Scale Purification

State-of-the-art Review and Guidelines

Wieland, Andreas; Kock, Florian; Josiassen, Alexander

Document Version
Accepted author manuscript

Published in:
International Journal of Contemporary Hospitality Management

DOI:
10.1108/IJCHM-11-2017-0740

Publication date:
2018

License
Unspecified

Citation for published version (APA):
Wieland, A., Kock, F., & Josiassen, A. (2018). Scale Purification: State-of-the-art Review and Guidelines. *International Journal of Contemporary Hospitality Management*, *30*(11), 3346-3362. https://doi.org/10.1108/IJCHM-11-2017-0740

Scale Purification: State-of-the-art Review and Guidelines

Andreas Wieland, Florian Kock, and Alexander Josiassen

Journal article (Accepted manuscript*)

Please cite this article as:
Wieland, A., Kock, F., & Josiassen, A. (2018). Scale Purification: State-of-the-art Review and Guidelines. International Journal of Contemporary Hospitality Management, 30(11), 3346-3362. DOI: 10.1108/IJCHM-11-2017-0740

DOI: 10.1108/IJCHM-11-2017-0740

* This version of the article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the publisher's final version AKA Version of Record.

Scale Purification: State-of-the-art Review and Guidelines

Andreas Wieland, Florian Kock & Alexander Josiassen

Abstract

Purpose

The authors identify scale purification criteria for both uni- and multidimensional reflective scales and apply these criteria to an evaluation of the methodological status quo of the hospitality literature.

Design/methodology/approach

Based on a literature review, the authors develop a taxonomy of statistical and judgmental criteria across scale levels, from which best practice methodologies are derived. Recent publications in leading hospitality journals are then evaluated based on these scale purification steps.

Findings

The authors uncover a lack of transparency when reporting scale purification practices. Moreover, methodological steps are often entirely omitted or insufficiently followed, especially when it comes to judgmental scale purification practices.

Research limitations/implications

The authors focus on reflective scales in the hospitality discipline. Methodological traditions in other fields might lead to other results if the chosen approach were to be repeated there.

Practical implications

The authors provide a set of suggestions that will help researchers in hospitality and adjacent disciplines to greater consensus and consistency of application regarding the methodological steps when carrying out scale purification in reflective scales.

Originality/value

Research on scale purification in hospitality research has been scarce. The authors extend existing research and provide the most comprehensive study so far of present and best scale purification practice, using both statistical and judgmental criteria.

Introduction

Hospitality research has evolved into a largely empiricism-driven discipline (Morosan *et al.*, 2014) and survey methods are used for data collection, and regression or structural equation modeling for data analysis (Assaf *et al.*, 2016; Kock *et al.*, 2016). Among the challenges of applying such methods in the social sciences is that most constructs used in empirical models are operationalized as latent variables, thus, they represent "phenomena of theoretical interest which cannot be directly observed and have to be assessed by manifest measures which are observable" (Diamantopoulos *et al.*, 2008, p. 1204). The latent nature of phenomena studied in hospitality research creates challenges for measurement (Hwang and Seo, 2016). As a consequence, a structured and comprehensive process of developing measurement models (referred to as "scales" for brevity in this article) is essential (Liu and Arendt, 2016; Pijls *et al.*, 2017). Although scale purification – the justified removal of items from multi-item scales – has been acknowledged across disciplines as an important step towards creation of any scale (e.g.; Churchill, 1979; Hardesty and Bearden, 2004; Homburg *et al.*, 2015), researchers have noted that "there is little discussion of how to apply [criteria] to make decisions about which items to omit to purify the scale" (MacKenzie *et al.*, 2011, p. 311). Additionally, while such criteria exist for unidimensional scales, we found virtually no attempt to develop such criteria for higher-order (i.e., multidimensional) scales (cf. Jarvis *et al.*, 2003).

The goal of this research is to address these challenges and contribute to the existing literature in three important ways. First, by building on recent developments in the existing methodological literature, we present a dualistic approach to scale purification, which takes into account not only statistical criteria, but also judgmental criteria. These criteria are applied at different hierarchical measurement levels for unidimensional reflective scales. Second, we enhance this methodological framework by developing statistical and judgmental criteria also for multidimensional reflective scales. Third, we use the two resulting sets of criteria for uni- and multidimensional scales to evaluate the methodological status quo of the hospitality literature. Specifically, these criteria can also help readers of academic literature evaluate the quality of the scales, guide academics themselves when carrying out scale purification, provide reviewers with tools to identify methodological omissions, and support editors when making review decisions. An *ad-hoc* analysis conducted as part of this research has revealed that hospitality is dominated by unidimensional measures and, when higher-order measurement is applied, it is almost always in the form of *second*-order models, where both levels are specified in a *reflective* manner.

In the following sections, we distinguish the types of criteria (statistical and judgmental) for scale purification decisions and three qualities (reliability, validity and parsimony). Then, two frameworks are presented that systematically guide scale purification decisions for (1) unidimensional and (2) multidimensional scales. In order to provide an overview of the current state-of-the-art of scale

purification in hospitality, we reviewed relevant articles that have recently been published in leading hospitality journals. Finally, results of the evaluative review are presented and conclusions are drawn.

Developing a dualistic taxonomy of scale purification through criterion type and quality lens

Criterion type

A distinction can be made between *statistical* and *judgmental criteria* to assist scale purification decisions (Wieland *et al.*, 2017). Statistical criteria usually relate to statistical heuristics or tests. These criteria often, though not always correctly, involve "cutoff criteria" (Lance *et al.*, 2006). While often being essential to evaluate the quality of scales, statistical criteria are not appropriate to evaluate how these scales relate to the realm of ontology. Most radically, Rossiter (2008) argues that statistical procedures are often "inappropriate 'empirical crutch' procedures" (p. 380) and claims that such procedures are unable to provide the evidence needed to establish the validity of a scale. Even though Rigdon *et al.* (2011) criticize "revolutionary" conclusions by Rossiter to overcome these issues by rejecting any type of psychometric (i.e., statistical) technique, Rigdon *et al.* (2011) still agree that "content considerations are seriously undervalued in contemporary measure development" (p. 1591). As Borsboom *et al.* (2004) note: "[T]ables of correlations between test scores and other measures cannot provide more than circumstantial evidence for validity" and "the problem of validity cannot be solved by psychometric techniques or models alone. On the contrary, it must be addressed by substantive theory" (p. 1062). Where the criticism is right is that statistical criteria, indeed, ignore what Bagozzi and Yi (2012) call the "theoretical meaning", as these criteria exclusively operate on the level of "empirical meaning". Therefore, it may be rather surprising that the literature related to scale development and validation is, with some exceptions, dominated by a focus on statistical criteria (Hardesty and Bearden, 2004).

A second – as we argue complementary, not alternative – category of criteria exists that can be used to assess the distance between theoretical constructs and their associated scales. These criteria therefore bridge the gap between the theoretical and empirical meanings. These criteria are called "judgmental", as it is not a heuristic or a statistical test that produces a quality indicator, but the indicator is the subjective product of a judgment (Stanton *et al.*, 2002). In the hospitality literature, judges have only rarely been involved in the scale development process (a notable exception is Khan and Rahman, 2017). Among the implementations of judgmental scale purification is a procedure described by Moore and Benbasat (1991) who suggest that items are sorted by judges to establish which of the items should belong to which of the constructs. This procedure is the judgmental counterpart to a statistical factor analysis as, unlike the latter, it involves both the theoretical and empirical meanings. Note that, although judges, today, are human actors, it is not out of the question that, due the rapid

pace of advancements in the area of machine learning, non-human actors could serve as such judges in the future. The use of judgmental criteria is particularly important for higher-order measurement models, both reflective and formative ones. A higher-order manifestation of a given construct derives first and foremost from conceptual reasoning of its dimensions, and is therefore amenable to judgmental criteria.

Quality lens

The recent methodological literature suggests that scale purification decisions can be informed through three distinct lenses: the validity, reliability and parsimony of a scale (Netemeyer *et al.*, 2003, p. 57; MacKenzie *et al.*, 2011). First, *validity* has traditionally been understood as "the extent to which [scales] measure what they purport to measure" (Buckingham, 1921, p. 274). Borsboom *et al.* (2004) demonstrate that if a realist view is taken and when claiming that a scale is valid, the ontological position is taken that the attribute that is measured exists and that it affects the outcome of the procedure of measurement and, therefore, validity refers to whether the test developer has been successful in developing a test that senses variations in these attributes. Second, *reliability* is used to cover the precision of measurement (Mellenbergh, 1996). This view can be interpreted both statistically, e.g., in terms of a high item–factor correlation, and judgmentally, e.g., if judges come to the conclusion that an item is representative of the factor. Finally, following the idea that parsimony can be understood as a *ratio* (cf. Bacharach, 1989), we suggest that *parsimony* relates to the ratio between the amount of data (statistical criteria) or information (judgmental criteria) that are/is used for measurement and the complexity of the construct to be measured. In that sense and extending Wieland *et al.*'s (2017) view of parsimony for one-dimensional scales, parsimony aims to minimize the amount of data (or information) that is necessary to cover all relevant aspects of a construct. It is particularly noteworthy that for multi-dimensional scales the lower bound of items is restricted by the number of dimensions of the construct.

A systematic approach of purifying scales

Unidimensional scales

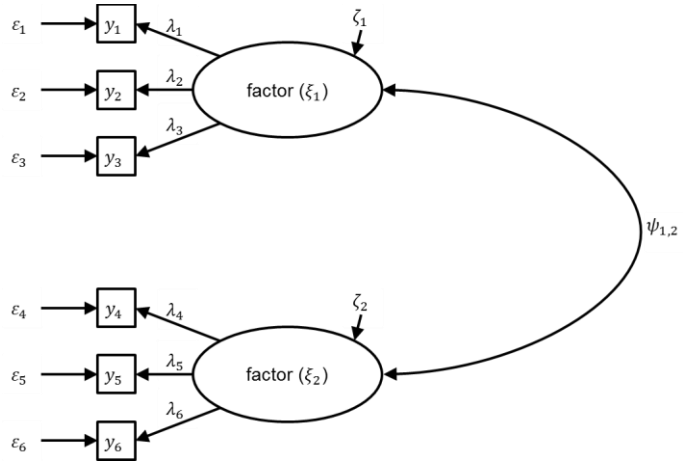

Fig. 1: Example of two reflective unidimensional scales.

In the following, drawing on Wieland *et al.* (2017), we present criteria for taking scale purification decisions in the context of reflective scales. We start by presenting criteria for developing unidimensional scales (see Fig. 1 for an example). Scale purification decisions for such scales should, in principle, consider two key elements: item and factor. Particularly, decisions are based on the qualities (i.e., reliability, validity and parsimony) of items and factors, but also on the qualities of the relationships among and between these elements (cf. Carpenter *et al.*, 2016). Combinatorically, this leads to five different units of analysis: intra-item, intra-factor, inter-item, inter-factor and item–factor. The *intra*-item and *intra*-factor units of analysis can be analyzed in terms of the quality of the respective element of the scale itself, i.e., only the item or the factor, respectively, are analyzed in isolation. For example, mean, standard deviation, skewness and kurtosis (Dawes, 2008) all represent indicators of the intra-item quality. Then, the *inter*-item and *inter*-factor units of analysis indicate the quality of the relationship among two or more items and factors, respectively. Here, representatives of the same element of a scale are juxtaposed, allowing pairwise comparisons. The correlation of two items that purport to represent the same factor (Bearden *et al.*, 2011) is an example of an indicator of inter-item quality. Finally, the *item–factor* unit of analysis indicates the quality of the relationship between these two different types of elements. For example, the factor loadings between an item and its designated factor (Bagozzi and Yi, 2012) represent and indicator of item–factor quality. We will now investigate these units of analysis looking at the three quality lenses sequentially (summarized in Table I).

Table I Statistical and judgmental criteria for purifying unidimensional reflective scales

Validity	Statistical criteria	Judgmental criteria
intra-item	• Mean and skewness do not relate to the underlying distribution (Dawes, 2008).	• Item formulation is not sufficiently sharp.
inter-item	• Correlation between items is too low for items representing the same factor (Bearden *et al.*, 2011). • Correlation between items is too high for items representing different factors (Bearden *et al.*, 2011).	• Item formulations are not equivalent for items representing the same factor. • Item formulations are equivalent for items representing different factors.
item–factor	• Convergent validity is too low (Campbell and Fiske, 1959), as indicated by low factor loadings between item and designated factor (Bagozzi and Yi, 2012) and values of the goodness-of-fit indices of the CFA model (Anderson and Gerbing, 1988). • Discriminant validity is too low, as indicated by high factor loadings between item and non-designated factor, i.e. cross-loadings (Bagozzi and Yi, 2012).	• Items are rated as "clearly representative" or "somewhat representative" of the non-designated factor (Zaichkowsky, 1994). • Items are not rated as "clearly representative" or "somewhat representative" of the designated factor (Zaichkowsky, 1994). • Q-sort procedures do not demonstrate that the item represents the designated factor (Moore and Benbasat, 1991). • Q-sort procedures demonstrate that the item represents the non-designated factor (Moore and Benbasat, 1991).
intra-factor	• Mean and skewness do not relate to the underlying distribution.	• Conceptualizations and definitions based on item formulations does not represent construct properly (Moore and Benbasat, 1991).
inter-factor	• Discriminant validity on the factor level is too low (Campbell and Fiske, 1959), as indicated by heuristics such as AVE–SE comparison (Fornell and Larcker, 1981) and HTMT approach (Henseler *et al.*, 2015), inferential tests such as the constrained-ϕ approach (Jöreskog, 1971) or non-significant χ^2-difference test between constrained and unconstrained model (Bagozzi *et al.*, 1991), high correlations (Dietvorst *et al.*, 2009), and values of the goodness-of-fit indices of the CFA model (Anderson and Gerbing, 1988).	• Conceptualizations and definitions of different factors are equivalent.

Reliability	Statistical criteria	Judgmental criteria
intra-item	• Kurtosis and standard deviation do not relate to the underlying distribution (Dawes, 2008).	• Item formulation is ambiguous (Puri, 1996).
inter-item	• See respective table field under validity.	• Pairwise comparison of item formulations reveals potential sources for ambiguity (cf. Moore and Benbasat, 1991).
item–factor	• Individual item reliability is too low (Bagozzi and Yi, 1988). • Item–total correlation is too low for the designated factor (Hair *et al.*, 1998). • Item–total correlation is too high for the non-designated factor (Hair *et al.*, 1998).	• See respective table field under validity.
intra-factor	• Internal consistency is too low, as indicated by tau-equivalent reliability ρ_T (Cronbach, 1951) and congeneric reliability ρ_C (Jöreskog, 1971). • Average variance extracted is too low (Fornell and Larcker, 1981). • Kurtosis and standard deviation do not relate to the underlying distribution.	• Factor conceptualization and definition is ambiguous (Gilliam and Voss, 2013; Podsakoff *et al.*, 2016).
inter-factor	• Correlation between theoretically unrelated factors is too high (Kline, 2005).	• Pairwise comparison of factor conceptualization and definitions reveals potential sources for ambiguity.

Parsimony	Statistical criteria	Judgmental criteria
intra-item	• Number of characters or words of item formulation is too high.	• Number of morphemes of item formulation is too high (Johnson, 2004).
inter-item	• Redundancy among items, as indicated by too high inter-item correlations.	• Semantic redundancy between items, as indicated by qualitative inter-item comparisons (Rossiter, 2002).
item–factor	• Removing an item would further increase the adjusted goodness of fit (AGFI) index (Voss *et al.*, 2003) or comparable indices (Frohlich, 2002). • Removing an item would not or not substantially decrease the explained variance in dependent variables.	• Measurement made with an item does not prove to be essential to capture the construct's meaning (Lawshe, 1975).
intra-factor	• Number of items per factor is too high (Stanton *et al.*, 2002).	• Number of items is not based on qualitative considerations (Rossiter, 2002).
inter-factor	• Redundancy among factors exists, as indicated by too high inter-factor correlations.	• Semantic redundancy among factors exists, as indicated by qualitative comparisons of conceptualizations and definitions of factors.

The first quality lens is *validity*. First, for the *intra-item* unit of analysis, statistical criteria that justify the removal of items are that the mean and skewness values (cf. Dawes, 2008) indicate undesired floor or ceiling effects: For most cases where a 7-point scale is used, items should be formulated in a way that the average mean value comes close to the fourth (i.e. central) scale point. From a judgmental perspective, this can be assessed by analyzing whether the formulation of an item reflects the desired degree of "sharpness". Second, for *inter-item* comparisons, too high or too low correlation between two items can be used as statistical criteria if the items represent different factors or the same factor, respectively (Bollen and Lennox, 1991; Bearden *et al.*, 2011). The corresponding judgmental criterion is concerned with whether the meaning (as indicated by formulations) of such items are equivalent or not equivalent, respectively. Third, statistical criteria for *item–factor* comparisons often build on factor analysis: Especially, a lack of convergent validity (Campbell and Fiske, 1959) on the item level is indicated by the values of the goodness-of-fit indices of the confirmatory factor analysis (Anderson and Gerbing, 1988) and low factor loadings between an item and its designated factor (Bagozzi and Yi, 2012). If, however, *high* factor loadings occur between a factor and a *non*-designated item, this can be interpreted as a lack of discriminant validity (Bagozzi and Yi, 2012). Due to the conceptual independence among unidimensional factors, an orthogonal rotation (e.g., Varimax) and principal component analysis should be used (Hair *et al.*, 1998). Criteria presented by Moore and Benbasat (1991) and Zaichkowsky (1994) constitute the "judgmental side of the coin" of the aforementioned statistical criteria. They relate to the degree to which an item is judged as representative of a factor. Fourth, similar to the intra-item unit of analysis, mean and skewness could also be evaluated statistically for the *intra-factor* unit of analysis. Moore and Benbasat (1991) describe a step of their procedure where judges are asked to categorize items and to create labels for these resulting categories; if such a label does not properly reflect the conceptualizations and definitions of the construct, this can be used as a judgmental criterion for item removal on the factor

level. Finally, when it comes to *inter-factor* relationships, several criteria exist that are based on a lack of discriminant validity (Campbell and Fiske, 1959): items should be eliminated if this is indicated by heuristics (e.g.; AVE–SE comparison, Fornell and Larcker, 1981; HTMT approach, Henseler *et al.*, 2015), inferential tests (e.g.; constrained-ϕ approach, Jöreskog, 1971; non-significant χ^2-difference test between the constrained and unconstrained model, Bagozzi *et al.*, 1991), high factor correlations (Dietvorst *et al.*, 2009), and the values of the goodness-of-fit indices of the CFA model (Anderson and Gerbing, 1988). Judges could also evaluate to what extent conceptualizations and/or definitions are equivalent when comparing the factors.

The second quality lens we can take focuses on *reliability*. First, for the *intra-item* unit of analysis a solution is to drop an item if its kurtosis and standard deviation do not sufficiently relate to the underlying distribution (Dawes, 2008). A judgmental criterion is that the formulation of the item is ambiguous (Puri, 1996). Second, for *inter-item* relationships we did not identify any distinct statistical criterion, but the criteria under "validity" also apply here. Judgmentally, a pairwise comparison of item formulations can help to reveal potential sources of ambiguity (cf. Moore and Benbasat, 1991) and high ambiguity serves as a criterion for item removal. Third, for *item–factor* relationships, an item should be removed when the individual item reliability is too low (Bagozzi and Yi, 1988) or when the item–total correlation is too low (or high, respectively) for the (non-)designated factor (Hair *et al.*, 1998). For the former criteria modern software for structural equation modelling (e.g., SPSS Amos) provides estimates, calling them "squared multiple correlations". Judgmentally, reliability on the item–factor level is typically already covered when assessing the corresponding level under "validity". Fourth, when turning to the *intra-factor* unit of analysis, a common criterion for item removal is that the internal consistency is too low. This is indicated by tau-equivalent reliability ρ_T (Cronbach, 1951; traditionally called "Cronbach's α") and congeneric reliability ρ_C (Jöreskog, 1971; traditionally called "composite reliability").[1] Another statistical solution is to remove an item if the level of average variance extracted is too low (Fornell and Larcker, 1981). Also here, the standard deviation and kurtosis should be evaluated, this time on the factor level, and also here, if assessed judgmentally, ambiguity leads to item removal – here the ambiguity of factor conceptualization and definition (Gilliam and Voss, 2013; Podsakoff *et al.*, 2016). Finally, for the *inter-factor* unit of analysis, a statistical criterion for item removal is that the correlation between theoretically unrelated factors is too high (Kline, 2005). Here, a judgmental criterion for item removal is that a pairwise comparison of factor conceptualization and/or definitions reveals potential sources for ambiguity.

[1] In most cases when measurement models are used in hospitality, the assumption of homogenous factor loadings ($\lambda_{i,k} = \lambda_{j,k}, \forall i, j, k$) for tau-equivalent measurement models cannot be maintained. Although it is a widespread practice to report ρ_T also for the common case of congeneric measurement models, we do not recommend doing so. If ρ_C is already reported for such measurement models, simultaneously reporting ρ_T is of very limited value due to the unrealistic assumption (cf. Cho, 2016).

Unlike at the intra-factor level, the ambiguity is detected not by analyzing the definition of one factor, but via comparing the definitions of two factors.

The last quality lens is *parsimony*. First, for the *intra-item* unit of analysis, a statistical criterion for item removal is that the number of characters or words of the item formulation is too high. Similarly, but requiring a judgmental evaluation of meaning, the number of morphemes of the item formulation can be too high (Johnson, 2004). Second, when comparing items (i.e., *inter-item* unit of analysis), too high inter-item correlations serve as an indication for redundant items and, thus, as a statistical criterion. Correspondingly, a judgmental criterion for item removal is if judges identify such redundancies when comparing items (Rossiter, 2002). Third, for *item–factor* comparisons, a statistical criterion for item removal is whether removing an item would further increase the value of the adjusted goodness of fit index (AGFI; Voss *et al.*, 2003). Frohlich (2002) followed a similar approach, although relying on other indices.[2] Another criterion for item removal is that removing this item would not substantially decrease the explained variance in dependent variables. A common judgmental criterion that relates to judging loss of meaning is if the measurement made with the item does not prove essential to capture the meaning of the factor (Lawshe, 1975). Fourth, moving to the *intra-factor* unit of analysis, a simple but effective statistical criterion is that the number of items per factor should not be too high (Stanton *et al.*, 2002; Diamantopoulos and Siguaw, 2006). This criterion also reflects the common research practice to keep removing items until the number of items is lower than a chosen cutoff value, which often results in retaining three to five items. Finally, for the *inter-factor* unit of analysis, redundant factors should lead to the removal of factors. This is statistically indicated by too high inter-factor correlations and judgmentally by qualitative comparisons of conceptualizations and/or definitions of the factors.

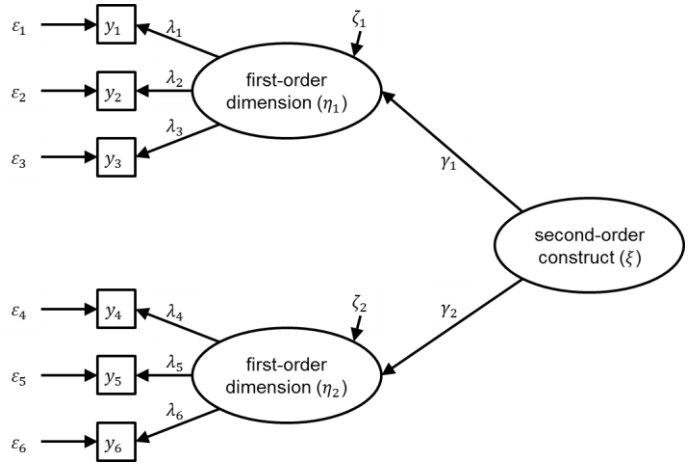

Fig. 2: Example of a multidimensional scale.

[2] We recommend that in each round also judgmental criteria should be incorporated when following such a procedure.

Multidimensional scales

So far we focused on criteria for unidimensional rather than multidimensional scales. To the best of our knowledge, no research has systematically addressed how scale purification decisions can be substantiated by applying statistical and judgmental criteria in order to increase the validity, reliability and parsimony of multidimensional scales (see Fig. 2 for an example). As concepts in hospitality research are occasionally operationalized as multidimensional scales (e.g., Hyun and Perdue 2017), understanding amenable scale purification criteria is important. Such models consist of at least two first-order dimensions that constitute a second-order construct. A key advantage of multidimensional scales is that by looking at dimensions, both managers and academics can spot important impact differences in structural models that would have remained hidden if a unidimensional scale was used. As such, multidimensional scales provide more diagnostics and are conceptually more appropriate than their unidimensional counterparts if the focal concept has significantly distinct facets.

As each of these dimensions is simultaneously a factor, the criteria for scale purification described for unidimensional models also apply on the dimension level. However, additional units of analysis need to be introduced to acknowledge the complexity of multidimensional scales. First, the inter-dimension unit is different from the general inter-factor unit of analysis, as it reflects the unique relationship between factors that manifest the same second-order construct. Second, the dimension–second-order construct unit of analysis can be analyzed in terms of the quality of the relationship between a first-order dimension and its designated second-order construct. Finally, the intra-second-order construct unit of analysis views the second-order construct as a "black box", hereby masking the contained dimensions and items. Similarly to the unidimensional case, we will now investigate these additional units of analysis, hereby again distinguishing between validity, reliability and parsimony (summarized in Table II).

The first quality lens is *validity*. For the *inter-dimension* unit of analysis, statistical criteria for scale purification relate to the different types of factor analyses: Item removal should be considered if an exploratory factor analysis (EFA) indicates cross-loadings and therefore does not support the conceptualized dimensionality. In order to acknowledge the conceptual integrity of dimensions under the same second-order construct, factors should be allowed to correlate (i.e., not constrained to be orthogonal). Thus, an oblique factor rotation (e.g., Promax or Oblimin) and maximum-likelihood estimation should be used. Item removal decisions should also consider whether a first-order confirmatory factor analysis (CFA) does not support convergent validity of the items for each dimension, and discriminant validity of items across dimensions. A corresponding judgmental criterion is that the conceptualizations and definitions of different dimensions are equivalent. For the *dimension–second-order construct* unit of analysis, a statistical criterion is that the second-order construct loadings are too low (Dietvorst *et al.*, 2009), which can be obtained through a higher-order

CFA to examine whether the dimensions satisfactorily load on a higher-order construct (high higher-order factor loadings). Here, a judgmental criterion is that the conceptual domain of the higher-order construct is not aligned with its manifest dimensions. Finally, for the *intra-second-order construct* unit of analysis, a statistical criterion for item removal is that the CFA goodness-of-fit indices indicate insufficient model fit, and a judgmental criterion is that the conceptual domain of the higher-order construct does not accommodate multidimensionality.

Table II Statistical and judgmental criteria for purifying multidimensional reflective scales

Validity	Statistical criteria	Judgmental criteria
inter-dimension	• EFA does not support the conceptualized dimensionality, indicated by cross-loadings. • First-order CFA does not support convergent validity of the items for each dimension, and discriminant validity of items across dimensions.	• Conceptualizations and definitions of different dimensions are equivalent.
dimension–second-order construct	• Second-order construct loadings are too low, obtained through a higher-order CFA.	• The conceptual domain of the higher-order construct is not aligned with its manifest dimensions.
intra-second-order construct	• CFA goodness-of-fit indices indicate insufficient model fit.	• The conceptual domain of the higher-order construct does not accommodate multi-dimensionality.

Reliability	Statistical criteria	Judgmental criteria
inter-dimension	• Co-variances between dimensions are too low or too high (Dietvorst *et al.*, 2009).	• Pairwise comparison of dimension conceptualizations and definitions reveals potential sources for ambiguity.
dimension–second-order construct	• See respective table field under validity.	• See respective table field under validity.
intra-second-order construct	• Internal consistency is too low, as indicated by multidimensional tau-equivalent reliability ρ_{MT} (Cho, 2016) and second-order factor reliability ρ_{SOF} (Cho, 2016).	• Second-order construct conceptualization and definition is ambiguous (Gilliam and Voss, 2013; Podsakoff *et al.*, 2016).

Parsimony	Statistical criteria	Judgmental criteria
inter-dimension	• Redundancy among dimensions exists, as indicated by high co-variances.	• Redundancy between dimensions exists, as indicated by qualitative comparison of conceptual definitions of dimensions.
dimension–second-order construct	• Explained variance in dependent variables does not or not substantially decrease when dimension is removed (Bagozzi *et al.*, 2017).	• The dimension is not essential to represent the conceptualization of the second-order construct.
intra-second-order construct	• Number of dimensions is too high.	• Number of dimensions is not based on qualitative considerations.

The second quality lens is *reliability*. As a statistical criterion for the *inter-dimension* level, it is required that the co-variance between two dimensions should neither be too low nor too high (Dietvorst *et al.*, 2009). If the covariance is too low, this can be interpreted as an indication that the two dimensions do not represent the same second-order construct; if it is too high, this puts the multi-dimensionality of that construct into question. Related, a pairwise comparison of dimension conceptualizations and definitions could reveal potential sources for ambiguity, which can be used as a judgmental criterion here. For the *dimension–second-order construct* unit of analysis we refer to the

criteria under "validity" before, as they also apply here. For the *intra-second-order construct* unit, a statistical criterion is that internal consistency should not be too low, as indicated by multidimensional tau-equivalent reliability ρ_{MT} (Cho, 2016) and second-order factor reliability ρ_{SOF} (Cho, 2016). However, some authors note that dimensions of multidimensional constructs are inevitably heterogeneous, as they represent different manifestations or facets of the construct (Edwards, 2011; Polites *et al.*, 2012). A judgmental criterion is that the conceptualization and definition of the second-order construct is ambiguous (Gilliam and Voss, 2013; Podsakoff *et al.*, 2016).

The remaining quality lens is *parsimony*. For the *inter-dimension* unit of analysis, scale purification criteria relate to redundancy among dimensions, which is statistically indicated by high co-variances between the latents of the dimensions, or judgmentally by a qualitative comparison of conceptualizations and definitions between them. Turning to the *dimension–second-order construct* relationship, a statistical criterion is that the explained variance in dependent variables does not or not substantially decrease when the dimension is removed (Bagozzi *et al.*, 2017). Judgmentally, a criterion is that the dimension is not essential to represent the second-order construct, which needs to be justified based on theoretical considerations. Finally, when taking an *intra-second-order construct* perspective, the number of dimensions can exceed a predefined cutoff value, which is a statistical criterion, or the number of dimensions lacks in sufficient qualitative considerations, which is a judgmental criterion.

Methodology

The developed framework enables us to analyze the current state of purifying scales of multi-dimensional constructs in the hospitality literature. In a first step, a sample of articles was identified that serves as the basis for the analysis. In this step we focused on two journals, which are generally considered the most influential journals in hospitality[3]: *International Journal of Contemporary Hospitality Management* (IJCHM) and *International Journal of Hospitality Management* (IJHM). To represent current practice, the analysis was further limited to the most recent (prior to October 2017) published five issues of each of these journals. As this research is about scale purification, only those articles were included that made use of multi-item scales. To identify such articles, the first author evaluated for each of the articles whether or not to include it and, in case of doubt, consulted the other authors. Particularly, editorials, case studies and articles that applied other types of unrelated methods were excluded, but also quantitative studies that exclusively used single-item scales. This yielded 154 (= 94 [IJCHM] + 60 [IJHM]) articles. For each journal, 30 of these articles were randomly selected for further analysis. This allowed us to have two sub-samples that are comparable in size and that are

[3] For example, as evidenced by the two highest 2016 impact factors of all hospitality journals in the InCites Journal Citation Reports.

each simultaneously large enough to be analyzed and recent enough to allow inference to the current methodological state in hospitality, as methodological standards tend to change over time. Thus, this step led to a sample of 60 articles to be considered. In a second step, the identified articles were independently coded by the second and third authors, hereby using a coding scheme. This scheme was based on the two frameworks for uni- and multidimensional scale purification described before and allowed coders to identify methods that were used in the respective articles. In addition, it was evaluated whether the article used uni- vs. multidimensional constructs and whether it was reported that scale purification was applied. The coding scheme also explicitly urged coders to evaluate whether the article should actually belong to the sample. The coders initially applied the coding scheme to a limited number of articles. In order to avoid any inconsistency, this step ensured that pending issues could be resolved and the decision-making process between the coders be aligned before the actual coding process started. Differences among coders were resolved through discussions among all three authors.

Results

The analysis of the hospitality literature allows us to make several observations. A first observation is that a group of authors do not report the criteria they used to make a scale purification decision. Particularly, in several cases the explanation was vague, simply referring to “statistical analysis results” rather than specifying the type of statistics used to drop items. Importantly and potentially concerning, various studies selected a subset of the original scale items without justifying this selection. In many of these incidents of implicit scale purification, the omission of items was not highlighted explicitly. Thus, the omission was only revealed after we compared the reported item list with the original study that the examined papers referred to. In those cases where the reporting makes a details analysis possible, the results indicate that scale purification decisions are almost exclusively justified based on statistical criteria for reliability (9 instances) and validity (7 instances) that relate to the item–factor unit of analysis. In specific, authors in hospitality turn out to report low factor loadings and factor cross-loadings as their dominating scale purification criteria. In addition, only one of the articles of our sample used multidimensional scales and in this single case cross-loading items were removed, too. It becomes apparent that parsimony is rarely reported on to justify the removal of scale items. Arguably, parsimony plays an implicit role in any such decision. The discrepancy between parsimonious measurement and explicitly reporting the motivation why the number of items should be reduced, therefore, comes with a bit of surprise. Interestingly, we also identified several articles that reported statistical results which should have led to scale purification, e.g., very low factor loadings. However, these results did not lead, without providing any justification, to scale purification although this would arguably have been the right decision. Finally, only very few articles took judgmental criteria into consideration. These results can be interpreted as a potential cause for

concern, as the unilateral domination of statistical criteria might be helpful to improve the empirical measurement properties but at the same time could harm the theoretical meaning of empirical findings.

Discussion

Conclusions

Our literature review of the hospitality literature offered clear evidence of the lack of tailored measurement models for key hospitality concepts. For example, Khan and Rahman (2017) motivate their scale development by writing: "In the absence of a scale that measures experiences of visitors evoked by hotel brand-related stimuli, hoteliers have had to rely on the general brand experience scale which may not be a very accurate measure of brand experience in context of the hotel industry" (p. 269). Similarly, Liu and Arendt's (2016) scale development research starts with the notion that there "appears to be no measurement tool specifically targeted toward identifying individuals' motives for choosing hospitality jobs" (p. 701). These exemplary quotes indicate that hospitality research and management cannot solely rely on scales which have been developed in other disciplines, such as marketing. Instead, efforts have to be made to at least adapt existing scales or to develop new ones. Either way, there is a need for scale development and purification. As DeVellis (2003) writes: "Even if a poor measure is the only one available, the costs of using it may be greater than any benefits attained" (p. 12). If scale purification is an important building block when adapting existing or developing new scales, a rigorous scale purification process is crucial to increase the trustworthiness of research results.

If not applied properly, any methodology loses its power to generate reliable and valid results. The identified omission of important methodological steps in many hospitality research articles should therefore make us consider whether enough efforts have been taken to create valid, reliable and parsimonious scales. But how can these findings be used to help hospitality research to thrive methodologically? We present several suggestions that are grounded in our results and that show how the identified negligence can be overcome in future hospitality research.

Theoretical implications

First and foremost, researchers in hospitality need to pay attention to judgmental criteria. It is only these criteria that explicitly link the theoretical meaning of a concept with the empirical meaning of a scale. By no means should the dominance of statistical criteria be interpreted in a way that prevalent scales do not make this connection. However, our study indicates that researchers overly rely on other steps of the scale development process than purification. But what happens if a statistical approach suggests to remove item a and to keep items b, c and d, whereas a judgmental approach would

indicate that this item a is the only item that sufficiently represents what it purports to measure? If researchers skip a judgmental approach, as so often happens in the analyzed articles herein, the improvement of the statistical measurement properties risks to take the scale away from the theoretical meaning that is was originally aimed to represent.

Another suggestion is to cover the full range of units of analysis when purifying scales. So far, hospitality researchers almost exclusively focus on the item–factor relationship. This relationship is certainly important for scale purification decisions, as each item that might or might not be removed belongs to a factor. However, it is not only the relationship with that factor but also the relationship of the item to other items or the characteristics of the item itself which make it a candidate for removal. Furthermore, even an analysis of statistical and judgmental measurement properties on a higher level, i.e., for the intra- and inter-factor units of analysis, can reveal psychometric shortcomings that might best be remedied by removing an item.

The fact that it is validity- and reliability-, not parsimony-related criteria that dominate scale purification in hospitality research indicates that many researchers put efforts into developing tests that sense variations in attributes and cover the precision of measurement. It might even be the case that validity and reliability should be the dominant quality lenses researchers should take when purifying a scale. However, given the behavioral constraints to capture the complexity of a survey questionnaire and also, much more trivially, the time constraints of survey participants and the space constraints of journal publications, researchers are also advised to limit the number of characters, words, items and dimensions. This advice closely links to the notorious principle of Ockham's razor.

The dominance of scale purification criteria that are concerned with unidimensional scales – compared to multi-dimensional scales – leads to our fourth suggestion. Researchers in hospitality are advised to more explicitly address scale purification issues for second-order constructs. This includes all three additional units of analysis that do not exist in unidimensional models (i.e., inter-dimension, dimension–second-order construct, intra-second-order construct). The higher one moves in the systemic hierarchy of a construct, from the item level up to the level of the second-order construct, the less obvious the issue of item removal might be linked to the unit under study. This, however, does not release the researcher from considering the consequences for the items to be derived from statistical and judgmental defects of the scale.

Our fifth suggestion pertains to the identified dearth of studies that involve scale development and purification of formative measurement models. Only rarely have studies in the area of hospitality or the related tourism discipline investigated phenomena through the conceptualization and operationalization of a formative index (e.g.; Murphy *et al.*, 2009; Josiassen *et al.*, 2016; Kock *et al.*, 2016). Unlike reflective scales that rest on the contention that co-variance is shared among items that

are manifestations of the same underlying latent construct, formative scales are formed through conceptually independent and statistically uncorrelated items that are not interchangeable (Petter *et al.*, 2007; Josiassen *et al.*, 2016). Against this background, scale development and purification implications derive largely from conceptual considerations, consequently referring to judgmental criteria. While statistical criteria exist, such as thresholds of the variance inflation factor, ongoing debates exist on whether and how to use them (e.g., Bagozzi and Yi, 2012). Future research, particularly in the area of hospitality, is needed in order to provide much needed conceptual and operational clarity for formative measurement models. Such research could significantly enhance studies that rely on concepts that are better conceptualized as formative rather than reflective, such as guest or resident satisfaction (e.g., Woo *et al.*, 2015). Using a formative measurement approach could leverage the validity of research results, and impact the study itself.

Based on steps described by Voss *et al.* (2003, p. 313) and Frohlich (2002) and based on our observations regarding judgmental criteria, we suggest the following scale purification procedure, which we believe integrates elements of the aforementioned suggestions. For each scale, an item can potentially be removed following both statistical and judgmental decision making. Voss *et al.* (2003) used the statistical criterion to remove the item with the lowest item–factor correlation. However, this could also be the item that judges identify as the item that represents the designated factor least (cf. Zaichkowsky, 1994). Moreover, other criteria than the one suggested by Voss *et al.* (2003) can be chosen, especially the ones on the intra-item level. Ideally, several statistical and judgmental criteria point to the same item to be removed but, if this is not the case, a balanced approach must be agreed upon. Now, a χ^2-difference test is conducted between the CFA models of the original and reduced scale. A non-significant result of this test indicates that both models do not differ substantially. Item removal routinely leads to an increase of GFI values, but not necessarily of AGFI values, which might even decrease. But if also the AGFI value increases, the reduced scale can be accepted as being better than the original scale – given that the difference test led to a positive result. The aforementioned steps are iterated as long as the χ^2-difference test, indeed, shows a difference and the AGFI value increases. If that is not the case, this procedure leads to a purified scale. As mentioned, the efforts by Frohlich (2002) are extended and generalized, who deletes items selectively using repeated CFA runs; after each run an item is identified for removal based on the standardized residuals, estimated improvements in the χ^2 value with corresponding degrees of freedom, the magnitude of modification indices, the normed fit index (NFI), the value of the comparative goodness of fit index (CFI), and the overall interpretability. What is important after all is to include both statistical and judgmental criteria in each iteration.

Practical implications

The results of this research are based on an analysis of survey research published in academic journals. However, surveys are also widely used for collecting data in service industries like hospitality. Therefore, the results also have implications for surveys used in hospitality organizations. For example, hotels use surveys to gather data about the degree of satisfaction of their guests (Pizam and Ellis, 1999). Although the purpose and, consequently, content and layout differs between academic and business surveys, practitioners, just like researchers, aim to reliably and validly measure indicators. And, in terms of parsimony, the length of questionnaires has often to be reduced to fit the attachment of an email or a one-page form which is left in the guest's hotel room. Practitioners may not always follow the methodological steps described above as rigorously as academics, but the results of this research can still inspire them to consider both statistical and judgmental criteria when designing surveys or when evaluating the quality of purchased survey-based data.

Limitations and future research

Our research can spur interest in the ontological and epistemological basis of research. One could argue that that part of the methodological literature that has exclusively used statistical criteria to make scale purification decisions has operated mainly in the realm of epistemology, whereas judgmental criteria explicitly take ontological considerations into account. In that sense, statistical criteria would be closer to a positivist perspective, whereas judgmental criteria would be closer to a *realist* perspective. It becomes evident from our discussion that, in general, in order to develop, and in particular, to purify a scale, it is not sufficient to limit one's own perspective to either positivism or realism (or any other perspective). Therefore, by demonstrating the benefits of combining statistical and judgmental criteria, our framework builds a bridge between different research philosophies. This insight has the potential to pave the way for better and more rigorous scale development efforts, and thus research that, more explicitly than now, integrates ontological, epistemological and methodological approaches.

References

Anderson, J.C. and Gerbing, D.W. (1988), "Structural equation modeling in practice: A review and recommended two-step approach", *Psychological Bulletin*, Vol. 103 No. 3, pp. 411–423.

Assaf, A.G., Oh, H. and Tsionas, M.G. (2016), "Unobserved heterogeneity in hospitality and tourism research", *Journal of Travel Research*, Vol. 55 No. 6, pp. 774–788.

Bacharach, S.B. (1989), "Organizational theories: Some criteria for evaluation", *Academy of Management Journal*, Vol. 14, pp. 496–515.

Bagozzi, R.P., Batra, R. and Ahuvia, A. (2017), "Brand love: Development and validation of a practical scale", *Marketing Letters*, Vol. 28 No. 1, pp. 1–14.

Bagozzi, R.P. and Yi, Y. (1988), "On the evaluation of structural equation models", *Journal of the Academy of Marketing Science*, Vol. 16 No. 1, pp. 74–94.

Bagozzi, R.P. and Yi., Y. (2012), "Specification, evaluation, and interpretation of structural equation models", *Journal of the Academy of Marketing Science*, Vol. 40 No. 1, pp. 8–34.

Bagozzi, R.P., Yi, Y. and Philips, L.W. (1991), "Assessing construct validity in organizational research", *Administrative Science Quarterly*, Vol. 36 No. 3, pp. 421–458.

Bearden, W.O., Netemeyer, R.G. and Haws, K.L. (2011), *Handbook of Marketing Scales: Multi-Item Measures for Marketing and Consumer Behavior Research*, SAGE Publications, Thousand Oaks, Calif.

Bollen, K. and Lennox, R. (1991), "Conventional wisdom on measurement: A structural equation perspective", *Psychological Bulletin*, Vol. 110 No. 2, pp. 305–314.

Borsboom, D., Mellenbergh, G.J. and van Heerden, J. (2004), "The concept of validity", *Psychological Review*, Vol. 111 No. 4, pp. 1061–1071.

Buckingham, B.R. (1921), "Intelligence and its measurement: A symposium", *Journal of Educational Psychology*, Vol. 12 No. 5, pp. 271–275.

Campbell, D.T. and Fiske, D.W. (1959), "Convergent and discriminant validation by the multitrait-multimethod matrix", *Psychological Bulletin*, Vol. 56, pp. 81–105.

Carpenter, N.C., Son, J, Harris, T.B., Alexander, A.L. and Horner, M.T. (2016), "Don't forget the items: Item-level meta-analytic and substantive validity techniques for reexamining scale validation", *Organizational Research Methods*, Vol. 9 No. 4, pp. 616–650.

Cho, E. (2016), "Making reliability reliable: A systematic approach to reliability coefficients", *Organizational Research Methods*, Vol. 19 No. 4, pp. 651–682.

Churchill, G.A. (1979), "A paradigm for developing better measures of marketing constructs", *Journal of Marketing Research*, Vol. 16 No. 1, pp. 64–73.

Cronbach, L.J. (1951), "Coefficient alpha and the internal structure of tests", *Psychometrika*, Vol. 16, pp. 297–334.

Dawes, J. (2008), “Do data characteristics change according to the number of scale points used?”, *International Journal of Market Research*, Vol. 50 No. 1, pp. 61–77.

DeVellis, R.F. (2003), *Scale Development: Theory and Applications*, 2nd ed., SAGE Publications, Thousand Oaks, Calif.

Diamantopoulos, A., Riefler P. and Roth, K.P. (2008), “Advancing formative measurement models”, *Journal of Business Research*, Vol. 61 No. 12, pp. 1203–1218.

Diamantopoulos, A. and Siguaw, J.A. (2006), “Formative versus reflective indicators in organizational measure development: A comparison and empirical illustration”, *British Journal of Management*, Vol. 17 No. 4, pp. 263–282.

Dietvorst, R.C., Verbeke, W.J.M.I., Bagozzi, R.P., Yoon, C., Smits, M. and van der Lugt, A. (2009), “A sales force–specific theory-of-mind scale: Tests of its validity by classical methods and functional magnetic resonance imaging”, *Journal of Marketing Research*, Vol. 46 No. 5, pp. 653–668.

Edwards, J.R. (2011), “The fallacy of formative measurement”, *Organizational Research Methods*, Vol. 14 No. 2, pp. 370–388.

Fornell, C. and Larcker, D.F. (1981), “Evaluating structural equation models with unobservable variables and measurement error”, *Journal of Marketing Research*, Vol. 18 No. 1, pp. 39–50.

Frohlich, M.T. (2002), “E-integration in the supply chain: Barriers and performance”, *Decision Sciences*, Vol. 33 No. 4, pp. 537–556.

Gilliam, D.A. and Voss, K. (2013), “A proposed procedure for construct definition in marketing”, *European Journal of Marketing*, Vol. 47 No. 1/2, pp. 5–26.

Hair, J.F., Anderson, R.E., Tatham, R.L. and Black, W.C. (1998), *Multivariate Data Analysis*, 5th ed., Prentice Hall, Upper Saddle River, NJ.

Hardesty, D.M. and Bearden, W.O. (2004), “The use of expert judges in scale development: Implications for improving face validity of measures of unobservable constructs”, *Journal of Business Research*, Vol. 57 No. 2, pp. 98–107.

Henseler, J., Ringle, C.M. and Sarstedt, M. (2015), “A new criterion for assessing discriminant validity in variance-based structural equation modeling”, *Journal of the Academy of Marketing Science*, Vol. 43 No. 1, pp. 115–135.

Homburg, C., Schwemmle, M. and Kuehnl, C. (2015), “New product design: Concept, measurement, and consequences”, *Journal of Marketing*, Vol. 79 No. 3, 41–56.

Hwang, J. and Seo, S. (2016), “A critical review of research on customer experience management: Theoretical, methodological and cultural perspectives”, *International Journal of Contemporary Hospitality Management*, Vol. 28 No. 10, pp. 2218–2246.

Hyun, S.S. and Perdue, R.R. (2017), “Understanding the dimensions of customer relationships in the hotel and restaurant industries”, *International Journal of Hospitality Management*, Vol. 64, pp. 73–84.

Jarvis, C.B., MacKenzie, S.B. and Podsakoff, P.M. (2003), "A critical review of construct indicators and measurement model misspecification in marketing and consumer research", *Journal of Consumer Research*, Vol. 30 No. 2, pp. 199–218.

Johnson, J.A. (2004), "The impact of item characteristics on item and scale validity", *Multivariate Behavioral Research*, Vol. 39 No. 2, pp. 273–302.

Jöreskog, K.G. (1971), "Simultaneous factor analysis in several populations", *Psychometrika*, Vol. 36 No. 4, pp. 409–426.

Josiassen, A., Assaf, A.G., Woo, L. and Kock, F. (2016), "The imagery–image duality model: An integrative review and advocating for improved delimitation of concepts", *Journal of Travel Research*, Vol. 55 No. 6, pp. 789–803.

Khan, I. and Rahman, Z. (2017), "Development of a scale to measure hotel brand experiences", *International Journal of Contemporary Hospitality Management*, Vol. 29 No. 1, pp. 268–287.

Kline, R.B. (2005), *Principles and Practice of Structural Equation Modeling*, 2nd ed., Guilford Press, New York, NY.

Kock, F., Josiassen, A. and Assaf, A.G. (2016), "Advancing destination image: The destination content model", *Annals of Tourism Research*, Vol. 61, pp. 28–44.

Lance, C.E., Butts, M.M. and Michels, L.C. (2006), "The sources of four commonly reported cutoff criteria: What did they really say?", *Organizational Research Methods*, Vol. 9 No. 2, pp. 202–220.

Lawshe, C.H. (1975), "A quantitative approach to content validity", *Personnel Psychology*, Vol. 28 No. 4, pp. 563–575.

Liu, Y.-S. and Arendt, S.W. (2016), "Development and validation of a work motive measurement scale", *International Journal of Contemporary Hospitality Management*, Vol. 28 No. 4, pp.700–716.

MacKenzie, S.B., Podsakoff, P.M. and Podsakoff, N.P. (2011), "Construct measurement and validation procedures in MIS and behavior research: Integrating new and existing techniques", *MIS Quarterly*, Vol. 35, pp. 293–334.

Mellenbergh, G.J. (1996), "Measurement precision in test score and item response models", *Psychological Methods*, Vol. 1 No. 3, pp. 293–299.

Moore, G.C. and Benbasat, I. (1991), "Development of an instrument to measure the perceptions of adopting an information technology innovation", *Information Systems Research*, Vol. 2 No. 3, pp. 192–222.

Morosan, C., Bowen, J.T. and Atwood, M. (2014), "The evolution of marketing research", *International Journal of Contemporary Hospitality Management*, Vol. 26 No. 5, pp. 706–726.

Murphy J., Olaru, D. and Hofacker, C.F. (2009), "Rigor in tourism research: Formative and reflective constructs", *Annals of Tourism Research*, Vol. 36 No. 4, pp. 730–734.

Netemeyer, R.G., Bearden, W.O. and Sharma, S. (2003), *Scaling Procedures: Issues and Applications*, SAGE Publications, Thousand Oaks, Calif.

Petter, S., Straub, D. and Rai, A. (2007). “Specifying formative constructs in information systems research”, *MIS Quarterly*, Vol. 31 No. 4, pp. 623–656.

Pijls, R., Groen, B.H., Galetzka, M. and Pruyn, A.T. (2017), “Measuring the experience of hospitality: Scale development and validation”, *International Journal of Hospitality Management*, Vol. 67, pp. 125–133.

Pizam, A. and Ellis, T. (1999), “Customer satisfaction and its measurement in hospitality enterprises”, *International Journal of Contemporary Hospitality Management*, Vol. 11 No. 7, pp. 326–339.

Podsakoff, P., MacKenzie, S. and Podsakoff, N. (2016), “Recommendations for creating better concept definitions in the organizational, behavioral, and social sciences”, *Organizational Research Methods*, Vol. 19 No. 2, pp. 159–203.

Polites, R.L., Roberts, N. and Thatcher, J. (2012), “Conceptualizing models using multidimensional constructs: A review and guidelines for their use”, *European Journal of Information Systems*, Vol. 21, pp. 22–48

Puri, R. (1996), “Measuring and modifying consumer impulsiveness: A cost–benefit accessibility framework”, *Journal of Consumer Psychology*, Vol. 5 No. 2, pp. 87–113.

Rigdon, E.E., Preacher, K.J., Lee, N., Howell, R.D., Franke, G.R. and Borsboom, D. (2011), “Avoiding measurement dogma: A response to Rossiter”, *European Journal of Marketing*, Vol. 45 No. 11/12, pp. 1589–1600.

Rossiter, J.R. (2002), “The C-OAR-SE procedure for scale development in marketing”, *International Journal of Research in Marketing*, Vol 19 No. 4, pp. 305–335.

Rossiter, J.R. (2008), “Content validity of measures of abstract constructs in management and organizational research”, *British Journal of Management*, Vol. 19 No. 4, pp. 380–388.

Stanton, J.M., Sinai, E.F., Balzer, W.K. and Smith, P.C. (2002), “Issues and strategies for reducing the length of self-report scales”, *Personnel Psychology*, Vol. 55 No. 1, pp. 167–194.

Voss, K.E., Spangenberg, E.R. and Grohmann, B. (2003), “Measuring the hedonic and utilitarian dimensions of consumer attitude”, *Journal of Marketing Research*, Vol. 40 No. 3, pp. 310–320.

Wieland, A., Durach, C.F., Kembro, J. and Treiblmaier, H. (2017), “Statistical and judgmental criteria for scale purification”, *Supply Chain Management: An International Journal*, Vol. 22 No. 4, pp. 321–328.

Woo, E., Kim, H. and Uysal, M. (2015), “Life satisfaction and support for tourism development”, *Annals of Tourism Research*, Vol. 50, pp. 84–97.

Zaichkowsky, J. (1994), “The personal involvement inventory: Reduction, revision, and application to advertising”, *Journal of Advertising*, Vol. 23 No. 4, pp. 59–70.