Statistical and Judgmental Criteria for Scale Purification

Wieland, Andreas; Durach, Christian F.; Kembro, Joakim; Treiblmaier, Horst

Document Version
Accepted author manuscript

Published in:
Supply Chain Management: An International Journal

DOI:
10.1108/SCM-07-2016-0230

Publication date:
2017

License
Unspecified

Citation for published version (APA):
Wieland, A., Durach, C. F., Kembro, J., & Treiblmaier, H. (2017). Statistical and Judgmental Criteria for Scale Purification. *Supply Chain Management: An International Journal*, *22*(4), 321-328. https://doi.org/10.1108/SCM-07-2016-0230

Statistical and Judgmental Criteria for Scale Purification

Andreas Wieland, Christian F. Durach, Joakim Kembro, and Horst Treiblmaier

Journal article (Accepted version)

CITE: Statistical and Judgmental Criteria for Scale Purification. / Wieland, Andreas; Durach, Christian F.; Kembro, Joakim; Treiblmaier, Horst. In: *Supply Chain Management*, Vol. 22, No. 4, 2017, p. 321-328.

DOI: 10.1108/SCM-07-2016-0230

Statistical and judgmental criteria for scale purification

Abstract

Purpose – "Scale purification" – the process of eliminating items from multi-item scales – is widespread in empirical research, but studies that critically examine the implications of this process are scarce. The goals of this research are threefold: (1) to discuss the methodological underpinning of scale purification, (2) to critically analyze the current state of scale purification in supply chain management (SCM) research, and (3) to provide suggestions for advancing the scale purification process.

Design/methodology/approach – A framework for making scale purification decisions is developed and used to analyze and critically reflect on the application of scale purification in leading SCM journals.

Findings – This research highlights the need for rigorous scale purification decisions based on both statistical and judgmental criteria. By applying the proposed framework to the SCM discipline, a lack of methodological rigor and coherence is identified when it comes to current purification practices in empirical SCM research. Suggestions for methodological improvements are provided.

Research limitations/implications – The framework and additional suggestions will help to advance the knowledge about scale purification.

Originality/value – This article demonstrates that the justification for scale purification needs to be driven by reliability, validity and parsimony considerations, and that this justification needs to be based on both statistical and judgmental criteria.

Keywords Scale purification; item elimination; survey; measurement; literature review; validity; reliability; parsimony

1. Introduction

Researchers conducting empirical studies regularly need to deal with latent variables – variables which cannot be directly observed. To achieve this, they use measurement instruments to operationalize those variables (Netemeyer *et al.*, 2003; DeVellis, 2012). This can be done in two ways: either reflective, where the direction of causality is from construct to measure (or item); or formative, where the direction is reversed. In the case of reflective measurement with multiple items, these items are expected to correlate, and, in order to improve psychometric measurement properties, the researcher may need to eliminate a number of them (Jarvis *et al.*, 2003).

Eliminating items from a reflective multi-item scale is commonly referred to as "scale purification" (Churchill, 1979; Frohlich, 2002) and is done to improve the measurement properties of newly developed or existing reflective scales. However, previous literature has identified a research gap when it comes to the scale purification process. MacKenzie *et al.* (2011) point out that "there is little discussion of how to apply [...] criteria to make decisions about which items to omit in order to purify the scale" (p. 311). Other authors emphasize the lack of accepted, objective judgmental criteria necessary to justify purification decisions (Hardesty and Bearden, 2004).

If carelessly done, the elimination of items may impair the measurement properties of scales to operationalize constructs – the essential building blocks of empirical models based on survey data (Suddaby, 2010). While the lack of coherent guidelines on the scale purification process is not specific to supply chain management (SCM), their absence poses a challenge to researchers in our discipline, and filling this gap is thus warranted. In this article we therefore seek to address three interlinked objectives. First, we synthesize the general management literature concerned with criteria for scale purification decisions, and develop a framework

with statistical and judgmental criteria to evaluate the reliability, validity and parsimony of a scale. Second, we screen the scale purification approaches commonly used in SCM research, and discuss how contemporary SCM research justifies the elimination of items from scales. Third, we compare the procedures currently used in SCM against the developed framework, with the goals of providing support to future researchers on scale purification and increasing the rigor of empirical SCM research.

2. A framework for scale purification decisions

In the following sections we propose a scale purification framework, which is based on a review of the relevant methodological literature. When discussing the issue of the length of survey study scales, Stanton *et al.* (2002) distinguish between three qualities of scale purification: internal item, external item, and judgmental item qualities. The latter quality is assessed judgmentally; the former two qualities are assessed statistically, although judgmental procedures exist that correspond to statistical procedures (e.g., Moore and Benbasat, 1991). In our research, the distinction between internal and external qualities is further refined by combinatorially providing a framework of qualities that refer to the construct and item levels as well as comparisons between these two elements. Finally, in addition to the considerations of *reliability* and *validity* that are prevalent in the scale development literature (Min and Mentzer, 2004; MacKenzie *et al.*, 2011), a third aspect, *parsimony*, is particularly relevant when it comes to scale purification (Netemeyer *et al.*, 2003, p. 57). While reliability refers to the consistency of measurement (Bryman and Bell, 2015, p. 169) and validity to "the issue of whether or not an indicator (or set of indicators) that is devised to gauge a concept really measures that concept" (Bryman and Bell, 2015, p. 170), we define parsimony as the principle that measurement is based on the least amount of information necessary (e.g., number of items, text per item). In Table I we present a summary

of our literature review on scale purification criteria, and in the sections that follow we will elaborate on these criteria in more detail.

Table I Statistical and judgmental criteria of scale purification decisions

Reliability	Statistical criteria	Judgmental criteria
Item level	Individual item reliability is too low (Bagozzi and Yi, 1988). Standard deviation and kurtosis do not relate to the underlying distribution (Dawes, 2008).	Item formulation is ambiguous (Puri, 1996).
Among items	see validity	Pairwise comparison of item formulations reveals potential sources for ambiguity (Moore and Benbasat, 1991).
Construct level	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).	Conceptual definition of the construct is unclear (Podsakoff *et al.*, 2016).
Among constructs	Correlation between theoretically unrelated constructs is too high (Kline, 2005).	Pairwise comparison of conceptual definitions reveals potential sources for ambiguity (see above).
Between item and construct	Item–total correlation is too low.	Items are not rated as "clearly representative" or "somewhat representative" of the construct under study (Zaichkowsky, 1994).

Validity	Statistical criteria	Judgmental criteria
Item level	Mean and skewness do not relate to the underlying distribution (Dawes, 2008).	Item formulation is not sufficiently readable (Richins, 2004).
Among items	Correlation between items is too low for items representing the same construct and too high for items representing different constructs (Bearden *et al.*, 2011).	Item formulations are not equivalent for items representing the same construct and equivalent for items representing different constructs.
Construct level	Insufficient evidence for criterion validity is provided (Bagozzi, 1981).	Conceptual definition based on item formulations does not represent construct properly (Moore and Benbasat, 1991). Range of the construct is limited (Busse *et al.*, 2017).
Among constructs	Convergent and discriminant validities are too low, as indicated by AVE–SE comparison (Fornell and Larcker, 1981) and HTMT values (Henseler *et al.*, 2015).	Conceptual definitions of different construct are equivalent.
Between item and construct	Substantial cross-loadings (via exploratory factor analysis) or the results of confirmatory factor analysis show that an item does not sufficiently represent the designated construct (Anderson and Gerbing, 1988).	Q-sort procedures do not demonstrate that the item represents the underlying construct (Moore and Benbasat, 1991).

Parsimony	Statistical criteria	Judgmental criteria
Item level	Number of words or characters of item formulation is too high.	Number of morphemes of item formulation is too high (Johnson, 2004).
Among items	Redundancy among items, as indicated by too high inter–item correlations.	Redundancy between items, as indicated by qualitative inter-item comparisons (Rossiter, 2002).
Construct level	Number of items per construct is too high (Stanton *et al.*, 2002).	Number of items is not based on qualitative considerations (Rossiter, 2002).
Among constructs	Redundancy among constructs exists, as indicated by too high inter-construct correlations.	Redundancy between constructs exists, as indicated by qualitative comparisons of conceptual definitions of constructs.
Between item and construct	Removing an item would further increase the adjusted goodness of fit (AGFI) index (Voss *et al.*, 2003) or similar indices (Frohlich, 2002).	Measurement made with an item does not prove to be essential to capture the construct's meaning (Lawshe, 1975).

2.1 Statistical and judgmental criteria

We categorize criteria for scale purification decisions into two separate groups: *statistical* and *judgmental*. *Statistical criteria* use quantitative data, with the purpose of comparing the results of a calculation to a cut-off value or conducting an inferential test (Guide and Ketokivi, 2015). In contrast, *judgmental criteria* are based on a qualitative assessment of the appropriateness of textual data, such as the wording of an item. Their application relies on methodological, theoretical and practical domain knowledge. Such criteria relate to what has previously been discussed as content validity (Nevo, 1985), and scholars have raised specific concerns over lack of consistency and guidance regarding item retention (Hardesty and Bearden, 2004). Moreover, Moore and Benbasat (1991) identify judgmental equivalents for several criteria that were traditionally assessed statistically. Building on this, we will demonstrate in the following sections that statistical criteria *always* have judgmental equivalents in terms of the underlying scale purification goals. However, it is important to note that, due to their different natures, statistical and judgmental criteria will not necessarily lead to the same conclusions. Statistical criteria assess quantitative data using standardized techniques, whereas judgmental criteria build on the intellectual interpretation of qualitative data (c.f. Churchill, 1979) – providing two complementary foci that are mutually supportive, but cannot fully replace each other.

Such criteria exist on the *item level* and *construct level* (Carpenter *et al.*, 2016). In addition, criteria also exist that reflect a comparison *among items* (e.g., do two items sufficiently correlate?), *among constructs* (e.g., does the definition of two constructs differ sufficiently?), and *between item and construct* (e.g., can items be assigned to a certain construct?).

2.2 Reliability, validity and parsimony

We begin with reliability to describe our framework. On the *item level*, statistical indicators to assess reliability include individual item reliability (Bagozzi and Yi, 1988) as well as standard deviation and kurtosis (Dawes, 2008). A judgmental procedure to evaluate the reliability of an item is to evaluate its potential ambiguity of meaning (Puri, 1996). When it comes to reliability *among items*, inter-item correlations can help to assess statistically whether two items actually cover the same domain or, in the case that the items stem from different scales intended to measure different concepts, are conceptually too similar. Judgmentally, potential sources of ambiguity can be revealed by comparing the content between items. On the *construct level*, statistical indicators include internal consistency reliability, which can be evaluated via tau-equivalent reliability ρ_T (often referred to as "Cronbach's α"; Cronbach, 1951), congeneric reliability ρ_C (often referred to as "composite reliability"; Jöreskog, 1971), or average variance extracted (AVE; Fornell and Larcker, 1981). A judgmental criterion on the construct level is conceptual clarity (Podsakoff *et al.*, 2016). Reliability *among constructs* can be assessed statistically by comparing correlations (Kline, 2005). Judgmentally, pairwise comparisons of conceptual definitions could reveal potential sources for misinterpretations. Finally, reliability *between item and construct* can be assessed by calculating item–total correlations. As an example, an item can be deleted based on the judgmental criterion that a panel of judges does not rate it as either "clearly representative" or "somewhat representative" (Zaichkowsky, 1994).

The second aspect of the framework is validity. When a study requires that item data is normally-distributed, statistical indicators to assess validity on the *item level* include mean and skewness (Dawes, 2008), whereas a judgmental criterion is the readability of items (Richins, 2004). Inter-item correlations provide a statistical indication of validity *among items* (Bearden *et al.*, 2011). Judgmentally, it is possible to compare the phrasing of pairs of

items, both within and across constructs. On the *construct level*, criterion validity is a possible statistical criterion (e.g., Bagozzi, 1981). Judgmentally, assigned judges can be asked to develop a conceptual definition based on the formulation of a set of given items (Moore and Benbasat, 1991). Judges could also evaluate the range of the construct (Busse *et al.*, 2017). Statistical criteria for validity *among constructs* include convergent and discriminant validity. For example, a commonly used heuristic is the average variance extracted–shared variance (AVE–SE) comparison (Fornell and Larcker, 1981). An alternative heuristic, whose applicability has been demonstrated previously (Voorhees *et al.*, 2016), is the heterotrait–monotrait criterion (Henseler *et al.*, 2015). By analogy, judgmental comparisons of construct definitions could help to identify those constructs that are conceptually "too similar". Researchers can also utilize first-generation methods such as exploratory factor analysis (EFA) to identify substantial cross-loadings (Treiblmaier and Filzmoser, 2010) or second-generation methods such as confirmatory factor analyses (CFA; Anderson and Gerbing, 1988) to assess convergent and discriminant validity *between item and construct*. Judgmentally, Moore and Benbasat (1991) describe two related q-sort procedures that form appropriate counterparts to the statistical approaches.

Although scale development procedures frequently focus on reliability and validity criteria, it is sometimes reasonable to remove items that are both reliable and valid, if this can help to reduce the length of a questionnaire and thus increase response rates (Yammarino *et al.*, 1991). Therefore, adding a third consideration to our framework – parsimony – is warranted. The length of an item formulation is a common criterion for parsimony on the *item level*, which can be quantified by the number of characters or words or by identifying the number of morphemes, which requires the involvement of judges (Johnson, 2004). It is possible to interpret excessively high values of inter-item correlations as an indication of redundancy, that is, a lack of parsimony *among items*. Similarly, judges can use qualitative inter-item

comparisons to identify items that are simply based on synonyms (Rossiter, 2002) and recommend their elimination. On the *construct level*, the number of items per construct can be interpreted as a statistical indicator of a lack of parsimony (cf. Stanton *et al.*, 2002). Rossiter (2002) suggests judgmental criteria for calculating the number of items required per construct. Also, the researcher can evaluate the length of a conceptual definition. *Among constructs*, researchers can again identify redundancies (statistically and/or judgmentally) and thus eliminate redundant constructs. Finally, a comparison *between item and construct* can help to optimize a scale's parsimony. Voss *et al.* (2003, p. 313), for example, describe an iterative procedure that involves CFA fit indices and χ^2 difference testing, eliminating the item with the lowest item–total correlation at each iterative step until the adjusted goodness of fit index (AGFI) does not increase and/or the χ^2 difference test between the original CFA model and the CFA model of the reduced scale shows no significant difference. Although this procedure does not involve any judgmental criteria, these could easily be integrated into each iterative step. Frohlich (2002) describes a similar approach. Lawshe (1975) suggests judgmentally assessing whether the measurement made with an item is essential to capture a construct's meaning, an approach that could be integrated in each iterative step.

3. Analysis of SCM literature

In order to analyze the scale purification approaches used in SCM research, we screened SCM journals for the most recent empirical studies that employ scale purification. The study selection was restricted to the highest ranked journals related to SCM, determined from the impact factors as reported in the Thomson Reuters Journal Citation Report (JCR). The discipline's leading six journals based on the 2014 and 2015 JCR lists included: (1) *Journal of Operations Management* (JOM); (2) *Journal of Supply Chain Management* (JSCM); (3) *Supply Chain Management: An International Journal* (SCMIJ); (4) *Journal of Business*

Logistics (JBL); (5) *International Journal of Physical Distribution & Logistics Management* (IJPDLM); and (6) *Journal of Purchasing & Supply Management* (JPSM). We limited our review to the ten most recent issues in these six journals in order to ensure that we captured the current methodological practice of the SCM discipline. The full text versions of a total of 360 articles were subsequently reviewed independently by two authors of this article. All primary studies that made use of reflective latent constructs using multi-item scales solicited via a survey were included in the final sample, which resulted in a total of 77 studies (JOM [16], JSCM [16], SCMIJ [13], JBL [16], IJPDLM [5], JPSM [11]). These articles were then further analyzed based on the following criteria: (a) application of scale purification, (b) presentation of statistical and/or judgmental criteria (using the criteria presented in Table I), and (c) discussion of the implications of scale purification decisions.

4. Results

Our analysis of the 77 SCM articles reveals that the authors of 44 of the measurement-related articles did not delete any items or did not discuss scale purification. Of the other 33 studies, in six articles the authors pointed out that they were aware of the importance of scale purification but did not delete any items, whereas the authors of 27 articles reported that items were deleted and explained why this was done (i.e., they identified their statistical or judgmental insufficiency for the study), however, the potential impact of scale purification was not discussed any further. Hence, not a single study both deleted items and discussed the potential consequences of this decision on the study's representativeness of the construct, as well as any further theoretical implications.

In the 33 articles that deleted items, we identified a total of 38 reasons for scale purification (see Table II). Of the items deleted, 31 were deleted due to statistical reasons. Most often – in 23 cases – this was due to low item loadings, which was demonstrated, for example, with the

help of EFA. Fifteen items were eliminated because of excessively low factor loadings (based on EFA); three items were eliminated due to excessively high factor cross-loadings (based on EFA); four due to excessively low factor loadings shown by CFA; and one item loaded on a completely different construct (based on EFA). Only a minority of scale purification decisions was based on, or included, judgmental criteria – and these few decisions were often vaguely described.

Table II Criteria applied for scale purification decisions in SCM publications

Reliability	Statistical criteria	Judgmental criteria
Item level	1	
Among items		1
Construct level		
Among constructs	1	
Between item and construct	3	
Validity	**Statistical criteria**	**Judgmental criteria**
Item level	1	
Among items		
Construct level		
Among constructs		
Between item and construct	23	
Parsimony	**Statistical criteria**	**Judgmental criteria**
Item level		
Among items		1
Construct level		
Among constructs		
Between item and construct		1
Unclear purification decision	2	4

Note: These criteria for scale purification were used by the authors of 33 studies to justify the elimination of items. Each of these studies applied between one and four statistical and/or judgmental criteria.

Summing up, four potentially troublesome findings arise from this review of SCM studies. First, of the various reasons why items should or should not be eliminated from reflective construct measurements (see Table I), SCM researchers make use of only a small subset (see

Table II). Second, the most common reason is still related to statistical criteria, highly skewed towards establishing validity between items and construct. Despite their demonstrated usefulness in this methodological step, judgmental criteria are rarely applied. Third, only in a few articles statistical and judgmental criteria are combined. Finally, despite our focus on the leading SCM journals, we still found a high proportion of scale purification decisions – over 15 percent – that are not underpinned by statistical and/or judgmental arguments.

5. Discussion

Based on our findings, we will now present four suggestions with the intention of helping researchers in SCM and related disciplines to conduct and justify their scale purification decisions. Our results show that currently only a small subset of criteria is applied to justify the majority of scale purification decisions. This raises concerns as to whether the SCM discipline is aware of all potential criteria and their applicability. Therefore, our first suggestion (S_1) is that *researchers should systematically apply all available statistical and judgmental criteria (as listed in Table I) when assessing the reliability, validity and parsimony of scales*.

At present, the most common scale purification criteria used in measurement-related SCM literature are based on quantitative arguments (i.e., statistical criteria). Therefore, our second suggestion (S_2) is to *ensure that judgmental criteria are incorporated in future scale purification decisions*. Instead of basing scale purification decisions solely on statistical criteria, a judgmental assessment performed by domain experts ensures that a scale covers the entirety of all relevant aspects that need to be measured. This way, the statistical procedure described by Voss *et al.* (2003, p. 313) could be supplemented by judgmental criteria in each iteration (e.g., Lawshe, 1975). Judgmental criteria on which to build scale purification decisions are available, and have been used in previous research. For example, Moore and

Benbasat (1991) set out a number of criteria that SCM researchers have applied repeatedly (e.g., Wieland and Wallenburg, 2012; Rojo *et al.*, 2016). Although Moore and Benbasat's (1991) original approach is used to reduce and reword the initial body of items rather than asking judges to identify potential missing items to represent all aspects of a construct, such a step could easily be added to their procedure.

Our third suggestion (S_3) is to *ensure that judgmental and statistical criteria are combined before making a scale purification decision*. Judgmental and statistical criteria cannot be used interchangeably and therefore researchers need to ensure that both types of criteria have been taken into account to improve the psychometric properties of a scale. This might sometimes tempt researchers to keep some of the items for judgmental reasons, even if a statistical criterion is not fully met. For example, in spite of a relatively low item–total correlation, Cambra-Fierro and Polo-Redondo (2008) decided to maintain an item in a scale, as it was "regarded as relevant from a theoretical perspective" (p. 216). Alternatively, before using a scale in a survey questionnaire, researchers might first consider increasing validity by adding "fresh" items *after* several items have been removed based on statistical criteria – an approach we did not find in the SCM literature.

Our results further reveal that some SCM studies still lack transparency regarding the methodological reasons for scale purification. Making this clear for the readers is critical, as this allows them to evaluate the reliability, validity and parsimony of the scales and, subsequently, of the final results. Therefore, our final suggestion (S_4) is to *make scale purification decisions transparent for readers*. Although many authors indicate that they eliminate items, the justification for this procedure sometimes remains unclear. Furthermore, to avoid problems of attenuation due to measurement error (Boyd *et al.*, 2005), researchers should transparently discuss the expected impact the elimination of scale items has on the estimates of the structural model and/or the theoretical results. This might require researchers

to estimate the same model using both scales – before and after purification – and present a comparative discussion.

6. Conclusions

The assessment of whether items should be retained or eliminated ("scale purification") is a vital part of the process of measuring theoretical constructs in empirical research that employs latent variables. Failures in this step – either by removing "good" items or by not removing "bad" items – result in constructs that are not sufficiently represented by the corresponding set of items. Quite importantly, this might severely impact the conceptual foundation of any research project. Due to a lack of guidelines on this important methodological issue and its profound impact on knowledge development in SCM and related disciplines, in this study we propose a framework to guide scale purification decisions. Specifically, we screened current scale purification approaches in SCM research, and provided suggestions for researchers to improve scale purification in the future.

The proposed framework provides a novel overview allowing researchers to more fully understand the structure and methods available and make better-informed scale purification decisions. The framework includes the reliability, validity and parsimony of scales with regard to statistical and judgmental criteria. Applying this framework to the SCM discipline, it is surprising to find that the vast majority of scale purification decisions are merely based on statistical criteria. Judgmental criteria, which help to ensure that the measurement of a construct reflects the underlying conceptual definition, were scarcely applied. Hardly ever are judgmental and statistical criteria combined before making scale purification decisions. The dominant use of statistical criteria without a parallel use of judgmental criteria can substantially inhibit the methodological discussion in our field, as scale items are eliminated without qualitative arguments. It has long been established that constructs are the foundation

of theory (Bacharach, 1989; Suddaby, 2010) and it is therefore crucial that the way in which we measure them is not inhibited by faulty judgments regarding scale purification. In particular, addressing the methodological idiosyncrasies that we observed in the reviewed SCM studies, we developed suggestions that will help empirical researchers to assure and improve the descriptive, explanatory and predictive power of their results.

Our research results are limited, as we analyzed only SCM research, focusing in this study on a set of leading journals. Other disciplines, and other journals, might have different methodological traditions. Our suggestions should therefore be critically reflected upon before applying them "out of the scope" of our research. We leave it to future researchers to update and expand the proposed framework, and to apply it in other fields. Furthermore, our research did not evaluate potential differences in the importance of individual criteria for ensuring a scale's quality. For example, an alternative interpretation of the dominance of statistical criteria that relate to the relationship between item and construct could be that these criteria are comparatively more relevant. Future research should investigate this further and also develop strategies to handle situations where statistical and judgmental criteria lead to conflicting scale purification decisions.

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.

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

Bagozzi, R.P. (1981), "An examination of the validity of two models of attitude", *Multivariate Behavioral Research*, Vol. 16 No. 3, pp. 323–359.

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.

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.

Boyd, B.K., Gove, S. and Hitt, M.A. (2005), "Construct measurement in strategic management research: Illusion or reality?", *Strategic Management Journal*, Vol. 26 No. 3, pp. 239–257.

Bryman, A. and Bell, E. (2015), *Business Research Methods*, 4th ed., Oxford University Press, Oxford, UK.

Busse, C., Kach, A.P. and Wagner, S.M. (2017), "Boundary conditions: What they are, how to explore them, why we need them, and when to consider them", *Organizational Research Methods*, in press.

Cambra-Fierro, J.J. and Polo-Redondo, Y. (2008), "Creating satisfaction in the demand-supply chain: The buyers' perspective", *Supply Chain Management: An International Journal*, Vol. 13 No. 3, pp. 211–224.

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.

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. (2012), *Scale Development: Theory and Applications*, 3rd ed., SAGE Publications, Thousand Oaks, Calif.

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.

Guide, V. and Ketokivi, M. (2015), "Notes from the Editors: Redefining some methodological criteria for the journal", *Journal of Operations Management*, Vol. 37, pp. v–viii.

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.

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), "Statistical analysis of sets of congeneric tests", *Psychometrika*, Vol. 36 No. 2, pp. 109–133.

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

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

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.

Min, S. and Mentzer, J.T. (2004), "Developing and measuring supply chain management concepts", *Journal of Business Logistics*, Vol. 25 No. 1, pp. 63–99.

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.

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

Nevo, B. (1985), “Face validity revisited”, *Journal of Educational Measurement*, Vol. 22 No. 4, pp. 287–293.

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.

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

Richins, M.L. (2004), “The material values scale: Measurement properties and development of a short form”, *Journal of Consumer Research*, Vol. 31, pp. 209–219.

Rojo, A., Llorens-Montes, J. and Nieves Perez-Arostegui, M. (2016), “The impact of ambidexterity on supply chain flexibility fit”, *Supply Chain Management: An International Journal*, Vol. 21 No. 4, pp. 433–452.

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.

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.

Suddaby, R. (2010), “Editor’s comments: Construct clarity in theories of management and organization”, *Academy of Management Review*, Vol. 35 No. 3, pp. 346–357.

Treiblmaier, H. and Filzmoser, P. (2010), “Exploratory factor analysis revisited: How robust methods support the detection of hidden multivariate data structures in IS research”, *Information & Management*, Vol. 47 No. 4, pp. 197–207.

Voorhees, C., Brady, M., Calantone, R. and Ramirez, E. (2016), “Discriminant validity testing in marketing: An analysis, causes for concern, and proposed remedies”, *Journal of the Academy of Marketing Science*, Vol. 44 No. 1, pp. 119–134.

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. and Wallenburg, C.M. (2012), “Dealing with supply chain risks: Linking risk management practices and strategies to performance”, *International Journal of Physical Distribution & Logistics Management*, Vol. 42 No. 10, pp. 887–905.

Yammarino, F., Skinner, S. and Childers, T. (1991), “Understanding mail survey response behavior: A meta-analysis”, *The Public Opinion Quarterly*, Vol. 55 No. 4, pp. 613–639.

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