One, few or many?
An integrated framework for identifying the items in measurement scales

Naresh K. Malhotra
Georgia Institute of Technology
Soumya Mukhopadhyay and Xiaoyan Liu
Nanyang Technological University
Satyabhusan Dash
IIM Lucknow

Churchill (1979) proposed a detailed procedure for the development of better multi-item measures that has become popular. Recently, however, many scholars have challenged this dominant paradigm. They argue that, in many marketing contexts where the target construct has a precise and concrete definition, long multi-item measures can be substituted by shorter measures with fewer items, or even single-item measures. This has resulted in the controversy around the relative superiority of single- versus multi-item scales. We review the extant literature to summarise various arguments in favour of (or against) multi-item and single-item measures, respectively. Moreover, we propose an integrated framework for developing a new scale, reducing long multi-item scales to shorter multi-item measures or to single-item measures, or to expand an existing short (single-item) scale. The significant contributions of this paper to the literature are identified.

Introduction

The importance of scaling and measurement is well known in any field of basic and applied research. The purpose of measurement is to provide empirical estimation for relevant constructs of interest. Therefore, from the perspective of theory development and validation, measurement issues play a crucial role in any scientific endeavour. Research in marketing is no exception to this rule (Kalwani & Silk 1982; Finn 1992; Laurent et al.)
The most influential research that shaped the present measurement practices in marketing was proposed by Churchill (1979), recommending a stepwise approach for multi-item scale development that ensures higher level of internal consistency and better captures all facets of the construct under study. However, this classical approach often requires a fine balance between scale length and quality of measurement, which is difficult to achieve. It was also argued that long multi-item scales often lead to respondent refusal, missing data, and higher cost of data collection and processing. Apart from these practical concerns, in recent times this approach has been questioned on more theoretical grounds (Smith 1999; Rossiter 2002; Bergkvist & Rossiter 2009). These arguments essentially stress that single-item scales are often sufficient for measuring marketing constructs that are ‘singularly concrete’ in nature (Rossiter 2002). In this situation, too much emphasis on the use of alphas as solitary evidence for scale reliability can lead to acceptance of redundant and even conceptually inappropriate items. Moreover, over-reliance on factorial unidimensionality to unravel the underlying structure of target construct can often lead to the deletion of conceptually necessary items. These counter-arguments have not only resulted in the controversy about the relative superiority of single- versus multi-item scales, but have also led to confusion about which types of scale should be used under what circumstances, and how to decide on the optimal number of items to be retained in the final scale. Therefore, there is a great need for a refined framework that provides guidance on the design and choice of single- versus multiple-item scales, and the long form of multi-item scales versus the short forms. Unfortunately, despite its significance, very little guidance exists on how to implement this process. In the current article, we try to reconcile these conflicting arguments.

This paper is divided into two parts. We begin by reviewing the extant literature, to summarise various arguments in favour of (or against) multi-item and single-item measures, respectively. More specifically, we focus on various fundamental issues that are universally applicable to any form of scale development, such as reliability, validity, nature of the construct, and critical empirical issues. In the second section, we propose an integrated framework that is suitable for new scale development (multiple items and/or single item), long multi-item scale reduction, and short (single-item) scale expansion. This framework, for a given marketing context, will assist researchers in deciding on the optimum scale length and
selecting the most appropriate items to improve the overall psychometric quality of the scale.

The major contribution of this paper is twofold. First, it helps in gaining more comprehensive understanding of the recent controversy between single-item and multi-item scales. Second, it also provides a systematic approach for researchers in terms of how to develop measurement scales with good psychometric properties as well as decide on the particular measurement form (single item/short form/long form with multi items). We believe the use of better measures will result in improved quality of the research findings and lead to accumulated knowledge over time.

**Review of the fundamental issues**

As the main purpose of scale development is to measure the true value of the property of an object, the scale should be precise and consistent. In other words, the scale should be reliable and valid. However, within the context of social and behavioural sciences, definitions of measurement constructs are often dependent on the research tradition within which the measurement theorist is working. Consequently there is a need to have a theoretically robust definition of reliability and validity. Classical test theory (CTT), one of the most widely accepted frameworks in the social and behavioural sciences, offers a solution. A measurement is considered to be reliable when random error does not bias the observed score and thus it can be consistent across various situations, whereas validity is achieved only when the observed score reflects the true characteristics of the object.

**Reliability**

In their widely cited papers, Peter (1979) and Churchill (1979) strongly advocated that internal consistency reliability (alpha coefficient) should be the most important measure used to assess the quality of a scale, not only because reliability is a necessary condition for validity, but also because unreliable measures attenuate the correlation between constructs. One of the fundamental objectives of classical test theory is to estimate and improve the reliability of psychometrical scales. Under this guidance, it is assumed that the items on a scale are a random sample from the universe of all possible items drawn from the domain (‘domain sampling approach’; Nunnally 1967; Churchill 1979). Consequently, it can be expected that the items coming from the same domain should be highly correlated with one another, which is labelled as internal consistency and usually calculated
using Cronbach’s (1951) alpha. However, despite strong acceptance in the psychometric and consumer psychology literature, the alpha measure has long been criticised on several grounds.

Although empirical evidence suggests that alpha reliability of the total scale always increases as the number of items increases (Churchill & Peter 1984; Peter & Churchill 1986; Peterson 1994), it is also recognised that the utility of adding more items might diminish quickly (especially when those items are highly inter-correlated; Simms & Watson 2007). In other words, adding more items does not mean adding more content information despite the increase in reliability of the scale (Drolet & Morrison 2001; Rossiter 2002; Bergkvist & Rossiter 2007). In fact, it has been argued that high internal correlation among items might lead to greater information redundancy. Therefore, in many marketing contexts where the target construct has a precise and concrete definition, long multi-item measures can be substituted by shorter measures with fewer items (Oshagbemi 1999; Stanton et al. 2002; Richins 2004; de Jong et al. 2009; Walsh et al. 2009). Moreover, for constructs that are narrowly defined, recent evidence suggests that single-item measures may suffice (Wanous et al. 1997; Drolet & Morrison 2001; Rossiter 2002; Bergkvist & Rossiter 2007, 2009). In fact, Simms and Watson (2007) argue that, for very brief scales, alpha may not be a sensible facet of generalisability at all. They cite examples of the recently developed short-form Big Five Inventory scale (Gosling 2003; Rammstedt & John 2007) that, despite having low internal consistency reliabilities, showed substantial test-retest reliability and nomological validity that predict hypothesised relationship with other constructs. Therefore it is important to ascertain whether the increase in the scale reliability (due to the addition of incremental items) can counter-balance the loss due to respondent fatigue and hence poor data quality. However, the current paradigm of multi-item scale does not provide any avenue for such assessment. Debates over the relevance and usefulness of inter-item consistency as the single most important measure of scale quality eventually led to the realisation that many issues related to quality of measurement are complex and go beyond the classical view of reliability. Of most importance, significance of the meaning and interpretation of measurements is crucial to measurement quality evaluation. Traditionally, issues of score meaning and interpretation are more closely related to the idea of validity and it is suggested that a measure with good construct validity may be qualitatively better, regardless of the size of the coefficient alpha (Peter & Churchill 1986). Accordingly, the idea of construct validity has attracted a lot of attention for a long time (Cronbach & Meehl 1955; Messick 1995) and it
is now widely recognised as one of the central concerns in psychological measurement (Kane 2004).

Validity

The most frequently mentioned types of construct validity are convergent validity (correlation between two different measures of the same construct), discriminant validity (unique and distinct from the measures of other constructs) and nomological validity (prediction of the relationships confirmed by previous theoretical framework). Peter and Churchill (1986) conducted a meta-analysis of marketing measures to determine factors that might affect construct validity and found that: reliability strongly impacts convergent validity, while measure development procedure significantly affects discriminant validity; nomological validity is primarily affected by reliability and convergent validity. Construct validity can be ascertained by employing methods such as the multitrait-multimethod (MTMM) matrix (Campbell & Fiske 1959), confirmatory factor analysis (CFA) and structural equation modelling (SEM) (Anderson & Gerbing 1991; Bagozzi et al. 1991).

The MTMM matrix is a two-dimensional cross-classification in which multiple traits are measured using distinct measurement methods. This technique requires the collection of data on at least two similar but different traits using two very different methods. In traditional MTMM design, convergent validity is indicated by significant and substantial correlations between the different methods for measuring the same construct. Discriminant validity is the extent to which an item does not relate to the measures of other constructs. Thus, discriminant validity is achieved when the correlation between two measures that assess the same construct is greater than the correlation between the construct and another measurement that assess a different construct. Later on, CFA was recommended to decompose the underlying factors in the MTMM matrix (Widaman 1985; Marsh & Hocevar 1988), due to its ability to test alternative models of MTMM matrix, take the latent variable into the model, and separate trait, method and unique variance for each measure (Gardner et al. 1998).

Confirmatory factor analysis enables researchers to examine pre-specified theoretical relationships using SEM; it takes sampling error into account more effectively, and results in more consistent and accurate factor solutions (Conway & Huffcutt 2003). Using CFA to establish measurement reliability and validity, and testing the structure model and research hypotheses...
through SEM, the two-step CFA–SEM approach is recommended as a useful method for scale development (Anderson & Gerbing 1991; Malhotra et al. 2004). More specifically, after conducting CFA, the researcher should assess the model fit first by looking at indices such as comparative fit index (CFI), goodness-of-fit index (GFI) and root mean square error of approximation (RMSEA). With sufficient model fit, the researcher can confirm the convergent validity and discriminant validity using related criteria such as factor loading and variance extracted (Fornell & Larcker 1981; Malhotra 2010). The significance level and magnitude of the relationship between the focal construct of interest and the other constructs from an established theoretical framework will provide evidence on the measurement’s nomological validity (Bagozzi 1980). Researchers can use SEM to test the causal propositions among the constructs.

The procedures discussed above are widely adopted when designing multi-item scales, yet several concerns were initiated over the implementation of validity tests. While proposing the C-OAR-SE procedure for scale development, Rossiter (2002) argued that convergent validity was not a good indicator for measuring instrument quality, as researchers cannot ascertain whether the new or the old scale is valid. He suggested that validity should be established for a scale independently and absolutely, not relatively via its correlation (or lack of correlation) with other measures. In this paper he also contended that predictive validity is relevant only as long as one knows the true construct-to-construct population correlation. Moreover, he strongly suggested content validity as the only validity needed for ‘singularly concrete’ constructs. The argument, that content validity (estimated by proportion of substantive agreement among experts or the substantive validity coefficient) is a necessary and useful criterion for scale development, indeed has merit. However, it cannot be accepted as the sufficient condition for good measurement because the data collection process is always vulnerable to various response biases such as uninformed response, social desirability, acquiescence and extremity bias (Finn & Kayande 2005). Although this kind of error can be limited by careful scale construction, it is too optimistic to say that it can be removed. Moreover, although it is true that the estimation of convergent and discriminant validity is dependent on the validity level of the paired constructs, they still can provide scientific and statistical evidence to support the construct’s validity. Finally, it might not be a good idea to assume there is a true correlation between two constructs, as the true correlation of the measure and the criterion can be attenuated by imperfect reliability or measurement errors (Hunter & Schmidt 1996). In a more recent article, Bergkvist and
Rossiter (2007), in contrast to Rossiter’s (2002) earlier claim, argued against content validity, and employ bivariate correlation analysis and multivariate regression analysis to compare the relative predictive abilities of single-item and multiple-item measures.

Nature of the construct

Another pertinent issue that is intrinsically related to most of the multi-item scales is the nature of the construct. The framework proposed by Churchill (1979) implicitly assumes that selected multiple items are reflective of a latent construct. More specifically, each item ($x_i$) corresponds to a linear function of its underlying construct ($\eta$) plus measurement error, which can be expressed as:

$$x_i = \lambda_i \eta + \epsilon_i$$  \hspace{1cm} (1)

where $x_i$ is the $i$th indicator of the latent variable $\eta$, $\lambda_i$ is a coefficient (loading) of the effect of $\eta$ on $x_i$, and $\epsilon_i$ is the measurement error. Therefore, all the items are supposed to be positively correlated and interchangeable with one another, and removing one of them will not change the exact meaning of the construct (Nunnally 1978; Bollen & Lennox 1991). It is this interchangeability that provides the basis for the ‘domain sampling approach’, as well as the justification for measurement of internal consistency coefficient (alpha).

However, the coefficient alpha is not suitable for another type of construct: formative construct (Curtis & Jackson 1962). In the case of formative constructs, the observed variables are the causes rather than the effects of the latent construct (Bagozzi 1994). Specification of the formative measurement model can be expressed as:

$$\eta = \sum_{i=1}^{n} \gamma_i x_i + \xi$$  \hspace{1cm} (2)

where $\gamma_i$ is a coefficient capturing the effect of indicator $x_i$ on the latent variable $\eta$, and $\xi$ is a disturbance term. Therefore, whereas reflective indicators are highly correlated and essentially interchangeable, each formative indicator contributes unique content to the latent construct, and omitting any indicator will omit a specific part of the construct (Bollen & Lennox 1991). Consequently, internal consistency is no longer relevant to the quality of scale because variables might be negatively related or
not related to one another (Nunnally & Bernstein 1994; Diamantopoulos & Winklhofer 2001). Furthermore, as formative measurement scales are based on multiple regression models, strong inter-item correlation might lead to multicollinearity issues. Specifically, if a particular indicator turns out to be a perfect linear combination of the other indicators, it might contain redundant information and should therefore be excluded from the scale (Bollen & Lennox 1991). Validity assessment is another controversial issue for constructs with formative indicators, because traditional CFA under reflective measures is not suitable for constructs that are not unidimensional. Some researchers even argue that no quantitative quality checks are applicable for assessing the appropriateness of formative indices (Diamantopoulos et al. 2008).

Proper identification of the construct’s measurement model as reflective or formative is important, not only because it determines the causal direction between indicators and latent construct, but also because it affects the methods of reliability and validity assessment (Diamantopoulos & Winklhofer 2001). A review of top marketing articles revealed that 29% of the constructs were incorrectly modelled, which severely biased the structural parameters estimation, and resulted in inappropriate relationships between constructs (Jarvis et al. 2003). Moreover, recent research revealed that measurement model mis-specification can inflate unstandardised structural parameter estimation by as much as 400%, or deflate it by as much as 80%, depending on whether the endogenous or the exogenous construct is mis-specified (MacKenzie et al. 2005).

**Empirical issues**

There is no doubt that reliability, validity and nature of construct play influential roles in measurement design and item selection when we are dealing with a single construct. And it has been well established that multiple items are needed for a complex and/or multi-dimensional construct (Churchill 1979; Peter 1979; Scarpello & Campbell 1983; Peterson 1994; Loo & Kelts 1998; de Jong et al. 2009), whereas a single item is sufficient for a construct that is concrete and/or unidimensional in nature (Wanous et al. 1997; Loo 2002; Poon et al. 2002; Rossiter 2002; Bergkvist & Rossiter 2007). However, situations become more complex when we conduct empirical study that puts many constructs together to measure their relationship to one another. In such cases, the criteria mentioned above are not sufficient for a best choice. Otherwise, multi-item measures, having a higher level of reliability and being able to tap in to
all the facets of the construct (Gorsuch & McFarland 1972; Churchill 1979; Peter 1979; Bruner & Hensel 1993; Peterson 1994), will always be a better choice than single-item measures. Therefore, the decision on measurement format, in reality, should be based on the integration of various requirements and practical constraints that go beyond the reliability, validity and construct nature considerations.

The nature of the study is one of the factors that should be established before the researcher chooses the scale format to satisfy the objectives of the current study. For different studies, the most widely accepted level of reliability is 0.5–0.6 for preliminary research, 0.7–0.8 for basic research and 0.9–0.95 for applied research (Nunnally 1967; Kaplan & Saccuzzo 1982; Murphy & Davidshofer 1988). In other words, single-item measures with modest to low reliability might be fairly good choices for preliminary research, while multi-item measures are necessary for applied research due to the requirement for higher reliability and managerial implications and guidance (Oshagbemi 1999; Loo 2002; Dolbier et al. 2005).

Practical constraints in time, monetary costs, respondent fatigue and survey refusal might be other factors that qualify researchers’ measurement choice. Churchill (1979) claimed that the payoff of multi-item measure development was substantial, despite the time and costs involved. However, many other research studies (Gorsuch & McFarland 1972; Wanous et al. 1997; Rossiter 2002; Dolbier et al. 2005; Bergkvist & Rossiter 2007) recommend single-item measures when time is highly constrained, or when costs of respondent recruitment and data processing are noticeable. In more recent studies, a group of scholars (Gardner et al. 1998; Drolet & Morrison 2001; Stanton et al. 2002; Malhotra et al. 2004) stress that a tailor-made short scale with a modest number of items might be a better choice as it balances the cost constraints and information needed to cover key facets of the construct.

The objective of the measurement is another important determinant. The simplicity of single-item measurement makes it more suitable for a comparative study across samples or situations (Wanous et al. 1997; Oshagbemi 1999; Dolbier et al. 2005). In contrast, multi-item scales are recommended for studies to determine the impact of all the facets of the construct (Gorsuch & McFarland 1972; Dolbier et al. 2005). In other words, if one merely focuses on changes in the focal constructs, single-item measures might be acceptable for simple comparison study, while multi-item measures can explain which dimensions of the construct impact the dependent variable.
Before sending out surveys, researchers still have to think about the characteristics of respondents. For empirical investigations that use public opinion polls as a data source, the simple and short single-item measures will help researchers and practitioners gain better survey responses (Oshagbemi 1999; Dolbier et al. 2005; Bergkvist & Rossiter 2007). In contrast, when respondents are students, corporate managers or people who have extensive survey experience and related training, multi-item measures will not cause big problems with survey fatigue or response bias (Churchill & Peter 1984; Peterson 1994). Recently, shorter multi-item scales have been recommended as an alternative format to balance the need for high-quality survey response and sufficient information for theory building and practical implication (Drolet & Morrison 2001; Richins 2004; de Jong et al. 2009).

We organised the current literature with regard to the specific measurement format and situations when it is recommended as a better choice than the others shown in Table 1. From the number of studies that support each of the three formats, it is not easy to conclude which format is to be preferred under different situations. Instead, we propose an integrated framework that will enable researchers to: (1) decide whether a scale should be single item or multi-item; (2) develop the new scale and establish its psychometric properties; (3) reduce the number of items in an existing scale while retaining the desirable psychometric properties; and (4) expand an existing short scale to enhance its psychometric properties.

**An integrated framework for scale development**

**Process overview**

The proposed framework, shown in Figure 1, is essentially an integration of four procedural frameworks for scale development contributed by Churchill (1979), Rossiter (2002), Stanton and colleagues (2002), and Maklan and Klaus (2011). In addition, we incorporate the methodological advances suggested by Anderson and Gerbing (1988), Malhotra et al. (2004), and Malhotra et al. (2006). Each of these frameworks starts with the creation of an item pool and ends with a full-length scale of desirable psychometric properties. The generation of item pool is preceded by construct conceptualisation and model identification, and followed by the sequential process of elimination of the least reliable and non-informative items so that the overall reliability of the resultant subscale does not fall below an acceptable level. Each of these sequential steps involves
simultaneous identification and elimination of those items that violate the essential expected norms to the greatest extent. The desired length of the final scale is in turn determined by the stringency with which different item selection criteria are employed. In applying this strategy, however, the test constructor should ensure that the final abbreviated scale conforms to
Figure 1 An integrated framework for new scale development (multi-item versus single-item), scale reduction and scale expansion.
the generally acceptable psychometric norms and the retained items are adequately representative of the theoretical construct(s) they are supposed to measure. It is important to note that our proposed framework can easily be applied for reduction of an existing full-length scale to a shorter form. In such cases the first three steps (construct conceptualisation, model identification and item generation) can be omitted if the researcher is satisfied with the original underlying arguments. The existing items used in the full-form scale constitute the initial item pool and can be subjected to further sequential item elimination steps. Likewise, this process can be used to expand a short-form scale by generating additional items to supplement the items of the existing scale and following the subsequent steps. In those cases where the researcher has enough theoretical ground to believe that the focal construct is ‘concrete singular’ (Rossiter 2002) in nature then s/he should follow the path leading to the creation of a single-item scale (path B in Figure 1).

An integrated framework

Construct conceptualisation
The conventional multiple-item scale development recommends a stepwise approach that begins with domain specification of the focal construct. This process of construct conceptualisation should ideally incorporate all known theoretical dimensions that are related to the research context. This specification must also be consistent with prior theoretical work and empirical evidence. An example of this approach is evident in the conceptualisation of exploratory consumer buying behaviour (Baumgartner & Steenkamp 1996), where the authors propose a two-dimensional structure based on prior theoretical frameworks (Berlyne 1963; Pearson 1970; Zuckerman 1979). A more recent example of this can be found in the conceptualisation and development of the Internet Users’ Information Privacy Concerns (IUIPC) scale (Malhotra et al. 2004). Drawing on social contract theory (Donaldson & Dunfee 1994), the authors characterise IUIPC in terms of three underlying factors: collection, control, and awareness of privacy practices. Moreover, on the basis of prior empirical evidence (Heide & John 1992; Stewart & Segars 2002), they operationalise the focal construct (IUIPC) as a second-order construct.

Model identification
Another very important precondition for scale development is the proper specification of the underlying measurement model. The measurement
model specification process follows the set of conceptual criteria that is suggested by Jarvis et al. (2003). More specifically, the target construct should be evaluated by a set of experienced coders against the following defining characteristics of a reflective measure: first, the indicators should be the manifestation of the construct; second, changes in the indicators should not cause changes in the construct, while changes in the construct should lead to changes in the indicators; third, the indicators should share a common theme and eliminating an indicator must not affect the conceptual domain of the construct; fourth, the indicators should not co-vary with one another; finally, the indicators must have the same antecedents and consequences. Any instance of disagreement among the coders must be resolved by mutual discussion and further clarification of the construct definition. In the current context, we assume that the measurement model is reflective in nature, and accordingly recommend different methods for reliability and validity benchmarking. This is the case for the vast majority of the constructs encountered in marketing.

**Item generation and initial item pool**
Based on the conceptual domain definition, an initial pool of several items is generated based on literature review, experience surveys, in-depth interviews, focus groups and critical incidents. The items generation should be based on how the construct has been defined and how many dimensions or components it has. These items should reflect the underlying dimensions of the focal construct. Care should be taken to ensure that wordings of items are as precise as possible. Moreover, the researcher should not use several items that are just slightly differently worded variations of one another. Such redundant items do not offer any construct-relevant information and generally contribute to the ‘attenuation paradox’ by increasing coefficient alpha while not enhancing validity (Loevinger 1957; Boyle 1991; Clark et al. 1995). For example, in their recent article on Customer Experience Quality (EXQ), Maklan and Klaus (2011) examined the perceptual attributes of experience through in-depth interviews using soft laddering (Grunert & Grunert 1995). These interviews were in turn transcribed, coded and analysed following a grounded approach (Strauss & Corbin 1998) to generate an initial pool of 58 items.

**Expert screening**
Once the initial item pool has been developed it should subsequently be subjected to a panel evaluation to assess the quality and clarity of the
individual items. This panel should consist of subject matter experts who have been clearly informed about the construct definition and its dimensions (Dagger et al. 2007). The ‘quality’ of items is reflected by their relevance, comprehensiveness and the degree to which individual scale items sample the content domain of the construct. The expert screening process should also be designed in a manner so as to identify items that are worded too similarly. An example of this process can be found in Maklan and Klaus’s (2011) article, where the authors subjected their initial pool of 58 items to a screening process where a panel of experts evaluated the items on the basis of their similarity, clarity and terminology. The panel also assessed the relevance of each item with respect to the construct domain, and suggested dimensions and sub-dimensions that evolved from the research model and items. The expert panel, in addition to evaluating each item in respect to its relevance for the construct domain, also suggested additional items, dimensions and sub-dimensions that comprehensively tap the construct’s domain of content. For example, in DeVellis (2003), this process resulted in 37 items representing five dimensions. Apart from ensuring the quality and relevance of the item pools, an ‘expert screening step’ could also be used to evaluate the degree of agreement among the experts regarding the true nature of the focal construct. Following the steps of the C-OAR-SE procedure for scale development (Rossiter 2002), the experts could be asked to classify the target object as either concrete singular or abstract collective object. The researcher can, thus, evaluate the possibility of developing a single-item scale from the item pool. Once a pool of revised items has been generated, these items can then be subjected to target group pretesting (Anderson & Gerbing 1991) to assess the content validity of individual items.

**Content pretesting**

In this stage, the revised pool of items are pretested with a smaller group of respondents to ensure that all items are clear and comprehensible to the target group (Reynolds & Diamantopoulos 1998). Moreover, this stage provides the ideal set-up for pretesting (Anderson & Gerbing 1991) the content validity of individual items prior to conducting confirmatory factor analysis. In this process a group of target consumers are provided with the descriptions of conceptual dimensions of the focal construct and then asked to complete an item-sorting task. Specifically, they read each item carefully and assign it to the dimension that, according to their individual judgement, it best reflects. Based on this sorting task, two validity indices can be computed: the proportion of substantive agreement ($p_{sa}$) and the
substantive validity coefficient \( c_{sw} \). Proportion of substantive index can be defined as the ‘proportion of respondents who assign an item to its intended construct’, while the substantive validity coefficient represents ‘the extent to which respondents assign an item to its posited construct more than to any other construct’ (Anderson & Gerbing 1991). Larger values of these indices indicate greater substantive validity. In order to identify the content-valid items, cut-off values of 0.5 for \( p_{sa} \) and 0.3 for \( c_{sw} \) could be applied. This validated item pool can now be subjected to quantitative scale purification and the abbreviation process. As an example, in their article related to the development of an experiential value scale in the catalogue and internet shopping environment, Mathwick et al. (2001) adopted a content pretesting approach to develop reduced length scales for various dimensions (e.g. economic value (10 to 4 items), excellence (16 to 4 items), etc.).

Scale purification
Once the item pool has been refined through initial content pretesting, the refined and reduced scale can be used for data collection following a sampling method well justified for the research context. The final respondent pool’s composition should be compared with population figures to reveal any major sample bias in terms of demographic characteristics so as to ensure the representativeness of the collected data. Moreover, the answers of early respondents (e.g. the first one-third) can be compared with those of late respondents (e.g. the last one-third) for any significant difference in group means on the constructs of interest. A significant difference generally indicates non-response bias (Armstrong & Overton 1977).

Structure identification: structure identification is driven by exploratory factor analysis (EFA), which has long been recommended by researchers for initial scale development (Cortina 1993; Russell 2002). The primary goal of this step is to identify a set of items that measured a single underlying latent dimension. Once identified, confirmatory factor analysis (CFA) can be used to validate the factor structures identified by EFA. This step is important from the standpoint of assessing factorial validity and unidimensionality of the scale (Allen & Yen 2002). Especially, unidimensionality is a crucial precondition for many of the assessment measures (e.g. IRT parameters ‘a’ and ‘b’) that are proposed in the following sections. This step begins by randomly splitting the data into two equally sized samples: one to be used for scale development and refinement, and one to be used for
evaluation and confirmation. The underlying structure among the items is then investigated with exploratory factor analysis using one of the split half samples. The underlying dimensions can be identified according to the item loading values. Items with very low factor loading in a one-factor solution can be removed, to ensure the unidimensionality of the final scale. Moreover, the researcher should identify and eliminate variables with ‘local dependence’ where the correlation between the two variables seems to be due not only to the underlying latent variable of interest but also a secondary content dimension. ‘Local dependence’ is indicated by large value of residual correlation (difference between the observed and model-reproduced correlation matrix) between two highly correlated variables. For example, Maklan and Klaus (2011) used EFA to identify four primary dimensions of ‘Customer Experience Quality’ from a pool of 19 corresponding items. These underlying dimensions, based on the corresponding item loadings, are labelled as Product Experience, Outcome Focus, Moments-of-truth and Peace-of-mind. Sequential application of EFA and CFA in the structure identification step should also enable the scale developer to confirm the possibility of developing a single-item scale from the existing item pool. Moreover this step can help identify the specific item that can be used in a single-item scale based on their respective factor loading values. Once the relevance of a possible single-item scale is ascertained and a small subset of suitable items has been selected, the scale development process could be bifurcated into two parallel assessment paths (paths A and B in Figure 1) to make a comparative assessment of both multiple- and single-item candidates on various consistency, reliability and validity criteria. Specifically, at a set time, the researcher should select one item and check its performance in terms of some specific reliability and validity criteria.

Distributional characteristics assessment: the scale abbreviation process begins with an assessment of each item with reference to the basic distributional statistics such as mean, variance, skewness and kurtosis. Following established scale-development guidelines (DeVellis 2003; Netemeyer et al. 2003), individual items with desirable qualities (i.e. means close to the centre of theoretical range, the observed item ranges corresponded to the theoretical ranges, item variances were relatively high, and all inter-item correlations were substantial and significant) should be retained, while any item with gross anomalies, such as low variability, extreme response bias or gross deviation from non-normality (Nunnally & Bernstein 1994), should be considered for elimination. The selected
item pool emerging out of the distributional assessment step can now be subjected to a more extensive assessment process.

**Reliability and dimensionality check:** once the items with desirable statistical properties have been identified, the different dimensions in the measurement model should be checked for adequate composite reliability (>0.60) and Cronbach’s alpha values (>0.70) to indicate the required level of reliability (Nunnally 1978; Bagozzi & Yi 1988). In addition to these basic indicators, following the suggestions of Stanton et al. (2002), we recommend that a graded response IRT (Samejima 1969) analysis be carried out for each facet of the original scale to calculate the ‘a’ and ‘b’ parameters for the respective items. IRT values such as the ‘a’ (discrimination) and ‘b’ (threshold) parameters essentially define the nature of the item characteristic curves (ICCs). These curves provide useful information about the relative contribution of an item towards a scale score at all different levels of the scale. IRT discrimination parameters (a) are interpreted as the strength of association between an item and the underlying trait; in many respects these parameters are similar to factor loadings or item–total correlations. Discrimination parameters above 1.70 are considered very high; those between 1.35 and 1.70 are high; and those between 0.65 and 1.34 are moderate. A high value (>1.35) of discrimination parameter implies that most items are contributing a relatively large amount of information to the measurement of focal trait, while a low value (<0.65) indicates a low level of information contribution. On the other hand, threshold parameters can be interpreted as the points on the latent trait continuum at which a respondent has a 50% (or above) probability of responding to an item in a certain response category and a 50% probability of responding in any other lower category (Embretson & Reise 2000). Consequently, the scale developer should examine whether the threshold parameters for the items belonging to one scale are evenly spread out around zero or not. For example, if the thresholds cover only certain areas of the continuum, with the best coverage provided in the below-average (above-average) portion of the trait continuum, then this suggests that the scale discriminates best among respondents who have below-average (above-average) trait levels, while it discriminates less well (or not at all) among respondents who have trait levels above (below) the mean. The researcher should also check if the coverage of the parameters is clustered in certain areas of the latent trait continuum. Generally, such clustering implies that there are narrow areas of the trait continuum at which the scale discriminates very well and other areas at which it does not discriminate well at all. Considering the
parallelism between the corrected item–total correlation (ITC) and the IRT ‘a’ parameter (Nunnally & Bernstein 1994), Stanton et al. (2002) suggested that the corrected item–total correlation (ITC) values be used in conjunction with the IRT parameters. The relative informational contribution of an item can also be assessed using the standard regression diagnostic for multicollinearity: variance inflation factor (VIF). Larger values (>5) of VIF usually indicate that an item is more redundant in the presence of other items in that facet. For example, Van Dam et al. (2010) have recently used IRT to analyse the response patterns and scale properties of the Mindful Attention Awareness Scale (MAAS). Their findings suggest that general statements of ‘automatic inattentiveness’ or ‘automatic pilot’ confer greater statistical information about the underlying latent trait. In addition to these two measures (IRT coefficients and VIF), we recommend another estimate of consistency and unidimensionality of scale items: Hunter’s similarity coefficient (Hunter 1973). This coefficient has often been used in the field of consumer research (e.g. Anderson & Gerbing 1982, 1988; Hunter & Gerbing 1982) and psychology (e.g. Hunter et al. 1982).

Hunter’s similarity coefficient (ϕ) is bounded between –1 and 1. At these extremes two items possess correlational patterns that perfectly satisfy the requirements of internal and external consistency, and hence constitute a unidimensional scale. On the other hand, a value of 0 on ϕ implies that the correlational patterns for two items are dissimilar and that the two items are not internally and externally consistent. Consequently, such items cannot be considered a part of the same construct and should not be placed together in the same scale. The factor structure – i.e. whether it is a first-order or second-order construct, etc. – should be ascertained using the procedure suggested by Rindskopf and Rose (1988), and illustrated by Mathwick et al. (2001). In case of a single-item scale, following the suggestions of Rossiter (2002), we recommend the use of Proportional Reduction in Loss (PRL) formula (Rust & Cooil 1994) as an alternative to alpha coefficient. The reliability assessment of the single item can be carried out following the estimation process recommended by Wanous and his colleagues (Wanous & Reichers 1996; Wanous et al. 1997). Apart from this reliability index, the scale developer should also use the IRT parameter estimates to assess the relative informative contribution of each item.

Convergent validity assessment: once the measures indicate an adequate level of reliability, appropriate distributional characteristics and unidimensionality, a CFA based on the conceptual construct specification should be conducted to examine the cross-loadings and correlated error
terms. CFA is the most widely accepted method for assessing convergent and discriminant validity for measurement models that are composed of latent factors. There are two different ways to apply CFA analysis. First, one can infer the estimation of construct validity based on the indexes of the CFA analysis results (for an example see Malhotra et al. 2004). The structural equation framework (SEM) provides a flexible and powerful way to evaluate the construct conceptualisation by examining the cross-loadings and correlated error terms, as indicated by the modification indices and expected parameter changes in a measurement model conforming to the construct conceptualisation (Malhotra 2010). Convergent validity is indicated when the observable indicators load significantly on their intended factor. In order to ascertain the convergent validity of the items, the observable indicators should indicate substantial (>0.6) and significant ($p < 0.01$) loadings on their intended factors (Chin et al. 1997). In the present context, the item set that consistently leads to similar factor solutions under different factor-retention methods (e.g. the scree test, parallel analysis) should be retained. Moreover, these factor solutions should be more interpretable and able to account for a greater proportion of variance (Conway & Huffcutt 2003). Specifically, the researcher should check whether the paths from the higher-order factor to the lower-order dimensions are positive and significant, and if the average variance extracted (AVE) from different constructs is above the critical value of 0.5 (Fornell & Larcker 1981; see Malhotra et al. 2004 for an illustration).

The measurement model should be assessed using various fit indices to determine how well the conceptualised construct fits the sample data (McDonald & Ho 2002). The most frequently used indices in this category are the chi-square statistic, RMSEA, GFI, AGFI, RMR and the SRMR. Malhotra (2010) provides a detailed account of different fit measures and their usage. In the case of a single-item measure, convergent validity can be assessed by showing high and significant correlation with other related constructs. The MTMM approach can be adopted for greater rigour. The researcher should check and adjust for common method variance bias using the methodology proposed by Malhotra et al. (2006).

**Discriminant validity assessment**: once the measures indicate an adequate level of factor loadings, average variance extracted and model-fit indices, the measurement model should also be assessed for discriminant validity by running a series of nested CFA model comparisons, and checking whether various constrained and unconstrained models were significantly different
from one another (Netemeyer et al. 2003; Algesheimer et al. 2005). Discriminant validity is generally measured using a $\chi^2$ difference test that is conducted for all pairs of constructs to determine their distinctiveness. In this approach each pair of constructs is collapsed into a single-construct constrained model, which is then compared with a two-construct freed model in terms of model fit (Anderson & Gerbing 1988). Discriminant validity is evidenced when a two-factor model with the freed coefficient shows a superior fit compared to the model with constrained or fixed coefficient. This result would indicate that the two constructs involved are empirically distinct. However, this actually turns out to be a weak test as significant fit differences may be obtained even when the correlations between the two constructs are very high. An alternative test of discriminant validity is based on the logic that a construct should explain its observed variables better than it explains any other construct. This test is conducted by showing that the average variance extracted is greater than the square of the correlations. Equivalently, discriminant validity is achieved if the square root of the average variance extracted is larger than the correlation coefficients (Fornell & Larcker 1981; for an annotated illustration refer to Malhotra 2010). Researchers most frequently use three different fitness indices in this context: the comparative fit index (CFI), incremental fit index (IFI) and non-normed fit index (NNFI). Each of these indices should have a value more than 0.90 to indicate a good fit of the model to the data. However, fit indices often get inflated due to the effect of freeing more parameters to be estimated from the data. For a single-item measure, discriminant validity can be established by showing low and non-significant correlations with other constructs that are theoretically unrelated, and demonstrating that correlations with related but distinct constructs are less than 1.0 in absolute magnitude. Here again, the MTMM approach can be used.

**Nomological validity assessment:** in the present context, multiple external referents should be used for the nomological validity evaluation. Nomological validity should be assessed in several ways. The researcher should examine the simple bivariate correlation (validity coefficient) between the predictor and the criterion, and compare the $R^2$ value derived, by running a multivariate regression. Nomological validity is generally indicated by a moderate level of correlation between predictor scores and criteria scores. However, too high a correlation might indicate the possibility that the item is tapping an external criterion to a higher degree than the facet it was intended to measure. Following Stanton
et al. (2002), the researcher should check if the corrected item–total correlation for each item is substantially higher than any of its item–criterion correlations. Cases with item–criterion correlations in excess of 0.60 should be carefully reviewed. As for the multiple regression method it is important to avoid under-specification of the causal model. Such conditions might result in an inflated validity coefficient because the single predictor is likely to include the effects of other causal variables. In addition, known theoretical relationships should be transformed into hypotheses that can be empirically tested using SEM. Nomological validity is established to the extent that these hypotheses are supported in the SEM analysis. The estimated parameter for a hypothesised relationship should be statistically significant and have the correct sign. For single-item scale development, each of the items should be subjected to a sequential assessment of their respective predictive criteria to finally identify the item that should be included in the final single-scale form. Predictive validity in this case can be assessed by two methods. One compares the simple correlation between the predictor and the criterion. Bivariate correlation is the usual statistic for ‘predictive validity’ in the psychometric test literature for concurrent or, if the criterion is delayed, predictive validity (Cronbach 1951). However, Fishbein and Middlestadt (1995), among others, argue that the validity coefficient (correlation) of a predictor with a criterion will be inflated if the model of causes of the criterion is underspecified because the single predictor is likely to include the effects of another causal variable or variables. Consequently, Rossiter (2002) suggests the use of multivariate regression as an alternative method for predictive validity assessment.

Conclusion

We have conducted a comprehensive review of the arguments for and against full-length, short-form and single-item scales. Providing any isolated set of guidelines for single- versus multi-item scales, long form versus short form, would be incomplete and insufficient; hence we have presented an integrated framework for scale development, reduction and expansion, and single-item generation. Our proposed framework integrates the procedural frameworks for scale development suggested by Churchill (1979), Rossiter (2002), Stanton and colleagues (2002), and Maklan and Klaus (2011), and incorporates the methodological advances suggested by Anderson and Gerbing (1988), Malhotra et al. (2004) and Malhotra et al. (2006).
In conclusion, this paper makes a significant contribution to the literature in many ways. It helps to resolve the recent controversy between single- and multi-item scales. It presents a general framework that can be applied for developing a new scale (multi-item or single-item), shortening an existing scale or expanding a short scale. Our framework will not only guide the selection of an appropriate scale form but should also result in scales with desirable psychometric properties. The use of better measures will result in improved quality of the research findings leading to accumulated knowledge over time.

References


**About the authors**

Dr Naresh K. Malhotra is Regents’ Professor Emeritus, Georgia Institute of Technology. He is listed in Marquis Who’s Who in America, since 1997, and Who’s Who in the World, 2000. In 2010, he was selected as a Marketing Legend and his refereed journal articles were published in nine volumes by Sage with tributes by other leading scholars in the field. He is ranked number one based on several published research rankings. In a landmark study by Ford et al. (2010) examining publications in the top four marketing journals (*JMR, JM, JAMS* and *JCR*) over a 25-year period from 1977 to 2002, Professor Malhotra has three top-three rankings: ranked number three based on publications in all the four journals combined, ranked number three based on publications in *JMR*, and ranked number one based on publications in *JAMS*. He also holds the all-time
record for the maximum number of publications in the Journal of Health Care Marketing. He is ranked number one based on publications in the International Marketing Review since its inception to 2011. His book, Marketing Research: An Applied Orientation, Sixth Edition, published by Prentice Hall, Inc. has been translated into several languages and has received widespread adoption in the United States all over the world. He is also the author of Basic Marketing Research: Integration of Social Media, Fourth Edition, another global leader.

Soumya Mukhopadhyay is a third year doctoral student in the marketing department at Nanyang Business School (NTU, Singapore). His interests cover various aspects of online consumer behaviour and user content generation. His current research focuses on developing individual level statistical models for online product opinions (ratings and reviews) using Bayesian econometrics.

Xiaoyan Liu is a third year PhD candidate in the Division of Marketing and International Business at Nanyang Business School (NTU, Singapore). Her research interests focus on differences in cross culture consumers’ behaviours, especially in brand evaluation and product selection. Her current research is to understand whether consumers across cultures react to mortality salience differently and its implications on their consumption behaviour.

Dr Satyabhusan Dash is currently working as an Associate Professor, Marketing area and Chairperson in the Centre for Marketing in Emerging Economies at the Indian Institute of Management Lucknow. His co-authored research paper published in Journal of Indian Business Research has been chosen as an Outstanding Paper Award Winner at the Literati Network Awards for Excellence 2012. His co-authored book with Naresh K. Malhotra, Marketing Research: An Applied Orientation, Sixth Edition (Indian Adaptation) published by Pearson has received widespread adoption in India. Dr Dash has published in several major international journals including Academy of Marketing Science Review, International Journal of Market Research, Journal of International Consumer Marketing, Marketing Intelligence and Planning, and International Journal of Bank Marketing.

Address correspondence to: Naresh K. Malhotra, Georgia Institute of Technology, Scheller College of Business, 800 West Peachtree St. NW, Atlanta, Ga 30308-1149, USA.

Email: naresh.malhotra@scheller.gatech.edu