A Comprehensive Assessment of Measurement Equivalence in Operations Management

Abstract

This paper provides a comprehensive framework for treating equivalence both prior to data collection and during subsequent analyses, and assesses the extent to which equivalence is considered in survey research in six leading empirical Operations Management (OM) journals (*Decision Sciences*, *International Journal of Operations & Production Management*, *International Journal of Production Research*, *Journal of Operations Management*, *Management Science*, and *Production and Operations Management*). Measurement equivalence of latent variables in survey data is an important condition that should be met in order to meaningfully pool and/or compare data stemming from apparently heterogeneous sub-groups. We assess 465 survey articles from a six-year period from 2006 to 2011 and document these articles in relation to the four main stages of our comprehensive framework: identifying sources of heterogeneity; maximising equivalence prior to data collection; testing measurement equivalence after data collection; and dealing with partial and non-equivalence. We conclude that pooling of data from heterogeneous sub-groups is common practice in OM, but that awareness and testing of equivalence remains limited. Given these findings, we further elaborate the best practices detected in those few OM studies that do address equivalence in some way. We conclude that to improve the quality of OM survey research, authors, editors and reviewers should pay greater attention to equivalence, and we provide a pragmatic checklist of measurement equivalence issues across the four stages.

Keywords: Survey research, Measurement equivalence, Operations Management (OM).

1. Introduction

Over the past two decades, Operations Management (OM) has seen rapid growth in survey research (Malhotra and Sharma 2008; Shah and Goldstein 2006). For analyses from such research to have high power and to take advantage of techniques including structural equation modeling (SEM), the sample size required is often too large to make it practical for an individual researcher to undertake a study (Schmidt 1996). This has led to significant growth in collaborative multi-institution studies within and across countries, where researchers jointly develop a research model, survey instrument and design procedures, while each institution is responsible for a part of the data collection, sometimes adapting to local circumstances. Examples of such collaborative multi-institution studies include the International Manufacturing Strategy Survey (IMSS; e.g., da Silveira 2011), the High Performance Manufacturing group (HPM; e.g., Peng, Liu, and Heim 2011), the Global Manufacturing Research Group (GMRG; e.g., Kull and Wacker 2010), and the International Purchasing Survey (IPS; e.g., Karjalainen and Salmi 2013). It is common practice to subsequently pool the sub-groups of data to create a larger data set for statistical analyses.

Such collaborative studies are also driven from a desire by researchers to compare means and causal relationships across sub-groups of data from different settings defined for instance by firm characteristics, sectors, cultures and languages, data collection methods, or time periods. For example, do companies in the US have lower levels of trust in their suppliers than companies in Japan (Bensaou, Coyne, and Venkatraman 1999)? Is the impact of a firm's global operating strategy on its global supply chain structure less for small firms than for larger firms (Prater and Ghosh 2006)? Does the correlation between JIT purchasing and inventory turnover shift over time (Kaynak and Hartley 2006)?

However, a key consequence of collecting data from different sub-groups is that it is difficult for researchers to know if findings reflect 'true' similarities and differences between selected groups rather than the spurious effect of cognitive differences in the way questions are interpreted, reflected on, or answered (Mullen 1995). For example, in interpreting survey questions, lower level employees may have different degrees of familiarity with the concept of "competitive environment" than top-level employees (Koufteros and Marcoulides 2006). Equally, in responding to questions, the use of response scales in selecting extremes or neutral answers is partly determined by cultural norms (Saris 1988). Consequently, data from heterogeneous sub-groups of respondents should not be pooled and/or compared without first examining whether or not they are in fact equivalent (Cheung and Lau 2012). Rungtusanatham et al. (2008) show that ignoring equivalence issues may lead to conclusions that are ambiguous at best and erroneous at worst.

Within OM, there have been a number of studies highlighting the need to improve the quality of survey research, shifting the focus from encouraging the development of reliable and valid measures (Filippini 1997; Hensley 1999) to more sophisticated empirical refinement, with respect to analysis, data triangulation, sampling frames, missing data, and response bias (Malhotra and Grover 1998; Rungtusanatham et al. 2003; Tsikriktsis 2005). However, to our knowledge, there are no studies that assess the extent to which equivalence issues are considered in OM research with latent variables. Given the growing awareness of the negative consequences of non-equivalence in other fields; the complex trade-offs between equivalence maximization and pragmatic considerations in survey design; and the opportunities for more advanced equivalence testing through SEM, an assessment of equivalence application in OM appears to be opportune. To this end, we look to build on previous work in order to further promote awareness of the importance of measurement equivalence among OM scholars undertaking survey research.

Our paper builds on a number of important reference studies by making two key contributions for those wishing to gain a more comprehensive understanding of equivalence issues in OM. These contributions, and how they differ from key extant work, are illustrated in Table 1. First, this is the only study to undertake a systematic assessment of the extent to which OM scholars publishing in our leading discipline journals consider equivalence. Furthermore, this assessment examines equivalence using a comprehensive framework of the different stages of survey research: (1) identification of possible sources of heterogeneity among respondent groups that may threaten equivalence; (2) maximizing equivalence during the design stage of survey research; (3) testing equivalence post data collection and prior to pooling of data; and (4) dealing with partial- and nonequivalence. As such, it is the only study to provide detailed discussion of *all* phases as they relate to issues of equivalence. Second, we identify best practice and pragmatic decision issues across all four stages of our equivalence framework both within extant OM literature and more broadly within disciplines where equivalence issues are arguably more established. As such, our study provides a more complete and interrelated list of guidelines of how OM scholars can identify threats to equivalence, maximize equivalence during design, test for measurement equivalence during analysis, and deal with partial or non-equivalence in future research.

[Insert Table 1 about here]

2. Overview of Measurement Equivalence

Before assessing the extent to which various issues relating to equivalence are considered in empirical survey research within OM, we provide a brief overview of equivalence. A measurement procedure is equivalent if it produces measurements of a variable X with identical measurement properties in two or more groups that differ with respect to an attribute such as nationality, firm size, or respondent position, for example. In other words, data are equivalent when individuals in different sub-groups have similar response functions (Drasgow 1984). Conversely, if individuals across sub-groups interpret and/or respond to survey questions in systematically different ways, comparison and/or pooling of data is compromised. For instance, if US firms, on average, have a more liberal definition of lean manufacturing programs than their Japanese counterparts, this difference in perception would confound any true difference in the existence or impact of such programs across the two countries. Hence, it is not the differences between sub-groups that create problems for researchers, rather the uncontrolled effect of possible systematic differences in the way in which individuals within each group interpret and respond to the same set of questions.

Equivalence, therefore, is by no means a given. Answers to survey questions involve several stages of cognitive processing by respondents, with questions interpreted, experiences reflected upon, and answers subsequently selected (Podsakoff, MacKenzie, and Podsakoff 2003, 886). Each of these processes may be influenced by particular frames of reference (Vandenberg and Lance 2000) and therefore threaten equivalence across sub-groups. Assessments of equivalence have revealed significant flaws in survey research that pools data from different sub-groups (Bensaou et al. 1999; De Jong, Steenkamp and Fox 2007; Hult et al. 2008; Nye and Drasgow 2011; Rungtusanatham et al. 2008). For example, Rungtusanatham et al. (2008) show how failure to account for equivalence – in this case between data collected from top management and middle management echelons – can lead to conclusions that are the polar opposite of those when equivalence is considered.

Within the literature, we have identified seven key reference studies that provide guidelines for dealing with equivalence at one or more stages of the survey process (Bensaou et al. 1999; Cheung and Rensvold 1999; Douglas and Craig 2006; Hult et al. 2008; Steenkamp and Baumgartner 1998; Rungtusanatham et al. 2008; Vandenberg and Lance 2000). Whilst these papers provide useful insights on various issues concerning equivalence, the guidelines are somewhat fragmented and predominantly focus on testing for equivalence *after* data collection. Increasingly, however, it is clear that researchers should go beyond the testing of equivalence (Douglas and Craig 2006) to include actions at *all* stages of survey research, as illustrated in Figure 1. These four stages form the structure of our assessment of equivalence in OM research.

[Insert Figure 1 about here]

3. An Assessment of Equivalence in Published OM Research

Our assessment focuses on determining the extent to which various issues relating to equivalence are considered in empirical survey research published in leading OM journals. In this section, we detail the selection criteria for journals and articles, before discussing our approach to article coding.

3.1. Journal Selection

We considered all OM journals that are recognised as publishing high quality empirical research. We selected (in alphabetical order) *Decision Sciences* (*DS*), *International Journal of Operations & Production Management* (*IJOPM*), *International Journal of Production Research* (*IJPR*), *Journal of Operations Management* (*JOM*), *Management Science* (*MS*), and *Production and Operations Management* (*POM*) because they have been consistently ranked in the top echelon of OM journals in terms of quality and relevance (Barman, Hanna, and LaForge 2001; Goh et al. 1997; Rungtusanatham et al. 2003; Soterriou et al. 1999) .

3.2 Time Horizon and Article Selection

We decided to focus on the six-year period from 2006 to 2011, given that the first papers on the issue of equivalence have only emerged very recently in the OM community (Rungtusanatham et al. 2008). In addition, our initial assumption was that consideration of equivalence is not yet common practice in the field and that in some cases where equivalence is considered, it may not have been explicitly described as such. Therefore, rather than using specific search terms for article selection, we manually selected all survey articles in the six selected journals and did not exclude articles based on their research topic. Whilst this approach was relatively time-consuming, it improved coverage compared with a keyword search. Starting from an initial 4,006 research articles published in *DS*, *IJOPM*, *IJPR*, *JOM, MS,* and *POM* between 2006 and 2011, we identified 465 studies that satisfied our selection criteria. Of these, the majority came from *JOM* (144), *IJOPM* (131*), IJPR* (83)*,* and *DS* (60). Overall, fewer empirical papers are published in the other two journals; hence the relatively low number of survey articles in *MS* (29) and *POM* (18).

3.3. Coding the Articles

For each of the 465 survey-based articles, we noted basic descriptives such as title, authors, volume and issue, population and sample characteristics, unit of analysis, number and type of respondents, and the type of data analysis methods used. Three independent raters (members of the author team) then coded the articles based on the key issues concerning equivalence. First, we coded articles regarding four key sources of heterogeneity emphasised in literature – different countries/languages/cultures; different types of respondents; different data collection methods; different time periods of data collection. In addition, any other sources of heterogeneity indicated in the studies, such as industry, firm size, performance level, were coded. Second, we coded the extent to which equivalence is designed and tested for. For all of these aspects, except for the additional sources of heterogeneity, we used binary codes (present/absent). For the additional sources of heterogeneity, we adopted text coding. Finally, for a short list of papers that actually performed equivalence tests, we used open coding in order to capture best practices related to dealing with partial and non-equivalence.

Two pilot studies were conducted to maximize inter-rater agreement on the coding. First, all three raters coded the survey articles published in the 2006 issues of JOM (volume 24). This resulted in an acceptable but not optimal inter-rater agreement, with 80% of papers consistently classified by all three raters (percentage method; Boyer and Verma 2000). In a subsequent meeting with the three coders and the other co-authors, we discussed points of confusion or disagreement in the coding process. A second pilot was then carried out in which all three raters coded the survey articles published in the 2008 issues of JOM (volume 26). The inter-rater agreement in this round was over 90% (96% if only considering the crucial classifications (1) sources of heterogeneity and (2) presence and type of equivalence tests). Based on this high inter-rater agreement, the remaining articles from all six journals were assigned to and coded by one individual rater. Any remaining doubts were adjudicated on a case-by-case basis in a discussion between the three raters. Moreover, after finalising the coding, random checks were completed by two other members of the author team, leading only to minor changes.

4. Critical Issues in the Consideration of Equivalence for OM Research

Our discussion of how OM scholars address equivalence issues is structured into the four stages of survey research. Clearly, these stages are inter-related and decisions made prior to data collection have an impact on subsequent issues that are likely to emerge during analysis. Per stage, we will first provide a short description based on the broader management literature including the research methods literature, before discussing how the selected OM literature has addressed equivalence in that stage.

4.1. Detecting Sources of Heterogeneity – the Identification Stage

Our point of departure in considering equivalence is the identification of possible sources of heterogeneity between groups of survey respondents, and hence an increased likelihood of violated assumptions regarding identical response functions of respondents. Asking oneself: "Do respondents understand the question in the same way?" and "Do respondents express themselves in the same way?" helps identifying such sources (Saris and Gallhofer 2007). Whilst not an exhaustive list, four key sources of heterogeneity are most commonly emphasized within the broader literature: different countries/languages/cultures; different types of respondents; different data collection methods; and different time periods of data collection. First, different countries, cultures or languages may interpret questions and use scales in fundamentally different ways (Hult et al. 2008; Van Herk, Poortinga, and Verhallen 2004). For example, some cultures in Latin America tend to use extreme end points of a scale, whereas Asian cultures tend to favour neutral middle points of scales (Saris 1988). Likewise, different countries will have different degrees of familiarity with certain OM terminology (Rungtusanatham et al. 2005).

Importantly, pooling data from groups from different regions within the *same* country could generate similar or even higher threats to equivalence as compared to cross-country research. For example, pooling data from French-speaking and English-speaking Canada may carry significant threats to equivalence, as the regional cultures of these two areas are significantly different (Cannon et al. 2010). In other words, these two groups may understand survey questions differently and use the scales in different ways. Conversely, using Hofstede's (2001) cultural dimensions theory, we see that typical Swedish and Norwegian respondents tend to score similarly on Hofstede's cultural dimension. As such, the threats to equivalence posed by pooling data from these two countries may be relatively low.

Second, sub-groups defined by respondents that have transparently different demographics may constitute a threat to equivalence (Rungtusanatham et al. 2008). In social psychology, examples often refer to differences due to race, gender, or age (Nye and Drasgow 2011). Other examples could include sub-groups of senior managers versus lower level employees (Koufteros and Marcoulides 2006) and selling versus supplying functions (Hult et al. 2000). There are risks in simply pooling data from such sub-groups because their response functions (as affected by different frames of reference) may be fundamentally different.

Third, a threat to equivalence arises when data are collected from different sub-groups in different ways. Data collection may vary in a number of ways, for instance due to differences in the availability of resources to access different sub-groups (Kish 1994). The availability of different sampling frames may lead to different sampling errors in different groups. For example, there may be systematic differences in response functions between a group of respondents drawn from an alumni database and a group drawn from a professional body membership list. Data collection procedures may also vary in terms of the way respondents will be contacted, the administration mode of the survey (e.g. telephone, mail, web-based), and coverage comparability (Hult et al. 2008). Finally, and undesirably, sampling methods may vary (e.g. random versus convenience sampling).

Fourth, data equivalence may be threatened by different timings of data collection for different sub-groups. The meaning of concepts and the way their scale is used may change over the years (Ariely and Davidov 2011). Take for example a purchasing manager responding to questions about supply risk before the start of the global economic crisis in 2008 and another one responding right in the middle of the crisis.

Table 2 presents a summary of the possible sources of heterogeneity identified within the selected articles. We used information from the (at times very short) sections on data collection and information about author affiliation to deduce from which country/ies data were collected. Of the total sample of 465 survey articles, 88 studies (18.9%) pooled data from two or more countries/languages. It is striking that 59 survey articles did not report explicitly from which country/ies the data were collected.

[Insert Table 2 about here]

Of the total sample of 465 survey studies, 221 studies (47.5%) pooled data from two or more respondent types, referring to different functional areas and/or positions/levels in these functions. If we consider papers that did not explicitly state the number of respondent types, but where it is clear that there were respondents from different types of functional areas, we need to add another 177 studies to the list. Of the total sample of 465 surveys, 101 studies (21.7%) had data that came from more than one method of data collection and, 15 studies (3.2%) had data that came from multiple data collection time periods. In sum, of the total sample of 465 survey studies, 293 (63%) were found to have at least one of the four key types of potential data heterogeneity, and 104 (22.4%) had at least two forms of potential heterogeneity. The latter group of papers clearly represents a particularly high risk of non-equivalence when sub-group data is pooled for data analysis. Overall, the conclusion is that within OM survey research, potential data heterogeneity, which poses a substantial threat to data equivalence, is widespread.

In addition, nearly all studies had one or more other sources of potential heterogeneity besides the four main sources we have discussed. These sources include pooling data from different industries and different organizational sizes. Moreover, studies with a contingency approach identified potential moderating variables, such as degree of technology turbulence, degrees of uncertainty in the business environment, and ownership structures. The same contingencies or control variables might point to non-equivalent measurement models (Koufteros, Vonderembse, and Jayaram 2005). The studies differ in how far they considered measurement model equivalence prior to considering path model equivalence. We will come back to this point later.

Many of the potential sources of heterogeneity do not necessarily threaten measurement equivalence. Therefore, it is the responsibility of the researcher(s) to identify any sources of heterogeneity that may threaten equivalence and make an explicit decision: the researcher may decide to avoid the source of heterogeneity (for instance, gather data from one rather than multiple demographic groups). However, when this is not desired (for instance, when the research aim is to compare groups) or practically not possible, the researcher should subsequently consider ways to (1) minimise these threats prior to data collection and/or (2) test for equivalence across groups after data have been collected. We now turn our attention to the first of these two elements.

4.2. Maximising Equivalence Prior to Data Collection – the Design Stage

During the design stage of the survey, researchers should actively try to maximise equivalence across the previously identified heterogeneous sub-groups. There are a number of different ways that researchers may maximise equivalence prior to data collection. Here, we focus on maximising construct equivalence, translation equivalence, and data collection equivalence, as explained below. The extent to which these issues are considered in the assessed OM articles with at least one source of heterogeneity is presented in Table 3. Not every study with one or several potential sources of data heterogeneity should have addressed each of these design issues, but the low number of papers addressing any of these issues (with the exception of translation equivalence) strongly suggests that equivalence is not being considered sufficiently during the design of a survey.

[Insert Table 3 about here]

4.2.1. Maximising Construct Equivalence

The relevance and meaning of concepts, especially those related to attitudes or behaviours, may differ across groups leading to differences in cognitive processing of survey indicators (Douglas and Craig 2006). Construct equivalence relates to whether an object, concept, or behaviour is the same (i.e., serves the same purpose and achieves the same salience) across different groups, and can be evaluated pre-data collection in relation to three aspects: conceptual, category, and functional equivalence (Craig and Douglas 2000; Hult et al. 2008).

Conceptual equivalence is the extent to which individuals across different groups interpret and express a given object, concept or behaviour in the same way. In other words, it is the extent to which the domains of the concept/behaviour are the same across groups (Hult et al. 2008). For example, Kaynak and Hartley (2006) evaluate whether the domain of the just-in-time purchasing construct has remained the same over time. Another example refers to the domain of trust: trust in a salesperson may be a function of the seller´s company reputation or creditworthiness in China, but it may be a function of the seller´s individual expertise and product knowledge in the US (Douglas and Craig 2006).

Category equivalence is the extent to which the same classification scheme can be used for a given concept across different groups. For example, are the meanings of job categories consistent across groups (Bensaou et al. 1999)? Another example is related to marketing research and how products are assigned to categories: beer may be an alcoholic beverage in some countries, but a soft drink in others (Craig and Douglas 2000).

Functional equivalence is the extent to which a given object, concept or behaviour has the same *role* or *function* across different groups. For example, a bicycle serves a different function in China (predominantly a means of transport) than in the USA (predominantly a means of recreation). Similarly, the concepts of 'asset specificity' and 'reciprocal investments' may serve different functions in the Japanese versus the American auto industry context. Therefore, Bensaou et al. (1999) performed exploratory fieldwork in both countries to reveal potential differences in the functioning of these concepts. No differences were found, which the authors related to the globalization of the auto industry and its associated best practices.

Of the 293 OM survey articles with at least one of the four key forms of potential heterogeneity, only three articles reported on specific design steps for *construct equivalence* across groups, and of these some were not very explicit. In one of these three articles, Tan (2006) describes checking the consistency of indicators across respondent groups and across data collection formats in the pre-test. Although similar discussions regarding pre-tests were evident in some papers, none discussed this issue in relation to sub-groups within the sample. In other words, maximising construct equivalence across sub-groups was almost universally ignored within the articles assessed.

The lack of consideration for construct equivalence that we found in OM research ties into a general concern that critical evaluation of construct equivalence prior to data collection remains the exception rather than a rule in multi-group studies (Hult et al. 2008). Douglas and Craig (2006) state, "*There is often a tendency, particularly in replication studies, to adopt research instruments used in the original, or base, study, appropriately translated when necessary into the language of the other research context. If the instrument "works" and exhibits acceptable levels of internal reliability, it is considered to provide an adequate measure. Typically, little attention is given to its appropriateness in another setting or to whether it covers all aspects of the construct to be measured*" (p. 10). Whilst construct equivalence is partly maximised through the application of established practices of good survey research, such application should be at the level of each predefined group of respondents/informants. Thus, it is vital to examine literature that identifies similar groups; to adopt validated survey instruments used earlier for the same groups; and, to conduct qualitative fieldwork such as interviews, focus groups, and pre-tests in each sub-group.

4.2.2. Maximising Translation Equivalence

After maximizing construct equivalence, *translation equivalence* becomes important. Translation equivalence is demonstrated when the content and the format of a survey have been correctly translated to ensure that they tap into the same concept and that they provide the same response stimuli, in different groups. Although good translation does not assure the success of a survey, a bad translation ensures that an otherwise good project fails because of non-equivalent data across different language groups (Harkness, van de Vijver, and Mohler 2003).

The issue of *translation equivalence* is most applicable to OM researchers undertaking cross-country work. Of the 88 survey studies identified as having collected data from multiple countries/languages, 38 addressed translation equivalence[1]. We also found examples of "translations" within one language in order to reflect regional differences (e.g., Kull and Wacker 2010).

Until recently, translation/back-translation was considered as the most appropriate method for translating source questionnaires. An alternative approach is a team approach to translation that involves different roles (two independent translators, one reviewer, and one adjudicator) (Harkness et al. 2003). In this approach, the independent translators develop in parallel local versions of the same source questionnaire. At a reconciliation meeting, the translators and the reviewer go through the entire questionnaire discussing versions, errors of meaning, and agreeing on a final version. The adjudicator, who has the broadest set of capabilities related to translation and content, has the final vote in case of disagreement. This so called TRAPD (Translation, Review, Adjudication, Pretesting, Documentation) approach is increasingly seen as both theoretically and practically superior to the traditional translation/back-translation method (Douglas and Craig 2006; Harkness et al. 2003).

[1] Please note that 22 of these papers did not explicitly state that they addressed translation equivalence, but these papers use data collected through multi-country research collaborations (IMSS, HPM, and GMRG) for which translation equivalence actions have been described in other references.

Moving beyond translation of *content*, which remains the dominant area of focus for translation equivalence, are issues concerning translation of *form*. Researchers make many decisions – consciously or unconsciously – when specifying survey indicators. Arguably, the decisions made for response scales (i.e. the *form*) are equally if not more prone to different interpretations by the translators than the decisions made around the *content* of translated material (Saris and Gallhofer 2007). Therefore, it is important to work with a checklist of key choices regarding the *form* in order to maximise coherence between the source scale and the translated scales. For example, the source scale and the translated scales should be coherent regarding: symmetry of the labels; agreement between the unipolar or bipolar nature of the concept and the scale; the use or avoidance of a neutral or middle category; the use of "don´t know" options; the avoidance of vague quantifiers or numeric categories; the use of reference points; the use of fixed reference points; and, the measurement level (Saris and Gallhofer 2007: 119).

4.2.3. Maximising Data Collection Equivalence

Four key facets of data collection have to be considered: sampling frame comparability; administration mode equivalence; coverage comparability; and sampling method comparability (Hult et al. 2008). Within the OM articles assessed, we identified 101 survey studies with data collection method heterogeneity. However, only five articles explicitly discussed ways in which they considered or controlled for data collection equivalence. Thus, the explicit attention of OM survey researchers to data collection equivalence issues appears to be extremely limited.

In some circumstances, the different resources available across sub-groups may require flexibility in sampling and data collection approaches (Kish 1994). In such studies, maximising data collection equivalence during design may be impractical, so researchers will proceed to testing for equivalence across the different groups post data collection (See section 4.3 below). However, in other studies, it is possible for researchers to try to maximise data collection equivalence during the design stage in terms of the four described facets.

4.3. Testing Measurement Equivalence after Data Collection – the Analysis Stage

Having examined issues relating to equivalence in the design stage of assessed OM surveys, we now turn to the analysis stage. Here the focus is on the extent to which testing for equivalence among sub-groups occurs within the discipline. Whilst it is clear that equivalence can be and should be maximised in a number of ways prior to data collection, it is clear that avoiding all threats to equivalence is not always feasible or pragmatic (Rungtusanatham et al. 2008). Therefore, a variety of methods have been developed for examining measurement equivalence. Initially, studies suggested the use of *t*-tests, analysis of variance (ANOVA), multiple analysis of variance (MANOVA), exploratory factor analysis (EFA), or other statistical tests of observed score differences across groups (Nye and Drasgow 2011). Rungtusanatham et al. (2005) for instance, relied on visual inspection of the similarity of factor configurations, Cronbach's alphas, and regressors, as well as comparison of means by MANOVA.

The dominant approach since the early 1990's is multi-group confirmatory factor analysis (MGCFA) (Byrne, Shavelson, and Muthén 1989; Cheung and Rensvold 1999; Vandenberg and Lance 2000). Confirmatory Factor Analytic (CFA) models generally refer to concepts that are perceptually based, comprised of multiple manifest indicators, and which are reflective (Vandenberg and Lance 2000). MGCFA forms part of SEM and focuses on assessing the similarity of response functions specified through CFA models. As an illustrative example, Figure 2 shows the difference in response functions for two respondents. Respondent A's actual response always equals his/her true opinion. Conversely, despite having an extremely negative opinion, respondent B will opt for a less extreme response (intercept > 0), and in this case actual response increases less than 1 point for every 1 point increase of his/her opinion (slope < 1).

[Insert Figure 2 about here]

A baseline step of every survey analysis is the test of unidimensionality, validity, and reliability for each of the selected concepts. Within the MGCFA procedure, these analyses may be performed per group as part of the configural equivalence test described hereafter, or additionally, as a preparatory step, to the entire data set collectively (Bensaou et al. 1999; Koufteros and Marcoulides 2006).

Measurement equivalence can be expressed on a continuum, with many steps ranging from non-restrictive to very restrictive (Bollen 1989), but is most commonly tested in three steps: the configural, metric, and scalar equivalence tests (De Jong et al. 2007; Nye and Drasgow 2011; Meredith 1993). Focussing on these three tests avoids getting lost in "the bewildering array of types of measurement invariance that can be found in the literature" (Steenkamp and Baumgartner 1998: p.79). First, the weakest constraint refers to *configural* equivalence. Configural equivalence implies that in the different groups, the same measurement model fits the data. It is established when indicators load significantly on the same factors across groups (also called same-form equivalence, Bensaou et al. 1999) and the correlations between the latent variables are significantly less than one, guaranteeing discriminant validity (Steenkamp and Baumgartner 1998). Configural equivalence alone is not enough to proceed with substantial analysis. Second, metric equivalence should be tested. Metric equivalence implies that the concept can be used in comparisons of relationships across groups. This is established when the factor loadings (the slopes of the response functions) across the different groups are not statistically different (also called factorial equivalence, Bensaou et al. 1999). Third, the most severe constraint refers to scalar equivalence. Scalar equivalence implies that *means* for the concepts of interest can be compared across groups. This is established when slopes and intercepts of the response functions are not statistically different across groups. Respondents A and B from Figure 2 thus do not pass the metric and scalar equivalence test for the evaluated indicator, therefore the observed data relative to this indicator should not be compared or pooled.

The three steps are most commonly executed in a bottom-to-top approach, starting with the weakest and finishing with the most severe constraint. In the different steps, the standard fit indices (χ^2, χ^2/DF, RMSEA, NFI, CFI, SRMR) may be assessed to establish the quality of the overall model and the change in quality from one step to another (Hu and Bentler 1998). Just as for any SEM model, it is good practice to complement the standard fit indices with a procedure that iterates between the test of misspecifications and subsequent partial – theoretically justified - modifications of the model evaluating the change in parameter values (expected parameter change, EPC) and improvement of fit (modification index, MI) (Saris, Satorra, and Van der Veld 2009).

The OM papers we assessed rarely test for measurement equivalence. Of the 293 survey articles that had at least one source of data heterogeneity, only 17 tested for at least one type of measurement equivalence. Of the 17 papers, six papers performed only configural equivalence tests (using MGCFA or visual inspection of factor structures). We did not code *t*-tests and ANOVA as evidence of a configural equivalence test, because these tests evaluate individual indicators and not consistency of factor structures across sub-groups. A positive outcome of the configural equivalence test still does not indicate that researchers may use the data for subsequent statistical data manipulation. In line with this latter argument, eight other papers performed metric tests on top of the configural tests (Boyer and Hult 2006; Kaynak and Hartley 2006; Bou-

Llusar et al. 2009; Hult et al. 2006; Kaufmann and Carter 2006; Cannon et al. 2010; Birou, Germain, and Christensen 2011; Peng, Liu, and Heim 2011). These papers specifically aimed to contrast path models, and it is not necessary to proceed to the scalar equivalence test. The three final papers performed scalar tests on top of the metric and configural tests (Allred et al. 2011; Nyaga, Whipple, and Lynch 2010; Bagozzi and Dholakia 2006). These papers aimed to pool data or compare means in addition to comparison of causal relationships.

Table 4 shows the same 17 studies but now in terms of the source of heterogeneity identified. Two studies (Boyer and Hult 2006; Hult et al. 2006) appear twice in the table; they have performed equivalence tests in relation to two different sources of heterogeneity. There are other papers that also identify multiple potential sources of heterogeneity, but they perform only equivalence tests related to one criterion. Xu et al. (2010), for example, gather data before and after a critical event (the introduction of 3G technology) and point out that understanding of key concepts of the study will be different before and after that moment. Nonetheless, they only test for equivalence across different respondent profiles.

[Insert Table 4 about here]

Two studies adopted measurement equivalence tests due to cross-country data. Cannon et al. (2010) examined buyer-supplier relationships in the United States, Canada and Mexico. They clearly related the type of required equivalence (i.e. metric) with their research aim (i.e. compare causal relationships across countries). In a similar vein, Kaufmann and Carter (2006) compared path analyses related to international supply relationships in Germany versus the USA, after establishing metric equivalence.

Five studies performed measurement equivalence tests because they involved different respondent profiles. Johnston and Kristal (2008) and Nyaga et al. (2010) compared causal relationships across different functions within the supply chain (buyers versus suppliers). In this regard, the former study established configural equivalence, although this is not sufficient for proceeding to path analyses, and the latter study established metric equivalence. Bagozzi and Dholakia (2006) identified “level of experience” as a moderator of the relationship between Linux user groups’ social influence and its impact on the user’s participation. They went beyond the required metric equivalence test and established scalar equivalence before evaluating the moderating impact. Xu et al. (2010) tested configural equivalence across three different types of consumers in terms of the use of mobile platforms. Configural equivalence is not sufficient however, in order to proceed with path analyses. Finally, Wouters et al. (2009) hypothesized path differences between project leaders and cost analysts and conducted MGCFA, through which they (implicitly) established configural equivalence. However, we posit that also metric equivalence needs to be established before comparing paths across the two groups.

Three studies performed equivalence tests because of multiple data collection approaches. Hult et al. (2006) gathered data from two sampling frames (Council of Supply Chain Management (CSCMP) and Institute of Supply Management (ISM)) and established metric equivalence for 44 out of 58 indicators. They dropped the non-equivalent indicators from further analysis. Boyer and Hult (2006) varied both the sample selection (stratified versus random) as well as the administration mode (web-based versus paper-and-pencil), and tested metric equivalence before pooling the data. Allred et al. (2011) pooled data gathered through different sampling frames, after establishing scalar equivalence.

Two studies performed equivalence tests because of multiple time periods of data gathering. Kaynak and Hartley (2006) gathered data on JIT purchasing in 1995 and 2000 in order to evaluate potential shifts in validity, reliability, unidimensionality, and equivalence of factor structure. Peng et al. (2011) collected data in two waves, and multiple countries, and established metric equivalence of the sub-groups defined by the two time periods of data collection.

Finally, seven studies performed equivalence tests because of multiple company profiles. Configural equivalence was tested for sub-groups based on firm size (Prater and Gosh 2006); ISO-certification (yes/no) (Dowlatshahi 2011); and industry (Zhang, Viswanathan, and Henke 2011). But again, configural equivalence alone is not sufficient however, in order to proceed with path analyses. Metric equivalence was tested across groups from manufacturing and service industries (Bou-Llusar et al. 2009), and groups based on strategy types (Boyer and Hult 2006; Hult et al. 2006). Finally, scalar equivalence was tested by Birou et al. (2011) for Make-to-Order versus Make-to-Stock companies. These studies suggest that based on the research aim and setting, there can be specific sources of heterogeneity (in this case, differences in company profiles) that go beyond the key sources identified by extant literature. It is good practice if researchers provide reasoning as to why these sources might or might not create differences in response functions of different sub-groups for the particular concepts at hand.

4.4. Dealing with Partial and Non-equivalence – the Fine-tuning Stage

So far, our discussion of equivalence testing has in some sense implied a false dichotomy in relation to equivalence testing. The tests described are omnibus tests which establish whether data are equivalent or not at a certain step in the test sequence. The impact of the outcome of these tests is relatively straightforward: configural equivalence alone does not permit further use of the concept across groups; metric equivalence permits the use of the indicator in comparisons of relationships; and scalar equivalence permits the use of the indicator in comparisons of means. However, the impact of *non-equivalence*, which in practice is likely to occur, is less clear-cut. Therefore, we now consider how to fine-tune non-equivalent test outcomes.

Overall, the studies that we assessed present limited evidence of how researchers have dealt with non-equivalence. There were two exceptions. Hult et al. (2006) dropped the non-equivalent indicators (14 out of a total of 58 indicators). Albadvi, Keramati, and Razmi (2007), on the other hand, evaluated the number of non-equivalent indicators versus the total number of indicators (6 versus 89) and concluded that the non-equivalent portion was small. Consequently, they retained the non-equivalent indicators. The latter study was not included in Tables 2 and 3, however, given that it used ANOVA to test for configural equivalence – an approach we do not recommend.

Within extant literature, there are three potential actions in case of non-equivalence: (1) assess to what extent partial equivalence exists and execute substantive analyses that are acceptable with partially equivalent data, (2) make sense of non-equivalence, or (3) exclude the non-equivalent group(s) from substantive analyses. First, the logic behind the partial equivalence test is relaxing equivalence constraints where they do not hold. In other words, a parameter that is not equivalent across groups can be estimated for each group separately, increasing the probability that equivalence holds for the reduced set of indicators. A condition for evaluation of partial equivalence is that factors are configurally equivalent, so the problem first emerges when metric or scalar equivalence is imposed on the model (Byrne et al., 1989). When a researcher feels that there is no justification for partial equivalence, a second course of action is to make sense of non-equivalence. The differences between the groups can then produce relevant insights into the nature of the concept of which the indicator is reflective (Cheung and Rensvold 1999; Smith 2002). A third option is to exclude non-equivalent data from further substantive analysis, which may be the least desirable bearing in mind the resources invested in data gathering.

4.5. Summary of how Equivalence is Considered in OM Studies

Our assessment indicates that while many survey studies in the OM literature have at least one key source of heterogeneity in the data, the explicit attention to maximisation of equivalence in the design stage and testing for equivalence in the analysis stage is minimal (except some consideration of translation equivalence in the design stage). Even in cases where some steps have been taken to test for data equivalence, studies rarely describe how and why these tests have been carried out. Moreover, we have not identified a single OM study that was exemplary across all four stages. We did come across studies, however, that provide good, or even best practices related to one or two stages of our proposed model. In the following, we will summarize these practices per stage of the model. These insights from the OM literature complement the initial model we have developed based on the broader literature.

4.5.1. The Identification Stage

Our assessment indicates that when authors identify a threat to equivalence in OM research, they mostly do so because of heterogeneity in company characteristics, as highlighted in Table 4 (e.g., firm size, industry, position in the supply chain, performance level, customer-order-decoupling-point). This finding deviates from other bodies of literature that emphasise key threats emerging from heterogeneity in countries/cultures/languages, and, to a minor extent, from individual demographics, data collection methods, and time periods of data collection. We also find that when OM research discusses differences in individual demographics, this is generally limited to functional differences (for example, manufacturing versus purchasing managers, or general managers versus operations managers). This complements other bodies of literature highlighting differences in race, gender, or age (Nye and Drasgow 2011). Pagell, Krumwiede, and Shey (2007) provide a good example of identifying such a threat to equivalence before combining data from two sub-groups defined by functional differences. They argue that two transparently different sets of respondents, in this case manufacturing managers and purchasing managers, may perceive the competitive environment and plant performance differently (2011).

4.5.2. The Design Stage

In the design stage, equivalence regarding constructs, translation and data collection should be maximized.

Regarding construct equivalence – subdivided into conceptual, category, and functional equivalence - we identified three studies that considered the issue. Tan and Vonderembse (2006) evaluate the ratings on computer-aided design use items for consistency between respondents (user and manager in each company) during the pre-test stage. In their study of UK pharmaceutical supply chains, Cullen and Taylor (2009) use focus groups and pilot tests to refine their questionnaire with a particular focus on ensuring relevance and interpretability of key constructs and items by different respondent types (purchasers and sellers). Wouters et al. (2009) hypothesize ex ante that the two sub-groups in their sample, project leaders and cost analysts, will have different functional perspectives. Consequently, distinct pre-test samples of project leaders and cost analysts were used in the design stage. The study of Melnyk et al. (2009) was not included in the coding, because this study is not a survey article and the detected sub-groups are not present in their final dataset, but is interesting to add given their care for construct equivalence. They highlight that researchers and executives may have potentially different perceptions of importance and framing of major supply chain issues. Consequently, they employed the Delphi technique with both academic and practitioner respondents to maximize construct equivalence across these two groups. Overall, we observe that OM studies focus largely on the conceptual dimension of construct equivalence.

Regarding translation equivalence, we observed that the multiple studies that did something in this respect, employed the translation/back-translation approach. A more recent OM study, not included in our review, employs the superior TRAPD-approach for each language group (Karjalainen and Salmi, 2013).

Regarding data collection equivalence, we observed one study demonstrating good practice by explicitly pointing out how it maximized this form of equivalence. Gonzalez-Benito (2007) seeks to minimise threats to equivalence by sampling from industries that have similar characteristics (machinery manufacturers, electronic manufacturers, and transportation equipment manufacturers), pilot-testing in each of these to correct any ambiguous terminology, and using similar size samples in each SIC code (134, 140, and 143 respectively).

4.5.3. The Analysis Stage

From the studies assessed, we also identify a number of best practices in the consideration of equivalence during analysis. First, it is good practice to perform multiple equivalence tests, when multiple threats to equivalence are detected. Hult et al. (2006) provide an interesting example in this regard. They first conduct an MGCFA to rule out potential heterogeneity due to different sample frames (CSCM versus ISM). Full equivalence cannot be established; 14 out of the 58 items were non-equivalent. This is an interesting finding, especially taking into account that many studies pool data from these two databases without further consideration. The authors solve this situation by dropping the non-equivalent items from the dataset before testing equivalence of sub-groups differentiated by the second criterion (strategy type). Again, they drop three items that appear non-equivalent across different strategy types. Statistical evidence is provided that the MGCFA model with the reduced number of items has an acceptable fit. Boyer and Hult (2006) also develop multiple equivalence tests, related to responses from customers of online/home delivery grocers. They define four sub-groups based on two criteria: customer experience with online shopping and the grocer´s picking method (from store versus from distribution centres). In other words, they apply both heterogeneity criteria at the same time, rather than in a sequential fashion as exemplified earlier by Hult et al. (2006). Similar to Hult et al. (2006), Boyer and Hult (2006) exclude non-equivalent items from further analysis.

Second, it is good practice to question measurement equivalence of sub-groups that are hypothesised to have different path models. In this regard, Bagozzi and Dholakia (2006: p. 1109) hypothesise that “Linux user experience” moderates the path between actual participation in Linux user groups and some key antecedents. Consequently, they verify that the expert versus novice sub-groups demonstrate metric equivalence to justify comparisons of path model estimators.

Third, when performing equivalence tests, it is desirable to be explicit regarding the type of subsequent substantial analysis to be performed (comparison of causal relationships versus comparison of means). Cannon et al. (2010) provide an excellent example in this regard. Many OM studies, like Cannon et al. (2010), aim to compare relationships (in path models) and not means, and the most strict equivalence test – for scalar equivalence – is therefore not required. Related to this issue, is that readability increases greatly when researchers provide test statistics for each consecutive step, as in Kaufmann and Carter (2006).

4.5.4. The Fine-Tuning Stage

Finally, as so few OM studies test for data equivalence, there are not many examples of how to deal with non-equivalence or partial equivalence. The studies by Hult et al. (2006) and Boyer and Hult (2006) mentioned above, drop non-equivalent items from further analysis. In these cases, this amounted to 30% and 21% of items to be dropped respectively, in order to reach a model with acceptable fit in the MGCFA. It would be good to complement such statistical arguments with a substantive discussion on potential shifts of the theoretical coverage of the constructs after having reduced the number of items so considerably (Saris, Knoppen, and Schwartz 2013). Another approach to dealing with non-equivalence is to repeat the substantial analysis per group, present the results per group, and discuss inferences from a within-group perspective (Rungtusanatham et al. 2008).

5. Discussion and Conclusions

The main aim of our paper was to build on previous work in order to further promote awareness of the importance of measurement equivalence among OM scholars undertaking survey research. In many survey studies, the dataset contains data from apparently heterogeneous groups, defined for instance by differences in company characteristics, respondent demographics, country/culture/language, data collection method, or timing. If individuals across different groups have fundamentally different mental models and thus understand survey questions and respond to survey questions in systematically different ways, then data from such groups is not equivalent. Non-equivalent data should not be compared or pooled, and differences/similarities across groups may be illusory.

Our paper makes two key contributions. First, using a comprehensive framework of the different stages of survey research – (a) identification, (b) design, (c) analysis, and (d) fine-tuning – this is the first study to undertake a systematic assessment of the way equivalence is treated by scholars in extant OM. Based on this assessment, we show that to date, OM researchers give only limited attention to issues of equivalence when data from different groups is pooled or compared. More specifically, the content analysis of 465 survey research articles in six leading OM journals over a six-year period shows that almost two thirds of these articles were found to have at least one source of data heterogeneity and nearly one quarter had at least two sources of data heterogeneity. Our assessment also revealed that of the studies with at least one type of potential heterogeneity, only a limited number of studies subsequently intended to maximize equivalence prior to data collection: three studies considered maximisation of construct equivalence, 38 studies considered maximisation of translation equivalence and five studies considered maximisation of data collection equivalence prior to data collection. Our assessment also demonstrated that not more than 17 studies analysed equivalence once data were collected. Initial attempts to test for equivalence have typically been based on single indicator analyses and EFA rather than on the superior CFA. Only recently has the OM community started to appreciate and use the CFA method (Shah and Goldstein 2006). It is interesting to note that OM researchers in particular perform equivalence tests when heterogeneity is detected based on different company traits (e.g. different groups based on the degree of environmental uncertainty, as in Koufteros et al. (2005)). Other academic disciplines have instead focused on heterogeneity due to different countries/cultures/languages, and, to a minor extent on different individual traits, data collection methods and time periods of data collection. Finally, regarding the final stage of survey research, there were only three studies that indicated how they managed test outcomes showing non-equivalence and/or partial equivalence.

As our second key contribution, we provide details of best practices from literature in dealing with equivalence issues at different stages of the research process. Some of these stem from the OM literature whilst others are drawn from further afield where consideration of equivalence is arguably more advanced. We believe that by drawing together best practices in this way, we offer a more comprehensive perspective of equivalence across various stages of a survey project and provide coherent guidelines for (1) identifying possible sources of heterogeneity among respondent groups that may threaten equivalence; (2) maximizing equivalence during survey design; (3) testing equivalence once data are collected; and (4) dealing with partial- and nonequivalence. This arguably results in a richer understanding of the issues that should be considered when compared with referent equivalence studies (see Table 1 above). For example, when exploring the identification stage of research, our paper moves beyond the discussion of how cross-cultural and cross-country differences threaten equivalence to examine a broader set of differences that OM scholars should consider when designing surveys.

Given the increasing movement towards collaborative multi-institution research and comparative research (and the subsequent increases in data heterogeneity risk), as well as the growing appreciation of the importance of measurement equivalence, there is clearly scope within the OM field to pay more attention to issues of data equivalence. As a first step, awareness among authors, reviewers, and editors needs to be raised. Whenever a dataset contains data from transparently different groups, the issue of data equivalence would need to be discussed by the authors, and if not, signalled by editors or reviewers. In appendix A, we provide a checklist of issues in this regard. This checklist builds upon the proposed comprehensive model and the best practice and pragmatic decision issues we detected in the literature. The decision regarding what are transparently different sub-groups cannot be answered in general, but our list of dimensions of heterogeneity (country/culture/language, respondent profile, company profile, data collection method, data collection moment) is a starting point[2]. Considering that so few studies in the OM field with at least one source of heterogeneity discuss data equivalence, there is ample scope to re-analyse or replicate such studies. This call for replication is not to cast doubt on these earlier findings; replication is a recommended strategy anyway to continuously strengthen the validity of research findings (Goldsby and Autry 2011; Van Weele and Van Raaij 2014).

Overall, we hope our assessment of the literature and illustration will help to improve understanding of equivalence issues and further increase the methodological rigor of OM research.

[2] *In extremis*, each individual respondent can be considered to have a unique response slope (Schwartz, 2007). Consequently, data on his human values scale, which is widely employed in cross-cultural research, are commonly "centered" by respondent. We thank one of the reviewers for raising this point.

Appendix A. Checklist for authors, editors and reviewers involved in survey research with latent variables

Detecting Sources of Heterogeneity – the Identification Stage	1) Does the study pool or compare data from apparently heterogeneous groups? • Does the study pool subsets of data? • Does the study hypothesize different path models? • Does the study compare values of variables across groups? 2) Can you think of any reason why respondents across groups might understand the questions in different ways? And, can you think of any reason why respondents might express themselves in different ways? A non-exhaustive list of reasons is: • Does the study involve companies with different characteristics (e.g., firm size, industry, position in the supply chain)? • Does the study involve respondents from different countries/cultures/languages? • Does the study involve respondents with different individual demographics (e.g., functional differences, age, gender, race)? • Does the study involve multiple data collection methods? • Are data collected in different time periods?
Maximising Equivalence Prior to Data Collection – the Design Stage	3) Have there been efforts to maximize equivalence prior to data gathering, for each of the identified threats? • Does the study consider construct equivalence (with pre-tests in each sub-group), for each of the identified threats? o Conceptual equivalence o Category equivalence o Functional equivalence • Does the study consider translation equivalence, when the threat stems from multiple languages? • Does the study consider data collection equivalence, when multiple institutions are responsible for data collection? o sampling frame comparability o administration mode equivalence o coverage comparability o sampling method comparability
Testing Measurement Equivalence after	4) Does the study perform multiple equivalence tests, when multiple threats to equivalence are detected? 5) Does the study report test statistics for the subsequent testing steps?

Data Collection – the Analysis Stage	6) Does the study relate the type of test performed with the subsequent substantial study (comparison of relationships, comparison of means or pooling)?
Dealing with Partial and Non-equivalence – the Fine-tuning Stage	7) In case full-equivalence is rejected, does the study fine-tune results? • Assess to what extent partial equivalence exists and execute substantive analyses that are acceptable with partially equivalent data • Make sense of non-equivalence • Exclude the non-equivalent group(s) from substantive analyses 8) Alternatively, in case of non-equivalence, does the study repeat the substantial analysis per group, present the results per group, and discuss inferences from a within-group perspective?

References

Albadvi, A., Keramati, A., and Razmi, J. 2007. Assessing the impact of information technology on firm performance considering the role of intervening variables: organizational infrastructures and business processes reengineering. *International Journal of Production Research* 45 (12), 2697–2734.

Allred, C.R., Fawcett, A.M., Wallin, C., and Magnan, G.M. 2011. A dynamic collaboration capability as a source of competitive advantage. *Decision Sciences* 42 (1), 129–161.

Ariely, G., and Davidov, E. 2011. Assessment of measurement equivalence with cross-national and longitudinal surveys in political science. *European Political Science* (2012) 11, 363–377.

Bagozzi, R.P., and Dholakia, U.M. 2006. Open Source Software User Communities: A Study of Participation in Linux User Groups. *Management Science* 52 (7), 1099–1115.

Barman S., Hanna, M. D. and LaForge, R. L. 2001. Perceived relevance and quality of POM journals a decade later. *Journal of Operations Management* 19(3): 367-385.

Bensaou, M., Coyne, M., and Venkatraman, N. 1999. Testing metric equivalence in cross-national strategy research: an empirical test across the United States and Japan. *Strategic Management Journal* 20, 671-689.

Birou, L., Germain, R.N., and Christensen, W.J. 2011. Applied logistics knowledge impact on financial performance. *International Journal of Operations and Production Management* 31 (8), 816–834.

Bollen, K.A. 1989. *Structural Equations with Latent Variables*. US: John Wiley and Sons.

Bou-Llusar, J.C., Escrig-Tena, A.B., Roca-Puig, V., and Beltrán-Martín, I. 2009. An empirical assessment of the EFQM excellence model: Evaluation as a TQM framework relative to the MBNQA model. *Journal of Operations Management* 27 (1), 1-22.

Boyer, K.K., and Hult, G.T.M. 2006. Customer behavioral intentions for online purchases: An examination of fulfillment method and customer experience level. *Journal of Operations Management* 24 (2), 124-147.

Boyer, K.K., and Verma, R. 2000. Multiple raters in survey-based operations management research: A review and tutorial. *Production and Operations Management* 9 (2), 128-140.

Byrne, B.M., Shavelson, R.J., and Muthén, B. 1989. Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. *Psychological Bulletin*, 105 (3), 456-466.

Cannon, J.P., Doney, P.M., Mullen, M.R., and Petersen, K.J. 2010. Building long-term orientation in buyer-supplier relationships: The moderating role of culture. *Journal of Operations Management* 28 (6), 506-521.

Cheung, G.W., and Lau, R.S. 2012. A direct comparison approach for testing measurement equivalence. *Organizational Research Methods* 15 (2), 167-198.

Cheung, G.W., and Rensvold, R.B. 1999. Testing factorial invariance across groups: a reconceptualization and proposed new method. *Journal of Management*, 25(1), 1-27.

Craig, C.S., and Douglas, S.P. 2000. *International Market Research*. 2nd edn. Chichester: John Wiley and Sons.

Cullen, A., and Taylor, M. 2009. Critical success factors for B2B e-commerce use within the UK NHS pharmaceutical supply chain. *International Journal of Operations and Production Management*, 29 (1) 1156-1185.

Da Silveira, G.J.C. 2011. Our own translation box: exploring proximity antecedents and performance implications of customer co-design in manufacturing. *International Journal of Production Research*, 49(13), 3833-3854.

De Jong, M.G., Steenkamp, J.B.E.M., and Fox, J.P. 2007. Relaxing measurement invariance in cross-national consumer research using a hierarchical IRT model. *Journal of Consumer Research* 34, 260-278.

Dowlatshahi, S. 2011. An empirical study of the ISO 9000 certification in global supply chain of maquiladoras. *International Journal of Production Research* 49 (1), 215–234.

Douglas, S.P, and Craig, C.S. 2006. On improving the conceptual foundations of international marketing research. *Journal of International marketing* 14(1), 1-22.

Drasgow, F. 1984. Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. *Psychological Bulletin* 95, 134-135.

Filippini, R. 1997. Operations management research: some reflections on evolution. *International Journal of Operations & Production Management* 17 (7), 655–670.

Goh, C., Holsapple, C.W., Johnson, L.E., and Tanner, J.R. 1997. Evaluating and classifying POM journals. *Journal of Operations Management* 15 (2), 123-138.

Goldsby, T.J., and Autry, C.W. 2011. Toward greater validation of supply chain management theory and concepts: The roles of research replication and meta-analysis. *Journal of Business Logistics*, 32 (4), 324-331.

González-Benito, J. 2007. A theory of purchasing's contribution to business performance. *Journal of Operations Management* 25 901-907.

Harkness, J.A., Van de Vijver, F.J.R., and Mohler, P.P. 2003. *Cross-cultural Survey Methods*. New York: Wiley.

Hensley, R.L. 1999. A review of operations management studies using scale development techniques. *Journal of Operations Management* 17 (3), 343-358.

Hofstede, G. 2001. *Culture's Consequences, Comparing Values, Behaviors, Institutions, and Organizations Across Nations*. Thousand Oaks CA: Sage Publications.

Hu, L.T., and Bentler, P.M. 1998. Fit indices in covariance structure modeling: sensitivity to underparametrized model misspecification. *Psychological Methods* 3, 424-453.

Hult, G.T.M., Hurley, R.F., Giunipero, L.C. and Nichols Jr., E.L. 2000. Organizational learning in global purchasing: a model and test of internal users and corporate buyers. *Decision Sciences* 31(2), 293–325.

Hult, G.T.M., Ketchen, D.J., Cavusgil, S.T., and Calantone, R.J. 2006. Knowledge as a strategic resource in supply chains. *Journal of Operations Management* 24 (5), 458-475.

Hult, G.T., Ketchen Jr. D.J., Griffith, D.A., Finnegan, C.A., Gonzales-Padron, T., Harmancioglu, N., Huang, Y., Talay, M.B., and Cavusgil, S.T. 2008. Data equivalence in cross-cultural international business research: assessment and guidelines. *Journal of International Business Studies* 39, 1027-1044.

Johnston, D.A., and Kristal, M.M. 2008. The climate for co-operation: Buyer-supplier beliefs and behavior. *International Journal of Operations & Production Management* 28 (9), 875-898.

Karjalainen, K. and Salmi, A. 2013. Continental differences in purchasing strategies and tools. *International Business Review* 22 (1), 112-125.

Kaufmann, L., and Carter, C.R. 2006. International supply relationships and non-financial performance: A comparison of US and German practices. *Journal of Operations Management* 24 (5), 653-675.

Kaynak, H., and Hartley, J.L. 2006. Using replication research for just-in-time purchasing construct development. *Journal of Operations Management* 24 (6), 868-892.

Kish, L. 1994. Multipopulation survey designs: Five types with seven shared aspects. *International Statistical Review* 62 (2), 167-186.

Koufteros, X., and Marcoulides, G. A. 2006. Product development practices and performance: A structural equation modeling-based multi-group analysis. *International Journal of Production Economics* 103, 286-307.

Koufteros, X., Vonderembse, M., and Jayaram, J. 2005. Internal and external integration for product development: the contingency effects of uncertainty, equivocality, and platform strategy. *Decision Sciences* 36 (1), 97-133.

Kull, T.J., and Wacker, J. G. 2010. Quality management effectiveness in Asia: The influence of culture. *Journal of Operations Management* 28(3), 223-239.

Malhotra, M.K., Grover, V. 1998. An assessment of survey research in POM: From constructs to theory. *Journal of Operations Management* 16 (4), 407–425.

Malhotra, M.K., and Sharma, S. 2008. Measurement equivalence using generalizability theory: An examination of manufacturing flexibility dimensions. *Decision Sciences* 39 (4), 643-669.

Melnyk, S.A., Lummus, R.R., Vokurka, R.J., Burns, L.J., and Sandor, J. 2009. Mapping the future of supply chain management: A Delphy study. *International Journal of Production Research*, 47 (16), 4629-4653.

Mullen, M.R. 1995. "Diagnosing measurement equivalence in cross-national research." *Journal of International Business Studies* 26, 573-596.

Nyaga, G.N., Whipple, J.M., and Lynch, D.F. 2010. Examining supply chain relationships: Do buyer and supplier perspectives on collaborative relationships differ?" *Journal of Operations Management* 28 (2), 101-114.

Nye, C., and Drasgow, F. 2011. Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. *Journal of Applied Psychology* 96 (5), 966-980.

Pagell, M., Krumwiede, D.W., and Shey, C. 2007. Efficacy of environmental and supplier relationship investments-moderating effects of external environment. *International Journal of Production Research*, 45 (9), 2005-2028.

Peng, D.X., Liu, G.J., and Heim, G.R. 2011. Impacts of information technology on mass customization capability of manufacturing plants. *International Journal of Operations and Production Management* 31 (10), 1022–1047.

Podsakoff, P.M., MacKenzie, B., Lee, J.Y., and Podsakoff, N.P. 2003. Common method biases in behavioural research: a critical review of the literature and recommended remedies. *Journal of Applied Psychology*, 88 (5), 879–903.

Prater, E., and Ghosh, S. 2006. A comparative model of firm size and the global operational dynamics of US firms in Europe. *Journal of Operations Management* 24 (5), 511-529.

Rungtusanatham, M.J., Choi, T.Y., Hollingworth, D.G., Wu, Z., and Forza, C. 2003. Survey research in operations management: historical analyses. *Journal of Operations Management* 21 (4), 475-488.

Rungtusanatham, M., Forza, C., Koka, B.R., Salvador, F., and Nie, W. 2005. TQM across multiple countries: Convergence hypothesis. *Journal of Operations Management* 23 (1), 43-63.

Rungtusanatham, M., Ng, C. H., Zhao, X., and Lee, T.S. 2008. Pooling data across transparently different groups of key informants: measurement equivalence and survey research. *Decision Sciences* 39 (1), 115-145.

Saris, W.E. 1988. *Variation in response functions: A source of measurement error in attitude research.* Amsterdam: Sociometric Research Foundation.

Saris, W.E., and Gallhofer, I. 2007. *Design, Evaluation and Analysis of Questionnaires for Survey Research*. Hoboken (New Jersey): Wiley Interscience.

Saris, W.E., Knoppen, D., and Schwartz, S.H. 2013. Operationalizing the theory of human values: Balancing homogeneity of reflective items and theoretical coverage. *Survey Research Methods,* 7(1), 29-44.

Saris, W.E., Satorra, A., and Van der Veld, W. 2009. Testing structural equation models or detections of misspecifications? *Structural Equation Modeling* 16, 561-582.

Schwartz, S. 2007. Value orientations: Measurement, antecedents and consequences across nations, in *Measuring Attitudes Cross-nationally. Lessons from the European Social Survey*, R. Jowell, C. Roberts, R. Fitzgerald and G. Eva, Eds. London: Sage Publications, 169-203.

Schmidt, F.L. 1996. Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods* 1 (2), 115-129.

Shah, R., and Goldstein, S.M. 2006. Use of structural equation modelling in operations management research: looking back and forward. *Journal of Operations Management* 24 (2), 148-169.

Smith, L.L. 2002. On the usefulness of item bias analysis to personality psychology. *Personality and Social Psychology Bulletin* 28 (6), 754-763.

Steenkamp, J., and Baumgartner, H. 1998. Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research* 25, 78-90.

Tan, C.L., and Vonderembse, M.A. 2006. Mediating effects of computer-aided design usage: From concurrent engineering to product development performance. *Journal of Operations Management* 24 (5), 494-510.

Tsikriktsis, N. 2005. A review of techniques for treating missing data in OM survey research. *Journal of Operations Management* 24 (1), 53-62.

Van Herk, H., Poortinga, Y.H., and Verhallen, T.M. 2004. Response styles in rating scales: Evidence of method bias in data from six EU countries. *Journal of Cross-Cultural Psychology*, 35 (3), 346-360.

Van Weele, A.J., and Van Raaij, E.M. 2014. The future of purchasing and supply management research: about relevance and rigor. *Journal of Supply Chain Management*, 50 (1), 56-72.

Vandenberg, R.J., and Lance, C.E. 2000. A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3, 4-69.

Wouters, M., Anderson, J.C., Narus, J.A., and Wynstra, F. 2009. Improving sourcing decisions in NPD projects: Monetary quantification of points of difference. *Journal of Operations Management* 27 (1), 64-77.

Xu, X., Venkatesh, V., Tam, K.Y., and Hong, S.J. 2010. Model of migration and use of platforms: Role of hierarchy, current generation, and complementarities in consumer settings. *Management Science* 56 (8), 1304–1323.

Zhang, C., Viswanathan, S., and Henke, J.W., Jr. 2011. The boundary spanning capabilities of purchasing agents in buyer-supplier trust development. *Journal of Operations Management* 29 (4), 318–328.

Figure 1. Considering equivalence issues across four stages of survey research

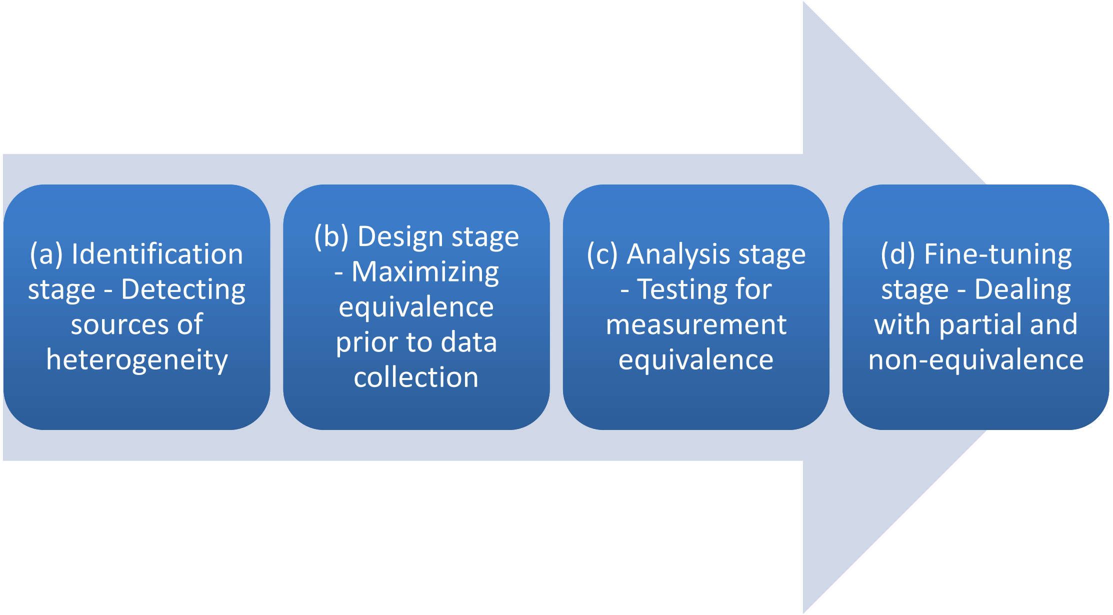
(a) Identification stage - Detecting sources of heterogeneity
(b) Design stage - Maximizing equivalence prior to data collection
(c) Analysis stage - Testing for measurement equivalence
(d) Fine-tuning stage - Dealing with partial and non-equivalence

Figure 2. Two alternative response functions

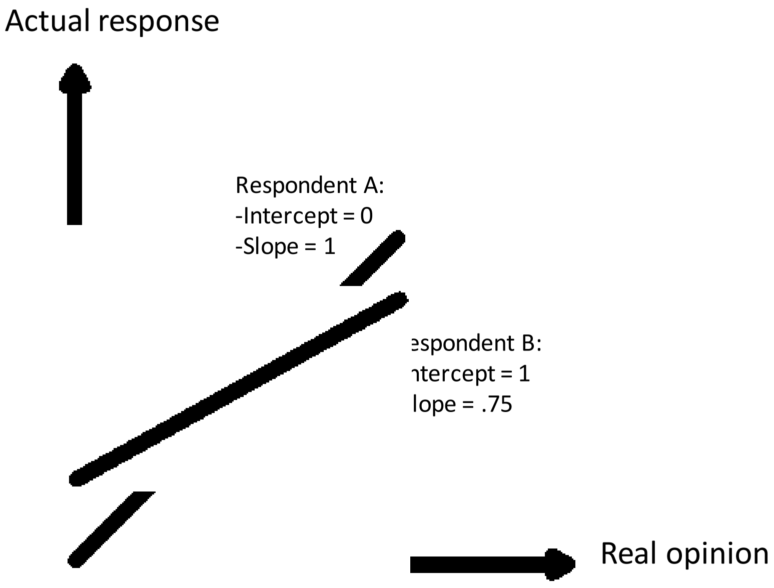

Table 1: Contributions and gaps of studies that provide guidelines for consideration of equivalence

	Domain	Systematic assessment of extant literature	A. Identification stage – Detecting sources of heterogeneity	B. Design stage – Maximizing equivalence prior to data collection	C. Analysis stage – Testing for measurement equivalence	D. Fine-tuning stage – Dealing with partial and non-equivalence
Steenkamp and Baumgartner (1998)	Consumer research.	No.	Only cross-cultural / cross-country differences.	Not discussed.	MGCFA procedure presented.	Suggestions on how to deal with lack of full equivalence.
Bensaou et al. (1999)	Strategy.	No.	Only cross-cultural / cross-country differences.	Only summarily discussed.	Brief MGCFA procedure presented (excl. scalar equivalence).	No suggestions on how to deal with lack of full equivalence.
Cheung and Rensvold (1999)	Management.	No.	Only cross-cultural / cross-country differences.	Not discussed.	MGCFA procedure presented.	Suggestions on how to deal with lack of full equivalence.
Vandenberg and Lance (2000)	Human resource management.	No.	Not discussed.	Not discussed.	MGCFA procedure presented.	Suggestions on how to deal with lack of full equivalence.
Douglas and Craig (2006)	Marketing.	No.	Only cross-cultural / cross-country differences.	Various techniques presented.	No suggestions.	No suggestions on how to deal with lack of full equivalence.
Hult et al. (2008)	International business.	Yes: 165 articles from 5 journals from 1995 to 2005.	Only cross-cultural / cross-country differences.	Various techniques presented.	MGCFA procedure and other procedures presented.	No suggestions on how to deal with lack of full equivalence.
Rungtusanatham et al. (2008)	Operations management.	No.	Only demographic differences.	Only summarily discussed.	Extensive MGCFA procedure presented.	No suggestions on how to deal with lack of full equivalence.
This paper	***Operations Management.***	***Yes: 465 articles from 6 journals from 2006 to 2011.***	***Discussion of nature of any potential differences.***	***Various techniques presented.***	***MGCFA procedure and other procedures presented.***	***Suggestions on how to deal with lack of full equivalence.***

Table 2: Sources of heterogeneity in OM survey research

	JOM	IJOPM	IJPR	DS	MS	POM	Total	Share in total number of survey papers
Total number of survey papers	144	131	83	60	29	18	465	
Multiple countries/languages	26	32	16	6	5	3	88	18.9%
Multiple respondent profiles	72	42	52	35	15	5	221	47.5%
Multiple data collection methods	28	18	27	18	8	2	101	21.7%
Multiple time periods of data collection	4	4	0	4	3	0	15	3.2%
At least one source of heterogeneity	98	68	59	41	19	8	293	63.0%
At least two sources of heterogeneity	26	19	28	19	10	2	104	22.4%

Table 3: Consideration of equivalence in the design stage of OM survey research

Equivalence issues in design stage	Measures	JOM	IJOPM	IJPR	DS	MS	POM	Total
Construct equivalence	Literature review, use of validated surveys, focus groups and/or pre-tests in each respondent group	1	2	0	0	0	0	3
Translation equivalence	Back translation, Team approach	11	16	8	2	0	1	38
Data collection equivalence	Similar/systematic selection of samples, data collection procedures, timing of data collection	5	0	0	0	0	0	5

Note: The numbers in the cells refer to the number of papers that addressed a specific issue; one paper may therefore appear in more than one cell

Table 4: Sources of heterogeneity and measurement equivalence tests performed per journal

	JOM	IJOPM	IJPR	DS	MS	POM
Multiple countries/cultures/languages	Kaufmann and Carter (2006); Cannon et al. (2010)					
Multiple respondent profiles (individual traits)	Wouters et al. (2009); Nyaga et al. (2010)	Johnston and Kristal (2008)			Bagozzi and Dholakia (2006); Xu et al. (2010)	
Multiple data collection methods	Boyer and Hult (2006); Hult et al. (2006)					
Multiple time periods of data collection	Kaynak and Hartley (2006)	Peng et al. (2011)		Allred et al. (2011)		
Multiple company profiles (company traits)	Boyer and Hult (2006); Hult et al. (2006); Prater and Ghosh (2006); Bou-Llusar et al. (2009); Zhang et al. (2011)	Birou et al. (2011)	Dowlatshahi (2011)			