .5. This point is shown for each item in Figure 1 as the intersection of the line representing p = .5 and the curve for each item. If one were to drop a line from each of these intersections to the baseline, the location estimates could be obtained. The location estimates for the 15 items in the shortened form of SERVPERF are depicted graphically in Figure 2..

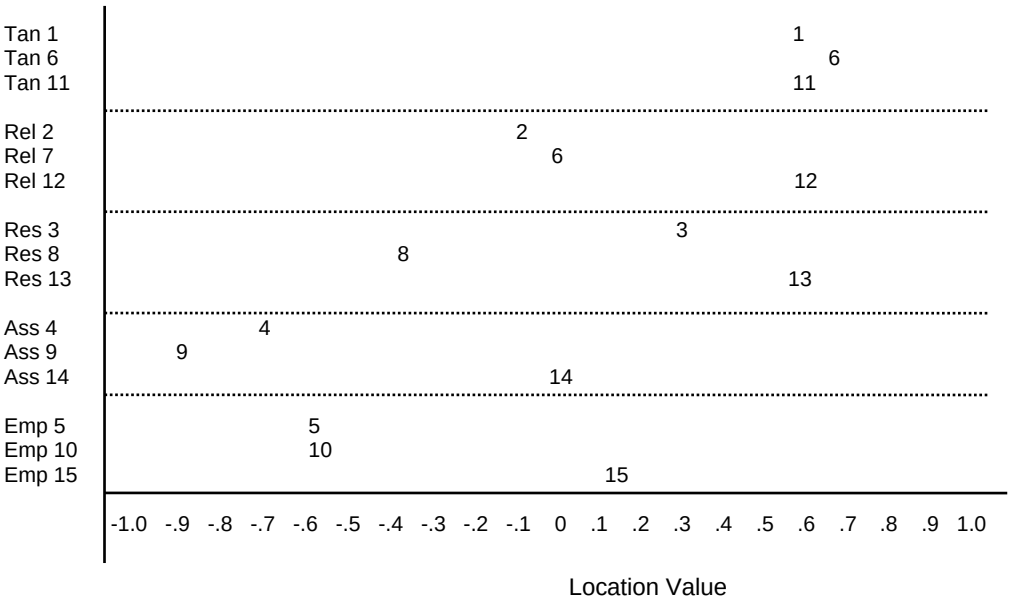

Figure 2
Location Order of Items: Dataset 1

The items have been grouped into factors to make it easier to see the areas of the service quality continuum tapped by the five factors. Two things are immediately apparent from this graph. The first is that all items have location estimates that fall between -1 and +1. It is possible for the items to range between -3.0 and +3.0 (and beyond); so the range is a little narrow here. The second point to note is that the factors tend to cover different parts of the continuum. Thus, Empathy and Assurance cover the lower end, Responsiveness the middle, and Reliability and Tangibles the upper end. In practical terms, this means that Empathy and Assurance are usually the first aspects of service quality experienced. It is possible for a company to rate highly on these but still fall short of true service quality. They are like the easy items in an ability test. Responsiveness, on the other hand, covers a broad span and is generally a more difficult quality to achieve. The second item in this scale (Item 8: "They provide a service at the time they promise to do so") is endorsed quite readily but the third item (Item 13: "Employees of XYZ are too busy to respond to customer requests promptly" - a reverse-scored item) is much more difficult. Tangibles is certainly the most difficult of the factors to achieve. According to the mathematical principles underlying item response theory, endorsement of a Tangible item implies endorsement of all other aspects as well.

Validation of SERVPERF-M

Pruning items in the manner demonstrated here will usually improve the properties of a scale in the immediate study because such changes target items that have not performed in that particular dataset. Having made such changes, it is important to show that the remaining items are robust and will not themselves become candidates for deletion in further validation studies. To check this possibility, SERVPERF-M was adminstered to three more business groups. These datasets were analysed separately before being combined to form a larger (N = 212) set for analysis.

Dataset 2

The sample for this dataset comprised 63 customers of a small retail business. Data was collected in the manner described earlier using SERVPERF-M. Descriptive statistics and reliability estimates are shown in Table 4.

Table 4
Descriptive Statistics and Reliabilities for Dataset 2

Scales	Mean	S.D.	Alpha
Tangibles	15.60	3.57	.85
Reliability	17.21	3.13	.89
Responsiveness	18.17	2.77	.91
Assurance	18.65	2.44	.91
Empathy	18.85	2.41	.92

Internal consistency estimates for all five scales were excellent, ranging from .85 to .91. Although the internal consistency estimates were high and all variables had high loadings on their respective factors, confirmatory factor analysis showed that the five factor structure did not fit this data set (χ^2 = 213.9, df = 80, $p < .001$; NNFI = .77; RMSEA = .16). The main reason for the misfit was undoubtedly the degree of overlap among the factors. The model tested allowed each item to load only on its own factor. To achieve this, some of the factors were pulled very close together and four of the resulting factor intercorrelations were above .90. Even then, a good solution could be obtained only if items were allowed to load on other

factors. A reasonable fit could also be achieved by dropping items seven and fourteen, but this solution still left the factors very highly correlated.

With the five factor model clearly inappropriate for this data set, exploratory analysis was employed to determine the factor structure with no constraints applied. The principal axis factoring (PAF) technique with oblique rotation was used for this analysis. Root one criterion and scree plots indicated that two factors were all that were needed to account for 78% of the variance in the matrix. The factors were correlated (r = .57) and the first factor accounted for 70% of the variance in the unrotated solution, suggesting that a single factor model would also have provided a reasonable fit to these data. The first factor in the two-factor solution included Reliability, Responsiveness, Assurance, and Empathy. The second factor was defined exclusively by the Tangibles items.

Dataset 3

The sample for this dataset comprised 75 customers of a small retail business. Data was collected in the manner described earlier using SERVPERF-M. Descriptive statistics and reliability estimates are shown in Table 5.

Table 5
Descriptive Statistics and Reliabilities for Dataset 3

Scales	Mean	S.D.	Alpha
Tangibles	17.05	2.91	.80
Reliability	16.71	2.70	.78
Responsiveness	17.61	2.73	.83
Assurance	18.29	2.50	.85
Empathy	18.03	2.54	.80

Again, the internal consistency estimates for all scales were very good. The LISREL fit indices for a five-factor model, however, were not satisfactory (χ^2 = 153.44, df = 80, p < .001; NNFI = .83; RMSEA = .12). Correlations among the Reliability, Responsiveness, Assurance, and Empathy factors were all above .95. Exploratory factor analysis using PAF with oblique rotation and root one criterion yielded a three-factor solution that explained 74% of the variance. The first factor was defined by items from the Empathy, Assurance, and Responsiveness scales. The second factor by the Tangibles scale. Items from the Reliability scale loaded on the first factor but - with the exception of item seven - also helped to define a third factor. The first factor was correlated with the second (r = .23) and the third (r = .38) but there was no correlation between factors two and three (r = -.01).

Dataset 4

The sample for this dataset comprised 74 customers of a small retail business. Data was collected in the manner described earlier using SERVPERF-M. Descriptive statistics and reliability estimates are shown in Table 6.

Table 6
Descriptive Statistics and Reliabilities for Dataset 4

Scales	Mean	S.D.	Alpha
Tangibles	17.24	2.87	.74
Reliability	17.17	3.74	.80
Responsiveness	17.45	4.01	.87
Assurance	18.42	2.92	.79
Empathy	18.60	2.86	.87

The situation was the same as for the previous two datasets. In fact, the descriptive statistics were remarkably stable across all four datasets with the alpha coefficients in each case indicating that the items in the scales were quite homogeneous. As with the previous two datasets, however, confirmatory factor analysis indicated that a five factor model was not appropriate (χ^2 = 244.84, df = 80, p < .001; NNFI = .70; RMSEA = .19). Two of the factor intercorrelations were above .90. A much better fit could be obtained by dropping items seven and fourteen but the fit was still not within acceptable limits. The same exploratory factor analytic routines used with the previous datasets yielded a three-factor solution that explained 75% of the variance. The three factors obtained were almost identical to those found in Dataset 3: Empathy, Assurance, and Responsiveness scales defined the first factor, Tangibles the second, and Reliability the third (once again, item 7 did not load on factor 3). The first factor was correlated with the second (r = .48) and the third (r = .47) and this time factors two and three were also correlated (r = .35).

Combined Datasets 2-4

The number of cases in the previous three datasets was somewhat smaller than desirable for factor analysis. We have chosen to present the three separately to check the reliability of the features of the SERVPERF-M scale noted in the first part of this study. These will be discussed later. In the final stages of data analysis, to create a larger dataset more suitable for multivariate analysis, data from samples 2-4 were combined to form a set of 212 cases and a psychometric analysis of SERVPERF-M was conducted. Conventional item analysis showed that internal consistency estimates for the scales were all at least .80 but could be improved further by deleting some items[2]. This was not done because internal consistency was not seen to be the major problem with the subscales of SERVPERF-M. As can be seen from the analyses above, internal consistency estimates can be quite misleading. Groups of items that tap the same underlying dimension can be used to form separate scales, each with high internal consistency estimates. There is little point, however, in having very similar sub-scales in the same instrument. The key question with SERVPERF-M is not whether its subscales are internally consistent but whether they measure separate constructs.

Principal axis factoring with oblique (OBLIMIN) rotation suggested that a two factor (root one criterion) or a three factor (scree plot) solution was all that was required to account for a major part of the variance in this matrix. The two factor solution accounted for 63% of the variance with the two factors corresponding almost exactly to those found in Dataset 2. The three factor solution accounted for 68% of the variance and corresponded to the solution found in Datasets 3 and 4 with Tangibles defining one factor, Reliability (minus item seven) a second, and Reliabity, Responsiveness, Assurance, and Empathy defining a third factor. If the

[2] The Tangibles scale by deleting item one, the Reliability scale by dropping item seven, item fourteen can be dropped from the Assurance scale, and item fifteen from the Empathy scale.

items that suppressed the internal consistency estimates for their respective scales were deleted, a one factor solution was obtained that accounted for 58% of the variance. Confirmatory factor analysis indicated that the five-factor model provided a poor fit to the combined dataset (χ^2 = 489.05, df = 80, p < .001; NNFI = .76; RMSEA = .16) with very high correlations among the resulting factors.

The last stage of data analysis consisted of a Rasch analysis of the 15-item SERVPERF-M scale using ASCORE (Andrich et al., 1991). The location values of the 15 items are shown in Figure 3.

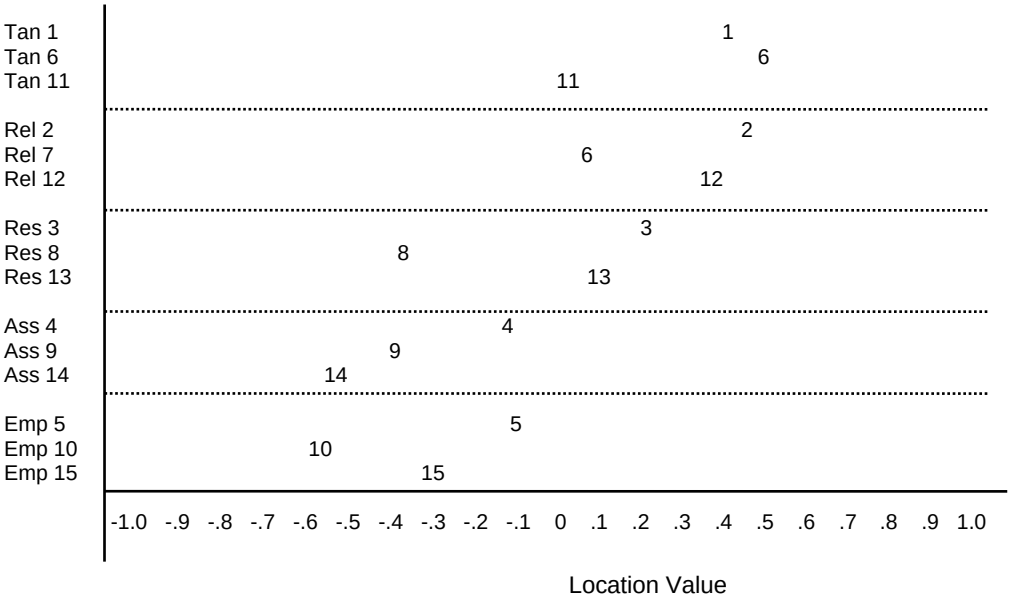

Figure 3
Location Order of Items: Datasets 2-4

It can be seen that the affectivity values of the items are much the same as those shown in Figure 2, although the spread is even more restricted. The correspondence between the two outcomes can be seen more clearly if the values from the two studies are used to form a scatterplot. If the location values were exactly the same, all points would lie in a straight line. If there was no correspondence between the outcomes, the points would be randomly spread in the two dimensional space formed by the axes of the scatterplot. Figure 4 shows the outcome.

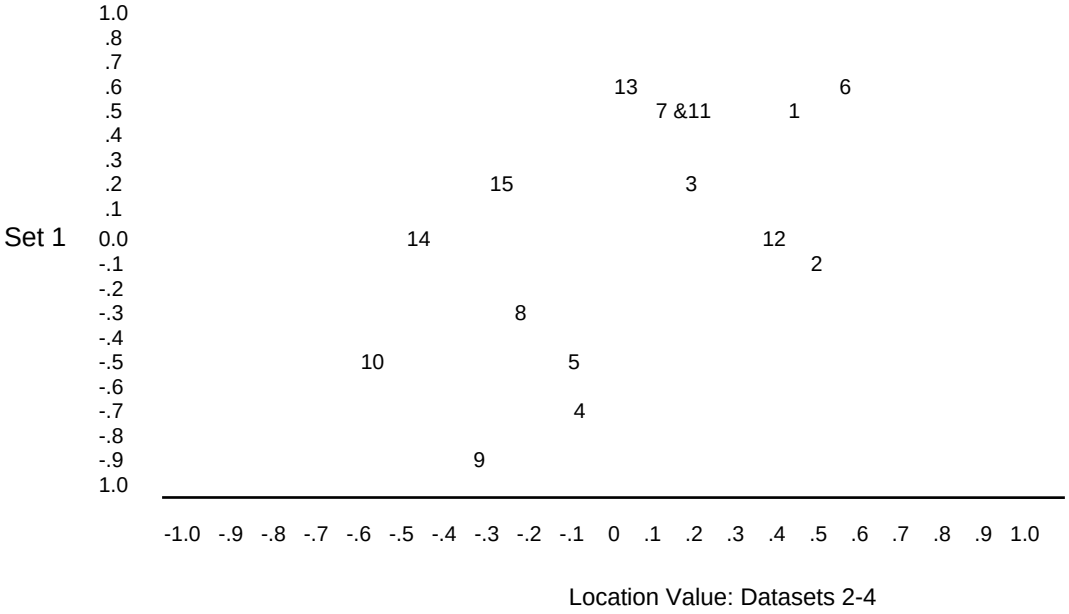

Figure 4
Correspondence between Item Locations Dataset 1 and Datasets 2-4

The points do not form a straight line but there is a strong linear relationship between the two sets of values. Items that had a high affectivity rating in Dataset 1 tended to have a high rating in the combined Dataset 2-4. Thus, the most difficult aspects of quality relate to the Tangibles dimension, the easiest relate to Empathy.

Discussion

The aim of the first part of this study was to explore the psychometric properties of the 22-item version of SERVPERF. This led to the formation of a 15-item version named SERVPERF-M and much of the early part of this paper concentrated on examing the features of this shortened scale in a sample of 113 respondents drawn from small businesses. There is no doubt that the shortened form had acceptable psychometric properties and measured the same underlying traits as the longer version. In fact, LISREL analysis suggested that it provided a better fit than the 22-item version. Aside from the question of fit, it was apparent that the factors extracted from the 22-item and 15-item versions of the scale were identical. Loadings for the items on their respective factors were very similar in both forms. The result of reducing the scale appears to have been a reduction in the size of the factor intercorrelations, rather than any change in the loadings per se. Very high factor intercorrelations are a worrisome feature of test instruments, so this change favours the use of the short version. The higher internal consistency estimates obtained with the short version also favour its use.

The second part of the study was basically a validation of the 15-item scale and an extension of the work undertaken in the first part. The three datasets analysed here revealed some interesting aspects of the SERVPERF-M scale. The first was that the five-factor structure was not validated in any of these datasets. The Tangibles factor did emerge in all three sets and there was some tendency for the Reliability factor to emerge but Responsiveness, Assurance, and Empathy could not be separated. When these factors were separated using confirmatory analysis, the resulting fit was poor. In order to achieve good fit, items had to be allowed to load on other factors with the result that the factors were highly

correlated. Parasuraman et al. (1991) made this same observation when noting that SERVPERF has a "diffused" factor pattern and high factor intercorrelations. They argued that the overlap among the dimensions is a function of a tendency for respondents to rate a particular company highly on all dimensions (Parasuraman et al., 1991, p.443). That is certainly what happened in this study with the means for all SERVPERF subscales towards the upper limit. This comment, however, still implies some criticism of the scale itself because it means that the items lack discriminability.

It is difficult to pinpoint the source of these problems using conventional item and test validation techniques. The internal consistency estimates do not tell us anything much about the dimensionality of a scale. As pointed out earlier, if a unidimensional scale were split into two subscales, both subscales should have good internal consistency but it would be wrong to regard them as representing separate dimensions. Exploratory factor analysis is commonly used to help resolve issues of dimensionality but it has serious limitations in the present situation. Thus, it may be possible to force a five-factor exploratory solution that yields a loading pattern approximating the hypothesised structure of SERVPERF-M but that is rather meaningless if the solution does not approach simple structure and if the factors themselves are highly correlated. Confirmatory factor analysis can prove very useful in this situation in that it allows researchers to describe an exact model and test its fit. The technique was employed here for that purpose and indicated that a five-factor model fitted the first dataset but not the other three. What the confirmatory approach cannot do is help determine why factors separate in one dataset but not in others. One possible reason is that the dimensions themselves vary from industry to industry. In some industries, for example, it may not make sense to distinguish among the dimensions of Assurance, Responsiveness, and Empathy. This suggestion, however, runs contrary to the basic purpose of SERVPERF which is claimed to measure *core* aspects of service quality.

A second reason is suggested by the Rasch analyses conducted in this study. One very important aspect of a Rasch analysis is that it locates items on a linear continuum that has a fixed zero point and equal units of measurement (logits) extending in either direction from this point. The location estimates are sample free, or nearly so in the present case (see Figure 4), and give an indication of the "ease of endorsement" of an item. It can be seen from Figures 2-4 that the subscales of SERVPERF-M, if arranged in the order shown in these figures, form a progression. When viewed in this light, service quality can be treated as a construct that has a number of more or less distinct stages. Empathy and Assurance are the first encountered; businesses should find it rather easy to achieve standards in this area. Responsiveness and Reliability are somewhat harder to achieve and Tangibles is the hardest area in which to be rated a success. This analysis puts the question of dimensionality in a somewhat different light: rather than emphasising qualitative aspects of the scale, such as the nature of the constructs tapped, it focuses concern on the continuum that is service quality and where businesses might be located on this continuum. This is surely a more fundamental concern.

One of the other purposes of Rasch analysis is to indicate the sections of the continuum that are not being tapped by existing items. This can be particularly useful for deciding where items need to be developed. In this study, the Rasch analysis confirms the rather narrow range of the continuum tapped by the 15 items in this scale. All items have affectivity values between -1 and 1. It would certainly be desirable to extend the range of the scale somewhat. This would overcome the problem of negative skewness and lead to a better distribution of responses with the possibility of better discrimination among the supposed five latent traits. This is certainly one of the recommendations of the study. It will always be

difficult to determine the factor structure of SERVPERF or SERVPERF-M whilst scores are clustered so tightly at the upper end of the distribution. There is a need for items that respondents will find more difficult to rate highly. It should be noted that none of the items discarded for the shortened version of SERVPERF had this quality. Although not reported, a Rasch analysis was conducted on the full 22-item scale. Six out of the seven discarded items had location values very close to zero. The remaining items did not extend the scale beyond the bounds covered by the shortened version.

In conclusion, the study set out to validate the 22-item version of SERVPERF in an Australian setting but, in the process, derived a shortened 15-item version of the scale. The shortened version performed as well as, if not better than, the longer 22-item version. The internal consistency estimates for the scale were actually better than those for the longer scale and, in the first dataset at least, the factorial structure of SERVPERF was more clearly distinguishable in the shorter version. Some concerns emerged, however, during the analysis of further datasets where it became apparent that there was a great deal of overlap among some of the factors of SERVPERF-M. This tendency has already been noted in the literature (Parasuraman et al., 1991). Two explanations were considered, both of them already discussed by these same authors. One possibility is that perceptions of service quality are not determined by five core dimensions but perhaps by two or three dimensions. Against this, a number of studies have reported a five-factor structure (see Parasuraman et al., 1991) and this model was supported by the exploratory (but not the confirmatory) factor analysis of the first dataset in this study. There is also the evidence provided by the Rasch analysis where the items, when grouped under subscale headings, appear to represent different areas of the affectivity continuum with Empathy at one end and Tangibles at the other.

A second possibility is that the items do not discriminate well among the dimensions. That is, there is a tendency for respondents to rate businesses highly on all dimensions and this makes it difficult to separate the dimensions in structural analyses. The high means for all subscales and the somewhat narrow range of the location estimates provided by the Rasch analysis indicates that this explanation is probably the correct one. According to Parasuraman et al. (1991), the SERVPERF scale is particularly susceptible to this response tendency. In the present study, this tendency might have been exacerbated by soliciting responses only from those who had actually made a purchase. It is possible that customers who were dissatisfied with the service left without making a purchase. In this case, the response bias could be a function of the sampling technique, something that may disappear if a different sampling strategy is used. Against this, the location estimates provided by Rasch analysis - as with all other item response models (IRT) - are meant to be sample-free, so this explanation may not hold the key to the skewness observed in item responses. It is also possible that the fault lies with the items themselves. The solution to this lies in the development of new items that do discriminate at the upper end of the service quality continuum. If this is not done, we would argue that it is more sensible to treat the SERVPERF scale as unidimensional and to talk about achieving stages of quality rather than separate qualitative states.

References

Andrich, D. (1975). The Rasch multiplicative binomial model: Applications to attitude data. Research Report Number 1, Measurement and Statistics Laboratory, Department of Education, University of Western Australia.

Andrich, D. (1981). Rasch's models and Guttman's principles for scaling attitudes. Paper presented at the International Conference on Objective Measurement. Chicago: The University of Chicago, Department of Education.

Andrich, D. (1982). An extension of the Rasch model for ratings providing both location and dispersion parameters. Psychometrika, 47, 105-113.

Andrich, D, Sheridan, B., & Lyne, A. (1991). ASCORE: Manual of procedures. Faculty of Education, University of Western Australia.

Brown, T. J., Churchill, G. A., & Peter, J. P. (1993). Research note: Improving the measurement of service quality. Journal of Retailing, 69, 127 - 139.

Browne, M.W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K.A. Bollen & J.S. Long (Eds.): Testing structural equation models. Sage Publications.

Cronin, J. J., & Taylor, S. A. (1992). Measuring service quality: A Reexamination and extension. Journal of Marketing, 56, 55-68.

Finn, D. W., and Lamb, C. R. (1991). An Evaluation of the SERVQUAL scales in a retail setting. Advances in Consumer Research, 18, 483 - 490.

Gagliano, K. B., & Hathcote, J. (1994). Customer expectations and perceptions of service quality in retail apparel speciality stores. Journal of Services Marketing, 8, 60 - 69.

Jöreskog, K.G., & Sörbom, D. (1993). New Features in LISREL 8. Chicago: Scientific Software.

McDonald, R.P., & Marsh, H.W. (1990). Choosing a multivariate model: Noncentrality and goodness-of-fit. Psychological Bulletin, 107, 247-255.

Parasuraman, A., Berry, L., & Zeithaml, V. A. (1991). Refinement and reassessment of the SERVQUAL scale. Journal of Retailing, 67, 420 - 450.

Parasuraman, A., Berry, L.L., & Zeithaml, V.A. (1991). Refinement and reassessment of the SERVQUAL scale. Journal of Retailing, 67 (4), 420-450.

Parasuraman, A., Berry, L. & Zeithaml, V. A., (1993). More on Improving Service Quality Measurement. Journal of Retailing, 69, 140 - 147.

Parasuraman, A., Zeithaml V.A., and Berry, L. (1988). SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. Journal of Retailing, 64, 12 - 40.

Pine, J. & Tingley, J.C. (1993). ROI of soft-skills training. Training, (February), 55-60.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Education Research.

Thurstone, L.L. (1928). The measurement of opinion. Journal of Abnormal and Social Psychology, 22, 415-430.

Wright, B.D., & Masters, G.N. (1982). Rating scale analysis. Chicago: Uni of Chicago Press.

Wright, B.D., & Stone, M.N. (1979). Best test design. Chicago: Mesa Press.