The problem with standardizing international market research: A case study from B2B services

M. Sajid Khan a,⁎, Earl Naumann a,1, Matti J. Haverila b,2

⁎American University of Sharjah, PO Box 26666, Sharjah, United Arab Emirates
1Thompson Rivers University, Kamloops, British Columbia, Canada

A R T I C L E   I N   P R E S S

A R T I C L E   I N   P R E S S

A U S T R A L A S I A N   M A R K E T I N G   J O U R N A L

A U S T R A L A S I A N   M A R K E T I N G   J O U R N A L

Contents lists available at ScienceDirect

Australasian Marketing Journal

journal homepage: www.elsevier.com/locate/amj

ARTICLE INFO

Article history:
Received 11 February 2013
Revised 4 March 2014
Accepted 22 March 2014
Available online xxxx

Keywords:
Cross-equivalence
Cross-cultural research
Customer satisfaction
Standardization versus adaptation

ABSTRACT

One of the key issues for multinational corporations (MNCs) is whether to standardize their marketing approach across all countries or adapt their practices to fit each country. In order to make this decision, MNCs must determine if their marketing approach is cross-culturally appropriate and equivalent from country to country. Unfortunately, recent research indicates that the vast majority of academic studies do not adequately address the cross-cultural equivalence issue. The primary purpose of this article is to illustrate the problem of using a standardized, global, B2B research approach. The second goal is to show how cross-cultural equivalence can be identified and managed. The firm in this study is a Fortune 100 MNC that provides facilities management services in over 100 countries. This article compares the cross-cultural equivalence of customer satisfaction survey data from the US and from Japan. The results show that about half of the items typically used to measure dimensions of product and service quality as drivers of customer satisfaction lack cross-cultural equivalence. The implication is that the use of a fully standardized approach to global research must be questioned.

© 2014 Australian and New Zealand Marketing Academy. Published by Elsevier Ltd. All rights reserved.

1. The importance of MNCs

Multinational corporations (MNCs) play a significant role in the global economy. The 500 largest MNCs control 70% of all world trade and are vital for economic growth (World Trade Organization, 2011). One of the key issues for an MNC is the extent to which they standardize their operations and business practices globally (Jain, 2007). By standardizing their operations and business practices, an MNC can achieve economies of scale, benefit from organizational learning, and maintain better operational control through the use of standardized metrics. However, others contend that national cultures exert a powerful influence on individuals and organizations (Hofstede, 1997), and, therefore, on customer perceptions and preferences. This implies that MNCs should adapt to fit each culture, “thinking globally but acting locally”. Despite over twenty years of research and debate, the standardization versus adaptation debate is far from over.

Our study examined the drivers of customer satisfaction from a US based Fortune 500 MNC that provides B2B services globally. The firm has broadly standardized its customer satisfaction research worldwide, using the same questionnaires and methodologies in all countries. The only research change from country to country was the translation of the questionnaire into the language of the host country. The firm implicitly assumed that business customers were essentially similar worldwide, having the same needs and performance expectations. Thus, the firm used a standardized customer satisfaction questionnaire worldwide to provide managers information about where internal, value creating processes needed to be improved. This study examined the appropriateness of using this standardized research approach for a global customer satisfaction program by examining whether cross-cultural equivalence of questionnaire items existed.

While the MNC operates in over 100 countries, two of largest economies in the world were chosen for this cross-cultural comparison. The US was chosen since it is the largest economy in the world, home for the MNC, and the US is where the standardized research program was designed. Japan was chosen as the comparison country for several reasons. First, Japan is the third largest economy in the world. The second reason was pragmatic, large samples in both countries were available. Third, a good deal of research is available comparing these cultures (Hall, 1987; Hofstede, 1997). These studies have found that the US and Japan differ on many measures of culture. Because of this cultural divergence, differences in response...
patterns to questions between these two countries should be more apparent in our research.

2. The global research approach

2.1. Customer satisfaction

Customer satisfaction may be the most researched issue in business, both by academics and by managers (Zeithaml et al., 2006). A major reason for this interest is that a wide variety of positive financial benefits accrue from high levels of customer satisfaction (Anderson et al., 2004; Grucia and Rego, 2005; Homburg et al., 2005; Williams and Naumann, 2011). Due to these positive financial outcomes, a large body of customer satisfaction research has attempted to identify the antecedents of high customer satisfaction levels. The implicit, or explicit, assumption by both managers and academics is that the antecedents can be managed to more effectively achieve high customer satisfaction.

Accordingly, the antecedents, or drivers, of customer satisfaction have received a good deal of research interest. Various dimensions of products, service, relationships, image, value, and price have been found to drive customer satisfaction (Anderson and Mittal, 2000; Coolit et al., 2007; Ganesan, 1994). In many business-to-business (B2B) situations, product and service dimensions are closely intertwined as drivers of satisfaction (Tuli et al., 2007; Vargo and Lusch, 2011). Ulaga and Reinartz (2006) recently used the term “hybrid services” to describe the product-service integration often found in B2B situations. The implication is that specific dimensions of products such as quality, reliability, and innovativeness could drive satisfaction in B2B relationships (Gill and Ramaseshan, 2007; Homburg and Garbe, 1999; Homburg and Rudolph, 2001; Zolkiewski et al., 2007).

Additionally, various dimensions of service have been identified as key drivers of satisfaction in B2B relationships (Homburg and Rudolph, 2001). Woo and Ennew (2005) found that B2B services typically involved dyadic exchanges at the exchange points of personal interaction between suppliers and customers. The most common touch-points of personal interaction suppliers and customers involve a supplier’s account representatives, technical support, call handling and maintenance personnel (Jackson and Cooper, 1988; Patterson and Spreng, 1997). These touch-points were investigated in this research.

While dimensions of product and service convey benefits to customers, these benefits are balanced against the supplier’s price level. Since B2B customers are generally viewed as rational decision makers, price would be expected to have an influence on customer satisfaction (Cronin et al., 2000; Gale, 1994; Reichheld, 1996). Various studies have found that price is significantly related to customer satisfaction (Bolton and Lemon, 1999; Dawes, 2009; Gill and Ramaseshan, 2007; Lye and Hamilton, 2000). In most studies, price has been found to be negatively related to satisfaction.

Although there is substantial evidence that various dimensions of products, service, and price drive customer satisfaction, cross-cultural studies on this subject are rare. Most customer satisfaction studies are based on results from a single country. The firm in this study based the global research program based on what had worked well in the US. The home country research design was then simply extended to other countries with only translation into the host country language allowed. The senior executives explicitly assumed that all customers were essentially the same worldwide in their performance evaluation and decision making processes. The executives expected to capture customer based perceptions that would help local managers improve process performance and achieve high customer satisfaction levels. The purpose of this paper was to examine whether the drivers of satisfaction were consistent across countries with significant cultural differences. The following sections of this paper examine some of the cultural differences that could constrain a standardized global approach.

3. Cross-cultural literature review

3.1. To standardize or adapt?

The standardization/adaptation (ST/AD) debate can be approached on several organizational levels. At the highest level, the issue could be applied to a firm’s strategies and structure (Jain, 2007). At an operational level, the issue could be applied to organizational procedures, processes, and business practices such as supply chain relationships and financial performance metrics. However, the majority of the ST/AD debate has focused on the marketing mix, and products and promotion have probably received the most interest in the ST/AD debate (Theodosiou and Leonidou, 2003). The general contention is that products and promotion should be adjusted to fit local preferences (Walters, 1986). Walters also indicated that activities and techniques such as data collection, statistical analysis, and the managerial use of market data hold potential for standardization.

One argument supporting standardization is that, in order to make valid cross-cultural comparisons of research results, a firm should use the same research methodologies worldwide. This approach would remove potential variations in data by country that could be caused by purely methodological differences. A standardized approach would use exactly the same questions, scales, data collection methods, and analytical techniques worldwide, thereby facilitating direct comparison. Theoretically, this standardization would result in directly comparable results, unless there is a significant cultural effect at work.

However, a cultural effect may cause respondents to conceptualize constructs, individual items, and response scales differently across cultures (De Jong et al., 2009). Thus, variation in results across countries, the US and Japan in our study, may be due to cultural nuances rather than real performance differences. If respondents in different countries conceptualize the items differently, the results should not be directly compared since there could be different underlying meanings for the constructs. In order to determine if research differences are caused by cultural variation, cross-cultural equivalence of constructs, items, and scales must be examined. Only those items that are cross-culturally equivalent should be compared across countries (Hult et al., 2008).

3.2. Cultural differences

National cultures shape individual values, beliefs, and decision-making processes. For example, Japanese culture may not only shape the needs and expectations of customers, but also shape how customers make decisions, both as individuals and organizations (Hofstede, 1997). In particular, service delivery should be adapted to fit local customers’ needs because service expectations are more directly influenced by aspects of culture (Malhotra and Peterson, 2001).

Barnlund (1975) described US and Japanese cultures as “polar extremes,” pointing to Japanese reserve and formality in comparison with American self-assertion and informal amiability. Hofstede (1991) found that the two cultures differ significantly on most of the dimensions of culture, with Japan being highly collectivistic and the US being highly individualistic. He suggested the individualism–collectivism divergence could have a powerful influence on behaviours, attitudes, and perceptions. These differences could lead to variation in the underlying meaning of constructs and their composition. In other words, Japanese respondents may conceptualize...
the constructs and items in a different way than their US counterparts, indicating a different underlying meaning.

Collective cultures, such as Japan, place a high emphasis on harmony and relationships, with an aversion to aggression and confrontation (Graham and Sano, 1986). This desire for unity makes courteous behaviour quite important in Japan to prevent discord and disagreement. In contrast, US culture favours informality, individualism, and assertiveness (Winsted, 1999). The cultural desire for harmony in Japan provides a pervasive framework for behaviour within society and organizations. This desire can be expressed as a need for consensus in group decisions and an emphasis on worker participation (Nonaka, 1988).

In order to maintain harmony in business relationships, Japanese have developed a tolerance for latent, perceived, and felt conflict. This implies that an individual may be frustrated but refrain from expressing their attitudes. To illustrate, a Japanese customer could be frustrated with a supplier’s performance, but hesitate to openly express the frustration. Shute and Carlante (1998) suggest that Japanese respondents may even attribute service failure to forces beyond the control of the service provider, mitigating the desire to express dissatisfaction. Overt displays of dissatisfaction would be a breach of cultural etiquette, causing a business partner to lose face. Because of the apparent importance of harmony as a dimension of culture, a “courtesy bias” may favour expressions of satisfaction (Steenkamp and Baumgartner, 2001). They argue that people in collective cultures are more likely to praise the service that they receive, even after a bad service experience.

It is clear that the collective nature of Japanese culture leads to a greater emphasis on relationships and harmony than the more individualistic American culture. Therefore, there is an apparent need to identify the impact of this cultural divergence on international research results. Specifically, we attempted to examine whether the questions and constructs are consistent across cultures, and, therefore, cross-culturally equivalent.

3.3. Establishing equivalence

International research can be affected by a cultural response bias between countries and/or cultures being compared (Van De Vijver and Leung, 2000). Cultural response bias is the systematic cultural tendency of individuals to respond to rating scales in a distinct way due to dimensions of their culture. To overcome cultural response bias, a researcher must establish the extent to which the research results are cross-culturally equivalent. Hult et al. (2008, p. 1027) defined data equivalence as “the extent to which elements of a research design have the same meaning, and can be applied in the same way, in a different cultural context”. A major concern is that, if results are used that lack cross-cultural equivalence, erroneous conclusions may be made, limiting knowledge advancement and theoretical development.

Although international business researchers have recognized the need to establish cross-cultural equivalence (Baumgartner and Steenkamp, 1998), the majority of international business research does not adequately address the issue. Hult et al. (2008) examined all cross-cultural research (167 articles) appearing in six leading journals that publish international studies over a ten year period (1995–2005). They found that the majority of cross-cultural articles failed to adequately address cross-cultural equivalence, even in the best journals. They concluded that the failure to properly establish cross-cultural equivalence limited the meaning of research results and also limited theoretical development in all areas of international business. Therefore, they felt that cross-cultural equivalence must be established in all business disciplines, such as marketing, management, and finance.

This issue of cultural response bias is particularly important when a standardized research approach is used worldwide. In a standardized approach, methodologies and questionnaires are used consistently across cultural groups. The underlying assumption for standardization is that the constructs, questions, and scaling have the same fundamental meaning in all cultures. However, if the items and constructs have different underlying meaning, the questionnaires should be modified accordingly.

The most commonly used technique to test for equivalence, or invariance, across cultures is factor analysis (Baumgartner and Steenkamp, 1998). Factor analysis examines if the factor loadings have the same configuration across cultures. If the factor loadings are different, the construct lacks conceptual equivalence, and subsequent direct comparisons are inappropriate. However, it is not unusual for some factor loadings to be completely equivalent across cultures while other factors may vary across cultures (Byrne and Campbell, 1999).

Two types of factor analysis are typically used to test for cross-cultural equivalence. Often, exploratory factor analysis (EFA) is used first, followed by confirmatory factor analysis (CFA) (Hult et al., 2008). EFA is used first to remove any preconceived bias that may exist in the organization of the data. For example, the grouping of questions on a questionnaire is typically based on assumptions by a researcher. The CFA results would indicate if these assumptions were valid. The EFA for each country produces a set of factors composed of individual items. Typically, some items on a questionnaire will not load significantly on the factors within a single country. These should be deleted from further analysis as they add little explanatory power to a factor (Hult et al., 2008). The reduced set of factors and items should then be introduced into the CFA. Only the factors that load identically in CFA for both countries have full cross-cultural equivalence. If the factors, and their composition, varied significantly across countries, the research would lack cross cultural equivalence. If at least partial conceptual equivalence is found (i.e. some items are the same) in the factor analyses, then further comparisons can be made with caution.

3.4. Hypotheses

Based on the previous discussions, we can develop a series of hypotheses. The first two hypotheses relate to the EFA, and the second two relate to the CFA. The first hypothesis indicates that we expect differences in the factors that will emerge from the EFA due to the significant cultural divergence between the US and Japan.

H1a: There will be significant differences between the US and Japan in the factors that emerge from the EFA analyses.

The second hypothesis indicates that we expect differences in the composition of factors between the US and Japan. Specifically, we expect differences in the individual items that comprise each factor.

H1b: There will be significant differences between the US and Japan in the item composition of each factor in the EFA analyses.

The third hypothesis indicates that we expect differences in the factors that emerge from the CFA. Specifically, the factors, and individual items comprising those factors, that emerged from the EFA will be introduced to the CFA. We expect to see differences in the factors between the US and Japan.

H2a: There will be significant differences between the US and Japan in the factors that emerge from the CFA analyses.

The fourth hypothesis indicates that we expect differences to emerge in the composition of each factor. Specifically, we expect...
the individual items to load differently on the factors between the US and Japan.

H2b: There will be significant differences between the US and Japan in the item composition of each factor in the CFA analyses.

4. Methodology

4.1. The firm

In an attempt to bridge the academia-practitioner divide, we examined “real world” market research data. The organization, which provided the data for our study, is a Fortune 500 firm in the facilities management industry operating in over 100 countries. The firm began developing a standardized customer satisfaction program in the mid-1990s. At least six different questionnaires were previously used by the various business units, which made direct performance comparisons difficult. Therefore, the firm completely redesigned and standardized their worldwide research program. Consequently, the only modification allowed from country to country was the translation of the questionnaires into the language of the host country. Data was gathered through telephone interviews conducted by a marketing research firm in the host country. The results of the customer interviews were a key input to the firm’s Six Sigma process improvement initiative. Therefore, the questionnaire structure was aligned with the firm’s value creating processes.

4.2. The questionnaire

All aspects of the questionnaire were designed in the US to fit the needs of US customers and to satisfy the objectives of the company. The original US questionnaire was developed by first conducting qualitative research (depth interviews) with random samples of customers. The intent was to capture the “voice of the customer” in the customer’s words. These qualitative results were used in the design of a draft questionnaire, which was circulated to an executive steering committee for review and modification. The question wording and scaling was also consistent with academic research in customer satisfaction (Barry et al., 2008; Tokman et al., 2007). The draft questionnaire was then piloted in a pre-test survey of customers for further refinement. This process was repeated every two years for each questionnaire used by the MNC to determine if US customer needs had changed. The original questionnaire was then translated into the Japanese language by a Japanese marketing research firm. It was then back translated by different individuals in the research firm for confirmation of meaning.

The total questionnaire included over 40 questions. However, some of these questions were global perceptions (overall satisfaction, recommend, met expectations, etc.) or demographic questions. As these were not relevant to measuring product and service quality, they were not included in this study. There were 30 individual items focussing on dimensions of product and service delivery that are typically viewed as key drivers of customer satisfaction (Table 1). There were multiple questions measuring customer perceptions of products, account reps, technicians, emergency services, call centre, price, and renewal services. The 30 questions were grouped into these seven categories. It should be noted that these categories were an operationalization of the “perceived service quality” construct in the loyalty model proposed by Garcia and Caro (2008). All questions used the same response scale of “Excellent—Very Good—Good—Fair—Poor”, and also included a “Do not Know” option. This five point scale is commonly used in both academic and real world market research.

4.3. Data collection

Telephone interviews were used to gather the data in both countries for three reasons. First, it was the uniform approach used globally by the MNC and generated high response rates. Second, telephone interviews were a fraction of the cost of personal interviews. Third, a telephone interview was viewed as less intrusive and took less of the respondent’s time. While telephone surveys are not widely used in consumer research in Japan, they are more commonly used in B2B research (Synodinos and Kobayashi, 2008).

As mentioned previously, sampling is a source of variation in cross-cultural research. Therefore, equivalence in sampling is important. The MNC that sponsored this research had a customer base consisting mostly of large Fortune 1000 type companies. About one third of the foreign customers were divisions of other US based MNCs who were also customers in the US. The remaining customers were divisions of foreign-based MNCs. There were no small customer firms, and few medium sized companies, as the MNC had a threshold contract amount that screened out small accounts. Therefore, the customer organizations in both Japan and the US were very similar in both countries.

The individual respondents, most commonly a facility manager, in all countries were identified as the “key decision maker” in their organization at the time the service contract was initiated. The intent was to have the samples as uniform as possible worldwide. The reason that key decision makers were interviewed was their role in decision making. The MNC wanted to track the perceptions of the most influential people in vendor selection in the customer organization. The service contract had a specific request for the key contact to participate in periodic customer satisfaction surveys.

Experienced U.S. and Japanese research firms gathered the data in each country. The proportion of customers who were contacted and actually participated in the survey was between 60 and 65 percent in both countries. Data was collected during the same time period in both countries. The sample size was 707 completed interviews in the US and 700 in Japan. Each respondent was interviewed no more than once every six months, and since the data was from two consecutive quarters, none of the respondents participated twice.

Consistent with Podsakoff, Mackenzie, Lee, and Podsakoff (2003), in order to reduce respondent’s evaluation apprehension and social desirability biases, respondents were reassured at the time of telephone interviews that there are no right or wrong answers, and the respondents answers will be kept anonymous. Hence, respondents have been explicitly asked to answer questions as honestly as possible.

In addition, statistical methods can be used to check for the possible effects of common methods bias after data is collected. One of the most commonly used tests is Harman’s single factor test (Harman, 1967). Generally, this test is used to examine if a single factor emerges from the exploratory factor analysis or if there is one general factor explaining the majority of the covariance in the independent and dependent variables (Podsakoff et al., 2003; Pavliou and Gelen, 2005). To follow this approach, all the variables from a study are loaded into an exploratory factor analysis and the un-rotated factor solution is examined across single factor to account for the variance in the variables (Aulakh and Kotabe, 1997). According to this test, if a single factor accounts for more than fifty percent of the variance in the items, methods bias is present (Mattila and Enz, 2002). We conducted separate tests for both samples of the US and Japan. The results revealed a factor structure for each sample with no one factor accounting for more than fifty percent of the variance. Therefore method bias per se, cannot explain our study results.
4.4. Analytical techniques

Two different types of analytical techniques were used. First, exploratory factor analysis (EFA) was used to determine the conceptual equivalence across the two samples (Hult et al., 2008). Then, consistent with Kumar (2000), confirmatory factor analysis was used on the factors that emerged at an exploratory level. These two levels of factor analysis would indicate whether further cross-cultural analysis of the data would be justified.

5. Results

5.1. Exploratory factor analysis

The first statistical test employed EFA to determine if the US and Japanese respondents conceptualized the 30 individual items in the same way (Anderson and Gerbing, 1988). The a priori model that the company had been using globally included seven factors that were viewed as drivers of customer satisfaction: technicians, emergency services, call handling, account representatives, products, price, and renewal services. The firm’s customer satisfaction model was virtually identical to that proposed by Garcia and Caro (2008). In EFA, the principal axis factoring was used to investigate the dimensionality of these constructs in both the US and Japan. The 30 individual items were used in two separate EFAs: one on the US dataset (n = 707), and second on the Japanese dataset (n = 700). The intent was to determine if the same constructs emerged consistently in both countries.

The latent root criterion was used to identify the number of items present in the factors (Hair et al., 2006). For this criterion, an Eigen value of 1.0 (the Kaiser criterion) was considered to be the threshold for extracting a factor. The rationale for the latent root criterion is that any individual factor should account for at least the expected variance of a single standardized variable if it is to be retained for interpretation (Hair et al., 2006).

The exploratory factor analysis suggested that the overall factor solution adequately accounted for the underlying structure of the data in both the US and Japan with Kaiser–Meyer–Olkin statistic achieving .945 for the US, and .891 for the Japan samples confirming the appropriateness of the data for factor analysis. Principal axis-factoring as the extraction method and varimax with Kaiser normalization as the rotation method were used to rotate the final components. Factor loadings of ±.50 were used as the threshold, and the items associated at this level were included in the factor (Hair et al., 2006). While our sample sizes justified the use of even lower factor loadings, we chose a threshold of 0.50 to be more conservative in our approach. Factor loadings of less than this amount were excluded. Cronbach alpha coefficients were then computed to quantify the construct reliabilities of the factors (Hair et al., 2006). All alpha scores were above the threshold level of 0.700 (Table 1).

Table 1
Exploratory factor analysis results: factor loadings and internal consistency.

<table>
<thead>
<tr>
<th>Construct name and items</th>
<th>US (n = 707)</th>
<th>Japan (n = 700)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Loading</td>
<td>Loading</td>
</tr>
<tr>
<td></td>
<td>Cronbach alpha</td>
<td>Cronbach alpha</td>
</tr>
<tr>
<td>Technicians</td>
<td></td>
<td></td>
</tr>
<tr>
<td>V9 Working safely</td>
<td>.518</td>
<td>.892</td>
</tr>
<tr>
<td>V20 Notifying prevent</td>
<td>.644</td>
<td></td>
</tr>
<tr>
<td>V21 Being courteous</td>
<td>.657</td>
<td></td>
</tr>
<tr>
<td>V22 Preventive maintena</td>
<td>.792</td>
<td></td>
</tr>
<tr>
<td>V24 Communicating effec</td>
<td>.726</td>
<td>.774</td>
</tr>
<tr>
<td>V26 Problem handling</td>
<td>.700</td>
<td>.777</td>
</tr>
<tr>
<td>V30 Diagnose and resolve</td>
<td>.815</td>
<td></td>
</tr>
<tr>
<td>V31 Willingness</td>
<td>.778</td>
<td></td>
</tr>
<tr>
<td>V32 Keeping informed</td>
<td>.565</td>
<td>.554</td>
</tr>
<tr>
<td>Emergency Services</td>
<td></td>
<td></td>
</tr>
<tr>
<td>V28 Quick response</td>
<td>.838</td>
<td></td>
</tr>
<tr>
<td>V29 Arriving within</td>
<td>.814</td>
<td></td>
</tr>
<tr>
<td>V30 Diagnose and resolve</td>
<td>.787</td>
<td></td>
</tr>
<tr>
<td>V31 Willingness</td>
<td>.801</td>
<td></td>
</tr>
<tr>
<td>V32 Keeping informed</td>
<td>.817</td>
<td></td>
</tr>
<tr>
<td>Call Handling</td>
<td></td>
<td></td>
</tr>
<tr>
<td>V25 Promptness</td>
<td>.750</td>
<td></td>
</tr>
<tr>
<td>V26 Problem handling</td>
<td>.768</td>
<td></td>
</tr>
<tr>
<td>V27 Scheduling service</td>
<td>.706</td>
<td></td>
</tr>
<tr>
<td>Account Reps</td>
<td></td>
<td>.922</td>
</tr>
<tr>
<td>V14 Technical knowledge</td>
<td>.767</td>
<td>.833</td>
</tr>
<tr>
<td>V15 Keeping in touch</td>
<td>.814</td>
<td>.869</td>
</tr>
<tr>
<td>V16 Arriving when</td>
<td>.812</td>
<td>.876</td>
</tr>
<tr>
<td>V17 Timely quotes</td>
<td>.813</td>
<td>.826</td>
</tr>
<tr>
<td>V18 Listening</td>
<td>.814</td>
<td>.866</td>
</tr>
<tr>
<td>V19 Proposals</td>
<td>.728</td>
<td>.854</td>
</tr>
<tr>
<td>Products</td>
<td></td>
<td>.857</td>
</tr>
<tr>
<td>V10 Product quality</td>
<td>.783</td>
<td>.932</td>
</tr>
<tr>
<td>V11 Product dependability</td>
<td>.796</td>
<td>.953</td>
</tr>
<tr>
<td>V12 Product innovativeness</td>
<td>.751</td>
<td>.950</td>
</tr>
<tr>
<td>V13 Availability of parts</td>
<td>.705</td>
<td>.920</td>
</tr>
<tr>
<td>Price</td>
<td></td>
<td>.787</td>
</tr>
<tr>
<td>V39 New system prices</td>
<td>.824</td>
<td>.924</td>
</tr>
<tr>
<td>V40 Replacement parts</td>
<td>.888</td>
<td></td>
</tr>
<tr>
<td>V41 System maintenance</td>
<td>.784</td>
<td>.877</td>
</tr>
<tr>
<td>Renewal Services</td>
<td></td>
<td>.762</td>
</tr>
<tr>
<td>V37 Renewal process</td>
<td>.841</td>
<td>.843</td>
</tr>
<tr>
<td>V38 Executing renewal</td>
<td>.818</td>
<td>.838</td>
</tr>
</tbody>
</table>

* Factor loadings across US and Japanese samples using principal axis-factoring as the extraction method and varimax with Kaiser normalization as the rotation method.

Please cite this article in press as: Khan, M.S., et al. The problem with standardizing international market research: A case study from B2B services. Australasian Marketing Journal (2014), http://dx.doi.org/10.1016/j.ausmj.2014.03.001
All 30 items were entered into EFA by using principal-axis factoring method for both datasets. For the US data, six factors emerged. Nine items loaded on the technicians factor, three items each on call handling and price factors, six items on account reps, four items on product, and two items on renewal services, but none on the emergency services factor. In Japan, six factors also emerged. Eight items loaded on technicians, six items on account reps, five items on emergency services, four items on product, three items on price, two items on renewal services factor, and none on the call handling factor were identified. While six factors emerged for each dataset, they were not the same factors. Since call handling emerged only in the US, and emergency services only in Japan, these factors lacked cross cultural equivalence and were removed from subsequent analysis.

There were three items in each country that failed to load on any factor. In the US, V20, V28, and V29 were weak and did not significantly contribute to the model. Hence, these three items were eliminated from further analysis. In Japan, V25, V27, and V32 were eliminated (Table 1). All six of these six items were eliminated in order to re-distribute and spread out the variance to those factors that had an Eigen value of equal to or greater than one (Hair et al., 2006).

In addition to this, three items were cross-loaded on constructs in the US and Japan (V26, V30, V31). Specifically, these three items loaded on different factors in the US and Japan. Hence, they were removed from subsequent analyses. These differences in item loadings indicated that some conceptual differences existed across countries.

Of the six factors that emerged from the two EFAs, four factors initially demonstrated full cross-cultural equivalence in terms of item composition across both countries. These four factors were account reps, products, price perceptions, and renewal services. These four factors included 21 items across the two datasets, and hence no cross-cultural differences emerged. However, the factor loadings were higher in Japan for all 21 items comprising these factors, indicating a better fit for the Japanese data.

The technicians factor had partial cross-cultural equivalence in the EFAs in terms of item loadings. For the US, the technicians factor had loadings on nine items. However, of the nine items in the US dataset, only six items loaded in the Japanese dataset. In addition to the six items, two other items also loaded on the technicians factor in Japan. Therefore, for the technicians factor, a significant, but not identical, overlap of items seemed to exist across two cultures. The technicians factor had partial cross-cultural equivalence.

Based on a priori model of product and service quality, we expected both emergency services and call handling to emerge as factors in both countries. The lack of conceptual equivalence suggested that the US and Japanese respondents conceptualize these constructs differently. Therefore, the account reps, products, price, renewal services, and technicians factors advanced to the CFA analyses.

Based on these results, we can say that H1a, which hypothesized that there will be differences between the US and Japan in the factors emerging from EFA, was supported. While four factors were consistent across both countries, each country had a factor that did not emerge in the other country. There is also support for H1b, that there will be differences between the US and Japan in the item composition of the EFA factors. The technicians factor had some differences, and call handling and emergency services were completely different.

5.2. Confirmatory factor analysis

Our next goal was to validate the final EFA model by using confirmatory factor analysis (CFA). We used CFA to establish cross-cultural equivalence of the refined constructs. Specifically, we wanted to see if the factors (account reps, products, price, renewal services, and technicians) emerged similarly across two different cultures when only these constructs were examined. In essence, CFA was used to refine and validate the items and factors that emerged from the EFA of the US and Japan datasets. Two separate CFA models were run, one for the US and another for the Japanese dataset.

Initially a series of CFA models were estimated using the US and Japanese data samples. The series included two separate uni-dimensional CFA models in which all items loaded on one factor and two separate correlated five-factor CFA models across two samples (Anderson and Gerbing, 1988; Bearden et al., 2006). Comparisons of these models revealed a significant chi-square difference suggesting that the five-factor correlated model was the best representation of the data ($\Delta \chi^2 = 2944.869, \Delta df = 10, p < .05$).

In particular, when performing the CFA, and we restricted each item’s loading to its a priori factor structure, and allowed each factor to correlate with all other factors. We found that the CFA model fit both the US and Japan data well. The measurement model was evaluated for uni-dimensionality, reliability, as well as for convergent and discriminant validity. These results of the correlated five-factor CFA are shown in Table 2. Overall, the fit statistics of the US and Japan analyses were good and acceptable (Byrne, 2001).

Convergent validity of the measurement model was examined by three measures: item reliability, composite reliability, and average variance extracted (AVE) (Fornell and Larcker, 1981). In both samples a total of six item loadings were higher than 0.60, the remaining were higher than 0.70 indicating that each measure is accounting for 50% or more of the variance of the underlying latent variable (Hair et al., 2006). Further, all factor loadings across the two measurement models were significant ($p < 0.001$) with t-values ranging between 13.873 and 31.619. Table 3 shows item loadings, composite reliability for all factors (ranging from 0.763 to 0.958), and the AVE as recommended by Byrne (2001).

For constructs to achieve adequate discriminant validity the square root of the AVE for each construct should be larger than the correlation between the construct and any other construct in the model (Fornell and Larcker, 1981). Therefore consistent with the guidelines of Fornell and Larcker (1981) and Baggozzi and Yi (1988) the AVE for each measure in both datasets exceeded 0.50. In particular, Table 4 shows that in each case the variance established by the construct was larger than its both average and maximum shared variance with other constructs thus indicating discriminant validity in the constructs for both samples. Therefore, we concluded that our measures were valid and reliable for further theoretical testing.

There were some significant differences in terms of factors and their loadings in CFA (Table 3). Three factors (products, price, and renewal services) included the same items in both the US and Japan datasets. The data did not generate significantly different factor loadings on these three constructs for different cultural contexts. These three factors had full cross-cultural equivalence. There were significant differences for other two factors, however.
There again appeared to be partial cross-cultural equivalence for the technicians factor. In particular, six items were included in the US and Japan CFA technicians factor that emerged from the EFA. In CFA for the Japan dataset, two items (V9 and V33) were deleted due to weak loadings from the analysis. Therefore, the final technicians factor included only four items in Japan, and six items in the US.

A similar finding appeared for the accounts reps factor, which consisted of the same six items in the US and Japan from the EFA. However, in the CFA, this was not the case for the Japanese respondents. Two items of the account reps factor were deleted in the Japan CFA (V15 and V16) from further analyses due to standardized residuals level. Hence account reps emerged consisting of only four items. All four of these items also appeared in the US CFA account reps factor. This pattern suggests, that although both the technicians and accounts reps factors did emerge as independent factors, their item content across two cultures was not entirely the same.

Based on the CFA analyses, H2a, that there would be differences between the US and Japan in the factors that emerged from the CFA, is rejected. The five factors were the same. This suggests that the EFA, alone, was adequate to identify the cross-cultural equivalence of the various factors. However, H2b, that there would be significant differences in the item composition of the factors, is supported. Two factors, account reps and technicians had significant differences in their composition.

6. Discussion

The goal of this research was to determine the appropriateness of the use of a fully standardized research effort by a large MNC. Specifically, we wanted to see if there were cultural dimensions that would influence the results, making direct cross-cultural comparisons inappropriate. Clearly, the standardized approach was not appropriate.
Of the 30 items that measured product and service quality in the standardized questionnaire developed and used by the MNC, about half (17) had cross-cultural equivalence between the US and Japan. Nearly half of the survey items either did not load on the factors or the factor, itself, was inappropriate cross-culturally. Since this MNC used the same base questionnaire in Europe, Asia, Australasia, North America, and South America, it is quite likely that even more items would lack cross-cultural equivalence if more countries were compared.

From our research, a blended standardized and adaptation approach seems most appropriate. Three of our constructs (products, price, and renewal services) had substantial conceptual cross-cultural equivalence since the same individual items loaded onto the constructs in roughly the same way in both countries. These three constructs could be standardized across both countries. The account reps construct had substantial equivalence with four of six items consistent across both countries. The technicians construct had partial equivalence since, of six items in the US, only four loaded in Japan. These two constructs should be slightly adapted for each country. The emergency services construct was appropriate for Japan, but not for the U.S. The call handling construct was appropriate for the US, but not for Japan. Therefore, each of these constructs was appropriate for one country, but not the other.

Our results illustrate the need to establish cross-cultural equivalence before making comparisons across cultures or countries. Fully standardized worldwide research, at least on the concepts addressed here, appears to be inappropriate. Of the seven original constructs of interest, only three had full equivalence, two had partial equivalence, and two had no equivalence.

If academic researchers wanted to test a theoretical model cross-culturally, they would first need to establish, which items had full, partial, or no conceptual equivalence before proceeding with further analysis. If an MNC wanted to make cross-cultural comparisons, which was the situation here, they should compare only the constructs that had full or partial conceptual equivalence. Comparing items or constructs that lack cross-cultural equivalence would be meaningless, much like comparing apples and oranges.

This study contributes to the literature on international research in a B2B services context. The study reported here is a contribution to the area because it has focused on actual B2B customers in two diverse populations. However, generalization of our study to other B2B situations and to wider variations of customer satisfaction research must be made with caution. A key task for marketing researchers and practitioners should be to determine if assumptions and models developed and established in a single country are appropriate in another country.

7. Conclusions

The research presented here bridges the practitioner/academic divide. Specifically, the data came from a very large MNC that operates in well over 100 countries. Because the company was in B2B markets, they explicitly assumed business customers were pretty much the same in all countries. Therefore, the VP in charge of research gave explicit directions to standardize the worldwide research, despite warnings by the market research firm. We suspect that other MNCs have made the same decision. As Hult et al. (2008) also noted, the vast majority of academics apparently make the same assumption, and mistake. Most international research fails to test for cross-cultural equivalence of the questions, categories, and response scales. Failing to test for this can lead to erroneous conclusions from the data. For example, call handling performance is important in the US, but is unimportant in Japan. Using the construct as a driver of customer satisfaction is simply inappropriate in Japan. Our results show that nearly half of the items typically used to measure product and service quality in B2B relationships simply do not travel well cross-culturally.

8. Limitations

One major limitation of his study is that it focused on just two countries, the US and Japan. Similar differences could exist between other countries. Investigating the response tendencies and conceptual differences from country to country would be important for a global company. Simply assuming that “customers are customers” and “data is data”, as done by the MNC in our study, is very ethnocentric and could lead to suboptimal foreign performance and difficulties in performance evaluation.

While we have noted that significant differences exist in the factors and their composition between the two countries, we have not examined what caused the differences. For example, we have not looked into respective firm’s marketing mix strategies in both countries. Nor did we analyse the specific dimensions of culture that may have caused the differences. To fully understand the meaning of data, researchers should also understand what is causing the differences. Likewise, understanding why the conceptualizations of the questions and constructs differ would improve our knowledge of the cultural differences.

References


