Market Lett (2008) 19:157–170
DOI 10.1007/s11002-007-9031-0

Nonparametric item response theory for investigating dimensionality of marketing scales: A SERVQUAL application

Leonard J. Paas · Klaas Sijtsma

Published online: 19 January 2008

Abstract Assessing scale dimensionality is an important issue in the marketing literature. In an exploratory context, principal axis factoring and principal components analysis receive emphasis, while other fields apply suitable alternatives. This article introduces a promising procedure known as Mokken scale analysis. Using an empirical data set, we demonstrate how Mokken scale analysis complements principal axis factoring and principal components analysis for gaining understanding of the dimensionality of the items in the SERVQUAL instrument.

Keywords Exploratory scale analysis · Mokken scale analysis · Construct dimensionality · Service quality · SERVQUAL

1 Introduction

The marketing research literature on measurement scales is characterized by an ongoing discussion on assessment of scale dimensionality (e.g., Finn and Kayande 2004; Rossiter 2002; Voss et al. 2000). Confirmatory factor analysis is often used for this purpose (Gerbing and Anderson 1988). This approach is well established in marketing and is based on extensive statistical theory and empirical experience from other disciplines. When little is known about a construct or when the expected structure is not found by means of confirmatory analysis, the researcher may rely on

December 2007, resubmission of a conditionally accepted paper.

L. J. Paas (✉)
Department of Marketing, Faculty of Economics and Business Administration,
VU University Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands
e-mail: lpaas@feweb.vu.nl

K. Sijtsma
Department of Methodology and Statistics, Faculty of Social and Behavioral Sciences,
Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands
e-mail: K.Sijtsma@uvt.nl

exploratory procedures, e.g., principal axis factoring (PAF) and principal component analysis (PCA). The exploratory approach is highly relevant, as marketing measurement scales are not always fully developed or may have to be adjusted to novel domains of application.

This article introduces Mokken scale analysis (MSA; Sijtsma and Molenaar 2002) for either (a) testing or (b) exploring the dimensionality and the scalability of marketing measurement scales. Dimensionality refers to the number of dimensions that account for the data structure. Scalability refers to the possibility of ordering respondents and items on one or more scales defined by the items that set up the dimensions. In this study, MSA is used next to PAF and PCA, to test the well-known five-dimensional structure of the SERVQUAL instrument (Parasuraman et al. 1988, 1994) in two novel applications.

We refrain from assessing superiority of one method over another. Instead, each procedure has its idiosyncratic properties, leading to different interpretations of data patterns. We argue that such alternative interpretations are required especially in an exploratory context, which is the main focus of this contribution. When a researcher develops new theories or when uncertainty exists about an established construct's dimensionality, exploratory scaling procedures can search for interesting and informative features of items. Such procedures do not posit an a priori chosen model serving as a null hypothesis, as in confirmatory scale analysis. Instead, in the absence of a clear-cut direction provided by strong theoretical guidelines, trait validation amounts to finding and identifying dimensions. The researcher labels the found dimensions using item content and relationships between items and dimensions. As weaker theoretical foundations often result in weaker empirical data patterns, exploratory scale analysis tends to capitalize on the idiosyncrasies of the analysis procedure that is used. Using multiple exploratory scaling procedures reduces such method bias.

Our contribution is threefold. First, the report introduces MSA to the marketing literature, and demonstrates its application in marketing-scale analysis. Second, the report demonstrates how the combined results of multiple exploratory scaling procedures improve knowledge about scale dimensionality. Third, the reported empirical studies provide novel insights into service quality measurement, with implications for other marketing measurement instruments. The next section introduces nonparametric item response theory (IRT), followed by one of its important members, which is MSA. Then we assess SERVQUAL dimensionality in two novel domains, using MSA, PCA and PAF. After this follow conclusions.

2 Nonparametric item response theory

Notation Let a questionnaire consist of J items. Let X_j (j=1, …, J) denote the random variable for the score on item j, with realization x_j; and let x_j=0, 1, …, m, represent $m+1$ discrete item scores indicating increasingly higher levels of agreement with the presented statements. A J-dimensional vector $\mathbf{X}$, with realization $\mathbf{x}$, represents the J item-score variables. Q latent traits drive the responses to the J items in the test, θ_q denotes the qth trait (q=1, …, Q), and a Q-dimensional vector $\boldsymbol{\theta}$ collects the Q latent traits.

Assumptions IRT is a family of statistical models, which explain the relationships among a set of J items from Q latent traits. The common core of IRT consists of assumptions on three classes of formal characteristics of items and latent traits. First, most IRT models assume that the J-variate distribution of the data $\mathbf{X}$, conditional on latent trait vector θ, is equal to the product of the J marginal, conditional distributions:

$$P(\mathbf{X} = \mathbf{x}|\theta) = \prod_{j=1}^{J} P\left(X_j = x_j|\theta\right). \tag{1}$$

This local independence assumption implies that the latent traits in θ fully explain individual differences. Excluding one or more latent traits from θ, leads to an inequality in Eq. 1, implying that the model fails to fully explain the relationships between the J measured items.

The second class of formal characteristics addresses the relationship(s) between each item and the latent trait(s). The monotonicity assumption expresses that individuals with higher latent trait values are expected to have higher item scores. For binary items (i.e., x_j=0, 1), the probability for respondents to score 1, $P(X_j=1|\theta)$, is known as the item response function (IRF). Most IRT models assume this function is nondecreasing (i.e., monotone) in each latent trait θ_q (i.e., increases or stays constant across intervals of θ_q) when the other $Q-1$ traits are held constant. Several functions are used to define monotonicity for polytomous items. A commonly applied function is: $P(X_j \geq x_j|\theta)$, for x_j=1, …, m (for x_j=0, by definition $P\left(X_j \geq 0|\theta\right) = 1$; this is a non-informative result that IRT models ignore). Here the probability of scoring at least x_j on the polytomous item j is coordinate-wise nondecreasing in θ. Function $P(X_j \geq x_j|\theta)$ (x_j>0) is the item step response function (ISRF). Formally, for two respondents with latent trait vectors θ_v and θ_w, such that $\theta_v<\theta_w$ in each coordinate, monotonicity means that:

$$P\left(X_j \geq x_j|\boldsymbol{\theta}_v\right) \leq P\left(X_j \geq x_j|\boldsymbol{\theta}_w\right), \ for\ x_j = 1, \ldots, m. \tag{2}$$

For m=1, Eq. 2 reduces to monotonicity for binary items.

The third class of formal characteristics addresses the number of latent traits. Most IRT models assume that one latent trait explains individual differences. The unidimensionality assumption requires each scale-item to be a different manifestation of the same latent trait θ.

Scale properties Let the sum score on a set of J items be defined as $X_+ = \sum_{j=1}^{J} X_j$. For binary items, Grayson (1988) proved that the assumptions of unidimensionality (θ=θ), local independence (Eq. 1 with θ=θ), and monotonicity (Eq. 2 with θ=θ) together imply stochastic ordering of latent trait θ by the sum score X_+. Specifically, let t denote an arbitrary value of θ, and for the sum scores of two respondents, v and w, let $x_{+v} < x_{+w}$; then stochastic ordering of θ by X_+ implies:

$$P(\theta > t|X_+ = x_{+v}) \leq P(\theta > t|X_+ = x_{+w}), \text{ for each } t. \tag{3}$$

For polytomous items the three assumptions do not necessarily imply stochastic ordering, but for empirical data sets stochastic ordering is closely approximated (Van der Ark 2005).

Parametric and nonparametric IRT Parametric IRT models employ a parametric function for modeling the IRF/ISRF, such as the two-parameter logistic function (Singh 2004). These models express articulated expectations about the shape of the IRFs/ISRFs and the latent trait structure underlying responses to the items. Thus, they are less appropriate for exploratory scale analysis, when the researcher has little knowledge about the latent trait structure. IRT models imposing only order restrictions on the IRFs/ISRFs are called nonparametric (Sijtsma and Meijer 2007). Nonparametric IRT models are flexible data analysis tools and are thus better suitable for exploratory scaling.

3 Mokken scale analysis

3.1 Definition of Mokken's monotone homogeneity model

Monotone homogeneity model Unidimensionality, local independence, and monotonicity define Mokken's monotone homogeneity model (MHM; Sijtsma and Molenaar 2002). The MHM implies stochastic ordering of latent trait θ by the sum score X_+ (Equation 3). Thus, the MHM is a measurement model for ordinal person measurement.

Item-scalability coefficients The slope of the IRF/ISRF provides information about the strength of the relationship between an item and the underlying latent trait(s). Assume that the items are a reasonable summary of the same latent trait. Coefficient H_j (Sijtsma and Molenaar 2002, pp. 56–58) expresses the strength of the relationship of item j with the latent trait as estimated by the other $J-1$ items in the test, correcting for the artifactual effects of the J different item-score distributions on inter-item covariances. The MHM implies $0 \leq H_j \leq 1$ (the value 1 is the maximum value possible, but negative values are possible when the MHM does not hold; Sijtsma and Molenaar 2002, pp. 58–60). For a fixed distribution of θ and a fixed location of an IRF/ISRF relative to this distribution, steeper IRFs/ISRFs lead to higher values of H_j (Mokken et al. 1986). In this *theoretical* situation, the properties of the test and the population are known but in *empirical* data analysis the researcher lacks this knowledge, and information about dimensionality (including local independence) and monotonicity has to come from the values of the J item coefficients H_j.

Two results relate dimensionality and monotonicity to the scalability coefficients H_j (Sijtsma and Meijer 2007). First, if different (subsets of) items measure different latent traits, together the items are a mixture of these latent traits. If item j measures one of these latent traits, coefficient H_j thus is expected to have a low value. However, if the item set can be divided into subsets, each measuring a unique latent trait, and H_j coefficients are computed with respect to only the items in the homogeneous subset to which item j also belongs, these H_j coefficients are expected to be relatively high. An important feature of MSA is the automated item selection procedure (discussed shortly) that selects these item subsets from a larger item set by means of the H_j coefficients, thus exploring the dimensionality of the complete item set. Second, although a higher positive H_j expresses a stronger relationship between item j and the latent trait as estimated by the other items, it only *supports* the hypothesis of monotonicity but the evidence is not conclusive. This is comparable with regression analysis in which a regression coefficient can be high but the regression may not be entirely monotonous (Fox 1997). Additional support comes from other methods that estimate the IRF/ISRF for each item and allow for a more direct assessment of monotonicity (Sijtsma and Molenaar 2002, Chapter 3).

Total-scale scalability coefficient The information in the J coefficients H_j is summarized by the total-scale scalability coefficient H. Coefficient H is a weighted mean of the item coefficients H_j (Sijtsma and Molenaar 2002) such that $\min(H_j) \leq H \leq \max(H_j)$, and expresses the average slope of the ISRFs. As a result of this property, H expresses the accuracy with which persons can be ordered on a single latent trait θ by means of the total score X_+ (Mokken et al. 1986). Higher values imply a more accurate ordering. Mokken (1971, p. 185) proposed rules of thumb for the interpretation of coefficient H: Under the MHM, $0 \leq H \leq 1$, but positive values close to 0 are insufficient for an accurate person ordering. Thus, the rules of thumb are: $H<0.3$, items are unscalable (this includes $H<0$); $0.3 \leq H<0.4$, weak scale; $0.4 \leq H<0.5$, medium scale; and $0.5 \leq H \leq 1$, strong scale. These rules have become generally accepted (Sijtsma and Molenaar 2002).

Definition of a scale Items form a *scale* (Mokken 1971) if, for inter-item product-moment (pm) correlations (ρ_{jk}, for items j and k) and for positive lower bound value c of H_j, the following two conditions have been satisfied: (1) $\rho_{jk}>0$, for all items pairs j, k; and (2) $H_j \geq c$, for all J items. The MHM implies the first condition. The second condition ascertains a minimum accuracy for ordering persons. For example, $c=0.3$ implies resulting item clusters are at least a weak scales.

3.2 Confirmatory Mokken scale analysis

When the dimensionality and the scalability of a questionnaire have been ascertained in one or more populations and when the questionnaire has been applied successfully in those populations, the use of this questionnaire in novel populations calls for a re-assessment of these properties. For example, a questionnaire that has been successful in measuring service quality a sample of the elderly (say, $H=0.55$, and other favorable psychometric properties hold as well) may or may not be successful in the population of young adults. This can be investigated by collecting a sample of data from the latter population, and then computing the H_j and H values, and estimating the IRFs/ISRFs to check for monotonicity. Thus, the scale is treated as an a priori scale and assessed as it stands without the purpose of deleting or replacing items.

3.3 Exploratory Mokken scale analysis

Item selection The automated item selection procedure uses coefficients H_j and H for dimensionality assessment (and by implication checking for local independence), by allocating items to unidimensional subsets each satisfying the definition of a scale. The selection procedure has been implemented in the MSP computer program (Molenaar and Sijtsma 2000). MSP starts by selecting from the item pool the item pair that has the largest positive significant H. Only when this value is at least equal to lower bound c, are the conditions (1) and (2) of the definition of a scale fulfilled, i.e., (1) $\rho_{jk}>0$, and (2) $H_j\geq c$. If none of the item pairs in the pool fulfill both conditions, the set of items is unscalable and the algorithm terminates without forming scales.

Alternatively, the item pair that has the highest significant $H(\geq c)$ is denoted as the starting pair, (j, k_1). Given this starting pair, the algorithm selects from the remaining $J-2$ items the item k_2, which (a) correlates positively with each item in pair (j, k_1)—condition (1); (b) has an H_{k_2} value with respect to the selected items j and k_1 greater than c—condition (2); and (c) maximizes the total-scale H coefficient of items j, k_1, and k_2. This second step is repeated for the selection of a fourth item, say, k_3, from the remaining $J-3$ items, etc. The iterative process terminates at the point that none of the remaining items satisfy both conditions (1) and (2). The selected items constitute the first scale. If the first scale does not contain all J items in the pool, the same selection algorithm searches for a second scale from the unselected items and, if possible, a third scale, and so on. Items excluded from all clusters are non-scalable. This procedure provides insight into item-pool dimensionality, as different item clusters or scales may represent different latent traits (Hemker et al. 1995).

Strategy for determining dimensionality The choice of lower bound value c influences the outcome of item selection, just as the outcome of a PCA is influenced by the choice of a lower bound for the eigenvalues (how many components are rotated?), the rotation method (orthogonal or oblique?), and the choice of a lower bound, which we denote a, for the factor loadings ($a>0.3$, $a>0.4$ or $a>0.5$?). Such arbitrariness in assessment is omnipresent in statistical analysis and calls for a sensible decision strategy. Based on extensive simulations, Hemker et al. (1995) recommend as a strategy running the item selection procedure for several c values, starting at $c=0$ and then increasing c with steps of, for example, 0.05 until, for example, $c=0.6$. Notice that $c=0$ is the smallest value permitted under the MHM, and although it does not lead to *useful scales* (this requires $c\geq 0.3$) it does contribute to finding the most likely *dimensionality* of the item set. Also, notice that $c=0.6$ is smaller than the maximum of c (i.e., $c=1$). Hemker et al. (1995) found that higher values were not informative anymore about dimensionality. Thus, $0\leq c\leq 0.6$ is an effective range for dimensionality analysis.

Hemker et al. (1995) distinguished two particularly relevant patterns of outcomes. First, when analyzing *unidimensional* data, for different values of c the typical outcome pattern was that for lower c values all items were allocated to the same cluster, and for higher values items dropped out of the cluster one by one. Second, for *multidimensional* data, $c=0$ or values close to 0 forced items in different, often large clusters. In both cases, the higher c values provided information which items form the core of the item clusters, and which items have a marginal position. Finally, in each identified item cluster the IRFs/ISRFs are estimated to check for monotonicity, using procedures described in Sijtsma and Molenaar (2002).

3.4 Comparing MSA with PAF and PCA

Three differences between MSA on the one hand and PAF and PCA on the other hand are relevant here. These differences imply that the different methods provide different interpretations of the data patterns. This report concentrates on differences between MSA and PCA, but the line of reasoning also applies to PAF.

First, MSA and PCA construct "item clusters" in different ways. MSA selects items one by one (i.e., sequentially), whereas PCA constructs a weighted linear combination of all J item scores (i.e., simultaneously), called a principal component. PCA explains the maximum variance possible from the item-score residuals after removing the influence of the previous components. Based on, for example, the eigenvalue-greater-than-1 criterion the retained principal components are rotated to a geometrical position in which they are more interpretable. After rotation the components are called factors. The researcher uses the loadings of the J items for interpreting factors, concentrating on items having their highest loading on a factor. Loadings must exceed the user-defined lower bound a.

Second, MSA and PCA both involve arbitrary but different choices that affect the outcomes of the methods and that necessitate multiple analyses of the same data to reach a stable conclusion. Both methods analyze the associations between the items – PCA often analyzes the pm-correlations, and MSA the H coefficients – and in both methods the vulnerability of the outcomes to choices made by the researcher is caused by the magnitude of the inter-item pm-correlations and H coefficients and the magnitude of the contrasts between these association measures (Hemker et al. 1995). Henceforth, we call inter-item pm-correlations or H coefficients associations. Consider an item pool in which all item pairs have associations of small magnitude. Also these associations show little contrast, i.e., the associations are of a similar magnitude. Under such conditions the item scalability coefficient H_j could be of a similar, small magnitude for each of the J items in the pool, resulting in, for example, H_j<0.3 (for all items). Such a result would suggest unidimensionality (while explaining little variance). This may not be picked up by c values higher than 0.3. Under these conditions MSA would not select one large item cluster, as the items do not fulfill the two conditions for scalability that were defined in Section 3.3. Similar problems occur in PCA when using high values on a. In this case the loadings of rotated factors are likely to be smaller than a. Likewise, matrices showing much contrast may lead to inappropriate interpretations of item pool dimensionality when using a low value on c. Consider an association matrix in which all associations are of a reasonable magnitude. However, the three associations among items 1, 2 and 3 and the three associations among items 4, 5 and 6 are higher than the associations between items from the two sets. Such a contrast is a clear sign of multidimensionality, in which items 1, 2 and 3 load on one dimension and items 4, 5 and 6 on another dimension. Nevertheless, it may be found that all H_j≥0.3, and if c=0.3 the item pool will be interpreted as unidimensional. Similarly, a PCA might very well suggest unidimensionality due to a large first eigenvalue. What these examples serve to show is that statistical analysis of complex data structures requires different analysis rounds trying different choices for c (MSA), and the number of components to be rotated, the particular rotation method, and the choice of *a* (PCA). Even better is to use both MSA and PCA (and other methods such as PAF), which we advocate.

Third, the MHM (which is the basis of MSA) is a *measurement model*, whereas PCA is a pure and highly efficient *data-summary method*. The MHM implies an ordinal person scale (Eq. 3) but this is much different in PCA and other factor models, in which a factor does not automatically have particular measurement properties. Such properties—here an ordinal person scale—only follow from the assumptions of unidimensionality, local independence, and monotonicity (Grayson 1988). These are typical assumptions of IRT, not of PCA or related factor models. Alternatively, PCA is identical to the eigenvalue decomposition of the inter-item covariance/correlation matrix, thus realizing an efficient data summary in as few dimensions as possible, assuming a linear relationship between observable item scores and the component or factor.

4 Empirical application

The SERVQUAL instrument assumes the following dimensions of perceived service quality: (1) Tangibles, (2) Reliability, (3) Responsiveness, (4) Assurance, and (5) Empathy. However, debate on the measurement of perceived service quality continues, particularly on dimensionality (e.g., Brady and Cronin 2001; Finn and Kayande 2004).

Application SERVQUAL was applied for measuring perceived service quality of two facilities at a large university in The Netherlands: the university restaurant and the student helpdesk. Following Cronin and Taylor (1992) and Brady et al. (2002) we employed performance-only measures, based on seven-point Likert scales. The report uses codes for each of the 22 SERVQUAL items: ‘t1’ refers to the first item for the tangibles dimension and ‘t2’ to the second tangibles item, etc. Furthermore, ‘rl’ refers to the reliability dimension, ‘rs’ to responsiveness, ‘a’ to assurance, and ‘e’ to empathy. A total of 223 students of the university’s faculty for business and economics were interviewed. Respondents only evaluated services, which they had personally used. This provided 209 usable response records for the university restaurant and 135 for the student helpdesk. Because the research concerns service quality of university facilities, students represent the population.

Analysis strategy We applied MSA, PAF, and PCA to investigate dimensionality of the SERVQUAL instrument. Because for this instrument a five-dimensional structure with fixed items is known, we first used confirmatory MSA: (1) For each of the five a priori scales the item H_j coefficients and the total-scale H were computed, and (2) for each item the ISRFs were estimated to assess monotonicity using a method contained in MSP and described in Sijtsma and Molenaar (2002). Then, for the purpose of illustration, we also used exploratory MSA: (1) The automated item selection procedure was applied to the complete item set using increasingly higher c values, and (2) in each identified scale, for each item the ISRFs were estimated to assess monotonicity. Because monotonicity assessment is of secondary interest to this study, we only mention main results. Furthermore, for the complete 22-item set PAF was used with oblique rotation (OBLIMIN), and PCA with orthogonal rotation (VARIMAX). Different rotation techniques provide more differing perspectives on the data structure. The PAF and PCA analyses retained factors and principal components with eigenvalues greater than 1 and also applied the scree test criterion.

5 Results

5.1 University restaurant results

Confirmatory analysis In the university restaurant application, for the four tangibles items H=0.35, and except for H_{t3}=0.23, the other item coefficients ranged from 0.37 to 0.42. For reliability, H=0.33, and except for H_{r15}=0.23, the other item coefficients ranged from 0.31 to 0.37. For responsiveness, H=0.37, and except for H_{rs1}=0.20, the other item coefficients ranged from 0.35 to 0.46. For the four assurance items, H=0.40, and except for H_{a4}=0.29, the other item coefficients ranged from 0.42 to 0.44. For the empathy items, H=0.30, and except for H_{e4}=0.29 and H_{e5}=0.14, the other item coefficients ranged from 0.31 to 0.41. Thus, the confirmatory MSA analysis showed some consistency but also differences with the theoretical SERVQUAL dimensionality. Except for some minor violations, monotonicity was supported for each of the five a priori scales (no details given).

Exploratory analysis Table 1 presents results for the four tangibles items (t1-t4) in one column, for the reliability items (rl1-rl5) in the next column, etc. The outcome of the scree test criterion for PAF and PCA suggested a one-factor solution in which all 22 SERVQUAL items have loadings of at least 0.3. This result supports unidimensionality.

Contrarily, the eigenvalue-greater-than-one criterion suggested a seven-factor solution. For factor loadings of at least 0.5 (i.e., $a \geq 0.5$), various items of the same SERVQUAL dimension loaded on the same PAF factor. The reliability items loaded on different factors. These results suggested mild consistency with SERVQUAL theory but other cut-offs for factor loadings suggested stronger inconsistency. For example, for $a \geq 0.3$ the items for tangibles, reliability, responsiveness and assurance blended together on the first factor. The seven-factor PCA solution resulting from the eigenvalue-greater-than-1 criterion also agreed mildly with SERVQUAL but again the reliability items were the main exception. Allowing loadings below 0.5 suggested that items of different theoretical SERVQUAL dimensions loaded on the same factor and other items loaded on multiple factors.

MSA provided a distinct pattern of clusters for varying values of the lower bound c. First, c=0 and c=0.1 yielded two clusters: (1) t1, t4, rl1, rl2, rl4, rs1, rs2, rs3, rs4, a1, a2, a3, e1, e2, e3, e4 and e5; and (2) t2, t3, rl3, rl5 and a4, whereas c=0.2 yielded four clusters, c=0.3 five clusters, etc. The clusters became smaller, suggesting a predominately two-dimensional construct (Hemker et al. 1995). Because the outcomes for c=0 and c=0.1 suggested that 17 of the 22 items loaded on the same

Table 1 SERVQUAL scales for the university restaurant

Criterion		Tangibles	Reliability	Responsiveness	Assurance	Empathy
MSP: c=0.0	Scale 1	t1, t4	rl1, rl2, rl4	rs1–rs4	a1–a3	e1–e5
	Scale 2	t2, t3	rl3, rl5		a4	
MSP: c=0.1	Scale 1	t1, t4	rl1, rl2, rl4	rs1–rs4	a1–a3	e1–e5
	Scale 2	t2, t3	rl3, rl5		a4	
MSP: c=0.2	Scale 1		rl1, rl2	rs2–rs4	a1–a3	e1–e4
	Scale 2	t1, t2, t4	rl5		a4	
	Scale 3		rl3, rl4			
	Scale 4			rs1		e5
MSP: c=0.3	Scale 1					e1–e4
	Scale 2		rl2	rs2–rs4	a3	
	Scale 3	t1, t2, t4				
	Scale 4				a1, a2	
	Scale 5		rl1, rl3-rl5			
MSP: c=0.4	Scale 1					e1–e3
	Scale 2			rs2–rs4		
	Scale 3				a1–a3	
	Scale 4	t1, t2, t4				
	Scale 5		rl3, rl4			
	Scale 6		rl1, rl2			
MSP: c=0.5	Scale 1					e1, e2
	Scale 2			rs2, rs3		
	Scale 3				a2, a3	
MSP: c=0.6	Scale 1					e1, e2
PAF[a]	Fact. 1	*t3, t4*	*rl1, rl2*	*rs3, rs4*	**a1–a3**	
	Fact. 2		*rl2*	*rs4*		**e1–e3**, *e4*
	Fact. 3			**rs2–rs4**	*a3*	
	Fact. 4		*rl1*,rl3,**rl4**,rl5		*a3*	
	Fact. 5	**t1,t2**,*t3*,**t4**	*rl2*		*a4*	
	Fact. 6		**rl2,rl3**			
	Fact. 7		*rl1,rl2*	rs1		e5
PCA[a]	Fact. 1					**e1–e3**, *e4*
	Fact. 2	t3	*rl1*		**a1–a3**,*a4*	
	Fact. 3			**rs2–rs4**		
	Fact. 4	**t1,t2**,*t3*,**t4**				
	Fact. 5		rl3, **rl4**, rl5	rs1	a4	
	Fact. 6		**rl2,rl3**			
	Fact. 7		rl1	rs1	−a4	**e5**

[a] Bold print implies a factor loading≥0.50; normal print implies: 0.40≤factor loading<0.50; italics implies: 0.30≤factor loading<0.40. Variance explained by the seven factor PAF and PCA solution=62%.

dimension, one could also argue in favor of a unidimensional solution. However, results for c=0.2 and c=0.3 contradict unidimensionality. For c=0.4 outcomes suggest consistency with the theoretical SERVQUAL dimensionality; that is: (1) t1, t2, t4; (2) rs2−rs4, (3) a1−a3, (4) e1−e3. A confirmatory MSA on the 17 items suggested to be part of the same dimension shows that the general dimension has insufficient scalability properties (i.e., H=0.19 and all H_j<0.30) and only seems to be relevant as a higher order dimension in addition to the specific SERVQUAL dimensions. These findings are consistent with the third-order factor model for perceived service quality (Brady and Cronin 2001), with a higher-order general factor besides more specific first-order factors.

5.2 Student helpdesk results

Confirmatory analysis For the student helpdesk application, a high level of consistency between the theoretical SERVQUAL dimensionality and the empirical data patterns for tangibles, reliability and assurance was found (i.e., for each scale $H>0.5$ and all $H_j>0.5$; three strong scales, no violations of monotonicity). For responsiveness, we found $H=0.42$, and for the items we found $H_{\mathrm{rs1}}=0.29$ (for rs1 one significant violation of monotonicity was found) and all other H_j are ranging from 0.43 to 0.49. For empathy $H=0.47$, and for the items $H_{e5}=0.29$ and all other H_j's ranging from 0.47 and 0.59 (no violations of monotonicity). Thus, except for minor details, responsiveness and empathy constitute medium scales.

Exploratory analysis The scree-test suggested a one-factor solution for PAF and PCA for 21 SERVQUAL items ($a\geq0.3$) except rs1. The eigenvalue-greater-than-1 criterion suggested five factors. The five-factor PAF solution (Table 2) showed some consistency with SERVQUAL theory. Minimum loadings of $a\geq0.5$ lead to the following patterns: (1) t1, t4, rl1–rl5; (2) rs2, e1–e4; (3) a1–a4; (4) t1–t4, rl2, and (5)

Table 2 SERVQUAL scales for the student's helpdesk

Criterion		Tangibles	Reliability	Responsiveness	Assurance	Empathy
MSP: $c=0.0$	Scale 1	t1–t4	rl1–rl5	rs2–rs4	a1–a4	e1–e4
	Scale 2			rs1		e5
MSP: $c=0.1$	Scale 1	t1–t4	rl1–rl5	rs2–rs4	a1–a4	e1–e4
	Scale 2			rs1		e5
MSP: $c=0.2$	Scale 1	t1–t4	rl1–rl5	rs2–rs4	a1–a4	e1–e4
	Scale 2			rs1		e5
MSP: $c=0.3$	Scale 1	t1–t4	rl1–rl5	rs2–rs3	a1–a4	e1–e4
MSP: $c=0.4$	Scale 1	t1, t2, t4	rl1–rl5		a1–a4	
	Scale 2			rs2, rs3		e1–e4
MSP: $c=0.5$	Scale 1				a1–a4	
	Scale 2					e1–e4
	Scale 3		rl1–rl5			
	Scale 4	t1–t4				
	Scale 5			rs2–rs4		
MSP: $c=0.6$	Scale 1				a1–a3	
	Scale 2					e1, e2, e4
	Scale 3		rl1–rl4			
	Scale 4	t1, t2, t4				
PAF[a]	Fact. 1	**t1**,t2,**t4**	**rl1–rl5**	*rs4*	a1–a4	*e1*, *e2*
	Fact. 2		*rl1*, rl2, *rl5*	**rs2**, rs3	*a1*, a3	**e1–e4**, *e5*
	Fact. 3		*rl2*, *rl5*	*rs3*	**a1–a4**	*e2*
	Fact. 4	**t1–t4**	rl1, **rl2**, rl3–rl5		*a1–a4*	
	Fact. 5		*rl1*, rl2, *rl4*	rs1, **rs2–rs4**	a3	e1, e2, *e4*, *e5*
PCA[a]	Fact. 1	*t1*, *t4*	**rl1–rl5**			
	Fact. 2			*rs2*		**e1–e4**,e5
	Fact. 3		*rl5*	*rs3*	**a1–a4**	−e5
	Fact. 4	**t1-t4**	*rl2*	*−rs1*		
	Fact. 5			**rs1–rs4**		

[a] Bold print implies a factor loading≥0.50; normal print implies: 0.40≤factor loading<0.50; italics implies: 0.30≤factor loading<0.40. Variance explained by the five factor PAF and PCA solution=67%.

rs2–rs4. For smaller loadings, however, most items loaded on multiple factors, and items from different SERVQUAL dimensions loaded on the same factor. The five-factor PCA solution is more consistent with SERVQUAL theory. For $a \geq 0.5$, the following dimensions were found: (1) rl1–rl5; (2) e1–e4; (3) a1–a4; (4) t1–t4 and, (5) rs1–rs4. Item e5 failed to load on the same factor as the other empathy items. For loadings smaller than 0.5 only few inconsistencies with SERVQUAL were found.

MSA again provided additional insight. Table 2 shows that $c=0$, $c=0.1$ and $c=0.2$ resulted in two clusters: (1) t1–t4, rl1–rl5, rs2–rs4, a1–a4, e1–e4; and (2) rs1 and e5. For $c=0.3$ the first cluster was still found. Higher c values produced smaller item clusters. For $c=0.5$ results were consistent with SERVQUAL theory. Again the MSA results support the relevance of a general service-quality factor besides several more-specific factors reflecting the theoretical five-factor definition of service quality. However, for the student helpdesk data the general factor seems to be stronger, as items are distributed over different scales no sooner than $c>0.3$ instead of $c>0.1$. The general signal for the student helpdesk data is strong enough for considering SERVQUAL as unidimensional in this application. Applying confirmatory MSA to 20 of the 22 SERVQUAL items (rs1 and e5 were excluded) leads to a weak scale based on Mokken's rules of thumb ($H=0.37$; all $H_j>0.27$ of which 18 H_j's are at least 0.30). Except for one violation for one ISRF, monotonicity was supported for these 20 items.

6 Discussion

6.1 Conclusions

This paper applied confirmatory and exploratory MSA next to PAF and PCA to assess dimensionality of the SERVQUAL instrument. Based on the scree test criterion, PCA and PAF results suggest that perceived service quality is a unidimensional construct. Contrarily, the eigenvalue-greater-than-1 criterion produces a multiple-factor solution, which is more consistent with the five-dimensional SERVQUAL instrument. In addition, MSA suggests a distinct pattern of clusters for varying values of the lower bound c suggesting one general perceived service-quality factor next to the five SERVQUAL dimensions. The general factor is not the same in the two applications reported in this paper. This precludes an active response tendency being responsible for this general factor (a result that would be unlikely given previous studies). More likely, our results are consistent with interpreting perceived service quality as a third-order factor model (Brady and Cronin 2001), with a general perceived service-quality dimension and dimensions that are related to specific aspects of perceived service quality. The general factor seems stronger for the student helpdesk than for the university restaurant. Relative importance of the general factor may explain different results of previous SERVQUAL applications. A highly dominant general factor may lead to a unidimensional interpretation, as for the student helpdesk; otherwise alternative solutions are more feasible, as for the university restaurant. This is an issue for further research.

In the analysis reported in the current paper, MSA, PAF and PCA do not lead to consistent results. This suggests that the theory underlying the SERVQUAL instrument requires further development for these applications. Less appropriate theories and instruments tend to result in weaker data patterns in which different analytical procedures may cause method bias. Because different analytical procedures also caused inconsistent results in simulation studies based on larger samples (Van Abswoude et al. 2004), the small sample sizes used here could be ruled out as explanation of our results.

The paper has broader implications. Theoretically, the role of the general factor in perceived service quality and for other marketing constructs requires further investigation. This report demonstrated that MSA is useful for detecting the general factor in an exploratory context. Therefore, we suggest MSA should be a component of the exploratory toolkit for assessing marketing measurement scales. The required time-investment would be small, as a highly user-friendly program implementing MSA is available (Molenaar and Sijtsma 2000). Another implication concerns a suggestion for future research on the use of other combinations of exploratory scaling procedures, such as nonlinear PCA and non-metric bilinear multidimensional scaling.

6.2 Guidelines for future applications

When MSA, PAF, and PCA or other exploratory scaling procedures are used in combinations for future studies, consistent results obviously lead to the most straightforward conclusions. Contrarily, inconsistent results may pose the challenge to combine them into a useful conclusion. Sometimes a leading theory is available, such as SERVQUAL in this study. However, researchers are often guided by weaker theory and have to rely even more on empirical outcomes of scale analysis. In this situation, we also suggest that exploratory MSA is applied first with varying lower bounds c, starting with $c=0$ and then using higher values until $c=0.6$ in order to assess whether the data are unidimensional or multidimensional and which items cluster on the same dimension (Hemker et al. 1995; Van Abswoude et al. 2004). For each dimension, monotonicity of the ISRF's (Eq. 2) is necessary to have an ordinal scale (Eq. 3). The next question is whether unidimensional, monotone item clusters are useful in practice for a sufficiently accurate ordering of respondents on a scale. For this assessment, Mokken's (1971, p. 185) rules of thumb prescribe that item sets for which $H<0.3$ are unscalable and increasingly higher values of H indicate stronger degrees of scalability. We suggest using the results of MSA next to the results of PCA and PAF for allocating items to the dimensions. When this allocation is consistent across scaling procedures, a high level of certainty about the found scales may be assumed. In other situations, the researcher should consider whether items should be deleted from scales or whether (s)he should conduct a follow-up study with newly formulated items.

The most challenging situation, when investigating a novel scale, would be similar to the results of the university restaurant application of SERVQUAL reported in this paper. If most items are found in the same cluster for low values of c, and then are split over multiple scales for higher values of c, this can be considered to be a weak signal reflecting unidimensionality but a scalable set of items (i.e., $H\geq0.3$) has not been obtained. Better items can possibly be formulated for measuring the intended construct on this single dimension. This would require additional empirical research. Despite these general suggestions, the development of scales will continue to depend strongly on the researcher's relevant theoretical knowledge and his/her expertise in analyzing complex data sets with respect to scale construction.

References

Brady, M. K., & Cronin, J. J. (2001). Some new thoughts on conceptualizing perceived service quality: A hierarchical approach. *Journal of Marketing*, *65*, 34–49.

Brady, M. K., Cronin, J. J., & Brand, R. R. (2002). Performance only measurement of service quality: A replication and extension. *Journal of Business Research*, *55*(2), 17–31.

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

Finn, A., & Kayande, U. (2004). Scale modification: Alternative approaches and their consequences. *Journal of Retailing*, *80*, 37–52.

Fox, J. (1997). *Applied regression analysis, linear models, and related methods*. Thousand Oaks, CA: Sage.

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

Grayson, D. A. (1988). Two-group classification in latent trait theory: Scores with monotone likelihood ratio. *Psychometrika*, *53*(3), 383–392.

Hemker, B. T., Sijtsma, K., & Molenaar, I. W. (1995). Selection of unidimensional scales from a multidimensional item bank in the polytomous Mokken IRT model. *Applied Psychological Measurement*, *19*(4), 337–352.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague: Mouton.

Mokken, R. J., Lewis, C., & Sijtsma, K. (1986). Rejoinder to "The Mokken scale: A critical discussion". *Applied Psychological Measurement*, *10*(3), 279–285.

Molenaar, I. W., & Sijtsma, K. (2000). *MSP5 for Windows*. Groningen: IEC ProGAMMA.

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

Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1994). Alternative scales for measuring service quality: A comparative assessment based on psychometric and diagnostic criteria. *Journal of Retailing*, *70*, 201–230.

Rossiter, J. R. (2002). The C-OAR-SE procedure for scale development in marketing. *International Journal of Research in Marketing*, *19*(4), 305–335.

Sijtsma, K., & Meijer, R. R. (2007). Nonparametric item response theory and related topics. In C. R. Rao, & S. Sinharay (Eds.) *Handbook of statistics, vol. 26: Psychometrics* (pp. 719–746). Amsterdam: Elsevier.

Sijtsma, K., & Molenaar, I. W. (2002). *An introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.

Singh, J. (2004). Tackling measurement problems with item response theory: Principles, characteristics, and assessment, with an illustrative example. *Journal of Business Research*, *57*(2), 184–208.

Van Abswoude, A. A. H., Van der Ark, L. A., & Sijtsma, K. (2004). A comparative study of test data dimensionality assessment procedures under nonparametric IRT models. *Applied Psychological Measurement*, *28*(1), 3–24.

Van der Ark, L. A. (2005). Stochastic ordering of the latent trait by the sum score under various polytomous IRT models. *Psychometrika*, *70*(2), 283–304.

Voss, K. E., Stem Jr., D. E., & Fotopoulos, S. (2000). A comment on the relationship between coefficient Alpha and scale characteristics. *Marketing Letters*, *11*(2), 177–191.