Summer 2004

An Experimental Examination of Equivalence Failures in Multi-Cultural Comparative Research

F. Mark Case
Old Dominion University

AN EXPERIMENTAL EXAMINATION OF EQUIVALENCE FAILURES IN MULTI-CULTURAL COMPARATIVE RESEARCH

by

F. Mark Case
B.A. 1990, Grand Rapids Baptist College
MBA 1996, Grand Valley State University

A Dissertation submitted to the faculty of
Old Dominion University in partial
fulfillment of the Requirements for the
Degree of

DOCTOR OF PHILOSOPHY

BUSINESS ADMINISTRATION

OLD DOMINION UNIVERSITY
August, 2004

Approved by:

John B. Ford, IV (Chair)

Earl D. Honeycutt, Jr.

Edward P. Markowski

ABSTRACT

As the global economy integrates, there is an increased need to understand international business phenomena. This forces a reliance upon multi-cultural marketing research, which is evidenced by a marked increase in international studies, particularly multi-cultural comparative research (Sin, Cheung & Lee 1999). Central to the usefulness of this research is the question of the validity and comparability of results, which is greatly affected by a property known as equivalence of test instruments; or the degree to which the scales and the items in them are seen to be the same across cultures.

To date, no research exists that approaches the problem of equivalence from a position of knowledge, or beginning with a known equivalence error and then tracing its psychometric effects. This dissertation fills that need by experimentally manipulating a translation error in a scale and then using conjoint analysis to decompose a respondent's choice patterns for items or attributes that cause equivalence failures.

Results from a probability sample of American consumers indicate that: 1) current techniques to diagnose equivalence failure can adequately identify items that are inequivalent, but also 2) that items failing equivalence have a pronounced tendency to attenuate other items in the scale. Conjoint results were similarly affected by translation error. This presents serious implications for international researchers and global marketing managers, including some question as to the usefulness of existing scales in multi-cultural contexts. In addition, theoretical development regarding response behaviors is needed to explain the differences between control and experiment groups with respect to non-manipulated items. Further research is also needed to systematically examine the effect of translation error across response formats and scale types.

AN EXPERIMENTAL EXAMINATION OF EQUIVALENCE FAILURES IN MULTI-CULTURAL COMPARATIVE RESEARCH

DEDICATION

This dissertation is dedicated to my daughters, Anna M. Case and Julie A. Case, who have unknowingly sacrificed a great deal for their father's success. Throughout the entire process, they have shown great courage, loyalty, good spirits, and most of all, love. I admire both of them for their ability to handle difficult circumstances with grace – to say I am fortunate to have them as daughters is an understatement. This dedication also comes with a statement of obligation on my part – to be as supportive of them in the pursuit of their goals as they have been for me.

In addition, this dissertation is dedicated to my Grandmother, A. Margaret Case. Without her support, I would not have been able to complete a doctoral degree. One statement she made to me has forever stuck in my mind, and I will live by it until my last days. She said, "it is alright if your goals change from time to time. It's *not* alright to not have goals." In following my life through various goal changes, she has been generous with her time, insight, emotional support, and even her finances. Evening phone conversations about everything from religion to university politics to jazz music have been a godsend to me, and the value of being able to talk honestly and openly with her about all of life's offerings cannot be understated. She has also been my most vociferous cheerleader, even at times when others had criticized me for being too hard-headed about grad school. Thank you, Grandma, for this degree.

ACKNOWLEDGEMENTS

It would be a real faux-pas to neglect acknowledging certain individuals who have been a godsend to me as I wrestled with grad school. In no particular order, I wish to have my sentiments permanently recorded regarding the following people:

My committee – John B. Ford, Earl Honeycutt, and Ed Markowski, who have all been excellent examples of what a faculty member should be. Each has exhibited leadership by example – by being competent, understanding, and encouraging. More specifically, John Ford and Earl Honeycutt have set high expectations for their students while simultaneously instilling a dread fear of falling short of these expectations. Some may say that this isn't much of a compliment, but I think that it is the highest compliment a professor can receive from a former student. I have learned much and prevailed in a program few enter and even fewer complete because of this. Thank you, gentlemen for your work, and I will strive to be sure that your faith has not been misplaced.

My "partner in crime," Tarek Mady, who wasn't quite sure if he would pursue the Marketing track at first. I'm so very glad he did. We sweated the comps together, colluded in the never-ending political ad experiment, shared the joys and frustrations together, and generally lived life in friendly competition during grad school (by the way, I think I'm still up two points). In fact, he has gone so far as to adopt this quasi-Christian into his family, and I will always be proud and grateful for my association with him.

My friend Yuhong Fan, who is someone I could always count on for good advice and encouragement. I've learned from her that there are benefits to being an optimist, and she has earned my respect by being level-headed and easy to talk to.

Father David LaSalle, who brought me into the Episcopal Church, and who has been a shining example of the good that people can be, if they really try. Somehow, someway, a gigantic bottle of single malt is headed your way (but don't tell Sheron!).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

Chapter 1

INTRODUCTION

STATEMENT OF THE PROBLEM

Part 1

"Paradoxical though it may seem, what we stand to gain most from comparative studies is not what they tell us about others, but what they force us to learn about ourselves in order to understand what we see abroad" – Reavis Cox (1965)

International businesses is said by many to be increasing in its importance. This is evidenced by the fact that a growing amount of manufacturing activity rests upon the relationships with international suppliers (Czinkota, Ronkainen, & Moffett, 1999). As the global economy integrates, there is an increased emphasis on the importance of understanding international business phenomena. As such, there has been a dramatic increase in the volume of academic literature that has been published in international research (Sin, Cheung, & Lee, 1999). One concern of multinational researchers is that even though there is an increase in the need for understanding of multinational and multicultural phenomena, the methods used to investigate these phenomena have not advance significantly (Sin, Cheung, & Lee, 1999). This problem has been in existence since multinational research began, and was highlighted by such researchers as Green & White (1976)

In marketing, researchers face distinct challenges when attempting to understand differences between heterogeneous groups of consumers in different cultural contexts.

The understanding of such constructs as image (e.g., Presley, 1972) is complicated when two basic research challenges are evident: 1) the failure to understand the interplay between latent constructs that may provide valuable insight for multi-national managers; and 2) the fact that measurement scales designed to capture these latent constructs do not take into account cultural influences that potentially distort the meaning of the items in the scale.

Difficulties not withstanding, multicultural studies have attempted to understand the nuances of many phenomena in comparative marketing (Boddewyn, 1981), which is defined as, "the systematic detection identification classification measurement and interpretation of similarities and differences among entire national systems or parts thereof" (p. 61). In keeping with Boddewyn's (1981) general criticism, multicultural researchers tend to measure attitudes with respect to these attributes using traditional research techniques developed in American market settings. These techniques normally involve a number of established scales, typically with Likert-style responses, which may or may not provide the information needed to develop accurate analyses. Attempts to gain meaningful knowledge from these techniques are frustrated in a multicultural setting by various influences rooted in cultural differences (Malhotra, Agarwal, & Peterson, 1996).

These influences have a serious impact upon a researcher's ability to make sense of data collected in international and multicultural markets. Of principal concern to researchers, which has been well documented, are validity and reliability. Reliability is concerned with the consistency of the data (Hair, et al., 1998), and validity is the degree to which research test instruments measure what they are supposed to measure (Carmines

& Zeller, 1976). These problems have been noted in numerous streams of research, including psychology (e.g., Nevid, & Spa Maria, 1999, Weinfurt & Moghaddam 2001), management (Lenartowitcz & Roth 1999, Cheung & Rensvold 1999), anthropology (Ember & Ember 2000), sociology (Arts & Halman 1999), public health (Small et al., 1999), marketing (e.g., Malhotra, et al., 1996; Green & White, 1976; Albaum & Peterson, 1984; Boddewyn, 1981; and Clark, 1990), business ethics (McDonald 2001), and political science (Lancaster & Montinola 2001). Early work that highlights difficulties in quantifying multicultural research notes that fundamental differences between cultures in human behavior make the job of quantifying marketing phenomenon extremely challenging. These differences include culturally based response styles, sampling errors, and various construct equivalences that attenuate the results.

The primary question addressed by this work is how international research can be improved; namely to be conducted in such a way as to detect and reduce the effects of equivalence failure. The importance of this issue cannot be understated, especially with regard to trying to assess real and supposed differences between cultural environments. The problem was well stated by Mullen (1995), as follows:

> "A fundamental, unresolved issue with multinational research is whether similarities or differences are in fact real [Barksdale & McTier-Anderson 1982]. If results are different than expected (that is, statistical significance is not achieved, items do not load in factor analysis as expected, or reliability assessment is low), researchers (e.g., Adler, Campbell and Laurent [1989]) often question whether measurement problems inherent in international research have attenuated the results, that is, whether the results are measurement and scaling artifacts or true cultural differences. Cross-national researchers must tackle the hard issues of measurement equivalence in order to reduce the threats to measuring reliability and validity (e.g., Adler et al. [1989]; Albaum and Peterson [1984]; Davis, Douglas and Silk [1983]; Nason [1989]; Aulakh and Kotabe [1993]). For instance, Aulakh and Kotabe [1993] recently noted that a major reason for lack of attention to equivalence issues is the insufficiency of existing

> techniques, and this "methodology issue is one area in need of immediate attention to make international research more rigorous" [Aulakh and Kotabe 1993, p. 24]" (p. 574).

This dissertation is designed to develop and test a technique for capturing and understanding with greater accuracy the effects of equivalence failure in multi-cultural comparative contexts. Namely, an experiment to manipulate purposeful equivalence failure will use a combination of existing procedures – multi-group structural equation modeling and conjoint analysis – to understand with greater clarity the effect of translation equivalence failures. In addition, a diagnostic tool for multi-cultural researchers will be proposed based upon the results of the experiment that will enable researchers to clearly identify major equivalence problems and

The purpose of this chapter is to establish the groundwork, justification for, and description of this research. As such, the remainder of this chapter is divided into three sections; the first establishes a working definition of the various equivalences. The second provides the justification for this dissertation by showing gaps in current research, and recommending a means to fill these gaps. The third section describes the approach that is used in the design and execution of the project.

BACKGROUND

As noted earlier, Boddewyn (1981) has been critical of international marketing researchers because of a general lack of comparability in the data. While some techniques have been developed over time to help with the problem, criticism still rings in the marketing literature (Malhotra, et al., 1996). In other words, twenty years have passed with no serious attempt to suggest improved methods for cross-cultural discovery.

Leading researchers point to equivalence failure as a major obstruction to both the validity and reliability of cross-cultural comparisons (e.g., Malhotra, 1996; Cavusgil & Das, 1997; Green & White, 1976; Mullen, 1995). Equivalence in multi-cultural research refers to the degree to which a construct or measurement instrument is seen to be the same across cultures. Malhotra, et al. (1996) provide a typology of multi-cultural research equivalences, classifying them into four distinct groups: functional equivalence, conceptual equivalence, instrument equivalence, and measurement equivalence (p. 19). In addition, there may be low literacy levels which obstruct respondent's ability to understand items in questionnaires.

Definitions

Functional Equivalence

Functional equivalence refers to the degree to which a given phenomenon relates to the same basic behavior, or assumes the same function between cultures (Malhotra, et al., 1996, p. 19). There is a widely used example of bicycles in China serving a fundamentally different function than those in the United States – which for multicultural comparative researchers may or may not present difficulties, depending upon the goal of the research. For example, if a researcher is interested in a question of potential differences between one population and another with respect to the functional role of an object or a concept, then functional equivalence is an item of discovery, and its existence is not required to gain valid data and conclusions. If, however, the goal of the research is to understand nuances of preferences vis-à-vis product or service features in a multi-cultural context, then functional equivalence is probably an assumption underlying the

research design. So, in the previous example, if the bicycle manufacturer wanted to discover Chinese attitudes about certain sports performance features of its products, the results would likely be spurious and not of genuine use to the researcher, due to the fact that usage of the bicycle is so completely different.

Conceptual Equivalence

Conceptual equivalence is defined by Malhotra, et al. (1996) as "whether the concept or construct is expressed on similar attitudes or behaviors across cultures" (p. 20)). Malhotra (1988) provides an excellent example of potential conceptual equivalence failure regarding marketing activities in developing economies. It is probable in these regions that the economic environment is one in which marketing is not a real concern, as the economy is driven primarily by manufacturing, and demand far exceeds supply making marketing related phenomena of relatively little interest. Obviously, in an environment such as this, personal sales to consumers or sales promotions are not likely to be viewed in the same light as they would be in more developed markets such as the U.S. or Western Europe.

Instrument Equivalence

Instrument equivalence refers to whether the test instrument or experimental treatment is interpreted the same across cultures. In the Malhotra, et al. (1996) typology, this is ascribed to a three types of equivalence influences; calibration equivalence, translation/linguistic equivalence, and scalar equivalence. Calibration equivalence refers to units of measure (the classic example is of temperature, or metric versus imperial

measurements). Translation equivalence is obviously whether the language of the test instrument is understood by the respondent. According to Malhotra et al. (1996), scalar equivalence refers to "whether the psychometric properties of the data from various cultures exhibit the same coherence or structure" (p. 20; c.f. Bhalla & Lin 1987). In other words, the researcher is interested in ensuring that the test instrument captures true impressions by virtue of the fact that respondents in different cultures will be prone to respond the same way, given that attitudes match and all other equivalences are satisfied.

JUSTIFICATION

It would appear from several researchers that since equivalences are a major problem (e.g., Green & White 1976; Boddewyn 1981; Sekaran 1983; England & Harpaz 1983; Mullen 1995; Malhotra, Agarwal, & Peterson 1996; Malhotra, Peterson, & Kleiser 1999), issues that need to be dealt with directly are not necessarily psychometric, but instead could be input related. While Mullen (1995) has described in detail the benefits of metric solutions to equivalence problems – specifically multi-group structural equation modeling and optimal scaling – these appear to be metric solutions to problems that may be either conceptual in nature or rooted in cultural or translation difficulties. More specifically, with functional and conceptual equivalence failures, the problem is not with the data, but with the input that generates the data. Mathematical methods of dealing with equivalences and response styles are certainly helpful (e.g., Clarke 1996), but true comparisons can only be made if the input is understood to truly reflect the views of the respondents.

In a multicultural context, this is extremely challenging. As Boddewyn (1981) noted, many cross-cultural researchers make use of techniques, especially scales, developed in the U.S. and exported to foreign research environments. When these scales are used in foreign markets, and there is a serious concern regarding equivalences, each scale item represents an opportunity for some sort of separation between the interpreted response (researcher analysis) and the actual intended response (subject attitudes). Hence, for a scale that contains twenty-five items, the probability of equivalence error (metric or otherwise) is multiplied by twenty-five, which constitutes a potentially serious threat to both validity and reliability. Techniques designed to assuage the problems presented by metric equivalence can then be seen as a *ex-post-facto* repair to a problem that originates in the collection of the data, not in the data itself (Burns & Bush 2002).

An example would be a test instrument containing descriptions that are not clearly understood by the respondent, who then is expected to metrically evaluate the item. One would expect that the data for a representative sample will have different psychometric properties to other scale items or constructs in the test instrument. Researchers may then be tempted to use some type of metric technique after data collection to transform the variables and turn them into statistically comparable scale items. Obviously, any analysis of such an item would be misleading. Even more difficult to deal with is when the scale item is not understood, and respondents consistently demonstrate a "neutral" position or preference on the item, which may be normally distributed, and the equivalence error goes undetected.

This research proposes a procedure to identify and analyze equivalence error in multi-cultural comparative studies by using the strengths of conjoint analysis to

decompose respondent's choice patterns for items or attributes that cause equivalence failures, and comparing these results with those of a structural equation model commonly used with scale items.

DESCRIPTION OF APPROACH

To examine the effects of equivalence failure in multi-cultural research, an experiment in cross-cultural advertising will be used. Advertising is chosen because of its importance in global marketing (citation), and also because it is particularly amenable to experimentation (citation), specifically with regard to translation errors. In this context, a translation equivalence failure will be purposefully introduced in a controlled experiment, which will be expected to produce a failure in metric equivalence (Malhotra, Agarwal, & Peterson, 1996). To accomplish this, sample selection will be conducted in cultures that are very similar, yet different enough to allow for a misunderstanding of manipulated items. The proposed analytical technique is in actuality a combination of existing techniques: conjoint and structural equation modeling. This proposal is based upon several unique strengths of conjoint that make it particularly well suited to multicultural comparative studies. First, the input is simple. Rather than assuming the risk that foreign respondents will understand/misunderstand a scale item in a survey, the respondent observes a card with the treatment on it, and provides input as an overall response. The advantage in this is that scale items are less likely to be misunderstood, and metrically misrepresented. In addition, one second advantage is that conjoint experiments can be established so that the items designed to be manipulated can be presented in a number of different ways; either with a typed description, or pictorially

(e.g., Craig & Douglas, 2000). This way, a researcher may be able to manipulate testable items in a way that reduces the potential for consumer confusion.

Third, responses from conjoint can be clustered (or Q-type factor analyzed) to understand better which groups of respondents are truly homogeneous, which can be very useful if there are confusing or conflicting responses with regard to a given item. A researcher can then examine similarities between respondents whose input appears to be spurious and better understand contextual influences (e.g., culture) on a respondent's choice patterns.

Fourth, since conjoint is a theory-driven technique (Hair, et al., 1998), all possible interaction effects must be considered before designing the conjoint experiment. Thus, output from a well designed conjoint experiment may be useful to understanding the measurement model of a structural equation analysis, and ultimately, a more soundly designed structural model.

Given the advantages noted above, it would seem that conjoint analysis would be very useful in multicultural comparative research. However, conjoint is not a method that can reasonably be expected to answer all research questions. For example, one serious limitation to conjoint analysis is that one cannot test theoretical models, or relationships between latent constructs. A proposal to bridge this gap is to include in the design of a given study both a structural/measurement model and a conjoint experiment. The conjoint experiment may be of immense value in terms of understanding the equivalence (both conceptual and functional) of items used in SEM. SEM is necessary for in-depth theory testing.

In light of the advantages of conjoint analysis in a multicultural context, and the need for structural equation modeling for in-depth theory testing and development, it is proposed that a multicultural comparative study be conducted in such a way as to construct separate conjoint experiments in which each observed variable of a measurement model is manipulated in the conjoint experiment using a factorial design appropriate to its complexity. Using separate samples, a test instrument is used that is compatible with standard structural equation models; namely that each observed variable is represented as an item on the test instrument. In this way, each item is tested both by scale item (survey) and by exposure (conjoint). These results are compared both within cultural groups (between methods) and between cultural groups to gain a better understanding of the effects of equivalence failure. The goal of the experiment is to note differences between valid conclusions drawn from the conjoint experiment and valid (statistically significant) results from the scaled measurement model. It is hypothesized that major differences between the two will be due to manipulated translation errors.

The remainder of this dissertation is divided into four remaining chapters. The second chapter contains a review of relevant literature in multi-cultural comparative research and discusses equivalences in depth, as well as includes a review of literature in multi-cultural advertising. Chapter three describes the experiment design and analytical methodology used. Chapter four discusses the results of the experiment, and the final chapter presents arguments for the usefulness of this research and discusses limitations, with recommendations for future research.

CHAPTER 2

Introduction

As noted in Chapter 1, there has been a high level of interest in multi-cultural research, specifically as it applies to marketing over the past 25 years (e.g., Green & White 1976, Boddewyn 1981, Douglas & Craig 1983, Cheng 1989, Mullen 1995, Malhotra et al. 1996, Sin et al. 1999). In the past 10-15 years, there has been an explosion of editorial interest in multi-cultural comparative studies. An electronic search of the ABI database revealed that since 1990, 243 studies of all disciplines have been multi-cultural comparative studies. A manual search of leading marketing and international business journals produced another 42 studies. In the broadest sense, multi-cultural research has been the subject of investigation in many disciplines over an extended period of time. The multi-disciplinary nature of the field indicates the importance of cross-cultural research, not just to marketing, or even to business, but to the entire academic community. To that end, numerous studies have examined the methodologies used, the problems inherent in these methods, and suggestions for research agendas to potentially repair those problems.

Unfortunately, after almost three decades, few studies have offered repairs to the methodological shortcomings unique to multi-cultural research. With the exception of Mullen (1995), who introduced multi-group analysis for use in multi-cultural contexts, and Singh (1995) who recommended econometric means of repairing measurement inferential errors, there have been few developments. This review begins with the larger scope of multi-cultural research and its uses, including multi-disciplinary interests. Next, a review of research that details the complexity of conducting such studies, including

specific challenges of using these scales in multi-cultural settings, followed by prominent scales used in cross-cultural marketing research. The focus will then narrow to the problem of cross-cultural scale equivalences, including current methods of detecting and/or adjusting for equivalence errors. All of this will demonstrate that in the totality of multi-cultural literature, including studies devoted to the econometrics and/or psychometrics of multi-cultural scales, there is no work that clearly identifies an error of equivalence and details the resulting statistical properties of the failure. Conjoint analysis will then be reviewed and introduced as a potential tool for gaining great understanding of consumer choice behaviors in multi-cultural contexts, and how this may be helpful toward the understanding of the usefulness of scale items across heterogeneous groups.

Multi-Disciplinary Cross-Cultural Research

Since questions of culture, specifically similarities, differences and conflicts between cultures are rooted in human interaction, most of the literature regarding culture and multi-cultural comparisons is steeped in the social sciences and makes heavy use of research techniques that are designed to measure constructs by using scaled responses. Furthermore, these scales are typically developed in one cultural environment, then exported to another for use with conclusions rightly or wrongly derived from the data these scales produce. Of course, when researchers are looking to answer questions of a theoretical nature, scales must be used to capture latent constructs; yet using scales across cultures presents serious questions of *emic* versus *etic* interpretations. Emic refers to a belief that behaviors and preferences are unique to one particular culture, and can only be investigated within the context of that culture (Douglas & Craig 200). Etic refers to the

idea that some behaviors and preferences are universal and transcend culture (Douglas & Craig 2000).

By default, when researchers compare one culture with another, there is an implied element of *emic* discovery, as comparisons between cultures can only be reasonably established if there are enough differences between cultures that the phenomenon of interest is distinct within each culture. An example would be discovering different propensities to portray women in advertising, depending upon cultural influences (Ford et al. 1998). On the other hand, multi-cultural research many times contains an inherent *etic* element. To compare cultures along theoretical dimensions, there is an assumption that there are universals in terms of the constructs being used. In other words, one cannot use a single theory to explain behaviors in heterogeneous groups of respondents if the constructs themselves are not culturally invariant. For example, to understand consumer perception of service quality, the theoretical constructs that comprise the theoretical framework must be understood by *all* of the respondents equally, otherwise any comparison to understand *emic* phenomenon will not be valid. This is what Douglas & Craig (2000) refer to as the "emic-etic dilemma" (p. 153).

Academic attempts to draw conclusions from multi-cultural research make for a very broad stream of literature. Disciplines represented include Anthropology (e.g., Ember & Ember 2000, Zao 2001, Freestone & Murphy 1998, Himes 1994), Education (e.g., Awasthi et al. 1997, Stone et al. 1996), Business Ethics (e.g., Cherry et al. 2003, Hirsch et al. 2003, Tsalikis et al. 2002, Polonsky et al. 2001, Kennedy & Lawton 1996), Communications (e.g., Sauer 1996, Deuze 2002, Nixon & Dawson 2002, Hampton & Emerson 2003), Information Systems (e.g., Peterson et al. 2003,, Palvia & Hunter 1996,

Igbaria et al. 1996, Mejias et al. 1996, Bryan et al. 1995), Public Health (e.g., Henderson & Ainsworth 2003, Fuhrer et al. 2002, Spira et al. 1998, Turnipseed & Turnipseed 1997) Psychology (e.g., Thu et al. 2002, Chu et al. 1999, Park 1998, Moore 1998, Schneider 1996), Sociology (e.g., Arts & Halman 1999, Lee 2001, Weber et al. 1998, Tomoko & Banks 1997, Scott et al. 1996, Crompton 1996), Political Science (e.g., Lancaster & Montinola 2001, Tuohy 1994), and Public Policy (e.g., Green 1998, Bond 1996, May & Burby 1996).

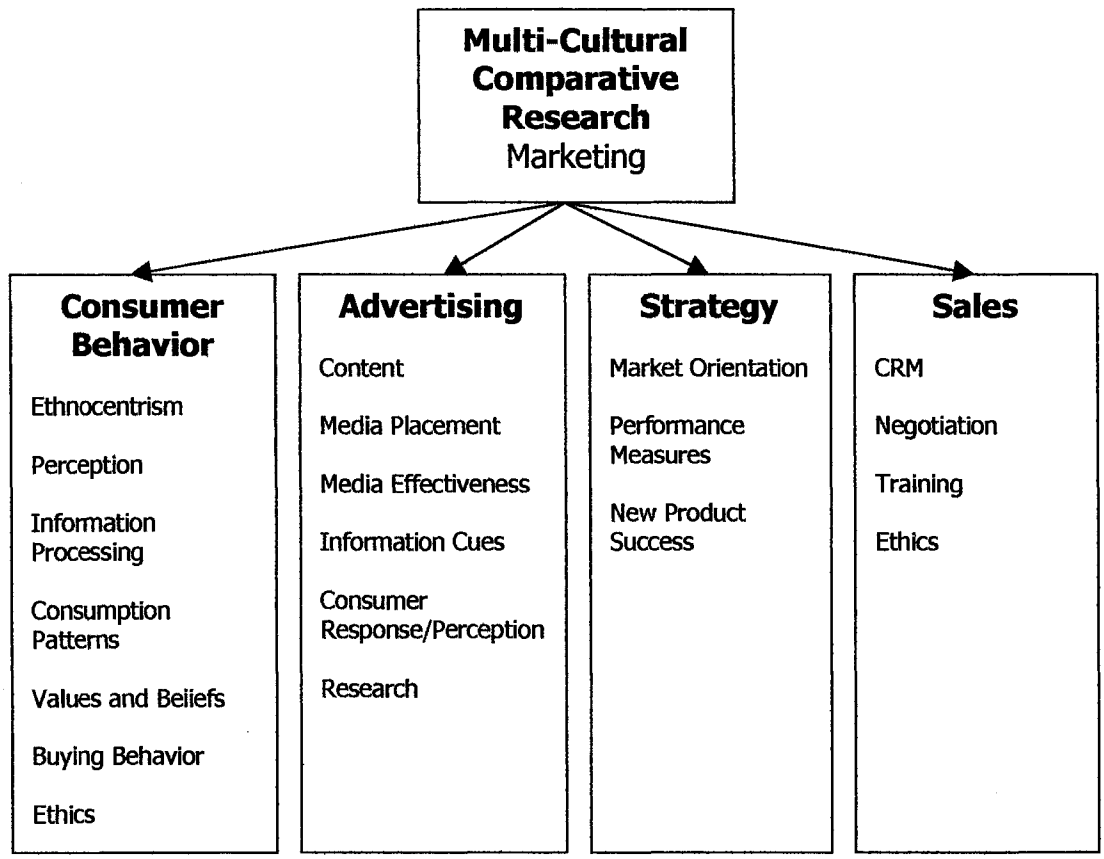
In addition to the disciplines listed above, multi-cultural comparative research has experienced increasing attention in the business literature, with Marketing and Management representing a heavy majority of the studies. Multi-comparative literature in Management clearly shows distinct streams in such areas as Organizational Behavior (e.g., Cherry et al. 2003, Groschl & Barrows 2003, Fraser & Fraser 2002), Human Resources (e.g., Pines 2003, Ho et al. 2002, Oliver & Cravens 2001, Smulders et al. 1996), Operations Management (e.g., Yen et al. 2002, Dayton 2001), and Strategic Management (e.g., Antoncic & Hirsch 2001, Johnson et al. 2001).

Multi-Cultural Marketing Research

Within the business literature, marketing by far represents the largest body of multi-cultural studies, with several distinct streams. Figure 2-1 shows a typology of multi-cultural studies within the marketing discipline, and clearly demonstrates that one of the largest areas of interest is consumer behavior. Given the interpersonal nature of culture, this is not surprising, and the field of consumer behavior presents a rich array of research topics including consumer ethnocentrism (e.g., Yu & Albaum 2002, Supphellen & Gronhaug 2003), consumer perception [not ethnocentrism] (e.g., Hui & Au 2001,

Hulland 1999), information processing (e.g., Liefeld et al. 1999), consumption patterns (e.g., Gnepa & Petrosky 2001, Eastman et al. 1997, Darley & Johnson 1993), and consumer values and beliefs (e.g., Alpert et al. 2001, Kropp et al. 1999, Mathur 1998).

Figure 2-1
Multi-Cultural Comparative Research Streams in Marketing



Consumer ethnocentrism is a phenomenon whereby consumers make judgments regarding the appropriateness of purchasing goods or services through foreign marketers, and has received much attention, primarily due to the development of the CETSCALE

(Shimp & Sharma 1987). Though few of these studies are multi-cultural *comparative* studies (by virtue of the fact that many of these studies focus on one specific culture and do not make direct comparisons between two or more cultures), what is striking about this line of enquiry is that the concept of ethnocentrism in and of itself carries with it an innate comparison of cultures. In psychological terms, it is difficult to envision a consumer making ethnocentric assessments without first making some comparison between the indigenous culture and the foreign culture in question. Therefore, each ethnocentric study contains at least an element of cross-cultural comparison in the responses generated by the instrument of measure.

In addition to consumer ethnocentrism, other cross-cultural consumer behavior topics may include the marketers' relationship to consumers (e.g., Laroche et al. 2002, Hui & Au 2001), buyer behavior (e.g., Malhotra & McCort 2001, Carlson et al. 1999), and consumer ethics (e.g., Erffmeyer et al. 1999).

Aside from consumer behavior, one other prominent stream of cross-cultural comparative literature is advertising. This stream consists of six distinctive areas of research; content (e.g., Ford et al. 1994, Okazaki & Rivas 2002, Jeon et al. 1999, Pak 1999, Carlson et al. 1996), media placement (e.g., Gould et al. 2000), consumer response/perception (e.g., Bridges et al. 1996, Andrews et al. 1991), and advertising research (e.g., Andrews et al. 1994, Albers-Miller 1996).

Methods of discovery in multi-cultural advertising research include one method unique to understanding content in advertising that is arguably the most objective by virtue of the fact that cultural inferences are made by multiple researchers, and without relying upon input from respondents. In content analysis, ads from the media of interest

are collected, examined, and coded for analysis. Similarities and differences between ads from disparate sources are compared, with conclusions drawn regarding the meanings of these differences. In addition, content analysis requires the involvement of more than one person to code the ads, which is designed to remove as much as possible subjective influences that affect the rating or interpretation of the material, which objectifies the process of evaluation. With specific regard to multi-cultural comparisons, people coding the ads may be from disparate cultures, which also helps to minimize cultural biases, unlike survey data that must be examined in an attempt to identify and isolate culturally based sources of variation.

Marketing strategy also contains a respectable body of work, driven primarily by a heavy editorial focus on market orientation. A number of studies use the MARKOR scale (Kohli & Jaworski 1993) that was developed to test executives' perception of firm market orientation. This area of study also includes a number of articles testing competing market orientation scales in multi-cultural, multi-industry contexts (e.g., Mavondo & Farrell 2000). Another area of high interest, especially regarding multi-national firms is the operationalization of, and contributions to firm performance (e.g., Styles 1988). Somewhat related to this is new product success (e.g., Calantone et al. 1996).

The last distinctive literature stream in multi-cultural marketing is sales, in which researchers examine questions of training (e.g., Yunxia 2000), negotiation (e.g., Palich et al. 2002), customer relationship management (e.g., Kivetz & Simonson 2003) and sales ethics (e.g., Parker & Pettijohn 2003, Bellizi & Hasty 2003).

The Complexity of Multi-Cultural Research

As important as multi-cultural research has become, it also proves to be an incredibly complex undertaking. All of the research streams noted above, and all of the studies in them have by necessity dealt with a wide variety of issues that seriously impact the validity and reliability of the scales used. Craig & Douglas (2000) state the case well:

> . . . the principles of marketing research are the same whether research is conducted in an international or a domestic context. However, international marketing researchers encounter greater difficulties than their domestic counterparts. These difficulties stem from operating across national boundaries and in a diverse range of socio-cultural environments. Examples of issues that may arise include how to obtain response from illiterate or semi-literate populations, how to develop a sampling frame in the absence of reliable census data or sampling lists, or simply how to find or train competent interviewers. Frequently, creativity and resourcefulness are required in coping with unexpected problems. In addition, an ability to manage and deal with and organize researchers of different cultural backgrounds and value systems is essential to successful international marketing research (p. xvii).

In addition, Craig & Douglas (2000) describe a research environment complicated by such issues as trying to design research when it is unclear how to define the target population (e.g., country as a proxy for culture), difficulty in collecting data and organizing research efforts, and from the practitioner's perspective, the intra-firm political strife that inevitably comes with major multi-national decisions (pp. 15-17).

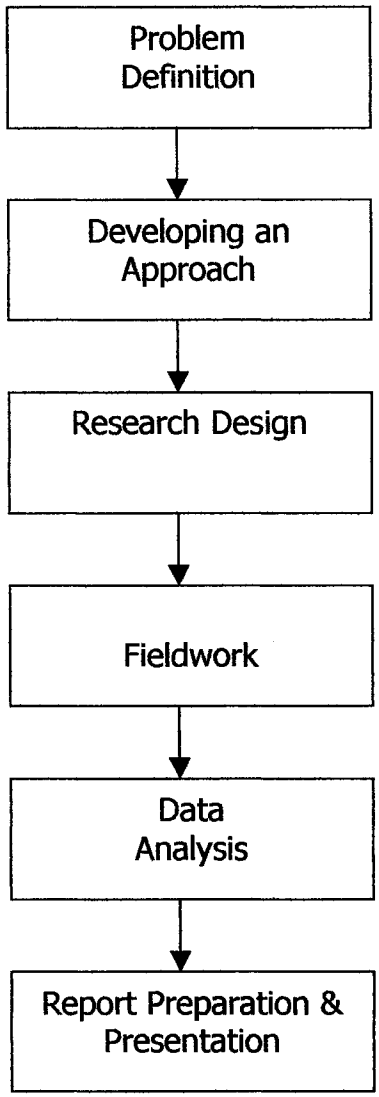
To manage this process, Malhotra et al. (1996) propose a step-by-step process to manage and control multi-cultural research, with the aim of minimizing as far as possible potential sources of bias and/or data contamination. Figure 2-2 contains a flow chart of recommended design procedures for multi-cultural research. The basic flow chart does not differ significantly from domestic research in that one begins with a definition of the research problem and ultimately concludes with report preparation and presentation, with

the steps in between being very similar. What is distinct in multi-cultural research is the meaning behind the descriptors used for each step, and/or the implications of trying to accomplish each step. The Malhotra et al. (1996) procedure contains six steps: problem definition, developing an approach, research design, fieldwork, data analysis, and report preparation and presentation. Others suggest more condensed conceptualization of the process; for example Cavusgil & Das (1997) classifies various steps of multi-cultural research into four broad categories – basic research design, sampling issues, instrumentation & data collection, and data analysis (pp. 73 & 74). For the purposes of this work, the Malhotra et al. (1996) framework is used because it is precise, yet parsimonious, and contains enough detail to clearly illuminate the challenge of multi-cultural research.

Problem Definition

Malhotra et al. (1996) maintain that problem definition in multi-cultural research is more complicated or involved than domestic research because the researcher must consider whether the phenomena of interest are comparable. This is supported by Douglas & Craig (2000) and Nasif et al. (1991), who add that the difficulty of defining the research problem also includes an element of criterion assessment – defined by the authors as a clear definition of culture and its various impacts. In addition, Malhotra et al. (1996) include an *a-proiri* judgment regarding *emic* versus *etic* research goals.

Figure 2-2
Flow Chart of Multi-Cultural Research Design



Source: Malhotra et al. 1996

Developing an Approach

According to Malhotra et al. (1996), approaches to multi-cultural research embody the philosophy of research. To this end, one could view research problems as anthropological, sociological, or psychological (p. 11). Other views include the degree of falsificationism where the researcher attempts to validate theory using stringent empirical techniques, or exploratory in nature which may involve more qualitative means of discovery.

Research Design

According to Malhotra et al. (1996), research design includes reliability and validity of secondary data, appropriateness of qualitative research, survey procedures, questionnaire design, and sample plans. While these issues are addressed in domestic research, they take on wholly new dimensions in a multi-cultural context. Namely, that if comparisons are to be made between heterogeneous cultures, each of these decisions must be made with several equivalences in mind (Mullen 1995, Nasif et al. 1991, Cavusgil & Das 1997, Craig & Douglas 2000). Equivalence essentially refers to whether elements of the research process are viewed to be the same cross-culturally. For example, the researcher must be highly concerned that input from focus groups are uniformly related to the research objectives, surveys must be similarly interpreted by respondents, and the samples must be equally representative. Equivalences are detailed in the following section, but serve as a serious concern to multi-cultural researchers. In addition, multi-cultural researchers must confront data collection problems not necessarily encountered in domestic research, such as whether telephone interviews are practical – the culture of interest may exist in a country with few telephones. Illiteracy

may also create conditions where the test instruments must be modified to be viewed instead of read.

Fieldwork

Fieldwork involves all activities associated with actually collecting data, and includes recruiting, training, supervising and controlling field workers. This is a process that is cumbersome in a domestic research environment, but becomes even more complex in a multi-cultural context. Multi-cultural research often involves recruiting indigenous fieldworkers, but the fieldworkers and the researcher may be separated by language and/or distance. If the researcher leaves the field, potentially serious control issues arise, including training of indigenous fieldworkers and supervision of data collection activities. Nasif et al. (1991) point to the fact that time constraints at times compel researchers to leave the field before proper controls can be established.

Data Analysis

One of the greatest sources of literary comment in multi-cultural research is analysis of data collected in cross-cultural environments. Essentially, there is agreement that analytical techniques have developed significantly over the past few years, yet remain somewhat limited (Nasif et al. 1991). Still, a growing need for multi-cultural research compels the use of techniques that are largely available. Malhotra et al. (1996) argue that the problem of data analysis can be broken down into five distinct issues: data preparation and standardization, sample comparability, equivalence, level of analysis, and methodological fallacies.

Data preparation involves examining data distributions, particularly outliers, to try to assess whether data entry errors had occurred, or whether the associated respondent may have not been from the population being studied. For example, a study in a developing country may include either expatriates or visitors from Western cultures that provide spurious scores to scaled questions.

Standardization of data is done by many researchers for ease of interpretation when making comparisons between heterogeneous groups of respondents. However, Singh (1995) argues very effectively that standardizing regression coefficients is not proper for making direct comparisons because of differences in data dispersion. This is supported by Malhotra et al. (1996) who note that standardized coefficients "have the same metric within a culture, but not across cultures," and that "standardized estimates eliminate any cross-culture differences on account of differences in variances" (p. 31).

Sample comparability is the degree to which the samples in the various cultures are representative of its members. A problem of multi-cultural research that complicates sample comparability is that some countries contain sub-cultures, which may be either large or decidedly different from the whole of the culture of interest. For example, a researcher studying food preferences in the U.S. may need to consider differences in taste between Hispanic-American, African-American and Native-American populations. Hispanic-American populations can be demonstrably prone to spicy (hot) dishes, while other groups will likely exhibit an aversion to these dishes in favor of others with different qualities. If this is the case, Malhotra et al. (1996) recommend tests of subcultures using multi-group analysis, and making statistical adjustments when practical. Regarding samples, one other issue of importance is equivalence of sample

sizes. This is important when analytical techniques are sensitive to differences in sample size, such as structural equation modeling. Equivalence (discussed in a separate section below) as described by Malhotra et al. (1996) in data analysis refers to measurement equivalence, and is generally tested using multi-group structural equation modeling.

Level of analysis refers to the scope of research question; whether the question is about individual respondents, within cultures, or across cultures. Research regarding individual responses may include cluster analysis traditional conjoint analysis, or qualitative techniques with the objective of understanding with great depth the perspective of each respondent. Within cultures analysis is also referred to as intra-cultural analysis, and may include such techniques as content analysis or multivariate methods. The objective is to gain insight into phenomena thought to be unique to each culture. Multi-cultural comparative research, however, falls into the final category, with the objective of finding similarities and systematic differences between cultural groups. Malhotra et al. (1996) note that these comparisons must not only include analysis of means, but also analysis of distributions and variances (p. 34). Methodological fallacy is meant to describe a situation where researchers may want to falsely apply generalizations to individuals, or vice versa. The example given is the use of Hofstede's cultural dimensions, and that some would apply these various dimensions to individuals, *not* to the societal contexts that they were meant to describe (p. 35).

Report Preparation & Presentation

This step is one that is largely left untreated in multi-cultural literature, yet important in that Malhotra et al. (1996) argue that interpreting the results of data analysis may be subject to cultural biases. This is due to the fact that the researcher must frame the

results in terms he/she is familiar with, which typically involves unintentional ethnocentric biases. In addition, some researchers may be tempted to reach *etic* conclusions, and generalize these findings to cultures not studied, which is obviously a questionable practice.

Biases in Multi-Cultural Research

The planning noted above must be conducted with a view toward mitigating, or at least minimizing many sources of bias in multi-cultural research. Craig & Douglas (2000) enumerate several biases relating to respondents that are extant in domestic research, but exacerbated in multi-cultural environments. These include social acquiescence, social desirability, topic bias, and item non-response.

Social acquiescence occurs when respondents answer survey questions in such a way as to attempt to be seen as desirable by the interviewer. Craig & Douglas (2000) argue that this type of bias is more common among less educated populations, particularly in developing economies. Social desirability bias is very close to social acquiescence, except that the respondent is also interested in projecting a positive picture of him/her self, as seen from the indigenous culture. An example given would be that a respondent reports the use of hygiene products regularly, when in fact this may or may not be true (p. 218).

Topic bias regards a willingness to respond to some subjects, but not to others. An example of topic bias would be questions of free trade in a controlled economy where there are also rigid policing activities, such as the former Soviet Union. Some cultures may be far more candid about personal behaviors than others, and thus the researcher must be aware of these propensities and adjust the survey instrument accordingly. Item

non-response is closely related – instead of recording inaccurate answers to questions that respondents are reluctant to answer, there simply is no response to that item.

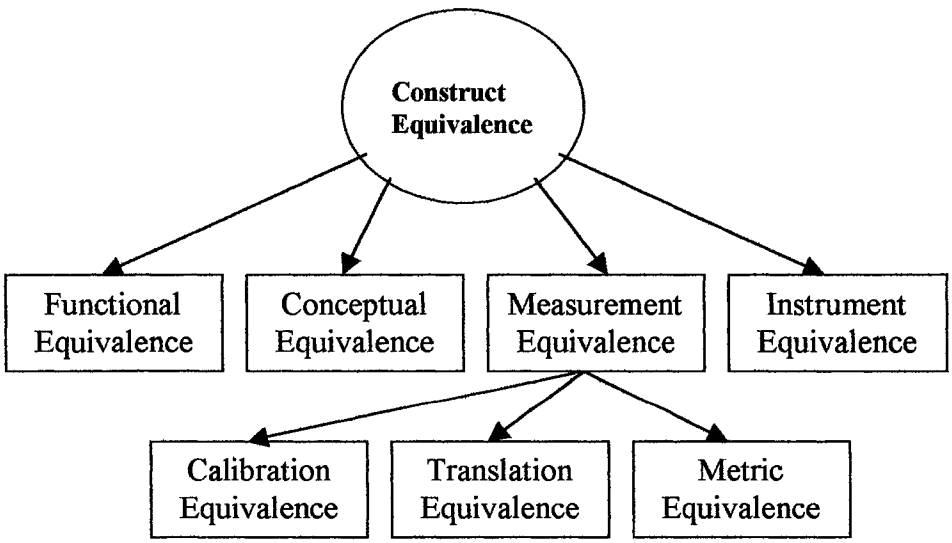
There are also potential biases that are a result of the design of the survey instrument; one of the most noted is the response format. For scaled responses, this includes issues such as the number of intervals in a scale and reverse coding. Research has shown that respondents of divergent cultures may prefer either more or less items in a Likert-type scale (Douglas & LeMaire 1974), or be confused by items that are negatively worded (e.g., Wong et al. 2003). Differences in interpretation of the poles of a scale may also result if a test instrument is administered in a country where reading is right-to-left rather than left-to-right, such as Arabic speaking countries (Craig & Douglas 2000). There are also potential difficulties related to pictorial stimuli, especially important in research among largely illiterate populations (Douglas & Craig 2000). An example would be to ensure that if pictorial or visual stimuli are used, that the correct meaning is projected to the respondent. While not directly related to scale surveys, an example from marketing baby food in Africa illustrates the problem. Gerber had produced and packaged baby food for sale in sub-Saharan Africa, but left the familiar smiling baby on the label. Illiterate consumers in the target market misunderstood the label to mean that the picture on the jar indicated its contents, which caused a rather strong avoidance of the product!

The Role of Equivalences

Equivalence in multi-cultural scales refers to the degree to which the scale or scale items are viewed as being the same across cultures. One difficulty researchers have routinely faced is the *etic/emic* dilemma (Douglas & Craig 2000), or the *imposed-etic* versus the *emic* approach (Ryan et al. 1999). The real impact for multi-cultural researchers is the degree to which a survey or test instrument can be standardized, or must be adapted to match local conditions.

Several types of equivalences exist; Figure 3 shows a typology offered by Malhotra et al. (1996). In this typology the overriding question of equivalence is

Figure 2-3
Typology of Equivalences in Multi-Cultural Comparative Research



Adapted from Malhotra et al. 1996

construct equivalence, which refers to the question of whether the construct being tested has the same meaning in divergent cultures (p. 19). There is general agreement in the literature regarding this typology, and a classic example of construct equivalence is whether a marketing construct such as "brand loyalty" means the same to a probability sample of American consumers as it would to a probability sample of Asian consumers. Construct equivalence is established through four types of equivalences: functional equivalence, conceptual equivalence, measurement equivalence and instrument equivalence.

Functional Equivalence

According to Green & White (1976), the most "concise expression of the problem of functional equivalence" is "obviously, if similar activities have different functions in different societies, their parameters cannot be used for comparative purposes" (p. 81). There is an ubiquitous example of bicycles; when researching consumer attitudes about bicycles, the responses are going to be affected by the fact that in some cultures, bicycles are used for recreation, while in others they are used for basic transportation (Green & White 1976, Malhotra et al. 1996, Singh 1995). Certainly, a researcher must be aware of these behavioral differences before engaging in research, and failure to do so creates a fundamentally unsound research project.

Conceptual Equivalence

Malhotra et al. (1996) describe conceptual equivalence as a question of whether a construct or concept is expressed similarly in attitudes or behaviors. In this sense, it is closely related to functional equivalence, except that instead of relating the equivalence

to a product or service, it is instead related to largely immeasurable latencies. The example in Malhotra et al. (1996) is of promotional sales, which are routine in American markets, but in developing economies where resources are exceedingly scarce, the marketing environment is a sellers' market, and consumers seriously question the product or the firm using the promotion (p. 20). Craig & Douglas (2000) point to the concept of "saving face" as being fundamentally different in Chinese cultures, while Western cultures appear to be far more individualistic in outlook. This places the construct of "self" in question, such that scale items measuring attitudes related to "self" may not be interpreted in a similar way (p. 158).

Measurement Equivalence

Craig & Douglas (2000) note that "construct and measure[ment] equivalence are highly interrelated insofar as the measure is an operational definition of the construct" (p. 160). Specifically, it is the question of whether scale items accurately measure a construct in a way that is culture invariant (Malhotra et al. 1996). Measurement equivalence is assessed across three equivalence dimensions: calibration, translation, and metric.

Calibration Equivalence

Calibration equivalence refers to whether units of measurement are equivalent, and are of concern when scale questions refer to some objective quantification such as weight, volume, temperature, or distances. In addition, Craig & Douglas (2000) note that calibration equivalence should also include perceptual cues such as colors, which have varying meanings in different cultures. For example, a restaurant chain interested in

looking into preferences for food presentation may ask about the temperature of the food; if the questionnaire is presented in Fahrenheit temperatures the responses are likely to be a bit lower than intended, and interpreted incorrectly.

Translation Equivalence

Translation equivalence simply stated is the degree to which scale items are linguistically equal, and understood equally well between divergent cultures. This is central to establishing other types of equivalences, since mistranslation can lead to conceptual errors, etc. Implicit in translation equivalence is the problem of translating *ideas* or word-pictures. Bearing this in mind, it is especially important that idioms are used with care, because equivalent use of whole phrases is exceptionally difficult. For example, an American confectioner, interested in preserving the traditional one-cent gumball and exporting them to Europe may ask a question like, "would you spend a penny on this?" The perception is fundamentally different, as the phrase has completely different meanings in the U.S. and U.K. cultures, common use of English notwithstanding.

Metric Equivalence

Metric equivalence is also commonly called scalar equivalence, and according to many sources can be safely assumed to exist if the psychometric properties of the data are invariant between or among cultures (Mullen 1995, Douglas & Craig 2000, Malhotra et al. 1996, Sin et al. 1999). This implies that distributions of data and data dispersions should not be significantly different between culture groups. According to Craig & Douglas (2000), metric equivalence can only exist if scoring procedures are equally

effective across cultures. For example, seven point Likert-type scales are common in U.S. research, while some countries may use a 10 or 20 point scale (p. 162). In addition, Craig & Douglas note that responses should also be equivalent – i.e., a score of (5) on a seven point Likert-type scale should equal a score of (5) from a respondent in another culture.

Instrument Equivalence

Malhotra et al. (1996) present the case that instrument equivalence "deals with whether the scale items, response categories and questionnaire stimuli such as brands, products, consumer behavior, and marketing effort are interpreted identically across cultures (p. 20)." This is to say that persons responding to a survey should have an equal understanding of the tasks presented. An example may be that a "cleaning aid" in the American market refers to a chemical solution designed to remove dirt or grease; whereas the same term in other countries may mean a person who helps clean. Any test instrument that seeks responses regarding a "cleaning aid" must make clarifications to the meaning of the term.

The Paradox of Equivalence

Sekaran (1983) pointed to an interesting problem regarding equivalences and the pursuit of them. In this study, the "paradox of equivalence" was described as follows:

> The paradox of equivalence as discussed by Sechrest, Fay, and Zaidi (1972), however, needs to be noted. The authors point out that it is entirely possible that important cultural differences would be obliterated, or at least obscured, by an attempt to achieve a rather misleading notion of equivalence. This implies that we should not be so obsessed by various types of equivalences that we preclude the cultural uniqueness of responses from surfacing. As we develop instruments in various languages through the help of linguists or through a process of decentering, we need to be sensitive to the paradox of equivalence, because we could easily fall into the trap of attaining excessive equivalence.

The effect of equivalences in multi-cultural research is that a violation of equivalence assumptions can be expected to alter response patterns of respondents for the item or construct (factor) in question. For example, if translation equivalence is not satisfied for a given item on a seven point interval scale, one would expect that respondents who do not understand the question may mark the item in the center (4), not wishing to present an extreme score, or to one extreme that may indicate a lack of preference or agreement due to not understanding, or to the other extreme in attempting to be complimentary. Very few responses would be rated in the intervening intervals 2,3,5, or 6. Obviously, there can be no valid interpretation of data generated by such responses. This drives the desire by researchers to minimize equivalence errors. In the case of translation equivalence, this is certainly necessary, but cases of conceptual or functional equivalence, modifying a scale to remove differences generated by inequivalence could mask important cross-cultural differences. To use the Asian/American concept of self as an example, ensuring that the scale exhibited conceptual equivalences would almost certainly lead to scale questions that would *not* appropriately measure attitudes of interpersonal relationships appropriately.

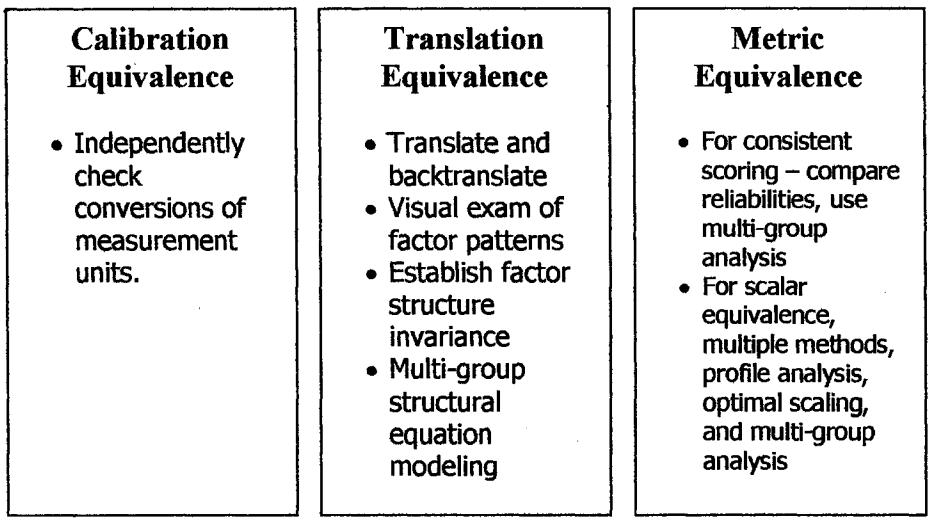
Current Equivalence Diagnostics

Multi-Group Structural Equation Modeling

Mullen (1995) summed up methods of establishing equivalences, as represented in Figure 4. According to Mullen (1995), translation equivalence can be tested through back-translation (a-priori), or through post-hoc methods such as examining factor patterns, and factor structure invariance as established through multi-group structural equation

modeling. One difficulty with this method is that factor structures may be dissimilar for a number of reasons, and isolating the cause of factor invariance may prove difficult. Also, factor invariance may indicate important *emic* insights as to why heterogeneous groups of respondents see things differently.

Figure 2-4
Diagnosing Measurement Equivalence



Source: Mullen (1995)

The example Mullen (1995) provides of scale inequivalence is a scale that is consistent, yet unreliable by overstating weight by ten pounds. The overstatement represents the systematic variance that is not related to the true weight, and thus the problem in multi-cultural research is to understand with greater clarity the ten pound systematic variance (p. 576).

Multi-group structural equation modeling establishes scale equivalence by isolating systematic variance to error terms associated with observed variables. By using a nested model approach, the error terms can be tested for invariance across groups, with the assumption that items that are invariant contain variance that is not systematic, and any remaining differences are artifacts of the cultures, which reveals the information the researcher is looking for.

One limitation of using this method is that it can be argued that some types of equivalence may not be exhibited in these error terms; in fact, a translation error in a semantic differential scale may predictably result in a concentration of centrally placed scores (central tendency error). This central tendency error may be strong enough to result in minimal error, which may pass muster in a multi-group analysis. Such a case presents a situation where the researcher believes that the concentration of central scores reflects a neutral attitude or opinion of an item that is not understood, resulting in mistaken conclusions.

Optimal Scaling

Optimal scaling establishes scalar equivalence when the researcher utilizes categorical (ordinal) data to measure latent constructs. This is accomplished by examining rank orders of input to ensure that scalar distances are equal between "a priori known populations" (Mullen 1995, p. 580). In optimal scaling, the ordinal data are transformed, iteratively, into interval data to examine the "underlying metric" of the data. This is done in such a way as to relieve potential differences between the conceptual model and the observed data.

Item Response Theory (IRT)

Item response theory (IRT) is a means of comparing scores across items and sources to detect differences in scores from different respondents (Barr & Raju 2003). Hui & Triandis (1985) describe the process as follows:

> The problem of selecting a relevant and unbiased criterion for judging an instrument is bypassed by the item response theory (IRT) approach . . . which uses item parameters derived "internally," thus avoiding the use of "external" criteria. In this approach, item characteristic curves (ICC), which represent the probabilities of responding to an item in a certain specified manner at different levels of the latent trait to be measured, are obtained from different cultures. Statistical tests are developed to examine the differences between ICCs. . . . Such differences can point to the lack of equivalence between the two cultures on a particular item. On the other hand, an instrument that has similar ICCs across cultures has, at least in part, demonstrated its item equivalence and scalar equivalence (p. 138).

The ICC curve referred to in the above quotation represents a non-linear relationship between the probability of a respondent selecting a particular response and the levels of a latent construct being measured. ICCs between different cultures are tested statistically for differences, with the assumption that a lack of statistical significance supports the assumption of equivalence.

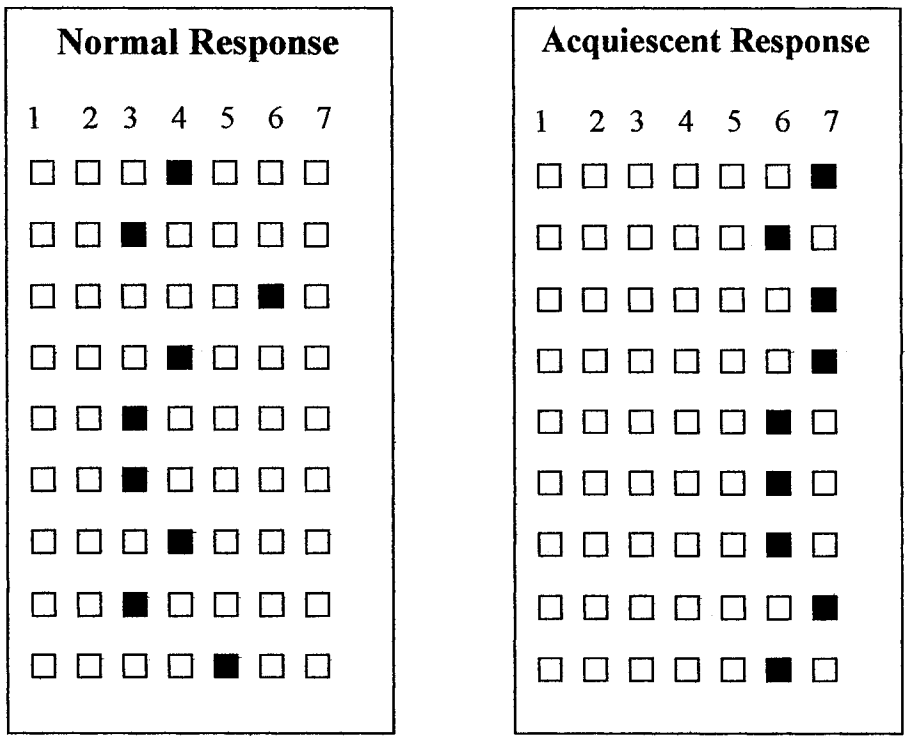
IRT has traditionally been used in education and psychological testing, and was specifically designed to aid in the administration of testing subjects and assessing skill levels of different test takers. It was modified to be useful in discovering differences between heterogeneous groups of people with respect to a common test instrument. Specifically, IRT has been applied to multi-cultural research to examine instrument equivalence, primarily in psychological literature, but rarely in international business.

The Role of Response Styles

Response style can greatly affect data analysis, and refers to the manner in which a respondent indicates his/her preference for an item given the type of scale and response format provided. This implies that responses to scale questions can be biased in a way that does not necessarily reflect true assessments, and that respondents make these indications based upon something other than what the scale or scale item was designed to measure (Baumgartner & Steenkamp 2001). In multi-cultural research, response styles are assumed to vary systematically as a function of culture; for example, if people in a given culture tend to be generally accommodating, response styles may reasonably expected to be "acquiescent" – or skewed toward more polite or complimentary responses. Figure (2-5) shows a comparison of scale responses; one is the expected response and variation indicative of accurate response and measurement (normal response), and the other is an example of acquiescent response bias – in this case consistent with more complimentary scores. One can visually detect that the acquiescent response scale contains systematically higher scores than the "normal" scale.

Baumgartner & Steenkamp (2001) typify seven response styles of interest to marketers; acquiescence, disacquiescence, net acquiescence, extreme response (also referred to as response range), response range, midpoint responding (also known as central tendency error), and noncontingent responding. Acquiescence response style, as illustrated earlier, occurs when a respondent attempts to provide responses that are assumed to please the researcher – so scores may be spuriously high or low. In other words, respondents become "yes men" when participating in a survey.

Figure 2-5: The Systematic Effect of Response Style



Disacquiescence is a tendency to disagree with all items without giving any particular thought to the substance of the questions. Baumgartner & Steenkamp (2001) refer to this as "nay-saying." While Baumgartner & Steenkamp (2001) present the case that "net acquiescence" is a separate response style, it appears to be a function of acquiescence and disacquiescence. Greenleaf (1992) established net acquiescence by generally taking the difference between acquiescence and disacquiesence, which leaves the question open as to the distinctiveness of the response style.

Extreme response style is a "tendency to endorse the most extreme response categories regardless of content" (Baumgartner & Steenkamp 2001, p. 145). With extreme response, respondents will tend to provide polarized responses with very little in the mid-range response. This may be related to response range, where responses are either tightly gathered or loosely dispersed around the mean.

Mid-point response is also referred to as central tendency error, and can be particularly problematic with semantic differential scales in multi-cultural settings (Yu et al. 2003). Mid-point response, or central tendency error occurs when responses are gathered around the midpoint of possible scale responses. For example, a seven-point semantic differential scale would see consistent scores of "4" as respondents would hesitate to indicate a preference toward either pole. Baumgartner & Steenkamp (2001) contend that this may be due to the fact that some respondents do not want to have their true opinions recorded, or that there is some measure of indifference. Yu et al. (2003) generally support this supposition, and define the phenomenon slightly differently as a "reluctance to give extreme scores" (p. 216).

Noncontingent response is a propensity for respondents to check items either randomly or carelessly, with no regard to the questions being presented. This phenomenon is not necessarily culturally bound, and can be seen in numerous U.S. and foreign market studies. The most probable explanation for this is that the respondent is not concerned with the study, and responds to it either out of politeness, or for some other motivation, at which point the motivation ends, and the responses are meaningless. An example of this would be university students who complete surveys for professors, making zig-zag patterns with responses. When this is the case, obviously the

questionnaire would be deleted from analysis, but if the responses are more random, they are more difficult to detect.

Current Response Style Diagnostics

Current methods of detecting and correcting for response style bias are rooted in an analysis of measures of central tendency and dispersion, and assume that response styles are systematic across all items and/or respondents. Table (2-1) presents the various response styles and methods of diagnosis.

Table 2-1: Response Style Diagnostics

Response Style	Recommended Diagnostic
Acquiescent Response Style	Correlation between items that are assumed to be uncorrelated, and between positively and negatively worded items.
Disacquiescent Response Style	Same as ARS.
Netacquiescent Response Style	ARS minus DARS; expressed as the mean response across many heterogeneous items.
Extreme Response Syle	Proportion of heterogeneous items the respondent endorses the most extreme (positive or negative) scale categories.
Response Range	Standard deviations of a person's responses across many heterogeneous items.
Mid-point Response	Proportion of heterogeneous items on which the respondent endorses the middle scale category
Noncontingent Response	Sum of absolute differences between responses to pairs of items, where the items in each pair are maximally correlated, have similar means across respondents, and are keyed in the same direction.

Source: Baumgartner & Steenkamp 2001

What is unclear from present research regarding response styles is whether an error of equivalence may trigger a particular response style that is systematic across a sample, but *not* across items. For example, if there were translation equivalence errors in

a semantic differential scale, one would expect that respondents would *systematically* select the middle option for items mistranslated. This would be due to a desire by the respondent to withhold their opinion for items not understood. However, these current diagnostic tools are designed to make assessments about response styles by considering entire samples and items, not by examining individual items. In a situation as described above, the error may go undetected. In fact, the item in question may appear to exhibit a high degree of consistency, and high level of significance. A researcher may be tempted to draw a conclusion of ambivalence among the sample with regard to that item, when in reality the problem is that the sample did not understand the item.

Scales Used in Multi-Cultural Marketing Studies

Having discussed the complicated nature of the use of scales in multi-cultural research environments, examples of uses of domestically developed scales in cross-cultural contexts are next presented. The complexity of multi-cultural research makes the development of scales in multi-cultural environments exceptionally difficult, leading to a tendency to rely upon previously developed scales, but modifying them to suit local conditions.

Three scales developed in the United States are commonly used in multi-cultural comparative studies, and have been used to make both emic and etic statements about various cultures. Each of these scales was developed in the American market and research environment, translated, and used in various multi-national studies. Table 1 provides an overview of studies conducted using these scales, including the samples involved, analytical techniques used, with emic and etic conclusions drawn as a result of

the analysis. These scales were chosen for inclusion in this review because they are representative of the theoretical questions of interest to multi-cultural researchers, and also of the techniques that are employed to seek answers to those questions.

One scale, the MARKOR scale, was designed to test marketers' perception of how well their firms generate, disseminate, and react to market intelligence. Two other scales, CETSCALE and SERVQUAL, were designed to measure consumers' attitudes. Specifically, CETSCALE measures consumers' tendencies to make value judgments about products (or the act of purchasing products) produced in foreign manufacturing firms. SERVQUAL was developed to measure consumer perception of the quality of services provided by a service provider. What is interesting about these scales is that the samples can be seen as systematically different, regardless of cultural environment. Namely, a survey using the MARKOR scale must be distributed to marketers or executives within marketing firms. These professionals are likely to be far more educated than the average citizen, and also more likely to have traveled abroad and been exposed to divergent cultures than the average citizens of their respective home countries. On this basis, we may expect to see the MARKOR exhibit a greater degree of cultural invariance than the CETSCALE or SERVQUAL, which are designed for representative samples of consumers.

Table 2-2

Study Main Question	Sample	Response Format	Analytical Method	Conclusions
CETSCALE				
Clarke 2001 Do extreme response styles bias results from CETSCALE?	1008 French, Austrian, Mexican, and U.S. university students.	Varied; 3,4,5,6,7,8,9 and 10 point Likert scales were used.	ANOVA	Significant differences exist between country samples and response formats regarding extreme response styles. ERS significantly biases the results of CETSCALE.
Klein, Ettenson & Morris 1998 Can a separate construct "animosity" impact a consumer's willingness to purchase and product adoption?	244 Chinese consumers/mall intercept	7 point Likert scale	Structural Equation Modeling	National animosity and consumer ethnocentrism are distinct constructs. These constructs together influence a consumer's willingness to purchase and adopt products.
Bailey, & Gutierrez de Pineres 1997 Do Mexican consumers vary in attitude toward foreign goods according to income and social class?	400 Mexican consumers	7 point Likert scale	ANOVA, Logistic Regression	Mexican consumers of higher social classes tend to view foreign products more favorably than those of lower social classes.
Durvasula, Craig & Netemeyer 1997 Do U.S. and Russian consumers vary in ethnocentric tendencies?	204 U.S. university students and 60 Russian university students	Response formats were mixed; 7 point Likert scales for some constructs, 7 point semantic differential scale for others.	Structural Equation Modeling	Russians tend to be far more open to the purchase of foreign products than Americans.
Netemeyer, Durvasula & Lichtenstein 1991 Is the CETSCALE reliable across national/cultural samples?	71 U.S., 73 German, 70 French, and 76 Japanese university students	7 point Likert scale	Structural Equation Modeling	CETSCALE appears to exhibit high international and/or intercultural structural reliability. Nomological validity was less strongly supported.

MARKOR				
Caruana 1999 Are the dimensions of the MARKOR scale consistent and stable across samples?	123 U.K. and 193 Maltese marketing directors	7 point Likert scale	Structural Equation Modeling	The MARKOR scale does not appear to fit the U.K. and Maltese environments; several items were not upheld in the Maltese sample. The authors suggest a different conceptualization of market orientation for the different markets.
Caruana et al. 1997 Is market orientation positively associated with organizational commitment in government departments in Australia?	134 government officials (department heads) in the Australian government.	7 point Likert scale	Multiple Regression	Market orientation has been found to be positively related to organizational commitment in the public sector of Australian firms.
Pitt, Caruana & Berthon 1996 Is market orientation related to business performance, and whether the market orientation-performance link is culture invariant.	106 U.K. and 193 Maltese executives	7 point Likert scale	OLS Regression	MARKOR is a reliable scale across cultural boundaries, and that there tends to be a relationship between market orientation and firm performance across cultures, although the evidence from this study is weak.
SERVQUAL *(Adapted from Smith & Reynolds 2001)*				
Imrie et al. 2000 Whether consumers in non-American markets evaluate service quality differently as a function of culture.	84 Taiwanese consumers	Qualitative study discussed 5 SERVQUAL dimensions	28 depth interviews and focus groups comprised of 7 members each	Findings revealed 5 SERVQUAL dimensions, but qualitative discussion shows different tolerances between cultures.

Stauss & Mang 1999 Do "culture shocks" alter perception of service quality?	48 Japanese, 64 U.S., and 108 German first and business class travelers	Qualitative study	Critical incidence analysis	Negative critical incidents do not prevail among inter-cultural encounters but amongst intra-cultural encounters, possibly due to attribution.
Caruana et al. 1998 Whether the expectations construct is equivalent between Australia and Singapore.	210 Australian and 104 Singaporean marketing executives	7 point Likert scale	SEM	The SERVQUAL measure is not stable cross-nationally.
Winstead 1997 Do U.S. and Japanese consumers evaluate service encounters differently?	200 U.S. and 176 Japanese university students	7 point Likert scale	Multiple regression	Significant differences between U.S. and Japan in the ability of SERVQUAL factors to predict encounter satisfaction.
Kettinger et al. 1995 Whether there are cultural effects that influence perceived service quality in information services.	87 Hong Kong, 148 Korean, 48 Dutch university students	7 point Likert scale	SEM	A potential "Asian" factor identified in Hong Kong and Korea.

MARKOR

The MARKOR scale was developed by Kohli, Jaworski & Kumar (1993) to measure a firm's market orientation, or sensitivity to market forces. More directly, this refers to its ability to gather, disseminate, and respond to market intelligence (Kohli et al. 1993). The final scale contained 22 items designed to measure 3 constructs: intelligence generation, intelligence dissemination, and organizational responsiveness (Appendix 2-A).

Intelligence generation is the organization's ability to gather meaningful information about the marketplace – external to the firm. An example would be whether the firm actively scans for competitor information or new developments in the marketplace. If instead a firm allows such information to "seep in" slowly from mainstream press or word of mouth, then the firm is not high in intelligence generation. Intelligence dissemination is the extent to which firm management ensures that all departments or business units are informed of market intelligence as it is made available. Departments that hold market intelligence proprietary, to the detriment of other departments, would not be seen as high in intelligence dissemination. Organizational responsiveness refers to the organization's ability to use the market intelligence to advantage; in other words, organizational responsiveness represents a firm's ability to make use of its market intelligence.

Since the development of this scale, it has been widely used in a number of international contexts. Pitt et al. (1996) specifically set out to test the reliability of MARKOR across divergent cultures, and also to test whether there is a relationship between market orientation and firm performance. The object of this study is to see whether the link is culture invariant. Pitt et al. (1996) tested the reliability of the scale and cultural invariance using multi-group structural equation modeling (Mullen 1995). The sample consisted of 106 U.K. and 193 Maltese executives who responded to a seven point Likert-type scale. Cultural invariance of the scale was tested only through reliability measures (Chronbach's alpha), with the conclusion that the scale is reliable across cultural boundaries.

Mavondo & Farrell (2000) tested the MARKOR scale against Narver & Slater's (1990) scale in heterogeneous samples from Australia. The object of this study is to understand which scale exhibits the best multi-cultural consistency when used in divergent industry and cultural environments. The results of this study show that the Narver & Slater (1990) scale appears to show relative stability in these environments. However, no mention of scale equivalences or testing for them is made.

Similarly, Caruana (1999) tested the MARKOR scale in samples of 123 U.K. and 193 Maltese marketing executives, and found that significant differences were apparent between the samples in the basic structure of the scale. It is important to note that scale equivalence was not addressed; the analysis was conducted using multi-group structural equation modeling, and traditional scale purification procedures followed.

In an interesting study, Caruana et al. (1997) studied market orientation as a function of organizational commitment in civic organizations in Australia. A sample of 134 government officials (heads of various governmental departments) were surveyed, with the analysis being conducted with multiple regression. These authors found that market orientation is positively related to organizational commitment in these public sectors, yet there is no mention in this study of testing the equivalence of the MARKOR scale in the Australian cultural environment.

CETSCALE

The CETSCALE was developed to measure consumer ethnocentric tendencies. This scale was developed by Shimp & Sharma (1997), with the final scale consisting of 10 items, as a unidimensional scale (Appendix 2-B). As noted earlier, consumer

ethnocentrism refers to a consumer's propensity to make ethical judgments about purchasing or using products or services that were produced in a foreign country.

Clarke (2001) tested whether CETSCALE is biased as a result of different response styles among divergent populations. With samples of French, Austrian, Mexican, and U.S. university students, Clarke (2001) found that response styles vary with response format according to culture, and that adjustments may need to be made to test instruments to ensure metric equivalence.

Martinez et al. (2000) sought to validate the CETSCALE in the Spanish market. The instrument was translated into Spanish by native speakers, and translation equivalence was checked by comparing the resulting translations with each other as well as with pre-existing translations. The final form was administered as a seven point Likert-type scale to 476 Spanish consumers. Validity of the scale was established via confirmatory factor analysis, which indicated a high degree of scale reliability.

Klein et al. (1998) used a sample of 224 Chinese consumers in Nanking to test whether a new construct "animosity" impacts a consumer's willingness to purchase and adopt products. The CETSCALE was translated and backtranslated into Chinese by Chinese nationals fluent in English and administered as a seven point Likert-type scale. Using structural equation modeling, the authors found that consumer ethnocentrism and animosity are distinct constructs for the Chinese. Establishment of scale equivalence was not addressed, and the analysis was conducted using structural equation modeling. Spurious responses are possible from this particular study due to the fact that the source country of "foreign" products was identified as Japan, notorious in the minds of many

Chinese consumers for war atrocities in the Second World War, placing the generalizability of the results in question.

Bailey et al. (1997) used an abbreviated CETSCALE to understand Mexican consumers' attitudes toward foreign goods, and to see if social class and income moderate ethnocentric tendencies. A sample of 400 Mexican consumers, 377 female and 23 male respondents were administered a ten item version of SERVQUAL, administered as a seven point Likert-type scale. Translation procedures were not mentioned, although structural equation modeling supported the factor structure of the scale.

Durvasula et al. (1997) investigated whether U.S. and Russian consumers vary in ethnocentric tendencies. In addition, one stated goal of this work was to validate the CETSCALE in a multi-cultural context. The scale was translated into Russian and back-translated into English by bilingual translators, and administered as a seven-point Likert-type scale. In addition to the CETSCALE items, other items were added as a seven point semantic differential scale. Equivalence was tested by examining residual item correlations, with no significant associations, indicating an achievement of scale equivalence. Analysis was conducted with structural equation modeling, and results indicate that Russians tend to be more open to the purchase of foreign products than Americans. Unfortunately, this study may suffer from unequal sample sizes, as the American sample numbered 204, while the Russian sample numbered only 60. In addition, university students were selected as the sample for both countries, which may not be representative of average consumers due to increased education and exposure to other cultures.

Netemeyer et al. (1991) conducted the first test of CETSCALE scale equivalence across national samples, with a sample of 71 U.S., 73 German, 70 French, and 76 Japanese university students. The test instrument was drafted in English and translated into each language by bilingual translators. Backtranslation was used for each instrument, and the sample sizes carefully controlled to avoid bias due to unequal sample sizes. Scale equivalence was tested first by examining factor invariance, which proved to be invariant across cultures. Further tests were conducted using multi-group structural equation modeling, with the conclusion that CETSCALE exhibits unidimensionality and invariant factor loadings across countries.

SERVQUAL

Parasuraman et al. (1988) undertook to operationalize service quality, which according to the authors is assumed to directly and indirectly affect customer satisfaction. In this effort, the authors argued that for manufactured goods, quality can be objectified through SPC or some other absolute measure. Service quality is instead a latent, unobservable construct that impacts customer satisfaction but is difficult to understand and measure. Specifically, the properties of the service sector present serious challenges in trying to do so; namely the dimensions of intangibility, heterogeneity, and inseparability (p. 13). The Parasuraman et al. (1988) article was written specifically to develop a scale that researchers could use that would capture the unobservable dimensions of service quality. The SERVQUAL scale was developed with a sample of 200 U.S. consumers, and resulted in 22 items (Appendix 2-C). These items measure five distinct constructs: 1) tangibles, which refers to physical facilities, equipment and appearance of personnel, 2) reliability – or the ability to perform the promised service

dependably and accurately, 3) responsiveness – the willingness to help customers and provide prompt service, 4) assurance – which represents knowledge courtesy of employees and their ability to inspire trust and confidence, and 5) empathy – or the caring and individualized attention the firm provides its customers (p. 23).

SERVQUQAL has been successfully used in domestic research, specifically in banking and healthcare industries. Multi-cultural studies, however indicate some variation in the reliability of the scale, and some have been critical of the validity of SERVQUAL, especially in a multi-cultural context (e.g., Lam & Woo 1997). Smith & Reynolds (2001) reviewed extant multi-cultural literature using the SERVQUAL scale, and concluded that SERVQUAL, in particular, is susceptible to response bias when used in multi-cultural or international contexts.

Qualitative studies designed to examine the multi-cultural equivalence of SERVQUAL concepts indicate that the five dimensions of SERVQUAL appear to exhibit varying degrees of conceptual equivalence among Taiwanese (Imrie et al. 2000), Japanese, German, and U.S. (Stauss & Mang 1999) consumers. These studies are useful in that they provide insight into how respondents (either focus group participants or interview subjects) view the concepts underlying the SERVQUAL scale, but due to the fact that quantitative methods were not used, there can be no testing for other types of equivalence. Even with the limitation of no statistical analyses, differences between cultural groups have been apparent – for example, Imrie et al. (2000) found that Taiwanese subjects seem to include "politeness" and "face" as concepts important to the evaluation of the quality of services, which is generally missing from North American samples. In addition, the tolerance for poor service performance appears to be greater in

ethnic Chinese culture. Stauss & Mang (1999) used critical incidence techniques to find that between Japanese, German, and U.S. air travelers, assessment of the quality of service appears to be different among the samples, principally with regard to the cultural affiliation of the service provider. For example, most respondents deem negative encounters as a "critical incident" if the service provider is culturally divergent, as opposed to those of the same culture. This is highly indicative of cultural influences that may not be explained by the SERVQUAL scale.

Uses of the SERVQUAL scale in multi-cultural contexts are numerous, with mixed results. Caruana et al. (1998) used multi-group structural equation modeling to evaluate Australian and Singaporean views of service quality, with the conclusion that there is an absence of construct equivalence, primarily due to differences in expectation levels. It must be noted, however, that this study was conducted using marketing executives as samples, which may not be representative of consumer populations.

Winstead (1997) used a modified SERVQUAL instrument to examine whether Japanese consumers evaluate service encounters differently than American consumers. One difficulty with this study is that tests for instrument equivalence appear to be missing; conclusions were drawn from regression results produced from factor scores that were generated from exploratory factor analysis. Tests for metric equivalence in this case are not extant, leaving the regression results in question. Kettinger et al. (1995) undertook to examine information services customers' views of service quality within Hong Kong, Korean, and Dutch markets using university students as a sample. Equivalence of the SERVQUAL scale was not tested; instead,

multi-group structural equation modeling was used to infer varying factor structures, which suggests that a separate factor distinct to Asian markets exists.

As mentioned earlier, these scales have been developed in U.S. markets, and "borrowed" for use in other cultural contexts. Douglas & Nijssen (2003) provide an excellent overview of the dangers of doing this; namely that the differing national and cultural environments provide contextual influences on respondents' perception of the instrument. The examples given by Douglas & Nijssen (2003) are vivid, including the use of the CETSCALE in the Netherlands, where the concept of nationalism is fundamentally different from that in the U.S. – or in many other countries in the world. This phenomenon was referred to as "contextual salience" (p. 632), and implies that scales developed in one cultural context may need to be either significantly revised or dispensed with all together in order to achieve an understanding of the phenomena of interest.

From the above discussion and related studies, it can be argued that the use of scales is fraught with difficulties that are difficult to assess until after the research had been completed – if then. Some of the studies referred to do not include any tests for equivalence of items; those that do principally use multi-group structural equation modeling, which is the very same method that is used to assess real differences between cultural samples. Techniques currently used to assess equivalence are based largely upon correlation – which is highly dependent upon assumptions that the responses that generate the correlations are true reflections of the attitudes and/or views of respondents. This leads to some question of whether other methods may be employed to gain a better understanding of respondents' true preferences, and if by doing, a greater understanding of scale equivalence can be achieved.

Conjoint Analysis

Given the difficulties noted of using scales in multi-cultural settings (namely the risk of an undetected equivalence error) and their impact on response styles and metric equivalence, it is reasonable to consider using a method of discovery that elicits true choice behaviors to compare stated (scaled) preferences with actual preferences. One leading candidate for this task is conjoint analysis, in which a respondent indicates by various means what his/her preferences are given a field of items (or attributes) that are presented simultaneously.

The purpose of conjoint analysis is to estimate respondent preferences given a certain span of choices – or as Gustafsson et al. (2001) state: "the goal of conjoint analysis is to explain and predict preferences that resulting an assessment of achievements (p. 7)." Hair et al. (1998) state the case slightly differently as, "a method that portrays consumers' decisions realistically as trade-offs among multiattribute products or services (p. 387)." It is the "multiattribute" quality of the procedure that gives it its name, conjoint – an acronym from CONsidered JOINTly. In estimating these preferences, respondents are shown a series of alternatives and asked to indicate in various ways which alternative appears to be most appealing. In this sense, conjoint is a real departure from previous methods because of the fact that respondents are not asked to evaluate scale items individually, or one at a time. The advantage in this is that a degree of realism is injected into the research process as consumers rarely evaluate actual product choices by examining various attributes singularly and in isolation. In addition, the response required of the respondent is simplified – a feature that is important to this research and will be revisited in a later section.

Gustafsson et al. (2001) provide an excellent picture of the growing prominence of conjoint methodology internationally, which reads as follows:

> The essay by the psychologist, Luce, and the statistician Tukey (1964) can be viewed as the origin of conjoint analysis (Caroll & Green 1995; Green & Srinivasan 1978). Since its introduction into marketing literature by Green & Rao (1971) as well as by Johnson (1974) in the beginning of the 1970s, conjoint analysis has developed into a method of preference studies that receives much attention from both theoreticians and those who carry out field studies. For example, Cattin and Wittink (1982) report 698 conjoint projects that were carried out by 17 companies included in their survey in the period from 1971 to 1980. For the period from 1981 to 1985, Wittink and Cattin (1989) found 66 companies in the United States that were in charge of a total of 1062 conjoint projects. As regard Europe, Wittink, Vriens and Burhenne counted a total of 956 projects carried out by 59 companies in the period from 1986 to 1991 (Baier and Gaul 1999; Wittink, Vriens and Berhenne 1994). A survey initiated in Germany in 1998 . . . shows that 52 institutions interested in the study design an average of 6 conjoint analyses per year (Melles and Holling 1999). If we project the number of analysis for the total period of five years, we get approx[imately] 1531 projects (p. 5).

While the last study noted above by Gustafsson et al. (2001) was conducted in a limited area (Western Europe), and primarily among universities, there are untold numbers of conjoint experiments being conducted by practitioners who desire to know more about customers or the competitive environment in a realistic way. Some of the most prominent uses for conjoint analysis in both practitioner and academic research include price sensitivity (e.g., Odekerken-Schroder et al. 2003), market share projection (e.g., Chakraborty et al., 2002), new product development (e.g., Steiner & Hruschka 2003), market segmentation (e.g., Moskowitz, Krieger, & Rabino 2002), and advertising (e.g., Gordon & Lima-Turner 1997).

Conjoint analysis can be viewed as a method of reverse engineering consumer decision-making, or preference patterns where respondents are asked to select their preferences from a group of potential choices. These choices vary along dimensions the

researcher is interested in investigating, which are known as "attributes." Examples of attributes commonly tested in conjoint include price of a product, color, shape, or some other quality of a product. Each attribute may have a number of different levels – or variations – for example, there may be several price points, or colors, etc. a respondent may choose from in conjunction with other attributes. A respondent is exposed to various combinations of attributes and levels (using different techniques as described below), and asked to indicate in some way his/her preference given the choices offered. By recording these preferences in a succession of evaluative tasks where attributes and levels of attributes are manipulated, it can be deduced what configuration of attributes and levels the respondent is most likely to prefer. This deduction process establishes with great accuracy the relative importance of the attributes as seen by individual respondents. These relative importance estimates are commonly referred to in conjoint as "part worth" estimates, and indicate the impact a given attribute has on the choices made by a respondent when considered in concert with the other attributes in the conjoint experiment.

The first noticeable entry of conjoint methods to marketing was made by Green and Wind (1975), in which conjoint was described, its benefits highlighted, and the then current mechanics of conjoint described. It is important to note that the venue for this publication was the Harvard Business Review – a respected publication that has wide readership among practicing marketers and academics. The examples given by Green and Wind (1975) included new product development (carpet cleaner products) and advertising (replacement tires) to highlight the practical application of the technique. It is

this practical dimension that has fueled acceptance among primarily practitioners, and use by academics because of the a-theoretical properties of the technique.

One limitation to the use of conjoint in academic work is that with conjoint, one cannot test the relationships between variables. Instead, as noted above, conjoint is used to make accurate predictions about choice behaviors. New developments in conjoint design and computer algorithms may serve as a basis to re-visit the issue. Specifically, the deduction method noted above allows for an estimation of choice preferences for *each individual respondent.* As detailed in the next section, however, newer techniques such as choice-based conjoint (CBC) allow for these estimations on an aggregate level, which provides the ability to apply statistical tests to whole samples. The following section describes the development of conjoint analysis with examples of its various uses.

The Development of Conjoint Analysis

Ratings Based Conjoint

Originally, conjoint analysis was designed to estimate as accurately as possible a consumer's preference structure within a given set of choices, and could accommodate only ordinal data (Gustafsson et al. 2001). These studies are referred to as rank-based conjoint, and use algorithms that calculate part-worth estimates in such a way as to predict as closely as possible the rank-order of attributes that the respondent would actually choose (Louvierre 1988). In rank-based conjoint, the part-worth (also commonly referred to as utility) estimate for a given choice is a predicted order of preference, or probability that a respondent would prefer attributes in a stated order.

Metric Conjoint

Metric conjoint was developed to allow respondents metric responses to choice sets by using scales. Part-worth estimates are derived in metric conjoint by ordinary least squares regression, with the part-worth estimates representing the relative importance of a given attribute. These estimates are also calculated for each respondent, but are not designed to be interpreted in the aggregate. For example, a respondent may be asked about a choice of brands given other attributes such as price, functions, etc. The metric input is used to estimate measurement of the relative importance of each attribute, but only for that respondent. In metric conjoint, there is no mechanism to aggregate these estimates so as to make inferences about populations – the part-worth estimates only reflect the choice characteristics of the individual, not the sample.

Adaptive Conjoint (ACA)

One drawback to both traditional ratings-based conjoint and metric conjoint is that if a researcher is interested in investigating multiple attributes and levels, the experiment quickly develops into an unmanageable number of possible outcomes, and a respondent cannot reasonably be expected to evaluate them all. For example, a conjoint experiment with ten attributes of three levels each produces 3^{10}, or 59,049 possible combinations of attributes and levels. Clearly, respondents will not be able to complete all of the evaluation tasks.

Adaptive conjoint addresses this problem by first asking respondents to rate the importance of the various attributes before presenting paired choice tasks. These evaluations are used to estimate a baseline of respondent preferences before presenting choice tasks. The baseline utility estimates are adjusted as the respondent makes successive choices in a series of paired comparisons.

Choice-Based Conjoint (CBC)

The conjoint methods described above all share one difficulty; each requires an orthogonal design, in which main effects only can be tested. Interactions must be taken into account in the experiment design and assumed not to exist. A method developed to address this is choice-based conjoint (CBC) in which the choice tasks are simplified into an "either or" set of choices. Part worth estimates (utilities) are estimated with multi-nomial LOGIT (MNL) estimation rather than OLS, and the analysis is conducted at the aggregate level. Appendix "2-D" shows the LOGIT model and its application in CBC software.

One would use CBC when interactions are unknown, and there are a limited number of attributes and levels. The input required from the respondent is simplified to a binary "this or that" response, and allows for a "none of the above" option.

Contemporary Multi-National Conjoint Studies

Table 2-2 shows a number of multi-cultural comparative conjoint studies. These studies show that in multi-cultural experiments, choice-based conjoint is becoming more prominent, yet metric conjoint is also widely used. The subjects studied are widely varied, and demonstrate that conjoint analysis can be used to answer research questions of a multitude of divergent cultures; from sub-cultural comparisons of a U.S. sample (e.g., Shepherd et al. 2002) to perception in developing countries such as Nigeria (Okechuku & Ohyemah 1999).

Non-Consumer Conjoint Studies

Conjoint analysis in multi-cultural research is not just used for consumer preference studies! Akaah & Yaprak (1988) used metric conjoint to discover marketing executive's attitudes regarding FDI and multi-national strategic decisions. This study was revealed the central role of political risk among MNC managers in making FDI decisions. In addition, Wetzels et al. (2000) used metric conjoint to clarify artifacts of SERVQUAL dimensions with respect to supplier selection and distribution channel decisions among Singapore wholesale managers. These authors discovered that differences exist between manufacturers and wholesalers regarding the definition of service quality, and the relative importance of the components of service quality. Other conjoint experiments that look into multi-national partnership dynamics include Hill & Shaw (1995), who employed metric conjoint to establish determinants of strategic alliances between travel industry businesses in the U.S., U.K., Japan, and Korea. Mummalaneri et al. (1996) used metric conjoint to test Chinese purchasing manager's criteria for choosing suppliers and evaluating supplier performance.

Baalbaki & Malhotra (1995) used adaptive conjoint to learn what inputs exist among managers of MNCs regarding standardization decisions. Arias (1996) conducted a segmentation study using metric conjoint among Spanish and English retailers, finding that preference structures revealed by conjoint could be coupled with demographic variables to simplify segmentation of customers for financial services.

Consumer Conjoint Studies

A wide range of research questions have been addressed with conjoint analysis, including the usefulness of choice-based conjoint in developing markets (Malhotra 1988), price tolerances (Sapede & Girod 2002), perceptual differences between subcultures (Shepherd et al. 2002), how different stimuli affect preference choices (Jaeger et al. 2001), customer satisfaction (Oppewal & Vriens 2000), ethnocentricity in developing countries (Okechuku & Onyemah 1999), consumer segmentation for European retail (Birtwehistle et al. 1998), optimizing direct mail design (Vriens et al. 1998), tolerance for Western marketing practices in Eastern Europe (Klenosy et al. 1996), and the effect of marketing standardization in Western Europe (Diamantopoulos et al. 1995).

The Interface of Conjoint Analysis and Traditional Research Methods

Of particular interest in this research are three principal studies that serve as a springboard for the experiment in this dissertation. The first study is that of Tsalikis, Seaton & Tomaras (2002), in which the authors purposefully demonstrate the usefulness of conjoint analysis in multi-cultural comparative research. In this study, two samples of Greek and U.S. consumers were asked to evaluate ethical situations in a conjoint experiment. The authors readily admit that there are potential sample biases (convenience samples; plus demographic differences), yet there remains a strong demonstration of the usefulness of conjoint as a research technique. While most multi-cultural researchers use some sort of means-based testing, conjoint allows for an interpretation of the impact (importance) of different variables on the phenomena of interest. In addition, this study clearly shows that conjoint analysis can illuminate differences in response tendencies.

The second study of interest is that of Shepherd, Tsalikis & Seaton (2002), in which a comparison of means (ANOVA) is evaluated concurrently with a conjoint experiment. A convenience sample of 209 respondents in a Southern U.S. city were divided into three groups of sub-cultures: Anglo, Hispanic (long-time U.S. residents who speak primarily English), and Immigrant Hispanic (recently moved to the U.S., speaks primarily Spanish), and asked to evaluate ethical scenarios based upon input manipulated in the conjoint experiment. The results showed that if only using traditional means-based analytical techniques, researchers may be tempted to draw conclusions that may or may not be accurate. Specifically, ANOVA produced no significant differences among the sub-groups, while metric conjoint analysis revealed that the preference patterns were different along several dimensions.

The third is a test of Malhotra (1988), in which the author seeks to understand how simplified input (binary choice) performs in developing markets. In this study, 208 U.S. homeowners served as respondents and selected between pictorial treatments in a choice-based conjoint experiment. The findings reveal that simplified input procedures (binary either-or choice responses and pictorial stimuli) perform very well in terms of predictive and convergent validity. It is striking that the experiment establishes choice tasks as a valid means of gaining accurate input from respondents who would obviously struggle with traditional scaled questionnaires.

These conjoint experiments show that in multi-cultural comparative research, there is value in the simplified system of collecting data from respondents. First, the choice function may provide an insulating effect from the laboriousness of questionnaire response. Specifically, this is clearly demonstrated by Malhotra (1988).

The Case for Conjoint Analysis in Equivalence Diagnostics

It is clear from a review of multi-cultural comparative research that a majority of researchers rely upon methods that make use of various scales. Methodological issues – including translation equivalence – inherent in the use of these scales bring to the forefront the question of whether scales used simultaneously in different cultural contexts are viewed similarly by the divergent cultural groups.

It is also clear that though the problem of equivalence error is routinely cited as a source of data contamination (e.g., Malhotra et al. 1996, Mullen 1995, Douglas & Craig 2000), there are *no* studies that examine the problem from a position of knowledge. By this it is meant that there currently is no research that begins with a known equivalence error and documents the resulting equivalence failure.

Given the importance of proper data collection, analysis, and reporting, one may wonder why no such study has been undertaken. Academics use data collected from divergent cultures to test theoretical models, while practicing marketing managers use similar data to make important decisions. Simply stated, bad data, bad analysis, and bad reporting lead to bad theory and bad decision making. Clearly, an experiment that shows unambiguously the psychometric consequences of equivalence error can be of great value.

It can be argued that a study of that type may rely upon data generated from *stated preferences* for analysis, but does not reveal true choice behaviors in conditions of scale inequivalence. Conjoint analysis provides the opportunity to gain great insight into respondents' true choice behaviors due to the following properties of conjoint; first that items for analysis (e.g., exogenous variables in a structural model) can be examined

concurrently with one another, rather than relying upon responses that are garnered for one item in exclusion. This is important because of the potential for translation error to impact other sources of bias, such as the self-reference criterion.

Second, conjoint analysis provides researchers with the opportunity to present research questions in such a way as to greatly reduce the evaluative tasks of the respondent (Malhotra 1988, Tsalikis et al. 2002). These authors have demonstrated that simplified tasks provide highly accurate responses, and have also shown that simplified (binary) input is an excellent way to gain information in a context where respondents are expected to have limited information. Though the objective of the Malhotra (1988) study was to discover whether simplified presentation of the choice tasks (pictorial versus written) produced equally reliable part-worth estimates, it can be argued that in multi-cultural contexts, simplified input is important where language barriers may exist..

Third, choice-based conjoint analysis may mitigate response styles due to the simplified input system (Tsalikis et al. 2002). It has been demonstrated that using conjoint analysis concurrently with traditional means-based analytical techniques affords great insight into response styles, and may point to areas of confusion or indecision in a way that multivariate techniques cannot.

Finally, conjoint analysis provides the researcher with a powerful tool to gain greater insight over the true intended responses of multi-cultural respondents (Shepherd et al. 2002). Even in instances when normal equivalence diagnostic tests provide evidence in support of construct equivalence, and no differences are noted between samples, conjoint analysis has been shown to provide key information that clearly demonstrates differences between stated preferences and revealed preferences.

Conclusion

Current multi-cultural research is a very complex process, and subject to a number of potential biases (Douglas & Craig 2000, Malhotra et al. 1996). It has also been demonstrated that scales used in multi-cultural marketing research have been developed mainly in U.S. contexts, and modified for use overseas. These scales exhibit varying degrees of reliability and validity in multi-cultural contexts, and the source of this variation of reliability and/or validity can be guessed, but remains unknown. For example, there is no evidence that clearly delineates translation equivalence error from any other type of equivalence error, such as calibration, metric, or functional equivalence. For illustrative purposes, if a researcher decided to use a scale to measure advertising perception in a developing economy, and the results from multi-group analysis indicate potential equivalence error, how would the researcher know *exactly* what the source of the problem is?

In addition, methods used to test for scale inequivalence rely only upon data generated from stated preferences, which if scale inequivalence is extant, that data is inherently biased and corrupt. There is no research that examines choice behaviors in conditions of a known equivalence error that would illuminate potential differences between choice behaviors and statements of preferences in such a case.

Therefore, the focus of this dissertation is to understand with greater clarity the effects of equivalence error; specifically, how scale inequivalence impacts the manner in which respondents record stated preferences, the psychometric effects of these response behaviors, and differences between stated and revealed preferences in conditions of scale inequivalence. This research will then allow researchers to distinguish with greater clarity

equivalence errors from true differences in perception between two or more cultures. In addition, this research provides a basis for improving scale development techniques in multi-cultural research. Chapter 3 formalizes the research questions and hypotheses related to the effect of equivalence errors on semantic differential scale responses and also details the experimental design employed.

Appendix 2A: MARKOR Scale Items

Kohli & Jaworski (1993) – as detailed by Olsen (2001)

Intelligence Generation

1. In this business unit, we meet with customers at least once in a year to find out what products or services they will need in the future.
2. In this business unit, we do a lot of in-house market research.
3. We are slow to detect changes in our customer's product preferences.
4. We poll end users at least once a year to assess the quality of our products and services.
5. We are slow to detect fundamental shifts in our industry (e.g., competition, technology, regulation).
6. We periodically review the likely effect of changes in our business environment (e.g., regulation) on customers.

Intelligence Dissemination

1. We have interdepartmental meetings at least once a quarter to discuss market trends and developments.
2. Marketing personnel in our business unit spend time discussing customers' future needs with other functional departments.
3. When something important happens to a major customer or market, the whole business unit knows about it in a short period of time.
4. Data on customer satisfaction are disseminated at all levels in this business unit on a regular basis.
5. When one department finds out something important about competitors, it is slow to alert other departments.

Organizational Responsiveness

1. It takes us forever to decide how to respond to our competitor's price changes.
2. For one reason or another, we tend to ignore changes in our customer's product or service needs.
3. We periodically review our product development efforts to ensure that they are in line with what customers want.
4. Several departments get together periodically to plan a response to changes taking place in our business environment.
5. If a major competitor were to launch an intensive campaign targeted at our customers, we would implement a response immediately.
6. Activities of different departments in this business unit are well coordinated.
7. Customer complaints fall on deaf ears in this business unit.
8. Even if we came up with a great marketing plan, we probably would not be able to implement it in a timely fashion.
9. When we find that customers would like us to modify a product or service, the departments involved make concerted efforts to do so.

Appendix 2B: CETSCALE Scale Items

Shimp & Sharma (1997)

1. Only those items that are unavailable in the U.S. should be imported.
2. American Products, first, last and foremost.
3. Purchasing foreign-made products is un-American.
4. It is not right to purchase foreign products.
5. A real American should always buy American-made products.
6. We should purchase products manufactured in America instead of letting other countries get rich off us.
7. Americans should not buy foreign products, because this hurts American business and causes unemployment.
8. It may cost me in the long run but I prefer to support American products.
9. We should buy from foreign countries only those products that we cannot obtain within our own country.
10. American consumers who purchase products made in other countries are responsible for putting their fellow Americans out of work.

The original scale utilized a 7 point Likert-type scale.

Appendix 2C: SERVQUAL Scale Items

Parasuraman, Zeithaml & Barry (1988)

Tangibility

1. They should have up-to-date equipment.
2. Their physical facilities should be visually appealing.
3. Their employees should be well dressed and appear neat.
4. The appearance of the physical facilities of these firms should be in keeping with the type of services provided.

Reliability

5. When these firms promise to do something by a certain time, they should do so.
6. When customers have problems, these firms should be sympathetic and reassuring.
7. These firms should be dependable.
8. They should provide their services at the time they promise to do so.
9. They should keep their records accurately.

Responsiveness

10. they shouldn't be expected to tell customers exactly when services will be performed.
11. It is not realistic for customers to expect prompt service from employees of these firms.
12. Their employees don't always have to be willing to help customers.
13. It is okay if they are too busy to respond to customer requests promptly.

Assurance

14. Customers should be able to trust employees of these firms.
15. Customers should be able to feel safe in their transactions with these firms' employees.
16. Their employees should be polite.
17. Their employees should get adequate support from these firms to do their jobs well.

Empathy

18. These firms should not be expected to give customers individual attention.
19. Employees of these firms cannot be expected to give customers personal attention.
20. It is unrealistic to expect employees to know what the needs of their customers are.
21. It is unrealistic to expect these firms to have their customers' best interests at heart.
22. They shouldn't be expected to have operating hours convenient to all their customers.

Appendix 2D: LOGIT Estimation for Choice-Based Conjoint

Adapted from Huber (2001)

A general choice model was described by Haaijer & Wedel (2001) in which a utility function assumes that the respondent will maximize the utility of a profile of attributes given a set of choices. According to Haaijer & Wedel (2001), This is expressed as:

$$U_{jkm} = X_{km}\beta + \ell_{jkm} \qquad (1)$$

"where X_{km} is . . . a vector of variables representing characteristics of the *m*th choice alternative in choice set *k*, β is a . . . vector of unknown parameters, and ℓ_{jkm} is the error term (p. 358)"

Sawtooth Software's CBC module uses multi-nomial logit (MNL) estimation methods for deriving the part-worth estimates (utilities). Huber (2001) details the LOGIT model embedded in the CBC program as follows:

$$P_i = \exp(U_i) / \Sigma_s \exp(U_j) \qquad (2)^*$$

where P_i is the probability of a respondent choosing an alternative (*i*) from a set of alternatives with given utilities. Exponentiation of the utilities (*U*)is performed both to ensure that the probabilities are always positive and to ensure that the probabilities do not change if all the utilities are increased by a constant (Huber 2001). The utilities (U) are an expression of the marginal impact of a change in an attribute, and expressed in the following function:

$$U_i = \Sigma_k \beta_k X_{ik} \qquad (3)^*$$

Which is simply the summation of the general utility function as expressed in equation 1. The marginal effect of differing levels of an attribute (X) is calculated by taking the derivative of P_i with respect to X, which produces the following function:

$$dP_i / dX_{ik} = \beta_k P_i * (1 - P_i*) \quad (4)*$$

where P_i* is the predicted probability of choosing i in the choice set provided in the model (p. F-7). With this expression, the marginal impact of a change in a variable is a function of the probability of choosing some alternative, and maximized when the respondent is undecided (when the probability is at .5). So, if the respondent is "sitting on the fence" with regard to a choice task, the impact of changing a level of an attribute is maximized, and this effect decreases as the respondent either adopts or rejects the attribute.

** Some notation has been modified to maintain consistency between sources.*

Chapter 3

METHODOLOGY

INTRODUCTION

This chapter describes an experiment that was designed to manipulate translation equivalence error in a scale, which allows for a far greater understanding of the effects of equivalence errors in multi-cultural research. Thus, the following sections present the research hypotheses, experimental design and methodology used to address the specific research questions. A review of literature regarding multi-cultural research has shown that scales developed in the U.S. and "exported" for use in divergent cultural environments exhibit tendencies to be unstable, presumably because of a failure in some sort of equivalence. The sources of equivalence failures remain largely unknown. The consequences of using information gained from scales with undetected equivalence errors cannot be understated. Academics stand in need of high quality data to make the proper inferences about cultures being studied. Practitioners absolutely must have information based upon high quality data which must be properly interpreted to make good investments and to avoid making mistakes that are measured in millions of dollars. Methods for diagnosing equivalence failures are currently restricted to two prominent techniques: multi-group structural equation modeling, and optimal scaling.

These methods are postulated to work based upon assumptions regarding the data generated by scales that have equivalence errors; namely that data distributions for inequivalent items will be significantly different, and in the case of multi-group analysis,

that error terms are likely to be disturbed when scale items or factors are not equivalent. In addition, for these techniques to work, there *must* be fundamental differences in the metrics of the items or factors in question. These metric differences must also be dependent upon different response patterns; this raises several relevant research questions. First, what are the relationships between equivalence phenomena such as translation equivalence, metric equivalence, and response styles? Second, is all of the information needed to truly assess scale equivalence inherent in the scale? Third, can assumptions underlying the use of multi-group analysis lead to erroneous conclusions? And finally, can an analysis of choice behaviors provide greater insight into the problem of scale equivalence?

RESEARCH HYPOTHESES

To answer the above research questions, the following relationships are posited to exist: equivalences (non-metric) affect response styles, which in turn impact metric equivalence. Figure 3-1 presents a graphic representation of these relationships. Here it is shown that various equivalences (non-metric) cause spurious responses (response styles), which in turn produce differences in the "psychometric properties of the data" (Malhotra et al. 1996) or metric equivalence.

It is important to distinguish between the response styles traditionally described in multi-cultural literature and the meaning as expressed here. Response styles have been defined in previous research in such a way as to describe a response pattern that is systematic to a test instrument and within a cultural group. As presented here, it is argued that "response style" can refer to spurious responses that are systematic within a culture, but *not* across all scale items. Instead, the response style is seen to be systematic

to particular items that fail to satisfy conditions of equivalence, either functional, conceptual, instrument, translation, or calibration.

Figure 3-1

The Relationship Between Equivalences, Response Styles, and Metric Equivalence

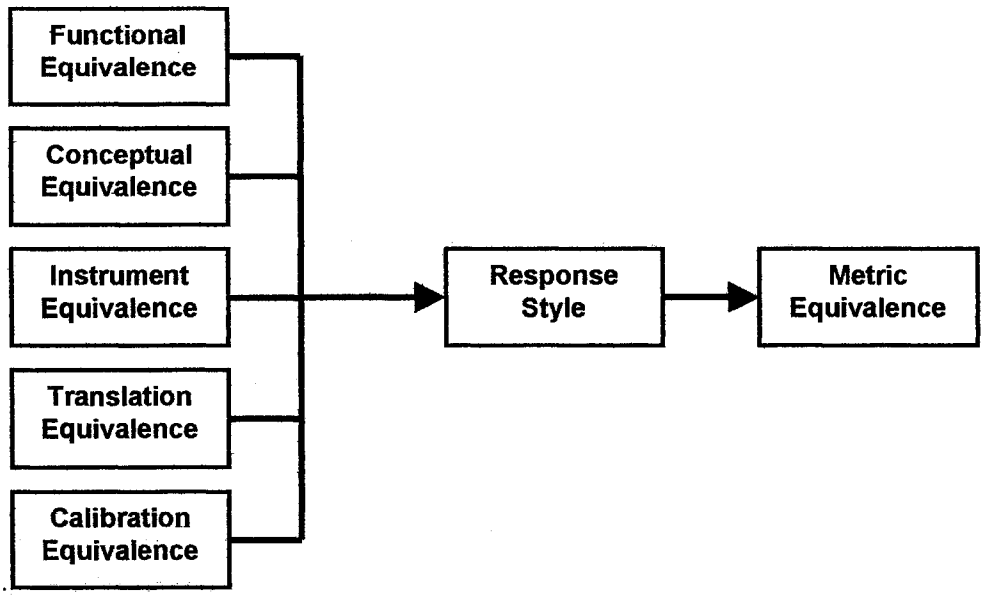


Figure 3-2 shows the focus of investigation in this study. Of particular interest is how translation errors affect the response styles for those items, and how these response styles affect metric or scalar equivalence. In addition, it is posited that stated preferences as indicated in a scale will differ from true preferences for items that exhibit translation error. To test the presented relationships, an experiment was designed to manipulate translation equivalence error and study its effect on responses, data distributions, and choice patterns. A semantic differential scale was used as the test instrument, leading to a series of hypotheses.

Figure 3-2
Focus of Research

The first step in the discovery process was to examine the effect of translation equivalence error on response patterns. With a semantic differential scale, one may reasonably expect that respondents will be unwilling to offer scores that lean too far to either pole due to an uncertainty of what is being asked. If this is true, distributions for manipulated translation errors may be expected to be somewhat concentrated toward the center and less widely distributed than for items that are clearly understood. Therefore the following hypotheses is offered:

H_1: Data distributions for non-manipulated items will be the same for the experiment and control groups.

Since metric equivalence is assumed to be affected by errors in translation equivalence, one could reasonably expect that items that are not equivalent will exhibit responses that are not consistent with items in which there is no equivalence error (Craig & Douglas 2000). If this is true, the following hypothesis should be upheld:

H_2: Distributions of data for manipulated items will differ between the experiment and control groups with respect to manipulated items.

If H_2 is supported, it is reasonable to assert that respondents would not be willing to indicate extreme views for items not understood. Therefore, with a semantic differential scale, the only response that indicates either no preference or no knowledge is the mid-point value, which leads to the following hypothesis:

H_3: Manipulated items will exhibit properties of central tendency bias within the experiment groups.

The second step in the experiment was to examine the factor structure of both the experiment and control groups, using exploratory factor analysis. If the response styles are disturbed with respect to manipulated items, metric equivalence should also be in question. Specifically, manipulated items should affect the factor structure of the experimental group, creating different factor solutions between experiment and control groups. In multi-cultural studies, researchers assume that the disturbances in the factor structure are due to items that are not equivalent. To be consistent with standard multi-cultural research techniques, the following hypothesis is offered:

H_4: The factor structure will be different for the experiment group and the control group.

When multi-group analysis is used to establish equivalence of scales, it is assumed that items that are not equivalent are responsible for any differences between groups (Mullen 1995). If this is the case, the predecessor to statistical tests of these differences is that the factor structure revealed in exploratory factor analysis should differ

with respect to those suspect items. To reflect this assumption of multi-cultural scale research, the following hypothesis is posited:

H_5: Manipulated items will be the only items to load differently between the control and experiment groups.

Traditionally, multi-cultural researchers would then test the invariance of the factor structure using confirmatory factor analysis. In multi-group analysis, the focus of discovery is in the theta matrix of error terms. Of particular interest are "systematic error variances," or variances that are systematic to one item, but not to other items in the factor. According to Mullen (1995), multi-group analysis would hold that if systematic error variance is not apparent, then there are real *perceptual* differences between the groups. Namely, equivalence is satisfied, and the manipulated items have real meaning among respondents in both groups. If a semantic differential scale is used, and a central tendency error is extant within manipulated items, it is probable that the distribution properties of the items will not necessarily create a significant difference in the error terms – noted by Mullen (1995) as systematic error variance. This would lead to invalid conclusions, namely that the manipulated items are grouped in separate factors, factors renamed, and either *emic* or *pseudo-etic* inferences made. It is this possibility that produces the third step, which is to compare the stated preferences with revealed preferences. It is important to note that if H_5 is supported, several invalid inferences may be made; that respondents are ambivalent regarding the item, or that respondents preferred a "medium" value, just to name a couple. This is a serious departure from the truth when in fact the respondents did not understand the item, therefore it is relatively unimportant to them. This phenomenon gives rise to the following hypothesis:

H_6: The statistically significant items in the scaled survey will be significant attributes in the conjoint experiment within the control group.

Items that are significant, if they had not been manipulated for translation equivalence error, can be expected to be invariant between control and experiment groups with respect to the relative importance of the items. Thus, non-manipulated items should have equivalent part-worth estimates between experiment and control groups in the choice-based conjoint experiment, which is the basis for the following hypothesis:

H_7: Part-worth estimates for non-manipulated attributes will be the same for control and experimental groups.

If items are manipulated for translation equivalence error, then it is reasonable to expect differences between the control and experiment groups with respect to the relative importance of items. Therefore, if H_7 is supported, any differences between control and experiment groups in the part-worth estimates of the items must be due to the manipulated translation error. Thus there is the following hypothesis:

H_8: Part-worth estimates for manipulated attributes will differ between experiment and control groups.

Since a major point of interest in this experiment is to understand if there are differences between stated and revealed preferences when there is a translation equivalence error, comparisons between scaled survey responses and conjoint preferences are necessary. Understanding the degree of similarity between stated and revealed preferences inherently involves making an assessment regarding whether the scale items and complementary conjoint attributes show signs of being similarly impactful from the respondent's point of view. Therefore, the following hypothesis is posited:

H_9: Factor loadings for scale items will resemble conjoint part-worth estimates for the complementary conjoint attributes in both direction and magnitude within the control group.

When items are manipulated to create a translation equivalence error, one would expect that similarities between stated preferences and revealed preferences may not maintain the same characteristics. However, items that are clearly understood and are important in the control group will remain so in the experiment group when they are not altered to be misunderstood. Hence the following hypothesis:

H_{10}: Factor loadings for non-manipulated scale items will resemble conjoint part-worth estimates for the complementary conjoint attributes in both direction and magnitude within the experimental group.

Consistency between scale response and conjoint preferences can be expected to deteriorate for items that are not understood. Specifically, items that are clearly understood and are important in the control group will become unimportant in the experiment group when altered to be misunderstood. However, information may be misleading in conditions of translation equivalence error; namely that with a semantic differential scale, misunderstood items may exhibit central tendency bias, which if consistent, provides concentrated mid-point responses that will appear to be highly significant. In contrast, conjoint analysis should reveal that the same item manipulated in CBC will be unimportant, and the associated part-worth estimation should not be reflective of coefficient estimates from structural equation modeling either in direction or magnitude. Therefore, the final hypothesis is presented:

H_{11}: Factor loadings for manipulated scale items will not resemble conjoint part-worth estimates for the complementary conjoint attributes in direction and magnitude within the experimental group.

THE TEST INSTRUMENT

The measurement model test instrument and conjoint questionnaire were designed to complement each other to capture true preference behaviors regarding perception of the usefulness of components of an ad. Advertising was selected as the scale subject for a number of reasons. First, advertising inherently contains an element of subjectivity that is routinely misinterpreted and/or misunderstood. It is this property that allows for the design of an experiment that truly mimics the process and outcomes of translation error.

Second, international advertising is fraught with mistranslated words and idioms, of which books are filled. From a research point of view, this presents a gilded opportunity to gain insight into phenomena that are both theoretical *and* practical. For example, the question of whether public perception is altered more by cognitive appeals or emotive appeals given a certain target audience and type of product or service may certainly be considered theoretical, but there is also direct practical application for advertising and public relations. This link (or tie) is less clear for such topics as market orientation, where the application of theoretical concepts looses clarity. In this sense, using an advertising context allows for a relatively easy manipulation of words and phrases, which impact both theoretical and practical implications of translation error.

Third, it is posited in this dissertation that mistranslated scales produce bad information, which in turn results in decision making that is based on faulty assumptions – the final outcome being poor business performance. It would be very difficult to imagine *any* subject that demonstrates this "trickle down" effect more clearly than advertising. Advertising decisions impact product positioning and branding, both of which are the basis for successful competition. Poor advertising decisions based on false

information will result in mis-positioned and poorly branded products which will fail in the face of competent competition. Thus, the value of this research is enhanced by using advertising as the contextual environment for the experiment.

For the measurement model, a semantic differential scale was used to more easily facilitate the use of the scale items as attributes in conjoint. In addition, it was hypothesized that a semantic differential scale is unique in that translation equivalence errors may not produce the psychometric effects commonly seen in Likert-type scales, which further highlights the value of revealed preferences analysis, as made possible by the choice tasks of choice-based conjoint.

Several extant scales may have been candidates for use in this experiment, however for ease of manipulation, a proprietary scale was used that was easily adapted to conjoint experimentation. In addition, a scale was needed that mimics the process of translation/back translation error, which can be better accomplished in semantic differential scaling because each pole must be translated and back-translated. Traditional Likert-type scales may have strings of words associated with the main term of interest, which complicates the manipulation of translation error. For example, a single item in the MARKOR scale reads, "in this business unit, we meet with customers at least once a year to find out what products they will need in the future" (Kohli & Jaworski 1993). This item presents a potentially serious question as to which words, if manipulated, would contribute to translation equivalence error.

Based upon the advertising literature, this scale provides the opportunity to realistically reflect errors of translation that go unnoticed; specifically a scale may be developed in one language or cultural setting, translated into another language, and back-

translated without correcting for differences of dialect. In this case, a questionnaire was developed in American English, with some items "mistranslated" into British English that an American sample would not be expected to understand.

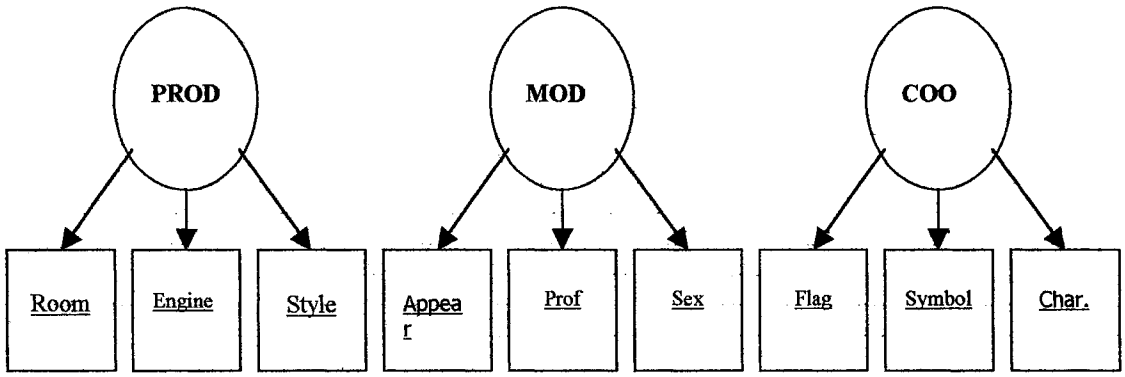
Measurement Model

The measurement model was designed to test respondents' perceived usefulness of various ad attributes along types of information cues. Numerous studies have shown that information cues are many and varied, and include everything from price to model attractiveness (Resnik & Stern 1977, Joseph 1982). For this experiment, each cue is assumed to constitute a single and distinct construct that captures the properties of a unique dimension of information. Since this experiment involves misinterpretation, one desirable property of these constructs is that they contain some latitude for misunderstanding; therefore, objective cues such as price were not considered.

Resnik & Stern (1977) identified several information cues in a seminal content analysis of television ads. These cues included price, quality, performance, components or contents, availability, special offers, taste, packaging, safety, nutrition, and new ideas. Later researchers added such cues as country of origin (i.e., Head 1988, Elliott & Cameron 1994, Moon & Jain 2001), product (e.g., Sengupta et al. 1997, Prashant et al. 1996), and model attributes (e.g., Joseph 1982, Solomon et al. 1992). Since country of origin, product, and model attributes have been extensively researched and are subjective, these cues are considered suitable for experimentation. Thus, the measurement model was designed to measure three distinct constructs related to the three cues: product (PROD), model (MOD), and country of origin (COO) (Figure 3-3). It is important to note

that in this experiment, the actual factor structure was not important; this scale is only constructed to facilitate the manipulations needed to imitate scale translation errors.

Figure 3-3
Simple Three Construct Measurement Model: Advertising Attributes.



Product Dimensions

Prashant et al. (1996) found that advertisements that emphasized product information affected information processing under circumstances where the product information is processed on an item-by-item basis. Specifically, these authors demonstrated that with a more involved purchase decision, people respond more positively to ads that combine direct and indirect cues (related directly to the product or indirectly via contextual stimuli). Therefore, one would expect that an itemized presentation of product attributes would be considered carefully in the context of an auto ad, especially when contextual cues are apparent. The product items were therefore specific and descriptive of the product including size (trunk size), engine (powerful versus economy), and style (sporty versus family).

Model Dimensions

The Beauty Match-Up Hypothesis states that "perceivers distinguish multiple types of beauty" (Solomon et al. 1992, p. 23) and that these qualities are determined by the viewer to be more or less suitable for the product being advertised (Solomon et al. 1992). In this experiment, the attributes of the model were represented according to body shape (heavy versus skinny), profession (white collar versus blue collar), and sex (male versus female), which can be expected to contribute more or less to a respondent's perception of whether the model is suitable for the ad, given the nature of the product being advertised. The items used to test these dimensions were drawn from ideas researched by Wells (1964), in which models are assumed to elicit an emotive reaction from viewers.

Country of Origin

Country of origin effects in advertising mirror to a degree ethnocentrism. This has been well documented, specifically with regard to country of origin information affecting judgments of products (e.g., Schmidt & Dube-Rioux 1989). In this experiment, the object was to test the contribution of symbols of national identity to the usefulness of the ad. To do so, country of origin is operationalized as "flag," "national symbol" and "national character." Each of these items is presented as either clearly visible or not visible as semantic differential anchors.

Of these factors, one can reasonably be expected to produce a cognitive reaction (product information), while the others should elicit an emotive response. The difference between the model and country of origin constructs is that emotive reactions are

associated with fundamentally different cues; one being a model and the other nationalism.

Scale

Each of the three factors - product, model dimensions, and country of origin [COO]) contained three items on a five-point semantic differential scale, one of which was manipulated for the experiment group (Appendix 3-A). For product dimensions, the size of the vehicle was inferred by using "trunk size," which for the experiment group was translated to, "boot size." The "engine" item had "powerful" versus "economy" as anchors, and the "style" item has either "sporty" or "family" as anchors.

For Model dimensions, the "appearance" item had either "very thin" or "very heavy" as anchors, and "sex" obviously had "male" and "female" as anchors. The "profession" item was manipulated in that the control group had "blue collar" versus "white collar" as anchors, while the experiment group had fictitious professions "tordelman" and "daveller" as anchors. All "country of origin" items are represented as either "shown clearly" or "not shown." The items were "flag," "symbol (e.g., Statue of Liberty)" and "national character" which was Uncle Sam for the control group and John Bull for the experiment group.

In addition, the item manipulations inherently assume that translation equivalence error occurs by degrees; specifically, respondents may understand, partially understand, or not understand items in the questionnaire. To test the effects of these different levels of understanding, the size item was manipulated to read, "large boot," in which respondents were expected to understand "large," and use it as a proxy for the size of the vehicle, while not understanding "boot." The item for national character, "John Bull"

offered the probability that a representative sample of American consumers would not understand, while some might. The role item contained fictitious words to ensure that respondents could not understand the item.

Manipulation Check

To test whether translation equivalence error has been achieved, a binary item, "all items in this survey are understood clearly: yes, no" was included. A chi-square test was used to test for significant differences between the control and experiment groups on this item. In addition, a scaled question was also used in which the respondents were asked to rate how well they understood the items in the scale. An independent samples t-test was used to test for significant differences between the control and experiment groups.

Covariates and Demographics

Information was collected for demographics that included age, education, and sex. Each of these items was presented in a categorical format, with the exception of age which was a ratio scale.

To control for the possibility that some respondents had been acculturated to U.K. English, and would therefore bias the result of the experiment, a nationality section was included in the questionnaire. The questions included, "do you consider yourself culturally American?", "how long have you lived in the U.S.?", "were you born in the U.S.?", and "if not, what age did you move to the U.S.?"

Further acculturation questions included potential exposure to British broadcast programs, which if either watched or listened to on a regular basis may lessen the

probability that manipulated items will be misunderstood. These questions included how often the respondent is exposed to U.K. (BBC) television and radio broadcasts. A question was also included regarding the frequency of travel outside of North America (specifically to Europe). All questions of acculturation were presented in a categorical format.

Pretest Results

A pretest with 22 observations provided preliminary indications of a clear three factor structure. Principal components extraction was used in exploratory factor analysis, with the factor loadings shown in Table 3-1, and Bartlett's test was significant at the .01 level. The only complex variable in the solution is the item for the sex of the model, which seemed to be identified most closely with the product cues. Given the extremely limited number of observations in the pre-test, this is very encouraging. In addition, even with the limited number of observations, the reliability of the factors shows promise; Chronbach's alpha for each factor, PROD, MOD, and COO were .67, .78, and .79 respectively.

Table 3-1
Exploratory Factor Analysis
Pretest n=22

Rotated Component Matrix[a]

	Component		
	1	2	3
ENG	5.907E-02	.496	.315
STY	-.318	.847	.132
RM	8.294E-02	.886	-.308
APP	.189	.299	.778
ROLE	8.381E-02	-8.73E-02	.808
SEXMOD	-.375	-.597	-.379
FLG	.857	6.613E-02	.388
SMBL	.851	.145	.424
CHAR	.787	-.210	-.369

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 6 iterations.

SAMPLES AND DATA COLLECTION

A probability sample of the consumer population of the U.S. was purchased from a professional sampling firm, Survey Sampling, Inc. This sample was a panel of participants who had agreed to participate in various surveys, and was divided randomly into control and experiment groups. For each group, holdout tasks and holdout samples were used to assess validity and reliability.

Survey Sampling, Inc. possesses all pertinent information regarding the profiles of the respondents, and has collected demographic data beyond that in the questionnaire. Respondents were offered a small incentive for participation, and upon agreeing to participate in the survey, were directed to log on to a secure website with unique

passcodes to respond to the test instruments. Data was then transferred electronically to a server, and formatted for analysis.

The sample size was set to satisfy requirements for both structural equation modeling and LOGIT estimation. Little is written about the impact of sample size on conjoint analysis, primarily because most conjoint models estimate utilities for individual respondents, not for an aggregate sample of respondents. Thus, for metric and traditional ratings-based conjoint, the principal issue is number of observations per respondent, not number of respondents. For choice-based conjoint, utility estimations are made at the aggregate level; therefore the sample of respondents must be sufficient for valid and reliable LOGIT estimations. The total sample specified for purchase was 300, with 150 in the control group and 150 in the experiment group. This would exceed the widely accepted standard of ten observations per parameter in structural equation modeling (Hair et al. 1998), and also be more than sufficient for LOGIT estimation, allowing for deletion of unusable surveys.

DATA ANALYSIS

Analysis for the data distributions and descriptive statistics and exploratory factor analysis was conducted with SPSS 11.0. Non-parametric tests were conducted to test equivalence of distributions. Confirmatory factor analysis and tests for factor invariance of the measurement model were conducted with AMOS 4.0.

The conjoint experiment used the Multi-Nomial Logit model embedded in Choice-Based-Conjoint from Sawtooth Software. Choice-Based-Conjoint (CBC) was chosen for a number of reasons; first, it is assumed in the experiment that one major problem of using scaled questionnaires is that respondents may incorrectly indicate a

degree of preference due to scale inequivalence. Since CBC offers an either/or – decompositional approach, the input for CBC is limited to either/or choices, which was expected to significantly lower the potential for scale/degree of preference bias.

Second, the Multi-Nomial Logit model allows for modeling of interactions, whereas OLS based metric conjoint only allows for one-way interactions. Choice-Based-Conjoint analysis (CBC by Sawtooth Software) was chosen as the analytical tool for the experiment to uncover "revealed" preferences of those who took the survey, and to discover whether there is a large and/or meaningful difference between the stated preferences expressed in the scale item questionnaire and those revealed through CBC.

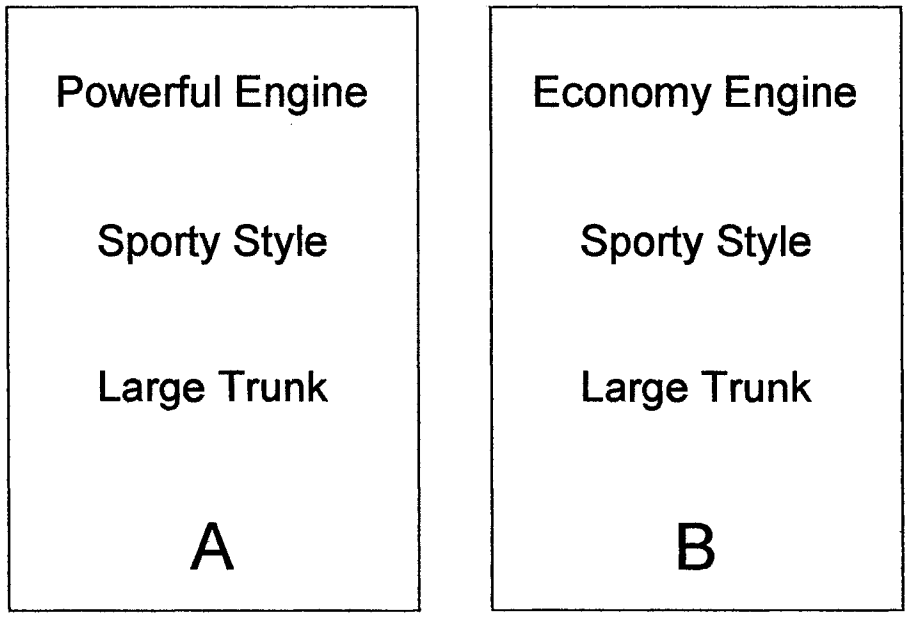
CBC was specifically selected over other conjoint models for a number of reasons: first, it is assumed in the experiment that one major problem of using scaled questionnaires is that respondents may incorrectly indicate a degree of preference due to scale inequivalence – not understanding the item. The decompositional approach of CBC allows for simplified "either/or" input, which may be seen to mitigate the potential for scale/degree of preference bias. Also, because potential interactions are unknown, Multi-Nomial Logit (MNL) is used to estimate the part-worth structure due to its superiority over metric (OLS based) conjoint methods in its ability both to accommodate aggregate level analysis and to, which are able to manage only main effects and one-way interaction.

While the advantages of Latent Class MNL and Hierarchical-Bayes algorithms are acknowledged for their usefulness in handling heterogeneity of respondents, neither is used in this experiment due to the pre-selection of control and experiment groups, which

provides a management mechanism for expected heterogeneity between groups. In addition, the thrust of investigation is at the aggregate level, not the respondent level, ergo these techniques would not appear to add significantly to an understanding of the phenomena under investigation.

Each attribute is binary, and the levels are indicated by the same terms that are used as poles on the semantic differential scale. Therefore, for the control group, attributes related to product dimensions are powerful/economy (engine), sporty/family (style), and large trunk/small trunk (size). Attributes related to model dimensions are heavy/skinny (body), male/female (sex), and (profession). The experimental groups have manipulated attributes that mirror the manipulations in the measurement model; namely that the attribute levels for car size for the U.S. experiment group will read "large boot" and "small boot." The attribute levels for the profession of the model will read, "tordelman" and "daveller," and the national character will read "John Bull." Below are examples of CBC cards:

Conjoint Treatment Cards



Please choose which of the above ads is most useful to you:

A ☐ ☐ **B**

The number of levels effect is not of concern due to the fact that all attributes are binary. With no variation among the attributes with regard to the number of levels, all part-worth estimations can safely be assumed to be unbiased with respect to varying attribute levels. Similarly, the range of levels effect was mitigated by uniform attribute ranges.

Some CBC procedures allow for a "none of the above" choice to stabilize the estimation of part-worth estimates, which may have a tendency to be over (under) stated. This was not done in this study for two reasons: first, the objective of the conjoint exercise is to mirror the semantic differential scale which has no "none of the above"

choices available to respondents. Second, even though the likelihood of seriously overstated part-worth estimates in this experiment is not high, it may be useful to note that the effect of manipulated items may be highlighted in this manner.

As noted above, the conjoint choice tasks were designed to mirror the hypothesized factors from the measurement model. Therefore, each respondent was asked to evaluate three separate full factorial models of three attributes with two levels each (2 x 2 x 2). For example, the PROD factor consists of three items; engine, style and size. The complementary conjoint experiment had three attributes; engine, style and size, with the previously described levels. In this way, each respondent will be able to evaluate all items with a minimum of choice tasks, which total 24.

Appendix 3A

Scale Items

PRODUCT

Engine

Powerful	□	□	□	□	□	□	□	Economy

Style

Sporty	□	□	□	□	□	□	□	Family

Room

Small Trunk	□	□	□	□	□	□	□	Large Trunk

MODEL

Appearance

Very thin	□	□	□	□	□	□	□	Very heavy

Role

Attorney	□	□	□	□	□	□	□	Janitor

Sex

Male	□	□	□	□	□	□	□	Female

COUNTRY OF ORIGIN

Flag

Clearly Visible	□	□	□	□	□	□	□	Not shown or mentioned

Symbol (e.g., Statue of Liberty)

Clearly Visible	□	□	□	□	□	□	□	Not Shown or mentioned

Uncle Sam

Clearly Visible	□	□	□	□	□	□	□	Not shown or mentioned

Chapter 4

RESULTS

INTRODUCTION

This chapter describes the data collection technique, psychometric properties of the data collected, and the analysis of the data. Five sections are used for this purpose, the first being a description of the data gathering technique and the sample. The second section tests the hypotheses related to scale responses. The third section tests hypotheses associated with the conjoint choice tasks, and the fourth section reports results from comparisons between the conjoint analysis and the structural model. Finally, the last section contains a summary of the findings and a brief discussion.

DESCRIPTION OF THE DATA COLLECTION AND SAMPLE

A probability sample of American consumers was purchased through Survey Sampling, Inc. Respondents were considered "qualified" to take the survey if they were American citizens. Administration of the data collection was conducted via a web-based panel, in which each respondent had previously indicated a willingness to participate in web-based surveys by responding to SSI with his/her consent and demographic data. Participants were invited to take part in the survey via email notification, of which 4,395 were sent. An incentive was provided where people who participated in the survey (either completed the questionnaire or were involuntarily screened out) were entered into a monthly sweepstakes with prizes that exceeded $10,000 per month. The invitations

specifically stated that the survey was a university project and not affiliated with any commercial organization. People wishing to participate in the survey were required to log into a secure SSI website (to track the various respondents), and were then redirected to the website that contained the questionnaire. Data generated from these responses were managed by an independent database manager. Sixteen respondents were screened out due to the fact that they were not U.S. citizens. The aggregate responses – responses that were collected and stored by the database manager - were 580, of which several incomplete surveys were recorded. 317 responses were deleted due to excessive missing data, leaving 263 useable surveys, and an overall response rate of 5.9%. This response rate may raise questions of non-response bias, yet later responses were compared with earlier responses with no significant differences in either the control or experimental group. The low response rate may be an artifact of web-based panels; in fact, given that the incentive was entry into a sweepstakes versus cash or other incentives, it is interesting to note that several of the deleted responses ended input either during or just before the conjoint choice tasks. This indicates that the conjoint choice tasks may have been more laborious than projected at the beginning of the project – yet the manner of data collection allows for this understanding. With conventional mail surveys, the researcher never sees questionnaires that are abandoned mid-stream.

Several options are available regarding demographic profiling of respondents; the sample purchased from SSI was specified to reflect census data demographics. Thus, the mean age of respondents who completed the questionnaire was 48.7, with a range from 18 to 82. Of this sample, 43% were male, and 57% were female. The median education level was high-school graduate with some college experience, and 44.5% of respondents

had at least a two-year college degree. Regarding advanced education, 13.2% of respondents had some sort of graduate degree.

Regarding control variables, the mean age of the respondents compared to the mean number of years living in the U.S. was 48.7 to 47.43, respectively. The range for "number of years living in the U.S." was 4 to 82. This clearly indicates that the sample is unbiased with respect to overseas exposure, and that those expatriates who are naturalized citizens had generally lived in the U.S. most of their lives. With respect to British media exposure, the mode for people viewing BBC television programming was "never" (52.9%), with less than 20% of respondents reporting having viewed BBC television at a frequency of once per month or more. BBC radio broadcasts showed even less exposure, with 95.1% of respondents listening to BBC radio broadcasts once per year or less.

TEST OF HYPOTHESES: SCALE RESPONSES

This study was designed to record the various effects of translation equivalence error in multi-cultural comparative research, and to discover any association between translation equivalence, response style, and metric equivalence. The manipulation check for effectiveness of the translation errors was represented to respondents in such a way as to provide both metric and binary responses. The interval scaled question read, "On a scale of 1 – 7 (1 being very low, 7 being very high) please rate how well you understood the items in the survey," and the binary question read as follows: "All items in this survey were clearly understood (y/n)." The metric response was tested with an independent

samples t-test; the mean for the control group was 5.36 and the experiment group mean was 4.18. This difference was significant at the .01 level. The binary manipulation check revealed that 10.1% of respondents in the control group answered, "no" to understanding all items in the questionnaire, compared to 35% in the experimental group. This difference was significant at the .01 level (z=5.00).

The first step in the analytical process was to examine the psychometric effects of translation error – or, to test the first two hypotheses regarding data distributions, which read as follows:

H_1: Data distributions for non-manipulated items will be the same for the experiment and control groups.
H_2: Distributions of data for manipulated items will differ between the experiment and control groups.

These hypotheses were tested using the Mann-Whitney U-test, the null hypothesis for which is that of distributional invariance. Results indicated that there were significant differences between the experiment and control groups with respect to Size, Social Role, and National Character (manipulated items), but not for the other items in the scale (table 4-1). This gives clear support for the first two hypotheses.

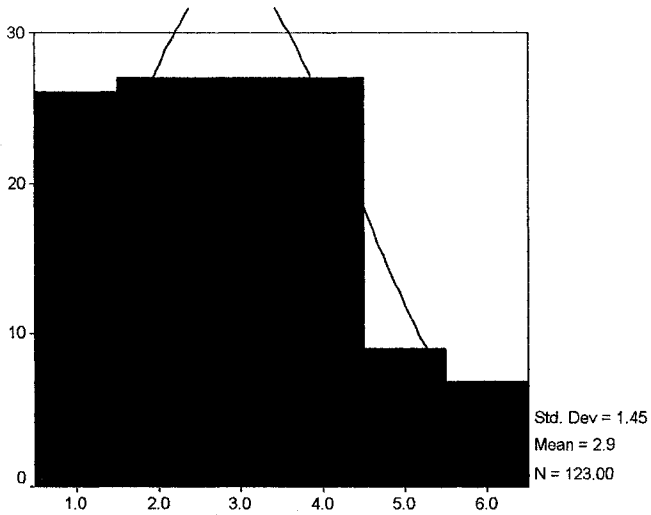
TABLE 4-1
Mann-Whitney U Test of Distribution Invariance Between Control and Experiment Groups

	ENGINE	STYLE	SIZE	APPEAR	ROLE	SEX	FLAG	SYMBOL	CHARACTER
Mann-Whitney U	8629.000	8216.500	6855.000	8468.500	7240.000	8741.500	7978.500	8226.000	4745.500
Wilcoxon W	16255.000	15842.500	17151.000	16094.500	17536.000	19037.500	18274.500	18522.000	15041.500
Z	-.268	-.936	-3.186	-.545	-2.792	-.087	-1.356	-.944	-6.672
Asymp. Sig. (2-tailed)	.789	.349	**.001**	.586	**.005**	.930	.175	.345	**.000**

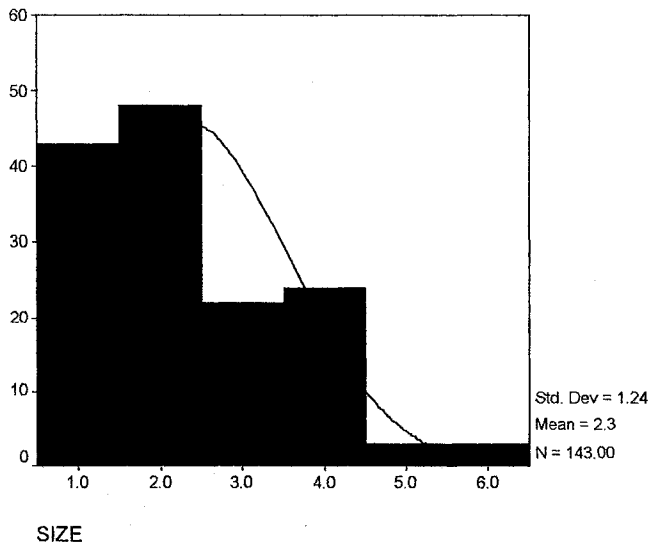
In addition, histograms with normal curves superimposed present a graphical view of the differences between the control and experiment groups with respect to the manipulated items (figure 4-2). From these it is clear to see that while some of the items may be slightly skewed, there is a pronounced tendency for experimental group responses to be concentrated toward the mid-point.

FIGURE 4-1
Histograms for Experiment and Control Groups

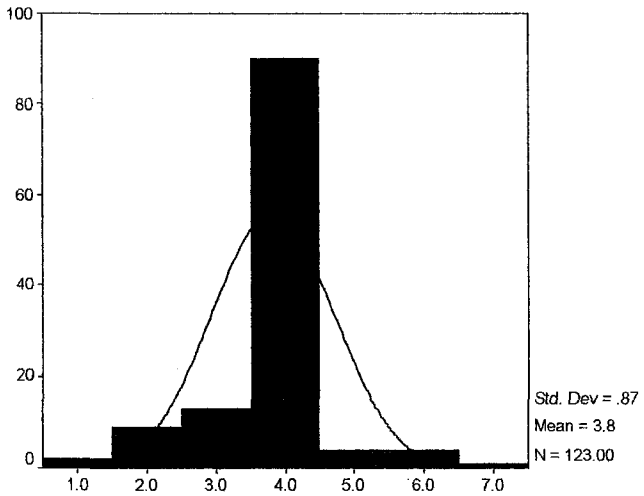
Size: Experiment Group



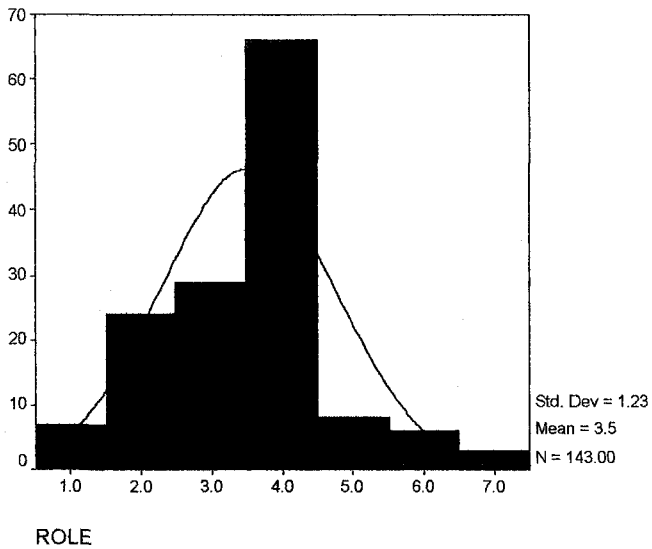
Size: Control Group



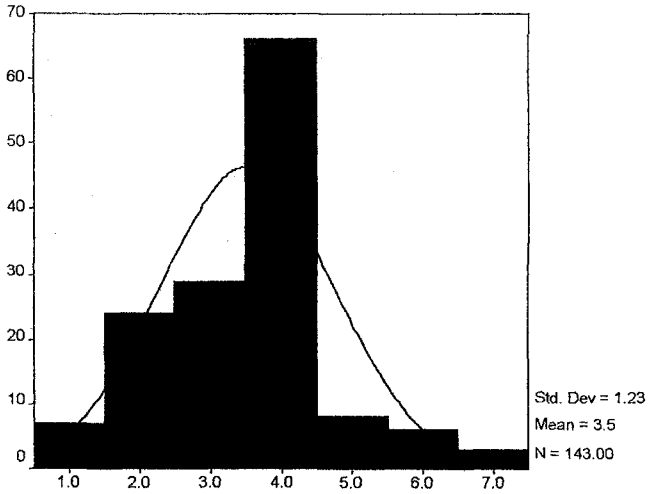
Role: Experiment Group



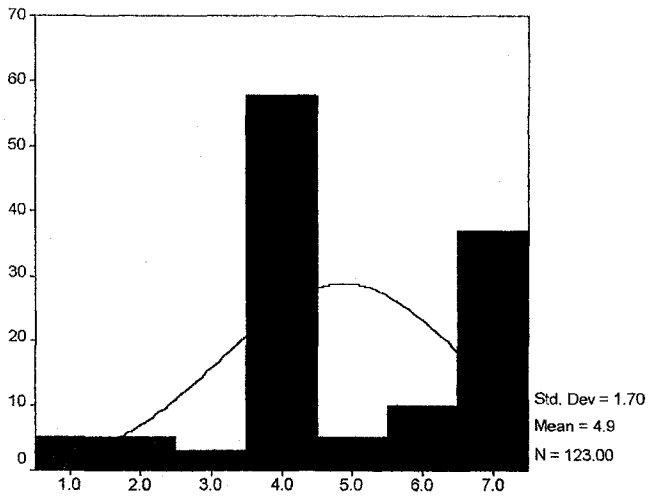
Role: Control Group



Character: Experiment Group



Character: Control Group



The next step in the analysis was to discover whether a response bias was present under conditions of translation inequivalence, or to test the third hypothesis which reads as follows:

> H_3: Manipulated items will exhibit properties of central tendency bias within the experimental group.

Central tendency bias has been conceptualized as, "a reluctance to give extreme scores" (Yu et al. 2003), or as "the proportion of heterogeneous items on which the respondent endorses the middle scale category (Baumgartner & Steenkamp 2001). For this study, midpoint response bias was re-defined to focus on particular items only, and not for an entire scale. Therefore, central tendency bias was operationalized as the proportion of responses that coincide with the midpoint of the scale for manipulated items. While the non-parametric tests above show that the distributions were different for the manipulated items, it is important to test the differences between the control and experiment groups with regard to the proportion of mid-point responses for these items. A frequency distribution of these items shows that the hypothesis of central tendency error is partially supported; table 4-2 provides the percent of midpoint responses for each item and group. Two of the three items show significant differences in midpoint response; role ($z=4.67$) and character ($z=3.94$). Size was not significant ($z=1.06$) – which may be attributable to the fact that it was designed with the weakest translation error manipulation. Since the words "large" and "small" were associated with the manipulated term "boot," it is reasonable that respondents were able to proxy "large" and "small" for a general meaning of size of the automobile.

TABLE 4-2
Proportion of Mid-Point Responses

	Experiment Group	Control Group	Z
Size	22.7%	16.8%	1.06
Role	73.2%	46.2%	4.67
Character	47.2%	24.5%	3.94

The analysis so far has demonstrated that the conceptualization of the relationship between translation error, response style, and equivalence error has been generally supported. Namely, only the manipulated items (translation error) show evidence of midpoint response bias and significantly different distributions (metric equivalence failure). Given the fact that distributions and the means of the manipulated items are different for the experimental and control groups, it stands to reason that factor analysis will reveal differences in the factors produced – this relates to the fourth hypothesis which reads as follows:

H_4: The factor structure will be different for the experimental group and the control group.

This hypothesis was tested using exploratory factor analysis (EFA) in SPSS version 11.0. Correlation matrices for the control and experimental group show a mixture of significant and insignificant correlations in varying strengths. Some differences can be seen between the correlation matrices for the control and experiment groups; namely that some items had lost significance – not just between manipulated items. Nevertheless, these

matrices demonstrate that minimum requirements for factor analysis have been met (see Table 4-3).

Table 4-3
Correlation Matrix: Control Group

	Engine	Style	Size	Appear	Role	Sex	Flag	Symbol	Character
Engine									
Style	.324								
Size	.037	-.123							
Appear	.014	.081	-.074						
Role	.241	.036	.065	**.249**					
Sex	.068	.064	.020	-.125	.000				
Flag	**.182**	.020	.072	.155	-.110	-.012			
Symbol	**.226**	.027	.095	.123	-.021	.034	**.903**		
Character	**.263**	.065	.031	.162	-.054	.127	**.790**	**.859**	

Bold items are significant at the .05 level

Table 4-4
Correlation Matrix: Experiment Group

	Engine	Style	Size	Appear	Role	Sex	Flag	Symbol	Character
Engine									
Style	.350								
Size	.322	.056							
Appear	.078	-.005	.010						
Role	.048	.172	.115	**.295**					
Sex	.077	.133	.035	.073	-.008				
Flag	.101	.139	.067	**.233**	.009	.081			
Symbol	.056	.133	.132	**.202**	.009	.059	**.895**		
Character	.077	.172	.011	.020	-.059	.041	**.387**	**.346**	

Bold items are significant at the .05 level

KMO statistics for the control and experimental groups were .647 and .542, respectively. Bartlett's test of sphericity was significant at the .001 level for both groups. While these may not be a ringing endorsement for conducting factor analysis with these data, minimum requirements for the appropriateness of using the technique have been established (Hair, et al. 1998).

Principal components factor analysis with varimax rotation was conducted for both the control and experiment groups. Rotated component matrices are shown in Figure 4-2. The initial solution shows differences between the groups in the factor structure. While both seem to provide a four factor solution, the variable loadings are different, supporting the fourth hypothesis. Since it has been demonstrated that

differences in the psychometric properties of the data were due to the manipulated translation error, one would expect that the fifth hypothesis should also be supported, which reads as follows:

> H_5: Manipulated items will be the only items to load differently between the control and experimental groups.

As can be seen in the rotated matrices, three items loaded differently: style, size, and sex. Since two of the three were NOT the manipulated items, the fifth hypothesis was not supported. The lack of support for this hypothesis brings to light substantial questions regarding the potential attenuation of other scale items when translation error is present. However, it can be seen that one of the manipulated items, "role" is a complex variable, loading on multiple factors. In "purifying" the scale by removing this variable from analysis, a three-factor solution appears (Figure 4-3). This solution yielded an improved KMO of .702, and factor loadings seemed to fall more squarely into place. "Size" was seen in the initial solution to be a "factor" of its own in the control group, yet loading strongly on the second factor in the experimental group. Since multi-cultural researchers strive for scale equivalence, comparisons are routinely made from exploratory factor analysis to eliminate variables that upset the factor structure.

FIGURE 4-2
Initial Rotated Component Matrices

Control Group

Rotated Component Matrix[a]

	Component			
	1	2	3	4
ENGINE	.203	.774	-3.00E-02	.112
STYLE	-2.07E-03	.697	-4.87E-02	-.416
SIZE	7.898E-02	-5.87E-02	-.148	.835
APPEAR	.153	.119	.787	-6.52E-02
ROLE	-.171	.483	.509	.460
SEX1	4.275E-02	.323	-.583	.110
FLAG	.947	-7.13E-04	5.654E-02	2.969E-03
SYMBOL	.959	7.263E-02	3.756E-02	7.030E-02
CHARACTE	.920	.143	-1.17E-03	1.466E-04

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 6 iterations.

Experiment Group

Rotated Component Matrix[a,b]

	Component			
	1	2	3	4
ENGINE	3.431E-02	.741	5.944E-02	.338
STYLE	.106	.237	.179	.749
SIZE	5.385E-02	.851	3.794E-03	-.152
APPEAR	.215	-1.14E-02	.765	-.174
ROLE	-.127	6.874E-02	.813	.159
SEX1	6.123E-02	-8.30E-02	-.128	.654
FLAG	.935	4.853E-02	.127	2.111E-02
SYMBOL	.923	8.119E-02	.111	-2.62E-02
CHARACTE	.601	-1.24E-02	-.140	.230

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 5 iterations.

b. Only cases for which GROUP = 1 are used in the analysis phase.

FIGURE 4-3
Rotated Component Matrix
Three Factor Solution; Control Group

Component Matrix[a,b]

	Component		
	1	2	3
ENGINE	.370	.631	.183
STYLE	.128	.827	-7.44E-02
SIZE	9.149E-02	-.343	.472
APPEAR	.223	6.537E-02	-.699
SEX1	7.975E-02	.232	.645
FLAG	.925	-.173	-4.84E-02
SYMBOL	.954	-.147	1.971E-02
CHARACTE	.926	-4.50E-02	3.130E-02

Extraction Method: Principal Component Analysis.

a. 3 components extracted.

b. Only cases for which GROUP = 0 are used in the analysis phase.

Rotated Component Matrix
Three Factor Solution; Experiment Group

Rotated Component Matrix[a,b]

	Component		
	1	2	3
ENGINE	3.140E-02	.838	.112
STYLE	.198	.489	.488
SIZE	5.843E-04	.701	-.167
APPEAR	.281	.183	-.614
SEX1	.126	6.261E-02	.645
FLAG	.936	6.768E-02	-8.75E-02
SYMBOL	.917	7.324E-02	-.112
CHARACTE	.603	-9.18E-03	.265

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 5 iterations.

b. Only cases for which GROUP = 1 are used in the analysis phase.

In the model with "role" removed, it can be seen that "size" is still a complex variable in the experiment group. Removing "size" from the solution provides a seemingly invariant factor structure that suits both the control and experimental groups (Figure 4-4). It is important to note that this procedure was successful in screening two of the manipulated items, "size" and "role," and that these were the only two items removed from the analysis. However, one of the manipulated items, "character" remained in the solution as part of the most cohesive factor. Chronbach's Alpha for each factor and group are shown in Table 4-5.

TEST OF HYPOTHESES: CONJOINT

The conjoint experiment was designed to gain insight into the revealed preferences of the respondents and to compare them with the stated preferences indicated in the scale responses. The first comparison to be made was to see whether items that were significant from the scale survey are significant attributes in the conjoint experiment – or to test the sixth hypothesis which is stated as follows:

H_6: The statistically significant items in the scaled survey will be significant attributes in the conjoint experiment within the control group.

A one sample t-test was used to determine the significance of the control group scale items. This test showed that within the control group, all items were significant at the .01 level. Table 4-6 shows the conjoint utility estimations with the associated t-ratios; here it can be seen that of the nine attributes involved, seven significantly impacted respondent choices, while the two attributes that were not significant were "national symbol" (Statue of Liberty) and "national character" (Uncle Sam). Thus, H_6 was partially supported.

When examining differences between the experiment and control groups, one would expect that utility estimations for non-manipulated items would be the same, hence the seventh hypothesis which reads as follows:

H_7: Part-worth estimates for non-manipulated attributes will be the same for control and experimental groups.

As can be seen from Table 4-6, this hypothesis was not supported for the "Product" dimension, as both "Engine" and "Style" (the two non-manipulated attributes) were significant for the control group, but not the experiment group. Conversely, for the "Model" and "National ID" dimensions, the non manipulated items were significant in

TABLE 4-6
Conjoint Utility Estimations and T-Ratios

Attribute	Control Group		Experiment Group	
	Effect	t-Ratio	Effect	t-Ratio
Product				
Engine				
Powerful	-0.129	**-5.68**	0.389	1.59
Economy	0.129	**5.68**	-0.389	-1.59
Style				
Sporty	-0.110	**-5.21**	0.041	1.83
Family	0.110	**5.21**	-0.041	1.83
Size				
Small Trunk (Boot)	-0.577	**-23.33**	-0.325	**-13.03**
Large Trunk (Boot)	0.577	**23.33**	0.325	**13.03**
Model				
Appearance				
Thin	0.433	**17.25**	0.340	**12.97**
Heavy	-0.433	**-17.25**	-0.340	**-12.97**
Social Role				
Attorney (Daveller)	0.296	**12.94**	0.105	**4.42**
Janitor (Tordelman)	-0.296	**-12.94**	-0.105	**-4.42**
Sex				
Male	-0.078	**-3.30**	-0.100	**-3.94**
Female	0.078	**3.30**	0.100	**3.94**
National ID				
Flag				
Clearly Visible	0.487	**19.09**	0.372	**13.97**
Not Shown	-0.487	**-19.09**	-0.372	**-13.97**
Statue of Liberty				
Clearly Visible	0.019	0.82	-0.010	-0.40
Not Shown	-0.019	-0.82	0.010	0.40
National Character (Uncle Sam – John Bull)				
Clearly Visible	-0.003	-0.14	-0.148	**-5.73**
Not Shown	0.003	0.14	0.148	**5.73**

each group, with the exception of the "Symbol" attribute (Statue of Liberty), which was not significant in either group. With respect to the degree of similarity of the estimates, they may be considered "the same" if one can reasonably draw the same conclusions from the estimates in either group. Thus, since the range of the LOGIT estimation is -1 to 1, and expresses the response as a probability of choosing one level of attribute over another, estimations can be considered similar if the likely result is that the same research or managerial decision would spring from them. For example, from the "Model"

dimension, "Sex" as an attribute appears to have similar estimations from both the experiment and control groups in that the estimated utilities were both significant, and were 0.1 and 0.07, respectively. A researcher can reasonably draw the same conclusion from both subsamples, ergo with respect to that item, the seventh hypothesis was upheld. The same may be said for the remaining non-manipulated attributes from the conjoint experiment, which allows for the conclusion that in total, H_7 was partially supported.

With regard to the manipulated items, using the same criteria as H_7, an examination of the manipulated items could be expected to generate different results for the control and experimental groups, as stated in the following hypothesis:

H_8: Part-worth estimates for manipulated attributes will differ between experimental and control groups.

With respect to the "National Character" attribute, the control group estimate was not significant, versus the experimental group which was significant. By default, the hypothesis was upheld for that item. The other two items show a difference in estimates of .25 for "Size," and .19 for "Role." Since these are probabilities, one would certainly expect that differences of .25 and .19 to be actionable – therefore, H_8 was supported.

TEST OF HYPOTHESES: COMPARISON OF CONJOINT ANALYSIS AND STRUCTURAL MODEL

Since the "purification" process of exploratory factor analysis had been successful in screening out two of the manipulated items, the focus of the comparison is now on the most cohesive factor, "National Identity." With a reliability coefficient of .94 for the control group and .79 for the experiment group, this factor appears to yield great insight,

given that one of the most cited terms for the qualitative question, "did not understand" was "John Bull" – the manipulated item from that factor. Confirmatory factor analysis was conducted with the two non-manipulated items from the "Product" factor, the two non-manipulated items from the "Model" factor, and the three items from the "National ID" factor. The model yielded a relatively poor GFI of .785. This is likely due to the fact that non-cohesive factors were included for model identification purposes, but still provide a basis for understanding the nature of the translation error. Table 4-7 shows estimated factor loadings from Amos 4.0 compared to the utility estimates from Choice-Based Conjoint analysis. This table provides insight into the ninth through eleventh hypotheses, which read as follows:

H_9: Factor loadings for scale items will resemble conjoint part-worth estimates for the complementary conjoint attributes in both direction and magnitude within the control group.

H_{10}: Factor loadings for non-manipulated scale items will resemble conjoint part-worth estimates for the complementary conjoint attributes in both direction and magnitude within the experimental group.

H_{11}: Factor loadings for manipulated scale items will not resemble conjoint part-worth estimates for the complementary conjoint attributes in direction and magnitude within the experimental group.

These hypotheses were clearly not supported, as most of the significant factor loadings were insignificant conjoint estimates, and vice versa. Upon reflection, this makes logical sense, as in order for LOGIT estimation to provide for a significant effect, the respondent must be relatively undecided concerning the attribute (please see Appendix 2D, page 90). Conversely, with factor analysis, a factor is considered a strong influencer if there is a strong association – in other words, the respondent must have a relatively strong opinion to create a correlation high enough to be clearly associated with any given factor. Ergo, the hypotheses regarding the regression weights and conjoint utility estimates were mis-

specified. Instead, the reverse can be expected when multi-nomial LOGIT estimation is used.

What is exceptionally interesting about the table shown below is that only one item, "National Character" was significant in both the confirmatory factor analysis and Conjoint analysis in the experiment group, but not the control group. Perhaps this is indicative of general confusion among the experimental group regarding this particular item. As noted earlier, one of the most cited items in the qualitative question "did not understand" was the manipulated term "John Bull."

In addition to the analyses provided above, traditional multi-group analysis was performed to test for the invariance of factor loadings and error variances. Table 4-8 shows the result of the nested model approach. In it, it can be seen that two factor loadings fail the test of invariance, "Engine" and "Character." It is interesting to note that one of these factor loadings was the manipulated item, "Character."

Regarding the test for invariance of error terms, only one item failed the test for invariance, the manipulated item, "Character." Mullen (1995) has correctly noted that this is indicative of an error in equivalence, however it is important to note that misleading conclusions may still be drawn from this test. Given the difference in distributions

TABLE 4-7
Conjoint Utility Estimations and AMOS Factor Loadings

Attribute	Control Group		Experiment Group	
	Utility	Factor Loading	Utility	Factor Loading
Product				
Engine		**1.86**		**2.97**
Powerful	**-0.129**		0.389	
Economy	**0.129**		-0.389	
Style*				
Sporty	**-0.110**		0.041	
Family	**0.110**		-0.041	
Size**				
Small Trunk (Boot)	**-0.577**		**-0.325**	
Large Trunk (Boot)	**0.577**		**0.325**	
Model				
Appearance*				
Thin	**0.433**		**0.340**	
Heavy	**-0.433**		**-0.340**	
Social Role**				
Attorney (Daveller)	**0.296**		**0.105**	
Janitor (Tordelman)	**-0.296**		**-0.105**	
Sex		-0.415		-0.142
Male	**-0.078**		**-0.100**	
Female	**0.078**		**0.100**	
National ID				
Flag*				
Clearly Visible	**0.487**		**0.372**	
Not Shown	**-0.487**		**-0.372**	
Statue of Liberty		**7.025**		**7.526**
Clearly Visible	0.019		-0.010	
Not Shown	-0.019		0.010	
National Character (Uncle Sam – John Bull)		**6.701**		**2.317**
Clearly Visible	-0.003		**-0.148**	
Not Shown	0.003		**0.148**	

** These items were constrained to a factor loading of 1 for model identification purposes*
*** These items were not included in the CFA analysis due to prior screening*
*Items in **BOLD** are significant at the .05 level*

between the control and experimental groups with respect to this item, it is perfectly reasonable to conclude that the group that shows a more normal distribution has understood the item, when in actuality, the reverse is true.

TABLE 4-8
Results of Multi-Group Analysis

Model Description	Comparative Model	χ^2	df	$\Delta\chi^2$	Δdf
Hypothesized Model (Model 1)	--	406.67	32	-	-
Factor loadings, error variances, factor covariances constrained to be equal	Model 1	491.89	44	**85.22**	12
Factor loadings constrained equal	Model 1	482.93	36	**76.26**	4
Factor Loading for item "Engine" constrained equal	Model 1	411.05	33	**4.38**	1
Factor loading for item "Sex" constrained equal	Model 1	410.04	33	3.37	1
Factor loadings for items "Sex" and "Symbol" constrained equal (Model 2)	Model 1	410.66	34	3.99	2
Factor loadings for "Sex," "Symbol," and "Character" constrained equal	Model 1	456.26	35	**49.59**	3
Factor loadings "Sex" and "Symbol," all error variances constrained equal	Model 2	441.91	40	**31.25**	6
Factor loadings "Sex" and "Symbol," error variance for "Style" constrained equal	Model 2	414.20	35	3.54	1
Factor loadings "Sex" and "Symbol," error variances "Style" and "Engine" constrained equal	Model 2	414.27	36	3.59	2
Factor loadings "Sex" and "Symbol," error variances "Style," "Engine," and "Appearance" constrained equal	Model 2	414.34	37	3.68	3
Factor loadings "Sex" and "Symbol," error variances "Style," "Engine," "Appearance," and "Sex" constrained equal	Model 2	415.31	38	4.65	4
Factor loadings "Sex" and "Symbol," error variances "Style," "Engine," "Appearance," "Sex" and "Flag" constrained equal	Model 2	417.96	39	7.3	5

*Items in **BOLD** are significant at the .05 level*

APPENDIX 4A: TEST INSTRUMENT AND CONJOINT TASKS

QUESTIONNAIRE

Thank you for participating in this study. Your input is very valuable. Please answer the following questions to the best of your ability. If there are items you don't understand, please answer the questions to the best of your ability, and don't leave any questions blank!

Are you a U.S. citizen? Yes □ No □

If yes, were you born in the U.S.? Yes □ No □

If no to either of the above, please exit the questionnaire, and thank you for your time.

The following questions assume exposure to an automobile advertisement. When considering these questions, please answer to the best of your ability with respect to how each item contributes to the usefulness of the ad from your perspective. If there are items you don't understand or recognize, please respond to the best of your ability from your own knowledge, and without input from others.

This section is designed to measure what product features best contribute to your assessment of the usefulness of the information in an ad. Please indicate in the box that best indicates how much one or the other attributes (for each item) means to you in terms of making the ad useful. Please be sure to mark every scale and do not omit any.

Engine

Powerful	□	□	□	□	□	□	□	Economy

Style

Sporty	□	□	□	□	□	□	□	Family

Room

Large Trunk	□	□	□	□	□	□	□	Small Trunk

This section is designed to measure how a model in an ad may contribute to your assessment of the usefulness of an ad. Please indicate the box that best represents how much one or the other attributes (for each item) means to you in terms of making the ad useful. Please be sure to mark every scale and do not omit any.

Appearance

Thin ☐ ☐ ☐ ☐ ☐ ☐ ☐ Heavy

Social Role

Attorney ☐ ☐ ☐ ☐ ☐ ☐ ☐ Janitor

Sex

Male ☐ ☐ ☐ ☐ ☐ ☐ ☐ Female

This section is designed to measure how nationality may contribute to your assessment of the usefulness of an ad. Please indicate the box that best represents how much the attribute to the right or the left of each item means to you in terms of making the ad useful. Please be sure to mark every scale and do not omit any.

Flag

Clearly Visible ☐ ☐ ☐ ☐ ☐ ☐ ☐ Not Shown Or Mentioned

Symbol (e.g., Statue of Liberty)

Clearly Visible ☐ ☐ ☐ ☐ ☐ ☐ ☐ Not Shown Or Mentioned

Uncle Sam

Clearly Visible ☐ ☐ ☐ ☐ ☐ ☐ ☐ Not Shown Or Mentioned

Finally, please provide the following demographic information:

Do you consider yourself "culturally" American?

Yes ☐ No ☐

If "no" to the above, which culture do you most closely identify with?

_____

How long have you lived in the U.S.?

_____ Years

Have you ever lived outside the U.S.?

Yes □ No □

If so, for how long and in which country?

_____ Years _____ Country

What is your age?

_____ Years

What is your sex?

Male □ Female □

What is the highest level of education you have achieved?

□ High School (did not graduate)

□ High School Graduate

□ Some College (did not graduate)

□ College Graduate (2yr)

□ College Graduate (4yr)

□ Some Graduate School (did not graduate)

□ Graduate Degree

□ Professional Degree

All items in this survey were clearly understood:

Yes □ No □

If no to the above, which items were not clearly understood?

_____

On a scale of 1 – 7 (one being very low, 7 being very high) please rate how well you understood the items in the survey:

Didn't Understand Everything	□	□	□	□	□	□	Understood Everything

Please indicate how often you view television shows from the U.K.:

□ Never

□ Once per year/almost never

□ Once per month

□ Twice per month

□ Once per week

□ Daily

Please indicate how often you listen to radio broadcasts from the U.K.:

□ Never

□ Once per year/almost never

□ Once per month

□ Twice per month

□ Once per week

□ Daily

How often have you traveled outside the U.S. (including neighboring countries):

Thank you for participating in this study! Your response is greatly appreciated.

CONJOINT CHOICE TASKS

PRODUCT:

	Ad A	**Ad B**
Engine	Powerful	Powerful
Style	Sporty	Family
Room	Small Trunk	Small Trunk
	☐	☐

	Ad A	**Ad B**
Engine	Powerful	Powerful
Style	Sporty	Sporty
Room	Small Trunk	Large Trunk
	☐	☐

	Ad A	**Ad B**
Engine	Powerful	Powerful
Style	Sporty	Family
Room	Small Trunk	Large Trunk
	☐	☐

	Ad A	**Ad B**
Engine	Powerful	Economy
Style	Sporty	Sporty
Room	Small Trunk	Small Trunk
	☐	☐

	Ad A	**Ad B**
Engine	Powerful	Economy
Style	Sporty	Family
Room	Small Trunk	Small Trunk
	☐	☐

	Ad A	**Ad B**
Engine	Powerful	Economy
Style	Sporty	Sporty
Room	Small Trunk	Large Trunk
	☐	☐

	Ad A	**Ad B**
Engine	Powerful	Economy
Style	Sporty	Family
Room	Small Trunk	Large Trunk
	☐	☐

	Ad A	Ad B
Engine	Powerful	Powerful
Style	Family	Sporty
Room	Small Trunk	Large Trunk
	☐	☐

	Ad A	Ad B
Engine	Powerful	Powerful
Style	Family	Family
Room	Small Trunk	Large Trunk
	☐	☐

	Ad A	Ad B
Engine	Powerful	Economy
Style	Family	Family
Room	Small Trunk	Large Trunk
	☐	☐

	Ad A	Ad B
Engine	Powerful	Economy
Style	Family	Family
Room	Small Trunk	Small Trunk
	☐	☐

	Ad A	Ad B
Engine	Powerful	Economy
Style	Family	Sporty
Room	Small Trunk	Large Trunk
	☐	☐

	Ad A	Ad B
Engine	Powerful	Economy
Style	Family	Family
Room	Small Trunk	Large Trunk
	☐	☐

	Ad A	Ad B
Engine	Powerful	Powerful
Style	Sporty	Family
Room	Large Trunk	Large Trunk
	☐	☐

	Ad A	Ad B
Engine	Powerful	Economy
Style	Sporty	Sporty
Room	Large Trunk	Small Trunk
	☐	☐

	Ad A	Ad B
Engine	Powerful	Economy
Style	Sporty	Family
Room	Large Trunk	Small Trunk
	☐	☐

	Ad A	Ad B
Engine	Powerful	Economy
Style	Sporty	Sporty
Room	Large Trunk	Large Trunk
	☐	☐

	Ad A	Ad B
Engine	Powerful	Economy
Style	Sporty	Family
Room	Large Trunk	Large Trunk
	☐	☐

	Ad A	Ad B
Engine	Powerful	Economy
Style	Family	Sporty
Room	Large Trunk	Small Trunk
	☐	☐

	Ad A	Ad B
Engine	Powerful	Economy
Style	Family	Family
Room	Large Trunk	Small Trunk
	☐	☐

	Ad A	Ad B
Engine	Powerful	Economy
Style	Family	Sporty
Room	Large Trunk	Large Trunk
	☐	☐

	Ad A	Ad B
Engine	Powerful	Economy
Style	Family	Family
Room	Large Trunk	Large Trunk
	☐	☐

	Ad A	Ad B
Engine	Economy	Economy
Style	Sporty	Family
Room	Small Trunk	Small Trunk
	☐	☐

	Ad A	Ad B
Engine	Economy	Economy
Style	Sporty	Sporty
Room	Small Trunk	Large Trunk
	☐	☐

	Ad A	Ad B
Engine	Economy	Economy
Style	Sporty	Family
Room	Small Trunk	Large Trunk
	☐	☐

	Ad A	Ad B
Engine	Economy	Economy
Style	Family	Sporty
Room	Small Trunk	Large Trunk
	☐	☐

	Ad A	Ad B
Engine	Economy	Economy
Style	Family	Family
Room	Small Trunk	Large Trunk
	☐	☐

	Ad A	Ad B
Engine	Economy	Economy
Style	Sporty	Family
Room	Large Trunk	Large Trunk
	☐	☐

MODEL:

	Ad A	Ad B
Appearance	Thin	Thin
Social Role	Attorney	Janitor
Sex	Male	Male
	☐	☐

	Ad A	Ad B
Appearance	Thin	Thin
Social Role	Attorney	Attorney
Sex	Male	Female
	☐	☐

	Ad A	Ad B
Appearance	Thin	Thin
Social Role	Attorney	Janitor
Sex	Male	Female
	☐	☐

	Ad A	Ad B
Appearance	Thin	Heavy
Social Role	Attorney	Attorney
Sex	Male	Male
	☐	☐

	Ad A	Ad B
Appearance	Thin	Heavy
Social Role	Attorney	Janitor
Sex	Male	Male
	☐	☐

	Ad A	Ad B
Appearance	Thin	Heavy
Social Role	Attorney	Attorney
Sex	Male	Female
	☐	☐

	Ad A	Ad B
Appearance	Thin	Heavy
Social Role	Attorney	Janitor
Sex	Male	Female
	☐	☐

	Ad A	Ad B
Appearance	Thin	Thin
Social Role	Janitor	Attorney
Sex	Male	Female
	☐	☐

	Ad A	Ad B
Appearance	Thin	Thin
Social Role	Janitor	Janitor
Sex	Male	Female
	☐	☐

	Ad A	Ad B
Appearance	Thin	Heavy
Social Role	Janitor	Attorney
Sex	Male	Male
	☐	☐

	Ad A	Ad B
Appearance	Thin	Heavy
Social Role	Janitor	Janitor
Sex	Male	Male
	☐	☐

	Ad A	Ad B
Appearance	Thin	Heavy
Social Role	Janitor	Attorney
Sex	Male	Female
	☐	☐

	Ad A	Ad B
Appearance	Thin	Heavy
Social Role	Janitor	Janitor
Sex	Male	Female
	☐	☐

	Ad A	Ad B
Appearance	Thin	Thin
Social Role	Attorney	Janitor
Sex	Female	Female
	☐	☐

	Ad A	Ad B
Appearance	Thin	Heavy
Social Role	Attorney	Attorney
Sex	Female	Male
	☐	☐

	Ad A	Ad B
Appearance	Thin	Heavy
Social Role	Attorney	Janitor
Sex	Female	Male
	☐	☐

	Ad A	Ad B
Appearance	Thin	Heavy
Social Role	Attorney	Attorney
Sex	Female	Female
	☐	☐

	Ad A	Ad B
Appearance	Thin	Heavy
Social Role	Attorney	Janitor
Sex	Female	Female
	☐	☐

	Ad A	Ad B
Appearance	Thin	Heavy
Social Role	Janitor	Attorney
Sex	Female	Male
	☐	☐

	Ad A	Ad B
Appearance	Thin	Heavy
Social Role	Janitor	Janitor
Sex	Female	Male
	☐	☐

	Ad A	Ad B
Appearance	Thin	Heavy
Social Role	Janitor	Attorney
Sex	Female	Female
	☐	☐

	Ad A	Ad B
Appearance	Thin	Heavy
Social Role	Janitor	Janitor
Sex	Female	Female
	☐	☐

	Ad A	Ad B
Appearance	Heavy	Heavy
Social Role	Attorney	Janitor
Sex	Male	Male
	☐	☐

	Ad A	Ad B
Appearance	Heavy	Heavy
Social Role	Attorney	Attorney
Sex	Male	Female
	☐	☐

	Ad A	Ad B
Appearance	Heavy	Heavy
Social Role	Attorney	Janitor
Sex	Male	Female
	☐	☐

	Ad A	Ad B
Appearance	Heavy	Heavy
Social Role	Janitor	Attorney
Sex	Male	Female
	☐	☐

	Ad A	Ad B
Appearance	Heavy	Heavy
Social Role	Janitor	Janitor
Sex	Male	Female
	☐	☐

	Ad A	Ad B
Appearance	Heavy	Heavy
Social Role	Attorney	Janitor
Sex	Female	Female
	☐	☐

COUNTRY OF ORIGIN:

	Ad A	Ad B
Flag	Visible	Visible
Statue of Liberty	Visible	Not Shown
Uncle Sam	Visible	Visible
	☐	☐

	Ad A	Ad B
Flag	Visible	Visible
Statue of Liberty	Visible	Visible
Uncle Sam	Visible	Not Shown
	☐	☐

	Ad A	Ad B
Flag	Visible	Not Shown
Statue of Liberty	Visible	Not Shown
Uncle Sam	Visible	Visible
	☐	☐

	Ad A	Ad B
Flag	Visible	Not Shown
Statue of Liberty	Visible	Visible
Uncle Sam	Visible	Visible
	☐	☐

	Ad A	Ad B
Flag	Visible	Not Shown
Statue of Liberty	Visible	Not Shown
Uncle Sam	Visible	Visible
	☐	☐

	Ad A	Ad B
Flag	Visible	Not Shown
Statue of Liberty	Visible	Visible
Uncle Sam	Visible	Not Shown
	☐	☐

	Ad A	Ad B
Flag	Visible	Not Shown
Statue of Liberty	Visible	Not Shown
Uncle Sam	Visible	Not Shown
	☐	☐

	Ad A	Ad B
Flag	Visible	Visible
Statue of Liberty	Not Shown	Visible
Uncle Sam	Visible	Not Shown
	☐	☐

	Ad A	Ad B
Flag	Visible	Visible
Statue of Liberty	Not Shown	Not Shown
Uncle Sam	Visible	Not Shown
	☐	☐

	Ad A	Ad B
Flag	Visible	Not Shown
Statue of Liberty	Not Shown	Visible
Uncle Sam	Visible	Visible
	☐	☐

	Ad A	Ad B
Flag	Visible	Not Shown
Statue of Liberty	Not Shown	Not Shown
Uncle Sam	Visible	Visible
	☐	☐

	Ad A	Ad B
Flag	Visible	Not Shown
Statue of Liberty	Not Shown	Visible
Uncle Sam	Visible	Not Shown
	☐	☐

	Ad A	Ad B
Flag	Visible	Not Shown
Statue of Liberty	Not Shown	Not Shown
Uncle Sam	Visible	Not Shown
	☐	☐

	Ad A	Ad B
Flag	Visible	Visible
Statue of Liberty	Visible	Not Shown
Uncle Sam	Not Shown	Not Shown
	☐	☐

	Ad A	Ad B
Flag	Visible	Not Shown
Statue of Liberty	Visible	Visible
Uncle Sam	Not Shown	Visible
	☐	☐

	Ad A	Ad B
Flag	Visible	Not Shown
Statue of Liberty	Visible	Not Shown
Uncle Sam	Not Shown	Visible
	☐	☐

	Ad A	Ad B
Flag	Visible	Not Shown
Statue of Liberty	Visible	Visible
Uncle Sam	Not Shown	Not Shown
	☐	☐

	Ad A	Ad B
Flag	Visible	Not Shown
Statue of Liberty	Visible	Not Shown
Uncle Sam	Not Shown	Not Shown
	☐	☐

	Ad A	Ad B
Flag	Visible	Not Shown
Statue of Liberty	Not Shown	Visible
Uncle Sam	Not Shown	Visible
	☐	☐

	Ad A	Ad B
Flag	Visible	Not Shown
Statue of Liberty	Not Shown	Not Shown
Uncle Sam	Not Shown	Visible
	☐	☐

	Ad A	Ad B
Flag	Visible	Not Shown
Statue of Liberty	Not Shown	Visible
Uncle Sam	Not Shown	Not Shown
	☐	☐

	Ad A	Ad B
Flag	Visible	Not Shown
Statue of Liberty	Not Shown	Not Shown
Uncle Sam	Not Shown	Not Shown
	☐	☐

	Ad A	Ad B
Flag	Not Shown	Not Shown
Statue of Liberty	Visible	Not Shown
Uncle Sam	Visible	Visible
	☐	☐

	Ad A	Ad B
Flag	Not Shown	Not Shown
Statue of Liberty	Visible	Visible
Uncle Sam	Visible	Not Shown
	☐	☐

	Ad A	Ad B
Flag	Not Shown	Not Shown
Statue of Liberty	Visible	Not Shown
Uncle Sam	Visible	Not Shown
	☐	☐

	Ad A	Ad B
Flag	Not Shown	Not Shown
Statue of Liberty	Not Shown	Visible
Uncle Sam	Visible	Not Shown
	☐	☐

	Ad A	Ad B
Flag	Not Shown	Not Shown
Statue of Liberty	Not Shown	Not Shown
Uncle Sam	Visible	Not Shown
	☐	☐

	Ad A	Ad B
Flag	Not Shown	Not Shown
Statue of Liberty	Visible	Not Shown
Uncle Sam	Not Shown	Not Shown
	☐	☐

Chapter 5

CONCLUSIONS, IMPLICATIONS, AND SUGGESTIONS FOR FUTURE RESEARCH

INTRODUCTION

This research was conducted to understand with greater clarity the effect of mistranslated scale items in multi-cultural comparative research. The conceptual approach involved the reversal of processes currently utilized in multi-cultural research projects. As it is, multi-cultural researchers select scales and attempt to adapt them to culturally divergent environments, and then attempt to understand what real differences are by deleting suspect items. This research instead moves from a known translation error, and then tracks the various effects of these errors on both scales and interpretations. A questionnaire was developed with three purposefully manipulated translation errors, and administered to a probability sample of American consumers via a web-based panel. Resulting data were analyzed for distributional differences using descriptive statistics, and then analyzed with correlational analysis. The same items were presented to the same respondents as choice-based conjoint choice tasks, and differences were noted between the scaled responses and conjoint utility estimations. The remainder of this chapter discusses the findings of these tests, the contribution of this research, and offers implications for both global brand managers and academic research, as well as limitations of the study.

CONCLUSIONS

In sum, the findings from this experiment document a clear and pronounced relationship between translation equivalence error and metric equivalence failure. Although this relationship was hypothesized, an unexpected finding was that the metric equivalence failure of a few items significantly impacts relationships among items where translation equivalence (and all other equivalences) had arguably been achieved. Interpretations of data corrupted by equivalence error will lead researchers to inaccurate conclusions about the real views of divergent groups – or pseudo-etic inferences. The seriousness of the implications of such a situation would be difficult to overstate, but implications are discussed in detail in the next section.

The problem of attenuation was revealed through the process of mimicking multi-cultural comparative research; namely, suspect items were deleted from analysis as traditionally accepted evidence of equivalence failure was presented. Exploratory factor analysis effectively screened out two of the manipulated translation errors – yet the remaining items in the respective factors showed unmistakable differences between the experimental and control groups with respect to factor loadings.

Confirmatory factor analysis was used to test the invariance of the factors. Using the nested model approach, factor loadings, factor covariances, and error variances were tested for invariance between control and experiment groups, with unexpected results. Just as the Nixon tapes provided clear evidence of obstruction of justice (the famous "smoking gun" tape), this experiment provided clear evidence that techniques currently employed to detect equivalence errors are effective. Consistent with the arguments of multi-cultural researchers, such as Mullen (1995), the "smoking gun" that revealed the

last of the manipulated items was the high significance of the error variances for that manipulated item. Mullen (1995) correctly asserted that this examination of the theta matrix in confirmatory factor analysis would reveal items that are not equivalent, however, this "smoking gun" was only mildly smoking, not belching the proverbial flame and cordite inferred by conventional multi-cultural research. Few researchers actually recommend taking a close look at distributions of data. In the case of the problem item (national character), an examination of data distributions would reveal that the responses of the experimental group for that item were normally distributed. The control group, however, had responses that had a pronounced negative skew, while at the same time appearing to be nearly bi-modal. Most researchers would be tempted to draw the conclusion that the *experimental* group understood the question and that the *control* group did not. Even more thought provoking is the fact that if it is assumed that the control group was the originating culture in which the scale was developed, the normal distribution of the responses from the experimental group may provide an impetus for researchers to ignore the evidence provided by the error variances. Thus, even following currently-accepted procedure, there is justification for ignoring the significant differences in error variance (which is a step considered unreasonably restrictive by some – e.g., Byrne, 1999), and accepting all other differences as reflecting culturally-influenced opinion.

The conjoint experiment was just as enlightening, but in a different way. While the LOGIT utility estimations seemed to be more stable, it would be disingenuous to state that the revealed preferences of choice-based conjoint clearly show which items are inherently inequivalent, or that any direct comparison between the utility estimations and

factor loadings are particularly enlightening. However, the study's results must be tempered by the fact that the test was conducted with only one "cohesive" factor. More cohesive factors *without* manipulated items may have revealed higher levels of consistency between the results. Given this, it is entirely possible that conjoint estimations from cohesive factors *may* be able to point to problems in the achievement of equivalence. It is interesting to note that the conjoint utility estimations were quite similar between experimental and control groups within the "national ID" construct, with the exception of the manipulated translation error. In addition, the largest difference between utility estimates for control and experimental groups within the second factor was the manipulated item "social role." What this suggests is an indication that with consistent factors, conjoint analysis allows a more stable comparison between groups when trying to establish equivalence. Unfortunately, the results from the first factor add inconsistency into the concept that revealed preferences are consistently able to discriminate between equivalent and non-equivalent items. Future research that employs established scales may provide more insight into such a possibility.

IMPLICATIONS

Implications for Multi-Cultural Researchers

The first implication of the findings of this study is that multi-cultural comparative research appears to be even more fraught with pitfalls than previously thought. In the past, researchers were advised to achieve scale equivalence as much as possible, and then use the scale. If researchers find items that fail equivalence, either due to translation or some other error, those items *only* are removed from analysis, and all

subsequent results are assumed valid. It is now known that removing only problematic items does not guarantee that the researcher had achieved accurate measures. Instead, the attenuating effect of items that are not equivalent is such that at least a factor or entire section of a questionnaire is almost certain to be corrupted. Current multi-cultural research techniques are not sufficient to handle this.

It appears that the answer to this problem is an iterative process of conducting scale purification and research. More directly, if there is a detected equivalence failure, there is a need to separate scale purification from data collection for hypothesis testing. While this increases the time and expense of conducting cross-cultural research, such a process appears to be necessary to guard against incorrect generalizations – which are now known to be highly likely. This process would require re-sampling and administration of questionnaires that are absent the problem items – in other words, a multiple iteration of a complete research project.

The second implication is that there is no fool-proof way of detecting equivalence failure. The process followed in this experiment involved non-parametric tests, exploratory factor analysis, confirmatory factor analysis, multi-group analysis, and conjoint analysis. The only technique used that was able to distinguish between manipulated items and non-manipulated items were the non-parametric tests. However, in multi-cultural research environments, it may not be altogether practical to rely upon these types of tests to check for equivalence failure; entire scales may have systematic response bias within a cultural sub-sample that creates significant distributional differences, which is not helpful. When these techniques are combined, it has been demonstrated in this research that at least the items that fail equivalence may be

identified. Thus, it would seem that multiple-techniques should be employed any time equivalence comes into question.

The third implication is that revealed preferences provide information that is useful for multi-cultural researchers in terms of gaining insight into confusion caused by inequivalent items. It may be that choice-based conjoint analysis insulates the respondent from response bias. Noting the relative stability of non-manipulated attributes in the conjoint experiment relative to regression weights from confirmatory factor analysis, one can see that these revealed preferences deliver insight closer to the true preferences than the scale (which was corrupted by the inequivalent items). Of course, this experiment has provided insight based only upon choice-based conjoint, and other forms of utility estimation were not investigated here. Nevertheless, there is evidence that conjoint analysis may be employed where inequivalence is suspected. For example, a researcher using an established scale may have reason to suspect an equivalence error – in this case, the scale items in question may be developed into a conjoint experiment to check the stability of part-worth estimates versus the stability of the factor structure for further direction, and possibly identification of the problem.

The fourth implication is that there is a distinct need for theoretical development regarding response behavior in conditions of inequivalence. The fact that relationships have been demonstrated between equivalence failure, response style, and metric equivalence is to be expected. The alteration of items that had not contained equivalence error, however, merits considerable thought. There had been no theoretical reason to believe that this would happen, but plainly it has. Therefore, work must be done to explain how people's perception changes when scale items are not understood.

Implications for Multi-National Marketing Managers

Implications for multi-national marketing managers are similar to those for multi-cultural researchers; except that the stakes are much higher for the marketing managers, because costly mistakes can be made if corrupt information is used. An excellent example of the types of decisions that hinge on multi-cultural comparative information would include any decision to standardize products or promotion. For instance, a confectioner wishing to decide whether the same product should be sold in overseas markets as is sold in the domestic market. This type of decision must be based upon information of consumer taste that is valid in its estimation of similarities and differences between divergent cultures in consumer preferences.

There may be a need for an iterative process of doing multi-cultural marketing research. Given that it is now evident that research should be conducted iteratively, the global marketing manager must be aware that the risk of making marketing decisions when time does not allow for an iterative research approach is likely to be higher than previously thought. If market conditions warrant speedy action (i.e., a unique perceived window of opportunity), and it is found that data from market research is corrupted by inequivalence, perhaps a stronger reliance upon advice from objective indigenous advisors may lead to better strategic action.

Conversely, when the potential consequences (such as market position, sales levels, or product positioning) of marketing decisions are high, and time is not a crucial issue, it would be advantageous for the global brand manager to ensure that data are not

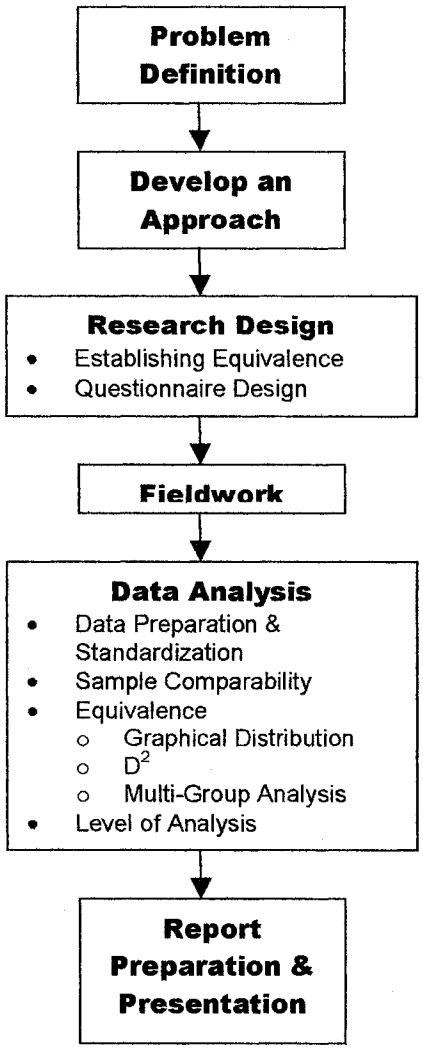
corrupted by inequivalence, even if that includes seemingly redundant sampling and data collection.

Proposed Checklist for Multi-Cultural Comparative Research

It is proposed that multi-cultural comparative research may be improved by revising generally accepted procedure for conducting comparative research. Malhotra, Agarwal & Peterson (1996) recommended a checklist for conducting such studies, as given in Figure 5-1. Of particular interest is the need to propose a revised view of data analysis procedures, and what action to take if equivalences fail. According to Malthotra, Agarwal & Peterson (1996), the "data analysis" procedure includes preparing data for analysis, ensuring sample comparability, checking equivalence, deciding on the level of analysis, and guarding against methodological fallacies. This experiment contributes to our understanding of the effectiveness of techniques designed to check for equivalence, and more importantly, the effect of equivalence error.

To check for equivalence, Malhotra, Agarwal & Peterson (1996) recommend examination of graphic distribution plots, as well as computions of Mahalanobis D^2 as a means of detecting outliers. These authors argue that outliers (responses that are far removed from the bulk of responses) distort the data, and create a situation where comparisons cannot be made reliably. Furthermore, it is accepted that once problem variables have been identified, *they be dropped from analysis and the process continue.* However, this experiment brings to the fore three critical issues; first, that outliers are not necessarily indicative of equivalence failure.

Figure 5-1
Checklist for Mutli-Cultural Comparative Research



Adapted from Malhotra, Agarwal & Peterson, 1996

It can be argued that in some cases, outliers may be a good indicator of an item that fails equivalence, but when using semantic differential scales, mid-point response bias has been demonstrated to be present when respondents do not understand the item.

Therefore, responses will be bunched in the center or mid-point of the scale, and not toward the poles, creating an absence of outliers. Second, because of the first point, calculating D^2 will be of limited use in identifying equivalence failures, depending upon the scale being used. Third, dropping an item from analysis because it fails equivalence may not lead to an accurate assessment of the views of at least one of the groups being examined. These taken together present a need for a revised data analysis checklist, in which a couple of steps are added to accepted procedure, and is embodied in four points: 1) ensure that data distributions are similar, 2) conduct exploratory factor analysis, 3) conduct confirmatory factor analysis or multi-group analysis, and 4) revise the test instrument if necessary and begin the fieldwork anew.

Checking Data Distributions

This experiment utilized a simple, yet effective technique to test invariance of data distributions; the Mann-Whitney U test. If an item exhibits significantly different distribution of its data, researchers should then examine whether a response bias exists that is systematic within a sample, but unique to a particular item. If this is the case, there is indication of an equivalence failure, which should be re-checked through qualitative means to see whether it is in fact an equivalence problem.

Exploratory Factor Analysis

Consistent with current practice, exploratory factor analysis should be conducted in order to identify items that load differently between samples. If items are loading differently, there may be real differences in perception between groups of people, or there may be an equivalence problem. Obviously, if data distributions are significantly different, one would expect differing factor solutions. However, these differences may be

an artifact of some misunderstood item or dimension – in which case further/remedial qualitative consultation is highly recommended.

Confirmatory Factor Analysis

This experiment has shown that with multi-group analysis, the test of interest in establishing equivalence is the test for invariance of error matrices. In this experiment, the test for invariance of error terms was the only multi-group analysis test that clearly revealed which item was inequivalent. Some authors argue that examination of the matrix of error terms is unduly stringent (e.g., Byrne, 2001); however, as advanced by Mullen (1995), this technique has been demonstrated as effective in diagnosing equivalence errors.

Revision of Test Instrument and Re-sampling

What is now known is that if there is equivalence error, and subsequent failure, the test instrument and data generated by it is not likely to be useful for making cross-cultural comparisons. In order to make valid comparisons, it is highly recommended that revision of the test instruments be conducted to eliminate inequivalent items. Afterward, it is highly advisable to re-sample the population affected, and conduct the fieldwork to gain a fresh data set that is not sullied by responses to inequivalent items.

LIMITATIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH

The results of this study are enlightening, but need to be tempered by the fact that an arbitrary scale was used in the study. However, for this study, the "contrived" nature of the scale helps to understand scale development in multi-cultural contexts, and the problems caused by translation equivalence error. Further research should be conducted

to understand the effects of translation error on established scales, such as CETSCALE or MARKOR.

The fact that the response rate was low for a web-based panel can be explained by the fact that the incentive for respondents was entry into a sweepstakes versus some other type of incentive, such as cash or tangible gifts. In fact, as noted in Chapter 4, it can be seen that respondents who abandoned the survey did so either at the beginning of, or just before the conjoint choice tasks. The method of comparing early and late responses to check for non-response bias was used due to the absence of another credible alternative, and while the procedure has detractors, seems to be fitting in this case (Armstrong & Overton, 1977).

Some may say that responses from those completing the survey may have been "forced." This is interesting; however in this particular study, it can be seen that each completed response was done so free volition – terminated questionnaires were not used in analysis. While it may be true that some respondents "guessed" answers to manipulated translation items, this is the phenomenon of interest, and analysis of these responses provides a greater understanding of the effects of entire samples of people doing exactly that.

It should also be noted that no actual translation from divergent languages was executed. While this is true, this question must be addressed in terms of control. The author is not bi-lingual, nor bi-cultural. Therefore, any attempt to select a foreign language for translation purposes unleashes a great deal of control from the experiment. It was the goal of this study to emulate a back-translation error (most of the manipulated items are British-English versus American-English). In other words, this type of

translation error can occur despite widely accepted means of preventing it, and can go undetected through the data collection process. Simulating this process requires an ability to manipulate the language in question, which in this case rules out the use of foreign interpretation. Yet, this is a potentially important observation, because the results of this experiment can only be applied to English speaking Americans. Future research should examine whether the same results occur in divergent cultures, such as those found in Asia or Latin America?

Results from this experiment can be directly applied to semantic differential scales, but extending these results to broader forms of research questions/response formats should be done with care. In addition, the scale used for the test was not a generally accepted scale in marketing research, and interpreting results from this experiment to more widely used and validated marketing scales will require further research. Thus, some future research questions include the following:

- How does translation equivalence error impact Likert-type scales?
- Is metric conjoint a better estimator of choice behavior in conditions of inequivalence?
- What may be the impact of equivalence failure in previously established and validated scales, such as MARKOR or SERVQUAL?
- Does the attenuating effect of equivalence failure hold in other cultures?
- What are the effects of other types of equivalence failure, such as conceptual or metric equivalence?
- Do response formats impact the effect of equivalence failure?

In conclusion, this experiment was a necessary step to develop a clearer understanding of multi-cultural comparative research. As globalization continues, the volume of cross-cultural research is likely to grow exponentially, which raises to the fore the importance of doing all that is possible to guarantee the quality of this research (Malhotra et al., 1996).

Some assumptions of multi-cultural research have been placed into question. First is the assumption that when a scale item exhibits equivalence error, it can simply be removed from the analysis, and conclusions drawn. This is clearly a risky assumption, as it has been demonstrated that one corrupt item will attenuate other items.

Second is the assumption that current techniques utilized to detect equivalence error provide needed diagnostic and repair tools. This study showed that currently employed techniques to detect equivalence error appear to work in the general sense, yet interpretation of results of tests for equivalence may still produce inaccurate conclusions. In a word, the process is not perfect, and a minor mistranslation still provides a basis for pseudo-etic inferences.

BIBLIOGRAPHY

Adler, Nancy, 1983. "A Typology of Management Studies Involving Culture," *Journal of International Business Studies*, 14(Fall):29-47.

Akaah, Ishmael P. and Attila Yaprak 1988. "Identifying Target Segments for Foreign Direct Investment (FDI) Attraction: An Application of Conjoint Methodology," *International Marketing Review*, 5(3):28-38.

Albers-Miller, Nancy D., 1996. "Designing Cross-Cultural Advertising Research: A Closer Look at Paired Comparisons," *International Marketing Review*, 13(5):59-77.

Algers, Staffan and Muriel Beser, 2001. "Modelling Choice of Flight and Booking Class – A study Using Stated Preference and Revealed Preference Data," *International Journal of Services Technology and Management*, 2(1,2):28-45.

Alpert, Frank, Michael Kamins, Tomoaki Sakano, Naoto Onzo, and John Graham, 2001. "Retail Buyer Beliefs, Attitude and Behavior Toward Pioneer and Me-Too Follower Brands: A Comparative Study of Japan and the USA," *International Marketing Review*, 18(2):160.

Andrews, J. Craig, Srinivas Durvasula, and Richard G. Netemeyer, 1994. "Testing the Cross-National Applicability of U.S. and Russian Advertising Belief and Attitude Measures," *Journal of Advertising*, 23(1):71-83.

----, Steven Lysonski, and Srinivas Durvasula, 1991. "Understanding Cross-Cultural Student Perceptions of Advertising in General: Implications for Advertising Educators and Practitioners," *Journal of Advertising*, 20(2):15-29.

Antoncic, Bostjan & Robert D. Hisrich, 2001. "Intrapreneurship: Construct Refinement and Cross-Cultural Validation," *Journal of Business Venturing*, 16(5):495.

Arias, Jose Tomas Gomez, 1996. "Conjoint-Based Preferential Segmentation in the Design of a New Financial Service," *The International Journal of Bank Marketing*, 14(3):30-32.

Armstrong, J. Scott and Terry S. Overton, 1977. "Estimating Nonresponse Bias in Mail Surveys," *Journal of Marketing Research*, 14(3):396.

Arts, Wil and Loek Halman, 1999. "New Directions in Quantitative Comparative Sociology," *International Journal of Comparative Sociology*, 40(1):1

Aulakh, Preet S. and Masaaki Kotabe, 1992. "An Assessment of Theoretical and Methodological Developments in International Marketing: 1980-1990," *Journal of International Marketing*, 1(2):5-29.

Awasthi, Vidya N., Chee W. Chow & Anne Wu. "Cross-Cultural Comparison – Sensitizing Accounting Students to Cultural Diversity: Report on a Teaching Experiment," *Issues in Accounting Education*, 12(1):77-81.

Baalbaki, Imad B. and Naresh K. Malhotra, 1995. "Standardization Versus Customization in International Marketing: An Investigation Using Bridging Conjoint Analysis," *Journal of the Academy of Marketing Science*, 23(3):182-195.

Bailey, William, Gutierrez de Pineres, and Sheila Amin, 1997. "Country of Origin Attitudes in Mexico: The Malinchismo Effect," *Journal of International Consumer Marketing*, 9(3):25-42.

Barr, Michael and Nambury S. Raju, 2003. "IRT-Based Assessments of Rater Effects in Multiple-Source Feedback Instruments," *Organizational Research Methods*, 6(1):15-44.

Baumgartner, Hans and Jan-Benedict E. M. Steenkamp, 2001. "Response Styles in Marketing Research: A Cross-National Investigation," *Journal of Marketing Research*, 38(May):143-156.

Bellizzi, Joseph A. and Ronald W. Hasty, 2003. "Supervising Unethical Sales Force Behavior: How Strong is the Tendency to Treat Top Sales Performers Leniently?" *Journal of Business Ethics*, 43(4):337.

Bensaou, M., 1999. "Testing Metric Equivalence in Cross-National Strategy Research: An Empirical Test Across the United States and Japan," *Strategic Management Journal*, 20(7):671

Bhalla, Gaurav, and Lynn Ys Lin, 1987. "Cross-Cultural Marketing Research: A Discussion of Equivalence Issues and Measurement Strategies," *Psychology & Marketing*, 4(4):275-286.

Birtwhistle, Grete, Ian Clarke and Paul Freathy, 1998. "Customer Decision Making in Fashion Retailing: A Segmentation Analysis," *International Journal of Retail & Distribution Management*, 26(4):147.

Boddewyn, Jean, 1981. "Comparative Marketing: The First Twenty Five Years," *Journal of International Business Studies*, 12(1):61-80.

Bond, John B., Jr., 1996. "Canadian Family Policies: Cross-National Comparisons," *Canadian Public Policy*, 22(3):306.

Bridges, Eileen, Renee A. Florsheim and Claudette John, 1996. "A Cross-Cultural Comparison of Response to Service Promotion," *The Services Industries Journal*, 16(3):265-287.

Bryan, Norman B., Ephraim R. McLean, Stanley J. Smits, and Janice M. Burn, 1995. "Work Perceptions Among Hong Kong and United States I/S Workers: A Cross-Cultural Comparison," *Journal of End User Computing*, 7(4):22-30.

Carlson, Patricia, Beverly K. Kahn and Frantz Rowe, 1999. "Organizational Impacts of New Communication Technology: A Comparison of Cellular Phone Adoption in France and the United States," *Journal of Global Information Management*, 7(3):19.

Caruana, Albert, B. Ramaseshan, Michael T. Ewing and Fariba Rouhani, 1998. "Expectations About Management Consultancy Services: Testing the Assumption of Equivalence Across Australian and Singaporean Firms," *Services Marketing Quarterly*, 18(1):1-11.

Carlson, Les, Stephen J. Grove, Norman Kangun, and Michael Polonsky, 1996. "An International Comparison of Environmental Advertising: Substantive Versus Associative Claims," *Journal of Macromarketing*, 16(2):57-69.

Caruana, Albert, 1999. "An Assessment of the Dimensions and the Stability of Items in the MARKOR Scale," *Marketing Intelligence & Planning*, 17(5):248.

----, B. Ramaseshan, and Michael T. Ewing, 1997. "Market Orientation and Organization Commitment in the Australian Public Sector," *The International Journal of Public Sector Management*, 10(4):294.

Cavusgil, S. Tamer and Ajay Das, 1997. "Methodological Issues in Empirical Cross-Cultural Research: A Survey of the Management Literature and a Framework," *Management International Review*, 37(1):71-97.

Chakraborty, Goutam, Dwayne Ball, Gary J. Gaeth and Sunkyu Jun, 2002. "The Ability of Ratings Based and Choice Conjoint to Predict Market Shares: A Monte Carlo Simulation," *Journal of Business Research*, 55(3):237.

Cheng, J.L.C., 1989. "Toward a Contextual Approach to Cross-National Organisation Research: A Macro Perspective," *Advances in International Comparative Management*, 4:3-18.

Cherry, John, Monle Lee and Charles S. Chien, 2003. "A Cross-Cultural Application of a Theoretical Model of Business Ethics: Bridging the Gap Between," *Journal of Business Ethics*, 44(4):359.

Cheung, Gordon W. and Roger B. Rensvold 1999. "Testing Factorial Invariance Across Groups: A Reconceptualization and Proposed New Method," *Journal of Management*, 25(1):1

Chu, P.C., Eric E. Spires and Toshiyuki Sueyoshi, 1999. "Cross-Cultural Differences in Choice Behavior and Use of Decision Aids: A Comparison of Japan and the United States," *Organizational Behavior and Human Decision Processes*, 77(2):147-171.

Clark, Terry, 1990. "International Marketing and National Character: A Review and Proposal for an Integrative Theory," *Journal of Marketing*, 54(4):66-80.

Clarke, Irvine III, 1996. "Extreme Response Style: The Impact of Culture and Liker Response Formats," Dissertation, Old Dominion University, Norfolk, VA.

Cox, Reavis, 1965. "The Search for Universals in Comparative Studies of Domestic Marketing Systems," *Marketing and Economic Development: Proceedings of the 1965 Fall Conference*, ed. P.D. Bennett, American Marketing Association, Chicago, pp. 143-162.

Craig, C. Samuel and Susan P. Douglas, 2000. International Marketing Research, New York: Wiley.

Crompton, Rosemary, 1996. "Paid Employment and the Changing System of Gender Relations: A Cross-National Comparison," *Sociology: the Journal of the British Sociological Association*, 30(3):427-446.

Darley, William K. and Denise M. Johnson, 1993. "Cross-National Comparison of Consumer Attitudes Toward Consumerism in Four Developing Countries," *The Journal of Consumer Affairs*, 27(1):37-55.

Davis, Harry L., Susan P. Douglas and Alvin J. Silk, 1981. "Measure Unreliability: A Hidden Threat to Cross-National Marketing Research," *Journal of Marketing*, 45(2):98-109.

Dayton, Nick A, 2001. "Total Quality Management Critical Success Factors, A Comparison: The UK Versus the USA," *Total Quality Management & Business Excellence*, 12(3):293.

Deuze, Mark, 2002. "National News Cultures: A Comparison of Dutch, German, British, Australian, and U.S. Journalists," *Journalism and Mass Communication Quarterly*, 79(1):134-150.

Diamantopoulos, A., B. B. Schlegelmich and J. P. Du Preez, 1995. "Lessens for Pan-European Marketing?" *International Marketing Review*, 12(2):38.

Douglas, Susan P., Maureen A. Morrin and C. Samuel Craig, 1994. "Cross-National Consumer Research Traditions," in Research Traditions in Marketing, Laurent, Lilien, & Pras (eds), Boston: Kluwer Academic Publishers.

Eastman, Jacqueline K., Bill Fredenberger, David Campbell, and Stephen Calvert, 1997. "The Relationship Between Status Consumption and Materialism: A Cross-Cultural Comparison of Chinese, Mexican and American Students," *Journal of Marketing Theory and Practice*, 5(1):52-67.

Elliott, Gregory R. and Ross C. Cameron, 1994. Consumer Perception of Product Quality and the Country-of-Origin Effect," *Journal of International Marketing*, 2(2):49-53.

Ember, Melvin and Carol R. Ember, 2000. "Testing Theory and Why the 'Units of Analysis' Problem is Not a Problem," *Ethnology*, 39(4):349.

England, George W. and Itzhak Harpaz, 1983. "Some Methodological and Analytic Considerations in Cross-National Comparative Research," *Journal of International Business Studies*, Fall: 49-59.

Erffmeyer, Robert C., Bruce D. Keillor and Debbie Thorne LeClair, 1999. "An Empirical Investigation of Japanese Consumer Ethics," *Journal of Business Ethics*, 18(1):35.

Ford, John B., Patricia Kramer Voli, Earl D. Honeycutt Jr., and Susan L. Casey, 1998. "Gender Role Portrayals in Japanese Advertising: A Magazine Content Analysis," *Journal of Advertising*, 27(1):113-125.

Fraser, Campbell and Anna Zarkada-Fraser, 2002. "An Exploratory Investigation Into Cultural Awareness and Approach to Negotiation of Greek, Russian and British Managers," *European Business Review*, 14(2):111-128.

Freestone, Robert and Peter Murphy, 1998. "Metropolitan Restructuring and Suburban Employment Centers: Cross-Cultural Perspectives on the Australian Experience," *Journal of the American Planning Association*, 64(3):286-298.

Fuhrer, R., M.J. Shipley, J.F. Chastang, A. Schmaus, 2002. "Socioeconomic Position, Health, and Possible Explanations: A Tale of Two Cohorts," *American Journal of Public Health*, 92(8):1290-1295.

Gibson, Lawrence D. 2001. "What's Wrong With Conjoint Analysis?" *Marketing Research*, 13(4):16-19.

Gnepa, Tahi and Alfred Petrosky, 2001. "Cross-National Comparison of Consumption Values: A Center-Periphery Perspective," *Journal of International Consumer Marketing*, 14(2,3):73.

Gordon, Mary Ellen and Kathryn DeLima-Turner, 1997. "Consumer Attitudes Towards Internet Advertising: A Social Contract Perspective," *International Marketing Review*, 14(5):362.

Green, Reginald Herbold, 1998. "A Cloth Untrue: The Evolution of Structural Adjustment in Sub-Saharan Africa," *Journal of International Affairs*, 52(1):207-233.

Green, Paul E. and Yoram Wind, 1975. "New Way to Measure Consumers' Judgments," Harvard Business Review, (July-August):105-116.

Green, Robert T., and Phillip D. White, 1976. "Methodological Considerations in Cross-National Consumer Research," *Journal of International Business Studies*, 7(3):81-87.

Greenleaf, Eric A., 1992. "Measuring Extreme Response Style," *Public Opinion Quarterly*, 56(3):176-189.

Groschl, Stefan and Clayton W. Barrows, 2003. "A Cross-Cultural Comparison of French and British Managers: An Examination of the Influence of Higher Education Management Style," *Tourism and Hospitality Research*, 4(3):228-247.

Gustafsson, Anders, Andreas Herrmann and Frank Huber, 2001. "Conjoint Analysis as an Instrument of Market Research Practice," in Conjoint Measurement: Methods and Applications, Second Edition (Gustafsson, Herrmann & Huber, eds.); Berlin; Springer.

Haaijer, Rinus and Michel Wedel, 2001. "Conjoint Choice Experiments: General Characteristics and Alternative Model Specifications," in Conjoint Measurement: Methods and Applications, Second Edition (Gustafsson, Herrmann & Huber, eds.); Berlin:Springer.

Hair, Joseph F. Jr., Rolph E. Anderson, Ronald L. Tatham, and William C. Black, 1998. Multivariate Data Analysis, Upper Saddle River, NJ: Prentice Hall.

Hampton, Fergus and Sarah Emerson, 2003. "Opinions and Loyalty Among European Motoring Journalists," *Journal of Communication Management*, 7(3):269-275.

Head, David, 1988. "Advertising Slogans and the 'Made-In' Concept," *International Journal of Advertising*, 7(3):237-253.

Henderson, Karla A. and Barbara E. Ainsworth, 2003. "A Synthesis of Perceptions About Physical Activity Among Older African American and American Indian Women," *American Journal of Public Health*, 93(2):313-318.

Hill, Tracey and Robin N. Shaw, 1995. "Co-Marketing Tourism Internationally: Bases for Strategic Alliances," *Journal of Travel Research*, 34(1):25-33.

Himes, Christine L., 1994. "Age Patterns of Mortality and Cause of Death Structures in Sweden, Japan, and the United States," *Demography*, 31(4):633-652.

Hisrich, Robert D., Branko Bucar and Sevgi Oztark, 2003. "A Cross-Cultural Comparison of Business Ethics: Cases of Russia, Slovenia, Turkey, and United States," *Cross Cultural Management*, 10(1):3.

Huber, Joel, 2001. "The Importance of Multinomial Logit Analysis of Individual Consumer Choices," Technical Paper, Sawtooth Software.

Hui, Michael K. and Kevin Au, 2001. "Justice Perceptions of Complaint-Handling: A Cross-Cultural Comparison Between PRC and Canadian Customers," *Journal of Business Ressearch*, 52(2):161.

Hulin, Charles L., and Laura J. Mayer, 1986. "Psychometric Equivalence of a Translation of the Job Descriptive Index Into Hebrew," *Journal of Applied Psychology*, 71(1):83.

Hulland, John S., 1999. "The Effects of Country-of-Brand and Brand Name on Product Evaluation and Consideration: A Cross-Country Comparison," *Journal of International Consumer Marketing*," 11(1):23.

Huo, Y Paul, Heh Jason Huang and Nancy K. Napier, 2002. "Divergence or Convergence: A Cross-National Comparison of Personnel Selection Practices," *Human Resources Management*, 41(1):31.

Igbaria, Magid and Moshe Zviran, 1996. "Comparison of End-user Computing Characteristics in the U.S., Israel and Taiwan," *Information & Management*, 30(1):1-14.

Imrie, Brian C., John W. Cadogan and Rod McNaughton, 2002. "The Service Quality Construct on a Global Stage," *Managing Service Quality*, 12(1):10-19.

Jaeger, Sara R., Duncan Hedderly and Halliday J. H. MacFie, 2001. "Methodological Issues in Conjoint Analysis: A Case Study," *European Journal of Marketing*, 35(11/12):1217-1237.

Jeon, Woochang, George R. Franke, Bruce A. Huhmann, and Joseph Phelps, 1999. "Apppeals in Korean Magazine Advertising: A Content Analysis and Cross-Cultural Comparison," *Asia Pacific Journal of Management*, 16(2):249.

Johnson, Jean L., John B. Cullen, Tomoaki Sakano, and James W. Bronson, 2001. "Drivers and Outcomes of Parent Company Intervention in IJV Management: A Cross-Cultural Comparison," *Journal of Business Research*, 52(1):35.

Joseph, W. Benoy, 1982. "The Credibility of Physically Attractive Commiunicators: A Review," *Journal of Advertising*, 11(3):15-25.

Kaynak, Erdener and Ali Kara, 2002. "Consumer Perceptions of Foreign Products: An Analysis of Product-Country Images and Ethnocentrism," *European Journal of Marketing*, 36(7/8):928.

Kennedy, Ellen J. and Leigh Lawton, 1996. "The Effects of Social and Moral Integration on Ethical Standards: A Comparison of American and Ukranian Business Students," *Journal of Business Ethics*, 15(8):901-912.

Kettinger, William J., Choong C. Lee and Sunro Lee, 1995. "Global Measures of Information Service Quality: A Cross-National Study," *Decision Sciences*, 26(5):569-589.

Kivetz, Ran and Itamar Simonson, 2003. "The Idiosyncratic Fit Heuristic: Effort Advantage as a Determinant of Consumer Response to Loyalty Programs," *Journal of Marketing Research*, 40(4):454.

Klein, Jill Gabrielle, Richard Ettenson, and Marlene D. Morris, 1998. "The Animosity Model of Foreign Product Purchase: An Empirical Test in the People's Republic of China," *Journal of Marketing*, 62(1):89-101.

Klenosky, David B., Suseanne B. Benet and Petr Chabraba, 1996. "Assessing Czech Consumers' Reactions to Western Marketing Practices: A Conjoint Approach," *Journal of Buisness Research*, 36(2):189-199.

Kohli, Ajay, Bernard Jaworski and Ajith Kumar, 1993. "MARKOR: A Measure of Market Orientation," *Journal of Marketing Research*, 30(4):467.

Kropp, Fredric, Anne M. Lavack, Steven J.S. Holden, 1999. "Smokers and Beer Drinkers: Values and Consumer Susceptibility to Interpersonal Influence," *The Journal of Consumer Marketing*, 16(6):536.

Lam, Simon S. K., and Ka Shing Woo, 1997. "Measuring Service Quality: A Test-Retest Reliability Investigation of SERVQUAL," *International Journal of Market Research*, 39(2):381-397.

Lancaster, Thomas D. and Gabriella R. Montinola, 2001. "Comparative Political Corruption: Issues of Operationalization and Measurement," *Studies in Comparative International Development*, 36(3):3-28.

Laroche, Michel, Frank Pons and Adree Turmel, 2002. "A Cross-Cultural Comparison of Direct Mail Receptivity," *Journal of International Consumer Marketing*, 14(4):5.

Lee, Hi-Kak, 2001. "A Cross-Cultural Study of the Career Maturity of Korean and United States High School Students," *Journal of Career Development*, 28(1):43.

Leeflang, Peter S. H. and Dick R. Wittink, 2000. "Building Models for Marketing Decisions: Past, Present and Future," *International Journal of Research in Marketing*, 17:105-126.

Lenartowicz, Tomasz and Kendall Roth, 1999. "A Framework for Cultural Assessment," *Journal of International Business Studies*, 30(4):781.

Liefeld, John P., Marjorie Wall, and Louise A. Heslop, 1999. "Cross-Cultural Comparison of Consumer Information Processing Styles," *Journal of Euro-Marketing*, 8(1/2):29.

Louviere, Jordan, 1988. Analyzing Decision Making: Metric Conjoint Analysis. Newbury Park, Sage Publications.

Malhotra, Naresh K., J. Daniel McCort, 2001. "A Cross-Cultural Comparison of Behavioral Intention Models: Theoretical Considerations and an Empirical Investigation," *International Marketing Review*, 18(3):235.

---- , Mark Peterson, and Susan Bardi Kleiser, 1999. "Marketing Research: A State-of-the-Art Review and Directions for the Twenty-First Century," *Journal of the Academy of Marketing Science*, 27(2):160-183.

---- , James Agarwal and Mark Peterson, 1996. "Methodological Issues in Cross-Cultural Marketing Research: A State-of-the-Art Review," *International Marketing Review*, 13(5):7-43.

---- , 1988. "A Methodology for Measuring Consumer Preferences in Developing Countries," *International Marketing Review*, (Autumn):52-66.

Martin, Ingrid M. and Sevgin Eroglu, 1993. "Measuring a Multi-Dimensional Construct: Country Image," *Journal of Business Research*, 28(3):191-211.

Martinez, Teodoro Luque, Jose-Angel Ibanez-Zapata, and Salvador del Barrio-Garcia, 2000. "Consumer Ethnocentrism Measurement – An Assessment of the Reliability and Validity of the CETSCALE in Spain," *European Journal of Marketing*, 34(11/12):1353.

Mathur, Anil, 1998. "Incorporating Choice Into an Attitudinal Framework: Cross-Cultural Extension and Additional Findings," *Journal of International Consumer Marketing*, 10(4):93-111.

Matsumoto, David, 2002. "Methodological Requirements to Test a Possible In-Group Advantage in Judging Emotions Across Cultures: Comment on Elfenbein and Ambady (2002) and Evidence," *Psychological Bulletin*, 128(2):236-243.

Mavondo, Felix T. and Mark A. Farrell, 2000. "Measuring Market Orientation: Are There Differences Between Business Marketers and Consumer Marketers?" *Australian Journal of Management*, 25(2):223-255.

May, Peter J. and Raymond J. Burby, 1996. "Coercive Versus Cooperative Policies: Comparing Intergovernmental Mandate Performance," *Journal of Policy Analysis and Management*, 15(2):171-174.

McDonald, Gael, 2000. "Cross Cultural Methodological Issues in Ethical Research," *Journal of Business Ethics*, 27(1-2):89-105.

Mejias, Roberto J., Morgan M. Shepherd, Douglas R. Vogel and Litva Lazaneo, 1996/1997. "Consensus and Perceived Satisfaction Levels: A Cross-cultural Comparison of GSS and Non-GSS Outcomes Within and Between the United States and Mexico," *Journal of Management Information Systems*, 13(3):137-162.

Moon, Byeong Joon and Shubhash C. Jain, 2001. "Consumer Processing of International Advertising: The Roles of Country of Origin and Consumer Ethnocentrism," *Journal of International Consumer Marketing*, 14(1):89.

Moore, Eric G., 1998. "Competitive Judgments in a Business Simulation: A Comparison Between American and Chinese Business Students," *Psychology & Marketing*, 15(6):547-566.

Moskowitz, Howard, Bert Krieger and Samuel Rabino, 2002. "Element Category Importance in Conjoint Analysis: Evidence for Segment Differences," *Journal of Targeting, Measurement and Analysts for Marketing*, 10(4):366-385.

Mullen, Michael R., 1995. "Diagnosing Measurement Equivalence in Cross-National Research," *Journal of International Business Studies*, 26(3):573-596.

Mummalaneni, Vankatapparao, Khalid M. Dubas and Chiang-nan Chao, 1996. "Chinese Purchasing Managers' Preferences and Trade-offs in Supplier Selection and Performance Evaluation," *Industrial Marketing Management*, 25(2):115-125.

Narver, John C. and Stanley F. Slater, 1990. "The Effect of Market Orientation on Business Profitability," *Journal of Marketing*, 54(4):20-36.

Nasif, Ercan G., Hamad Al-Daeaj, Bahman Ebrahimi and Mary S. Thibodeaux, 1991. "Methodological Problems in Cross-Cultural Research: An Updated Review," *Management International Review*, 31(1):79-91.

Nettemeyer, Richard G., Srinivas Durvasula and Donald R. Lichtenstein, 1991. "A Cross-National Assessment of the Reliability and Validity of the CETSCALE," *Journal of Marketing Research*, 28(3):320-328.

Nevid, Jeffrey S. and Nelly L. Sta Maria, 1999. "Multicultural Issues in Qualitative Research," *Psychology & Marketing*, 16(4):305-326.

Nixon, Judy C. and Gail A. Dawson, 2002. "Reasons for Cross-Cultural Communication Training," *Corporate Communications*, 7(3):184.

O'Cass, Aron and Kenny Lim, 2002. "Toward Understanding the Young Consumer's Brand Associations and Ethnocentrism in the Lion's Port," *Psychology & Marketing*, 19(9):759.

Odekerken-Schroder, Gaby, Hans Ouwersloot, Jas Lemmink and Janjaap Semeijn, 2003. "Consumers' Trade-Off Between Relationship, Service Package and Price: An Empirical Study in the Car Industry," *European Journal of Marketing*, 37(1/2):219.

Okechuku, Chike and Vincent Onyemah, 1999. "Nigerian Consumer Attitudes Toward Foreign and Domestic Products," *Journal of International Business Studies*, 30(3):611-622.

Oliver, Elizabeth Goad and Karen S. Cravens, 2001. "An International Comparison of Employee Welfare Plans," *Thunderbird International Business Review*, 34(4):501.

Oppewal, Harmen and Marco Vriens, 2000. "Measuring Percieved Service Quality Using Integrated Conjoint Experiments," *The International Journal of Bank Marketing*, 28(4):154-169.

Okazaki, Shintaro and Javier Alonso Rivas, 2002. "A Content Analysis of Multinationals' Web Communication Strategies: Cross-Cultural Research Framework and Pre-testing," *Internet Research*, 12(5):380-391.

Palich, Leslie E., Gary R. Carini, and Linda P. Livingstone, 2002. "Comparing American and Chinese Negotiating Styles: The Influence of Logic Paradigms," *Thunderbird International Business Review*, 44(6):777.

Palvia, Shailendra and M. Gordon Hunter, 1996. "Information Systems Development: A Conceptual Model and a Comparison of Methods Used in Singapore, USA and Europe," *Journal of Global Information Management*, 4(3):5-17.

Parameswaran, Ravi and Attila Yaprak, 1987. "A Cross-National Comparison of Consumer Research Measures," *Journal of International Business Studies*, Spring(18):35-50.

Parasuraman, A., Valerie A. Zeithaml and Leonard L. Berry, 1988. "SERVQUAL: A Multiple-Item Scale for Measuring Consumer Perception of Service Quality," *Journal of Retailing*, 64(1):12-41.

Park, Seong-Yeon, 1998. "A Comparison of Korean and American Gift-Giving Behaviors," *Psychology & Marketing*, 15(6):577-594.

Parker, R. Stephen and Charles E. Pettijohn, 2003. "Ethical Considerations in the Use of Direct-to-Consumer Advertising and Pharmaceutical Promotions: The Impact on Pharmaceutical Sales and Physicians," *Journal of Business Ethics*, 48(3):279.

Peterson, Dane K., Chung Kim and Joong H. Kim, 2003. "Perceptions of Information Systems Objectives: A Comparison of IS Professionals from the United States and Korea," *Journal of Global Information Technology Mangement*, 6(2):27.

Pines, Ayala Malach, 2003. "Occupational Burnout: A Cross-Cultural Isreali Jewish-Arab Perspective and Its Implications for Career Counseling," *Career Development International*, 8(2):97.

Pitt, Leyland, Albert Caruana and Pierre R. Berthon, 1996. "Market Orientation and Business Performance: Some European Evidence," *International Marketing Review*, 13(1):5-19.

Polonsky, Michael Jay, Pedro Quelhas Brito, Jorge Pinto and Nicola Higgs-Kleyn, 2001. "Consumer Ethics in the European Union: A Comparison of Northern and Southern Views," *Journal of Business Ethics*, 31(2):117-131.

Prashant, Malaviva, Jolita Kisielius and Brian Sternthal, 1996. "The effect of Type of Elaboration on Advertisement Processing and Judgment," *Journal of Marketing Research*, 33(4):410.

Resnik, Alan and Bruce L. Stern, 1977. "An Analysis of Information Content in Television Advertising," *Journal of Marketing*, 41(1):50.

Ryan, Ann Marie, David Chan, Robert E. Ployhart and L. Allen Slade. 1999. "Employee Attitude Surveys in a Multinational Organization: Considering Language and Culture in Assessing Measurement Equivalence," *Personnel Psychology*, 52(1):37.

Samiie, Saeed and Jeong, I., 1994. "Cross-Cultural Research in Advertising: An Assessment of Methodologies," *Journal of the Academy of Marketing Science*, 22(3):205-217.

Sapede, Claudine, and Isabelle Girod, 2002. "Willingness of Adults in Europe to Pay for a New Vaccine: The Application of Descrete Choice-Based Conjoint Analysis," *International Journal of Market Research*, 44(Special Issue):463-476.

Sauer, Beverly A., 1996. "Communicating Risk in a Cross-Cultural Comparison of Theoretical and Social Understandings in U.S. and British Mine Safety Training Programs," *Journal of Business and Technical Communication*, 10(3):306-330.

Schneider, Andreas, 1999. "The Violent Character of Sexual-Eroticism in Cross-Cultural Comparison," *The International Journal of Sociology and Social Policy*, 19(1/2):81-101.

----, 1996. "Sexual-Erotic Emotions in the U.S. In Cross-Cultural Comparison," *The International Journal of Sociology and Social Policy*, 16(9/10):123-144.

Scott, Jacqueline, Duane F. Alwin, and Michael Braun, 1996. "Generational Changes in Gender-Role Attitudes: Britain in a Cross-National Perspective," *Sociology: The Journal of the British Sociological Association*, 30(3):471-493.

Sekaran, Uma, 1983. "Methodological and Theoretical Issues and Advancements in Cross-Cultural Research," *Journal of International Business Studies*, Fall:61-73.

Shepherd, Philip L., John Tsalikis and Bruce Seaton, 2002, "An Inquiry Into the Ethical Perceptions of Sub-Cultural Groups in the U.S.: Hispanics Versus Anglos," *The Journal of Consumer Marketing*, 19(2/3):130-148.

Shimp, Terrance and Shubash Sharma, 1987. "Consumer Ethnocentrism: Construction and Validation of the CETSCALE," *Journal of Marketing Research*, 24(3):280.

Sin, Leo Y M., Gordon W. H. Cheung and Ruby Lee, 1999. "Methodology in Cross-Cultural Consumer Research: A Review and Critical Assessment," *Journal of International Consumer Marketing*, 11(4):75-96.

Singh, Jagdip, 1995. "Measurement Issues in Cross-National Research," *Journal of International Business Studies,* v(3):597-619.

Small, Rhonda, Jane Yelland, Judith Lumley, Pranee Liamputtong Rice, Valerie Cotronei, and Rosemary Warren, 1999. "Enhancing Data Quality: Cross-Cultural Research – Trying to Do it Better, Part 2," *Australian and New Zealand Journal of Public Health*, 23(4):390-396.

Smith, Anne M. and Nina L. Reynolds, 2002. "Measuring Cross-Cultural Service Quality: A Framework for Assessment," *International Marketing Review*, 19(4/5):450-482.

Smulders, Peter G W, Michiel A. J. Kompier and Pascal Paoli, 1996. "The Work of Environment in The Twelve EU-Countries: Differences and Similarities," *Human Relations*, 49(10):1291-1314.

Solomon, Michael R., Richard D. Ashmore, and Laura C. Longo, 1992. "The Beauty Match-up Hypothesis: Congruence Between Types of Beauty and Product Images in Advertising," *Journal of Advertising*, 21(4):23.

Spira, Alfred, Nathalie Bajos, Alain Giami and Stuart Michaels, 1998. "Editorial: Cross-National Comparisons of Sexual Behavior Surveys – Methodological Difficulties and Lessons for Prevention," *American Journal of Public Health*, 88(5):730.

Stauss, Bernd and Paul Mang, 1999. "Culture Shocks in Inter-Cultural Service Encounters?" *The Journal of Services Marketing*, 13(4/5):329.

Stone, Dan N., Vairam Arunachalam and John S. Chandler, 1996. "Cross-Cultural Comparisons: An Empirical Investigation of Knowledge, Skill, Self-efficacy and Computer Anxiety in Accounting Education," *Issues in Accounting Education*, 11(2):345-357.

Styles, Chris, 1998. "Export Performance Measures in Australia and the United Kingdom," *Journal of International Marketing*, 6(3):12-37.

Supphellen, Magne and Kjell Gronhaug, 2003. "Building Foreign Brand Personalities in Russia: The Moderating Effect of Consumer Ethnocentrism," *International Journal of Advertising*, 22(2):203.

Thu, Kendall M., Kelly Hattman, Vance Hutchinson and Scott Leuken, 2002. "Keeping the Game Close: 'Fair Play' Among Men's College Basketball Referees," *Human Organization*, 61(1):1-9.

Tomoko, Kanata and N. J. Banks, 1997. "A Comparison of "Social Issues Affecting Japanese and British Lone Mother Households," *The International Journal of Sociology and Social Policy*, 17(5):52-84.

Tsalikis, John, Bruce Seaton and Petros Tomaras, 2002, "A New Perspective on Cross-Cultural Ethical Evaluations: The Use of Conjoint Analysis," *Journal of Business Ethics*, 35(4):281-292.

Turnipseed, David L. and Patricia H. Turnipseed, 1997. "A Bi-Cultural Analysis of the Cost of Caring: Nursing Burnout in the United States and the Philippines," *Career Development International*, 2(4):102-120.

Tuohy, Carolyn, 1994. "Response to the Clinton Proposal: A Comparative Perspective," *Journal of Health Politics, Policy and Law*, 19(1):249-254.

Vriens, Marco, Hiek R. van der Scheer, Janny C. Hoekstra, and Jan Roelf Bult, 1998. "Conjoint Experiments for Direct Mail Response Optimization," *European Journal of Marketing*, 32(3/4):323-339.

Weber, Elke U., Christopher K. Hsee and Joanna Sokolowska, 1998. "What Folklore Tells Us About Risk and Risk Taking: Cross-Cultural Comparisons of American, German, and Chinese Proverbs," *Organizational Behavior and Human Decision Processes*," 75(2):170-187.

Weinfurt, Kevin P. and Fathali M. Moghaddam, 2001. "Culture and Social Distance: A Case Study of Methodological Cautions," *The Journal of Social Psychology*, 141(1):101

Wetzels, Martin, Ko de Ruyter, and Jos Lemmink, 2000. "Measuring Service Quality Trade-offs in Asian Distribution Channels: A Multi-Layer Perspective," *Total Quality Management*,11(3):307-318.

Winstead, Kathryn Frazer, 1997. The Service Experience in Two Cultures: A Behavioral Perspective," *Journal of Retailing*, 73(3):337-360.

Winter, Lothar G. and Charles R. Prohaska, 1983. "Methodological Problems in the Comparative Analysis of International Marketing Systems," *Journal of the Academy of Marketing Science*, 11(4):417-432.

Wong, Nancy, Aric Rindfleisch, and james E. Burroughs, 2003. "Do Reverse-Worded Items Confound Measures in Cross-Cultural Consumer Research? The Case of the Material Values Scale," *Journal of Consumer Research*, 30(1):72-92.

Yen, Hsiu Ju, Dennis W. Krumwiede & Chwen Sheu, 2002. "A Cross-Cultural Comparison of Top Management Personality for TQM Implementation," *Total Quality Management & Business Excellence*, 13(3):335.

Yu, Julie H. and Gerald Albaum, 2002. "Sovereignty Change Influences on Consumer Ethnocentrism and Product Preferences: Hong Kong Revisited One Year Later," *Journal of Business Research*, 55(11):891.

Yunxia, Zhu, 2000. "Building Knowledge Structures in Teaching Cross-Cultural Sales Genres," *Business Communication Quarterly*, 63(4):49-69.