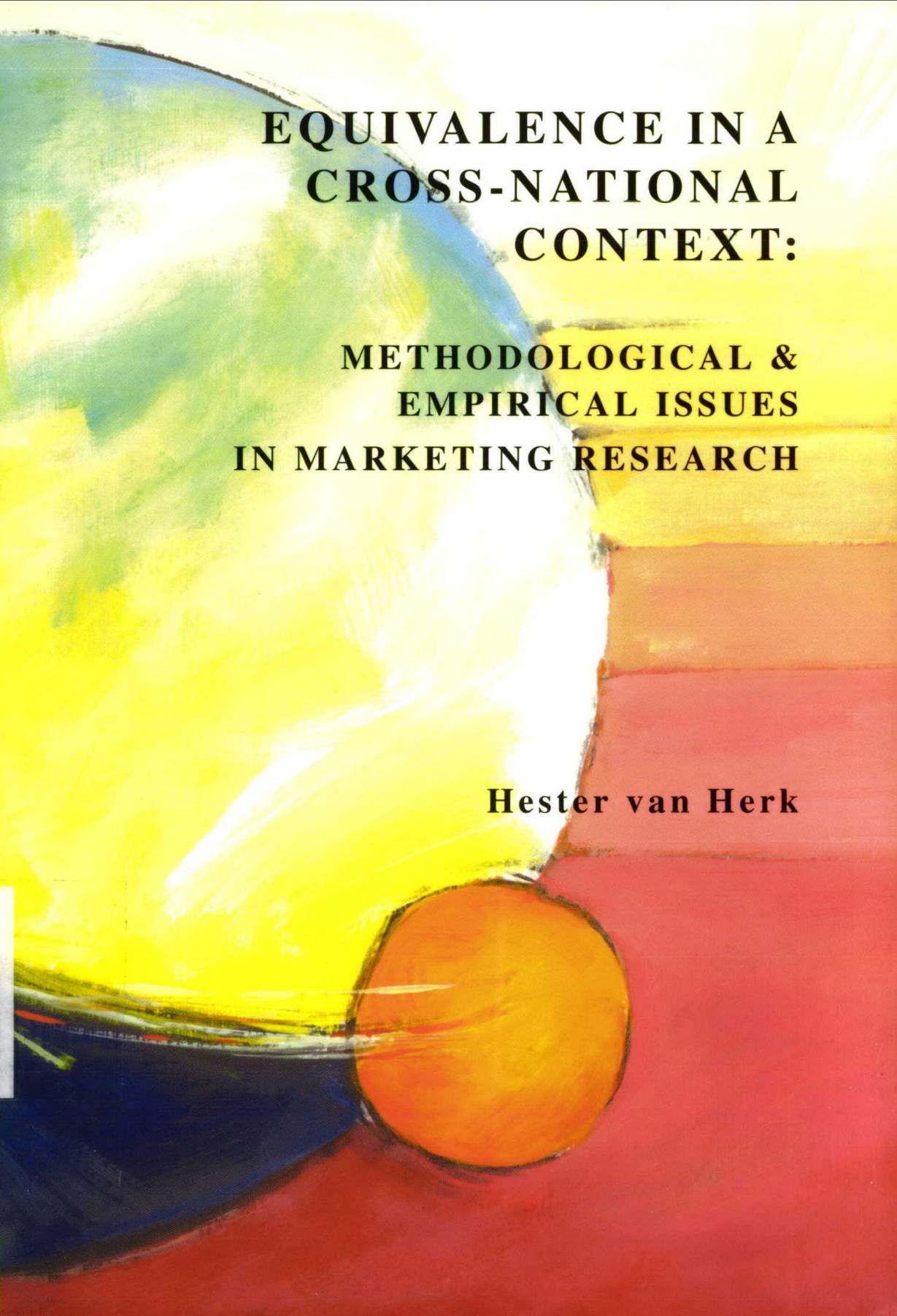
Equivalence in a Cross-National Context

van Herk, H.

Publication date:
2000

Document Version
Publisher's PDF, also known as Version of record


EQUIVALENCE IN A CROSS-NATIONAL CONTEXT:
METHODOLOGICAL & EMPIRICAL ISSUES IN MARKETING RESEARCH
Hester van Herk

EQUIVALENCE IN A CROSS-NATIONAL CONTEXT:

METHODOLOGICAL&EMPIRICAL ISSUES IN MARKETING RESEARCH

Hester van Herk

Cover design: Ella Steenmeijer, Zeewolde

ISBN 90-9013822-6

Printed by: Offsetdrukkerij Ridderprint B.V., Ridderkerk

Equivalence In A Cross-National Context:

Methodological & Empirical issues in marketing research

Proefschrift ter verkrijging van de graad van doctor aan de Katholieke Universiteit Brabant op gezag van de Rector Magnificus, prof. dr. F.A. van der Duyn Schouten, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de aula van de Universiteit op

woensdag 28 juni 2000 om 14.15 uur

door

Hesterina van Herk

geboren op 4 september 1961 te Gouda

Promotores:

Prof. dr Th.M.M. Verhallen
Prof. dr Y.H. Poortinga

Aan mijn ouders

Dit onderzoek is mede mogelijk gemaakt door financiële en/of materiële bijdragen van:

Katholieke Universiteit Brabant
Unilever Research
Philips DAP

Acknowledgments

Writing a dissertation has been my wish since I started working in marketing. In 1993, I was given the opportunity to start working on a large-scale international marketing research project, funded by Unilever Research, at Tilburg University. Although this was only for one day a week, it stimulated me to really begin writing a thesis. In practice I had seen that international marketing research was fascinating. However, I discovered that academic international marketing research was still in its infancy. By then few articles and only one book had been published on the subject. Working in this area has been a real challenge and still is.

A dissertation project is a project one cannot do on one's own. Therefore, I want to thank many people and some in particular for their help.

First, I would like to thank my thesis supervisors Theo Verhallen and Ype Poortinga. Theo gave me the opportunity to work at the Marketing Department in Tilburg. In the last seven years his interest in international marketing research and his enthusiasm have been catching. Ype stimulated me to strive for constant improvement. His constructive criticism and his eye for detail are exemplary. I hope that we will be able to continue our cooperation.

Second, I would like to thank the other members of my Ph.D. committee, Niels Noorderhaven, Jan-Benedict Steenkamp, Jean-Claude Usunier and Fons van de Vijver for carefully reading and judging my work.

Third, I would like to thank John Barzilay from Unilever Research in Vlaardingen and Leo de Vries from Philips Domestic Appliances and Personal Care in Groningen for providing the data used in this thesis. But, above all I recall the pleasant discussions we had about doing international marketing research in practice. The interaction between marketing theory and practice has stimulated me enormously.

Fourth, my family and friends have been a great help to me. Otto Wolthuis and Marianne Vorrink read my dissertation and provided helpful comments for increasing its readability. At home, Anneke Wolthuis deputized for me whenever necessary. I appreciate their help very much.

As in international research, in practice, the environment has an effect on how situations are assessed. The working environment at the Marketing Department has been very pleasant. Especially, I want to thank my colleagues Tammo Bijmolt, Roger Bougie, Paul Driessen, Ruud Frambach and Henk Roest for this.

Last but not least, I thank my husband Albert Jan for his patience, love, never-ending confidence in me and his 'being there'. Moreover, he spent quite some time making the layout of this dissertation, creating indices and doing the final editing. His support has been invaluable.

I dedicate this book to my parents. They taught me to be myself.

Hester van Herk

Zeewolde, April 2000

Contents

List of Tables

List of Figures

Chapter 1. Introduction

1.1 Introduction

The globalization of markets for goods and services has been one of the most talked about issues in the last few decades. Companies like Unilever, Philips, Ford, IKEA, Sony, and many others have taken a global view and structure themselves accordingly. Even for medium-sized enterprises, the international environment is important. They too export to and invest in other than the domestic market.

Global trade has grown 17-fold over the 50-year lifespan of the WTO[1,2]. In the last decade, a number of developments have facilitated the globalization process even more. These are, in particular, the establishment of the European Union (EU), the establishment of the North American Free Trade Association (NAFTA), the establishment of the Association of South East Asian Nations (ASEAN), and the disintegration of the Soviet Union. Other developments that have facilitated global trade are the growth of information technology (IT), the advent of the Internet, and faster international travel links.

As companies are increasingly engaging in global trade, global marketing has become a vital issue for them. Cultural, economic, legal, and geographic differences between the home market and the markets of other countries have to be taken into account. To gain insight into the complex international environment, many companies require international marketing information. The growing need for international marketing research information is shown in the worldwide turnover

---

[1] And its predecessor the GATT

[2] Source: http://europa.eu.int

for commercial opinion and market research; in 1998, this was 12 billion Euro, up 10% over the previous year (ESOMAR, 1999). In comparison with 1980, the increase was as much as 5-fold. The greatest part of world market turnover in opinion and marketing research is spent in the EU and the USA. To be more precise, in 1998, 43.5% of research was conducted in Europe (40.4% in the 15 EU countries), 38.6% in North America (36,8% USA), and 11.8% in Asia Pacific (6.7% Japan). The increase in worldwide information needs has led to international expansion of marketing research organizations and growth of revenues generated outside the home market. For example, in 1997, the top 25 global marketing research organizations derived on average 44.8% of their revenues from outside their home market (Craig & Douglas, 2000).

Marketing research information is requested by companies to help marketeers in making decisions. This information should help them decide which marketing mix strategies to deploy in other countries or which international markets to enter (Chee & Harris, 1998). In these decisions, a comparison between the outcomes of the research in different countries is made either implicitly or explicitly. A comparison is made if a manager wants to know whether a product can be marketed in a standardized way. Moreover, if the product has to be adapted, information is needed on the extent to which this has to be done. A comparison between countries is also made, if a manager wants to decide which country is the most interesting to enter first. Such comparisons are often difficult, since different languages and different customs are involved.

If comparisons between countries are to be made, 'comparability' or 'equivalence' of marketing information is a prerequisite. Equivalence exists if information can be compared directly across borders. For example, the number of inhabitants in Greece and Germany can be compared directly. However, establishing equivalence is not obvious; knowledge about local habits may be necessary to be able to decide whether a concept has the same meaning or not. The following example illustrates this. Superficially, a concept like 'deep frying' may have the same meaning in, for example, France and Greece. But 'deep frying' is not the same in France and Greece. The cooking process is totally different. The type of product and the quantity used for deep frying potatoes differs. In Greece, olive oil in relatively small quantities is used, whereas, in France, a large quantity of vegetable oil is used. Consequently, direct comparisons between France and Greece on attitudes towards deep frying cannot be made.

The outcomes of empirical comparative marketing research studies may be invalidated by differences between cultures. There may be differences in many aspects, for example, customs, language, wealth, and education level. In addition, a common data collection procedure in domestic marketing research may be uncommon in another country. Even in adjacent countries such as France and the United Kingdom, common data collection methods differ. In France, street interviews are common, whereas such interviews are almost never used in the United Kingdom (Kumar, 2000). Such differences between cultures or countries introduce bias to international research that is absent or less influential in domestic research.

Equivalence of international marketing research information is of vital interest, because inequivalent or biased information may lead to ambiguous or even erroneous conclusions. In spite of this, the issue of equivalence is little known in practice. In academic research, little has been published on international marketing research methodology in general and on equivalence in particular.

In the following section, the notions of culture, equivalence, and bias as used in this thesis are defined. Next, examples are given of the cultural differences researchers are confronted with in practice (in the international environment). Such cultural differences may introduce bias in international marketing research and thus hamper equivalence. This is followed by a brief overview of academic research on international marketing. The role of equivalence in these studies is explicitly addressed. Lastly, unresolved issues in international comparative studies on marketing are discussed; establishing equivalence is emphasized. An outline of this thesis is presented at the end of this chapter.

1.2 Terminology

1.2.1 Culture

In international research in marketing as well as in other disciplines, such as cross-cultural psychology and international business studies, the notion 'culture' is often used. The notion is abstract and many definitions of it exist in the literature. In a

review by Kroeber and Kluckhohn (1952), no less than 164 definitions of culture were listed. This may imply that culture is not a uni-dimensional concept and that the meaning of culture should be further specified. In studies in marketing as well as in other disciplines, the notion 'culture' is frequently equated with 'country'. This is understandable and practical, but not justified. Within one country, several cultural populations may exist. For example, the Hispanics in the USA and the Turks in Germany still have their own 'culture'. Moreover, students and business-people can be said to each have their own 'culture'. Therefore, the notion of 'culture' cannot be used without the concept being specified further. Researchers should specify which population is being studied, and what the characteristics of this population are. These characteristics exist at country level and at the individual level. At the country level, they include GNP, economic development, and legislation. At the individual level, they include language, nationality, ethnic group, religion, social class, family, and gender (Usunier, 1996).

In the theoretical Chapters of this thesis (2 and 3), we will mainly use the term 'cultural population'. By using this term, we want to emphasize that the methodology used in these Chapters can be applied in a comparison of different types of populations. It can be used to compare ethnic groups within a country as well as subjects from different countries. In the empirical Chapters (4 to 6), the term 'country' is used. The reason for this is that in the data from commercial marketing research organizations that were used, the 'cultural populations' were equated with 'country'.

1.2.2 Equivalence and bias

To explain equivalence, we adopt the categorization developed by Van de Vijver and Leung (1997a,b). They mention three kinds of equivalence, construct equivalence, measurement unit equivalence and scalar equivalence. Construct equivalence exists if a phenomenon has the same meaning across cultural populations. For example, the phenomenon 'temperature' has construct equivalence. Both measurement unit equivalence and scalar equivalence are concerned with the measurement of a phenomenon. Measurement unit equivalence exists if the phenomenon can be measured with scales consisting of equal units of measurement, but with a different origin. For example, 'temperature' can be measured in degrees Celsius and in degrees Kelvin. There is scalar equivalence if both units of measurement and the origin of the measurement scale are equal. For

example, this is the case if temperature is measured in degrees Celsius in two cultural populations. The three kinds of equivalence are nested. For instance, if there is no construct equivalence, measurement unit and scalar equivalence cannot be established.

Equivalence determines the extent to which research information from different cultural populations can be compared directly. Cultural differences hamper equivalence; they introduce bias. Bias implies that measurement results obtained in different cultural populations cannot be interpreted in the same way. Bias may emerge with respect to the construct, an item in the questionnaire, the data collection method, or the interaction between an interviewer and a respondent. In general, bias lowers equivalence.

1.3 Factors affecting marketing research in practice

In practice, international marketing research requires understanding of the international environment, because it is a well known fact that the environment people live in affects their behavior (Peter & Olson, 1996). General characteristics of the environment, such as economic and political-legal characteristics (e.g., Jeannet & Hennessey, 1993), determine in part what products are available, at what price they are sold, where these products are available, and how they are promoted. An example is the import restriction on Japanese cars that existed in several European countries (Jain, 1993). Even today, sales of Japanese cars are low in France and Italy.

In international marketing decision making, the needs and wants of consumers in different countries are compared. Similar needs are required if a company intends to pursue a global marketing strategy, where the same products are offered all over the world. For some products, needs are similar; products like portable CD players and watches can be largely standardized and still appeal to many people from different countries. For such products, companies may globalize their strategies, and benefit from economies of scale and increased competitive power (Levitt, 1983; Yip, 1995).

There may be similar needs and wants for high tech products (e.g., computers) or high touch products (e.g., Gucci bags) across the globe (Ohmae, 1989). However,

for most consumer products, needs and wants are, however, less similar. For largely culture-bound products such as food, adaptations to local preferences are often made to satisfy customers' needs (Maucher & Brabeck-Lethmathe, 1993). There are differences between countries in the degree to which family members influence food choice (Lenneräs, Fjellström, Becker, Giachetti, Schmitt, Remaut de Winter & Kearney, 1997) and in preferences for colors and design for clothing (Botschen & Hemetsberger, 1998). Moreover, seemingly simple habits may be different. For instance, habits concerning spreading a product on bread or toast differ from country to country. In France, people prefer butter; in Greece, people prefer margarine, and in Italy spreading is less common. These different habits may affect peoples' awareness of the various available margarine brands and thus affect product evaluation. Moreover, if products are evaluated, both the way in which the product is used (Eimers, Pieters & Verhallen, 1995) as well as its country of origin (Johansson, Douglas & Nonaka, 1985; Maheswaran, 1994) affect the evaluation.

Even if needs and wants are the same, there may be differences in the adoption and diffusion of products across countries (Gatignon, Eliasberg & Robertson, 1989; Helsen, Jedidi & DeSarbo, 1993). For instance, the penetration of VCRs was much faster in the US than in France (Kotabe & Helsen, 1998).

1.4 Academic international marketing research studies

Until the eighties, little attention was paid to research methodology in international marketing. Most earlier publications on international research stem from related fields like cross-cultural psychology (Berry, 1969; Brislin, Lonner & Thorndike, 1973), and comparative management research (Graves, 1973; Schöllhammer, 1973; Hofstede, 1980). In a review by Albaum and Peterson (1984), it is stated that most of the research in international marketing before 1982 was descriptive or exploratory rather than analytical or normative; no explicit attention was paid to research methodology. In 1983, several authors (Douglas & Craig, 1983; Sekaran, 1983) called attention to the need for greater methodological rigor. This had hardly an effect; in the period 1985-1990, compared to the period 1980-1984, there was no significant increase in the number of international comparative marketing research studies in which attention to equivalence was paid (Aulakh & Kotabe, 1993).

The importance of studies in several countries for theory building in international marketing is emphasized in several publications (Douglas & Craig, 1992; Samiee & Jeong, 1994). Yet, the number of cross-national studies involving two or more countries remains scarce. In the period 1980-1990, only 10% of the studies in international marketing were comparative studies involving two or more countries (Aulakh & Kotabe, 1993). In general, they can be divided into three types, namely (1) cross-national segmentation studies, (2) studies testing the universality of behavior models, and (3) studies of values and attitudes (Douglas & Craig, 1997). Between these types, there are large differences in the attention paid to equivalence.

1.4.1 Cross-national segmentation

In cross-national segmentation, two general approaches can be distinguished, namely (1) segmentation based on aggregate macro information, and (2) segmentation on the basis of individual consumer information. Studies following the first approach are usually country segmentations. In these studies, countries are grouped on the basis of geographic, political, economic, or cultural variables. The groupings can be based on a single variable or a combination of such variables. Country segmentation may also be performed on more marketing-linked information such as multinational diffusion patterns (Helsen, Jedidi & DeSarbo, 1993), or international purchasing patterns (Frear, Alguire & Metcalf, 1995). In a study by Kale (1995), seventeen European countries were grouped on the basis of the four Hofstede dimensions (Hofstede, 1980). Instead of grouping countries, regions within countries can be grouped. An advantage of regional segmentation is that cross-national segments may emerge. For example, a region in one country that has more characteristics in common with a region in a neighboring country can be grouped together with that region. An example is the study by Askegaard and Madsen (1998), who used aggregate data on food behavior and attitudes from 79 regions in the European Union to create segments. In none of these studies, attention was paid to equivalence.

In the second approach, individual consumer data are used to segment markets. The use of individual data may lead to consumer segments that transcend national borders. Emphasis in studies within this approach is on a direct comparison of consumers from different countries in order to identify cross-national market segments. Cross-national segments have been identified, based on psychographics

(Boote, 1982), values (Kamakura, Novak, Steenkamp & Verhallen, 1994), and consumer-product relations (Ter Hofstede, Steenkamp & Wedel, 1999). Only in the study by Ter Hofstede et al. (1999) was explicit attention paid to equivalence issues.

1.4.2 Universality of behavioral models

In studies on the universality of behavioral models, emphasis is on testing the validity of theories that were developed in one country. Frequently, theories originating in the US are tested (Öngel & Smith, 1994; Craig & Douglas, 2000). For instance, the Fishbein behavioral intention model has been tested in the US and Korea (Lee & Green, 1991). Findings suggest that this model has cross-cultural validity, since a good fit was found in the two samples. In this study, no explicit attention was given to equivalence of the various constructs in the model. Durvasula, Andrews, Lysonski, and Netemeyer (1993) tested a general model of attitudes towards advertising in five countries. They found the same relationships between constructs in all countries, suggesting universality of the proposed model. In the latter study, construct equivalence of the measures was established. Measurement unit equivalence was also mentioned, but this was not established.

1.4.3 Studies of values and attitudes

A key theme in international research has been the study of cultural values. The Rokeach Value Survey (Rokeach, 1973), the List of Values (Kahle, 1983), and Schwartz's eleven motivational domains (1992) are examples of value systems that have been compared across countries. Values are assumed to be universal and several researchers have argued that they are especially suited for cross-cultural comparisons (e.g., Kamakura & Mazzon, 1991; Steenkamp, 1992). As a consequence, researchers have examined the link between values and issues such as fashion and engagement in several activities (Kamakura & Mazzon, 1991), or female fashion leadership (Goldsmith, Freiden & Kilsheimer, 1993). In none of these studies, was equivalence of the concepts tested explicitly. Only in the study on the antecedents of consumer innovativeness (Steenkamp, Ter Hofstede & Wedel, 1999) was attention paid to construct equivalence of several measures.

1.5 Unresolved issues

Equivalence is hardly addressed in cross-cultural comparative studies. The reasons for this may be twofold. First, the literature seems to include confusing and unclear descriptions of the notion of equivalence (cf. Van de Vijver & Leung, 1997b). Second, there may also be a lack of unified treatment of cross-cultural research methodologies in the marketing literature (Malhotra, Agarwal & Peterson, 1996). To render equivalence more open to academic researchers as well as to practitioners, a new and clear framework is required.

In so far as equivalence has been addressed, the focus has been on construct equivalence. Other types of equivalence, such as measurement unit equivalence, have not been analyzed. This is strange because without measurement unit equivalence no comparisons between levels of scores can be made. Research on this latter type of equivalence is interesting from an academic as well as from a practical point of view. In particular, comparisons between levels of scores are made in practice. For instance, the highest level of scores or the 'top box' of an item is commonly employed as an indication of purchase intention (Morwitz, Steckel & Gupta, 1997). However, comparisons between 'top box' scores become ambiguous or even erroneous if subjects in a particular country tend to avoid, for example, the extreme categories of a rating scale, that is, if they display a particular response style.

Thus, an important issue in international marketing research is whether research instruments are affected by response styles (Van Herk & Verhallen, 1995b; Usunier, 2000). Response styles differ across countries (e.g., Chen, Lee & Stevenson, 1995; Yeh, Kim, Chompreeda, Rimkeeree, Yau & Lundahl, 1998). In the comparative studies in marketing discussed in Section 1.4, response styles were not addressed. Moreover, existing studies on response style have not included respondents from countries in the European Union. There is a need for insight into differences and similarities in response styles between countries. Research should indicate to what extent response styles hamper equivalence in international marketing research. Moreover, research should indicate the antecedents of response style.

Another problem related to differences in response style in rating scales is how researchers deal with those differences. Some researchers suggest ranking to avoid

response styles. However, a comparison between rating and ranking in an international context is still lacking.

1.6 Outline

This thesis consists of two parts. In the first part, a new framework for equivalence in culture comparative research is proposed. The framework aims at researchers in practice, as well as academics. In the second part, response style is studied.

In Chapter 2, methodological issues in cultural comparative studies are considered. The equivalence or comparability of data collected from different countries should be a basic issue in these studies (e.g., Douglas & Craig, 1983; Sekaran, 1983; Singh, 1995; Van de Vijver & Leung, 1997a,b). The establishment of equivalence is far from common (Aulakh & Kotabe, 1993). A reason may be that in studies that do not focus on construct equivalence, the procedures for establishing equivalence are unclear. The purpose of this chapter is to indicate how equivalence can be established in future studies. First, a conceptual framework to establish equivalence in marketing research studies is presented. In this framework, a key issue is the determination of sources of bias. In Chapter 3, the conceptual framework from Chapter 2 is used to indicate which inferences are allowed, given a specific level of equivalence.

In Chapters 4 and 5, response styles are studied. Response styles are an important source of method bias. If there is no construct bias, there is no guarantee that there is no other type of bias in the scores (Van de Vijver & Leung, 1997a). Bias does not only arise from the definition of the construct, but also from the measurement instrument that is employed. A form of bias that affects scores at the level of the instrument is method bias. A major source of method bias results from response styles (e.g., Hui & Triandis, 1989). The purpose of Chapter 4 is to explore whether response styles can be identified in six countries in the European Union (EU) across three behavioral domains, for items at three levels of abstraction and for two types of rating scales. In Chapter 5, the antecedents of response style are described. These can be found in socio-demographic characteristics and behavior of individuals in countries, as well as in characteristics at the level of regions within countries.

In Chapter 6, a method to deal with differences in response styles is discussed. An alternative measurement procedure, ranking, will be introduced. This Chapter intends to provide insight into the relative appropriateness of ratings and rankings used to measure values in cross-national research. First, it is investigated whether response behavior towards ratings and rankings of values is the same across five countries in the European Union. Second, it is investigated whether specific prototypical response patterns occur, that is, patterns based on differences in response style in ratings, and whether these can be explained by differences in socio-demographic characteristics. Third, it is assessed whether these response patterns and socio-demographic characteristics affect the correlation between ratings and rankings of the nine List of Values (LOV) items. Finally, the predictive validity is studied of ratings and rankings respectively. The result is related to the response patterns that subjects have displayed. It is determined whether ranking that is insensitive to yeasaying and extreme response style is an alternative for measurement in situations where rating scales induce these response styles.

Chapter 7 concludes with an integral discussion of Chapters 2 to 6, and points out a number of limitations, and suggests issues for further research.

Chapter 2. Methodological Issues In Culture Comparative Research

2.1 Introduction

In the last few decades, international comparative marketing research has become more important. Large multinationals are entering new, unknown markets. In this process, cross-national comparative marketing research is an essential tool. It is used by companies to determine market attractiveness, and to decide which markets are most interesting. Consequently, comparisons between cultures or countries are made, either implicitly or explicitly.

In general, three kinds of studies in international marketing can be distinguished: first, studies specifically designed for one country; second, studies in which adaptations are made per country to the design and instruments, and third, studies in which essentially the same design and instruments are applied, so-called culture comparative studies[3]. If studies are specifically designed per country, questionnaires are likely to be entirely different per country. This makes direct comparisons between countries almost impossible. If studies are adapted, some questions in the instrument are specifically changed to suit a certain country. As a rule, questions that have been changed cannot be compared across countries. Thus,

[3] A special case of the cross-culture comparative study is a direct replication of a study done in one country, in other countries. In a direct replication study, no adaptations are made, except translation of the questionnaire (if necessary). It should be noted that cultural bias can be high in such studies. For instance, a concept like 'healthy eating' may refer to other behavior in the US, and in Africa. In the US the item 'I am on a low-fat diet' may be appropriate, in Africa it is not. Proper translation of the item does not make it appropriate.

if a study entails adaptations per country, cross-national comparisons are complicated. Only the part that contains the same items (excepting translations) can be compared. Only culture comparative studies are specifically designed to make comparisons across cultural populations. Based on these studies, comparisons between cultural populations can be made, provided that equivalence is established. In this chapter, we focus on culture comparative studies.

The equivalence or comparability of the data collected across countries is regarded as a basic issue in culture comparative research (e.g., Douglas & Craig, 1983; Hui & Triandis, 1985a; Sekaran, 1983; Singh, 1995; Van de Vijver & Leung, 1997a,b). Despite its importance, the equivalence of data is usually not examined (Aulakh & Kotabe, 1993) and most culture comparative studies do not address equivalence issues. One of the reasons for this may be the confusing and unclear notions used to describe equivalence. To illustrate this, an overview will be given of equivalence frameworks used in cross-cultural psychology, marketing, and business studies in this chapter.

In addition to the confusion in terminology, there appears to be a lack in unified treatment of culture comparative research methodologies in the marketing literature (Malhotra, Agarwal & Peterson, 1996). In particular, it is not clear how 'measurement equivalence' should be established. For example, some researchers (e.g., Aulakh & Kotabe, 1993; Douglas & Craig, 1983; Kotabe & Helsen, 1998) mention that standardizing the data is a way to get measurement equivalence. Other researchers (e.g., Netemeyer, Durvasula & Lichtenstein, 1991; Steenkamp & Baumgartner, 1995) state that measurement equivalence should be established through advanced statistical techniques such as confirmatory factor analysis (e.g., Jöreskog, 1971). It should be noted that standardizing data and establishing invariance of a scale are different things. Standardization focuses on the levels of the scores while confirmatory factor analysis focuses primarily on structural relationships between items. Hence, establishing 'measurement equivalence' implies different approaches and different results.

In the remainder of this chapter, we define equivalence and its relation to bias. Next, we provide insight into the types of bias that affect successive steps in the marketing research process. Then, we discuss how researchers from various disciplines approach equivalence. Consequently, these approaches will be compared. Finally, we describe how Van de Vijver and Leung's equivalence framework (1997a) can be adopted to marketing.

2.2 Equivalence

Comparability or equivalence is defined by Douglas and Craig[4] (1983, p.131) as: 'Data that have, as far as possible, the same meaning or interpretation, and the same level of accuracy, precision of measurement, or reliability in all countries and cultures'. Singh (1995) does not give a definition of equivalence, but from his article it can be deduced that he refers to constructs and measurements that are directly comparable across countries. Van de Vijver and Poortinga (1997) mention various forms of (in)equivalence; constructs and their operationalizations can be (in)equivalent, but so can measurements or measurement scales. They define equivalence as 'free from bias' (p. 21), where bias 'is used as a generic term to indicate a lack of correspondence between the observed scores of subjects from different cultural populations and the domain of generalization' (p. 21). Equivalence or inequivalence may exist in various forms; it can be in the construct itself, in the operationalization of the construct, but also in the methods, or in the measurement scales.

In all definitions, equivalence is treated as a prerequisite in international comparative research. However, the elaboration of equivalence differs from one definition to the next. In marketing (Douglas & Craig, 1983), assessing equivalence comprises content, measurement, and execution of the research. In business studies, assessing equivalence is either broadly defined (Sekaran, 1983) as in marketing, or more specifically focused on content and measurement (Singh, 1995). In psychology (Van de Vijver & Poortinga, 1997), equivalence is defined within the precise context of (psychological) measurement. It is specified at the level at which comparisons can be made; for instance, at the structural level (e.g., 'the structure of inductive reasoning is the same in two countries') or at the measurement scale level (e.g. 'person 1 in country A' is *better in* inductive reasoning than 'person 2 in country B').

2.2.1 Bias

In the cross-cultural psychological literature (Van de Vijver & Leung, 1997a; Van de Vijver & Poortinga, 1997), equivalence refers to absence of bias. In this thesis, the same definition will be used. However, we emphasize a difference in degree

[4] In the updated edition of this book (Craig & Douglas, 2000) the same definition is used.

between the terms equivalence and bias. Equivalence is established after data have been collected. It refers to the extent to which the results of a study are comparable across cultural populations. Bias refers to measurement problems that emerge in the research process, and it lowers the level of equivalence that can be reached.

In this section, three kinds of bias are discussed, namely construct bias, method bias, and item bias. These three forms of bias represent different levels of bias: construct bias refers to bias in the specific topic or domain that is being measured, method bias to bias in the measurements or the measurement scales (response style, interviewer bias) and, finally, item bias refers to bias in certain items in the instrument.

Construct bias is likely to be present if the construct being studied differs across countries. For example, if studies are done in the Western world and are subsequently replicated in non-Western countries, construct bias may emerge. A well-known example showing that the definition of constructs that were being measured should have been broadened, are Hofstede's value dimensions (Hofstede, 1980). In the Western world, four dimensions were found but when a similar study was done in Asia (The Chinese Culture Connection, 1987) an additional factor emerged, Confucian Work Dynamism, which is unrelated to any of Hofstede's other dimensions. Construct bias can also be induced by a lack of overlap in behaviors associated with the construct studied. In marketing, for example, if a product does not have the same function in different countries, people will react differently to questions about these products. The use of bicycles for recreational purposes cannot be compared with the use of bicycles for transport, as butter used for baking cannot be compared with butter used for spreading (Van Herk, Verhallen & Barzilay, 1994). Thus, construct bias occurs if the construct measured is not the same across countries and/or if the behaviors associated with the construct are non-overlapping.

Method bias[5] refers to cases where all or most items are equally affected by a factor that is independent of the construct studied. Examples are different response styles across countries or cultures, such as yeasaying or extreme response style

[5] This definition refers to uniform bias. There is also nonuniform method bias; nonuniform bias refers to influences that are not equal for all score levels (for further explanation see Van de Vijver & Leung, 1997a).

(Hui & Triandis, 1989; Van Herk & Verhallen, 1995b). Method bias can be induced by the interviewer (interviewer-interviewee interaction), the research method (telephone, mail or personal interviewing), or background characteristics of respondents, such as age or social class (Greenleaf, 1992a). Method bias is difficult to detect, since all items in a questionnaire are likely to be equally affected by method bias. As a consequence, method bias and 'real' differences in, for example, attitude cannot be separated easily.

Both construct and method bias affect the entire research instrument. Item bias, on the other hand, occurs when one or a few items deviate from the expectations about the response patterns in two cultural groups derived from other items in the instrument. Or, as cross-cultural psychologists formulate it, an item is biased if 'persons from different cultural groups with an equal ability do not have the same probability of giving a correct answer' (Van de Vijver & Leung, 1997b, p. 273). An example from marketing research can illustrate this. Suppose we employ a scale on 'health consciousness' and an item is included on 'visiting a fitness club at least once a week'. It may well be that people in both groups have the same level of 'health consciousness'. However, if the availability of health clubs is low, the answer to this item will be 'no' instead of the expected 'yes,' given their attitude towards health consciousness. The other group will answer 'yes' as expected. In such instances, we say that the item is biased. Possible causes of item bias can lie in the translation of the instrument as well as in in-appropriate item content (as in the example). Item bias typically applies to multi-item scales (see e.g., Churchill, 1979). There are many multi-item measures in academic marketing research; examples include the CETSCALE (Shimp & Sharma, 1987), the CSI scale (Garlington & Shimote, 1964), and the EAP Scale (Baumgartner & Steenkamp, 1996). In applied marketing research, such multi-item measures are almost never used due to time and financial constraints.

2.3 Bias and the research process

Bias is a major threat to the validity of cross-national and cross-cultural comparisons. When data are not equivalent, no valid cross-cultural comparisons can be made. To obtain equivalent information in the final stage of the research process (e.g., Churchill, 1991; Dillon, Madden & Firtle, 1994), attention has to be paid to sources of bias in the previous stages. In each stage of the research process, sources of bias may emerge. These sources may lie in the construct, the

operationalization of the construct, the instrument, the translation, the sampling, the fieldwork, and the coding and editing of the data. Sources and types of bias that may emerge in various stages of the research process are given in Table 2-1.

At stage I of the research process in marketing, the problem is formulated and the objectives of the study are defined. In a cross-national study, a common first check is to determine whether the issue to be studied is relevant across countries. For example, based on desk research and consultations with fellow researchers, the issue can be defined in the cross-cultural context. At this stage, the equivalence of the issues studied is still assumed, and no measurements or tests have been administered to establish equivalence. If the issue studied appears incomparable in the problem formulation stage, it must be concluded that cross-national comparisons cannot be made.

At stage II, the research design stage, decisions are made concerning the type of study, the operationalization of the constructs, item selection, and the response format. If a construct cannot be operationalized in a similar way in the countries studied, construct bias emerges at this stage. Method bias is introduced at this stage if the response format of the items is more familiar to subjects in one country. For example, in the United States, a 5-point or a 7-point rating scale is most common, whereas in France, a 20-point scale prevails (Kotabe & Helsen, 1998). Item bias is introduced, if an item to measure a construct is appropriate in one country, but inappropriate in another. The same holds if the translation of one or more items is incorrect or impossible, because for certain words there are no words with an equivalent meaning in the other language.

Table 2-1 The research process and equivalence issues

	Stages in the marketing research process	Source of bias	Issues	Prevalent types of bias
I	Problem formulation	Concepts Category Function	Purpose of the study	construct
II	Research design	Operationalization	Type of study Type of questions	construct
		Instrument design	Item selection	method
			Type of response format	item
		Translation	Questionnaire design	
III	Method of data collection	Method	Personal, mail, telephone	method
IV	Sample selection	Sampling	Target population	construct
			Sampling frame	method
				item
V	Data collection		Procedures	
		Fieldwork	Interviewer selection	method
			Time frame	
VI	Data editing and coding	Editing	Data editing	item
		Coding	Data coding	
		Calibration		
VII	Analyzing and interpreting data		Statistical procedures	

At stage III, the data collection method, method bias can be introduced in the study if people are less familiar with a particular data collection method. For example, in Western countries, computerized personal interviewing or computerized telephone interviewing (CAPI method and CATI method; e.g., Wyndham & Goosey, 1997; Malhotra & Birks, 1999) is customary, whereas it is completely unknown in other parts of the world. Less familiarity with a research method will affect results (see e.g. Serpell, 1979).

At stage IV, the sample composition and the sampling frame are determined. The definition of the sample may introduce bias. Choosing samples that are alike with respect to demographic characteristics, such as students, can help reduce bias. However, even then bias can occur: students are certainly not representative for the total population of their country. In some countries, students come from all social classes, whereas in other countries students only originate in the upper classes. Their different social background affects their attitudes and may introduce construct and item bias. Bias may also be introduced if the sampling frame (e.g., Malhotra & Birks, 1999), such as a telephone directory, does not accurately represent the population of consumers due to unlisted or disconnected numbers.

At stage V, the data collection phase, method bias can be introduced by the procedures followed, for example, instructions to interviewers can be different. Method bias can emerge during interviews if respondents are more willing to talk about sensitive issues if a special interviewer is present. For example, women may be more willing to talk about violence to female interviewers than to male interviewers. Interviewer-respondent interaction can also introduce method bias. In some countries there is courtesy bias. This type of bias is caused by the fact that norms in some societies may demand that respondents should be courteous to strangers. Hence, their responses are often not indicators of their true feelings or opinions (Kumar, 2000).

At stage VI, coding and editing (i.e., assigning answers to response categories and correcting inconsistent answers in the questionnaires) may introduce item bias. This is especially true if coding and editing are done separately in each country.

Sources of bias may emerge at all stages of the research process, and they tend to accumulate. For example, if a computerized questionnaire (e.g., CAPI method) and a 5-point rating scale are used in a study, one of them, or both, may increase method bias. Bias at stage I that has not been addressed at stage I, cannot be

corrected at a later stage. At stages I through VI, various sources of bias can be indicated.

At stage VII, the analysis phase, it is possible to assess the absence of bias by means of statistical analysis procedures. In the preceding phases, comparability of a research topic can only be investigated by employing secondary data; an alternative may be to do qualitative research in order to gain insight into the problem. However, using both secondary data and conducting a qualitative (pre-) study do not ensure that the primary data collected later are construct equivalent. This information only makes presence of construct equivalence more likely; moreover, method and item bias cannot be assessed in this way.

Absence of construct, method or item bias can only be tested at phase VII. For the preceding phases, various sources of bias were identified. Statistical analyses that can be used to assess bias are given in Table 2-2. Construct bias can be assessed through techniques that focus on the structural relationships between items. In international marketing research, construct bias is the type of bias that is assessed most often. Examples include the use of Confirmatory Factor Analysis (e.g., Netemeyer, Durvasula & Lichtenstein, 1991; Durvasula, Andrews, Lysonski & Netemeyer, 1993; Steenkamp & Baumgartner, 1995; Kumar, Scheer & Steenkamp, 1995) and the use of Multi Dimensional Scaling techniques (e.g., Schwartz & Bilsky, 1987, 1990; Schwartz, 1992). Neither method or item bias are ever addressed in international marketing research literature.

Table 2-2 Techniques to assess bias in questionnaires

Bias	Statistical technique
Construct	Principal Components Analysis (PCA) + Procrustean rotation (e.g., Van de Vijver & Leung, 1997a) Confirmatory Factor Analysis (e.g., Jöreskog, 1971; Steenkamp & Baumgartner, 1998) Multi-Dimensional Scaling (e.g., Van der Kloot, 1997)
Method	Analysis of Variance (e.g., Hays, 1973) Analysis of Covariance (e.g., Hays, 1973) t-test (e.g., Hays, 1973) Hierarchical Linear Modeling (e.g., Bryk & Raudenbush, 1992)
Item	Item Response Theory (e.g., Lord, 1980) Analysis of Variance (e.g., Hays, 1973) Mantel-Haenszel procedure (e.g., Holland & Thayer, 1988)

Once bias has been analyzed, the level of equivalence that has been achieved can be determined. For example, if there is no evidence of construct bias, construct equivalence is accepted.

2.4 The equivalence framework by Van de Vijver and Leung

As mentioned before, Van de Vijver and Leung (1997a,b) distinguish three levels of equivalence: construct equivalence, measurement unit equivalence, and scalar equivalence. All three equivalence levels refer to issues regarding measurement in cross-national research. This means that Van de Vijver and Leung focus their equivalence framework on phase VII in the research process, analyzing and interpreting data.

Construct equivalence (or structural equivalence) refers to similarity of structural psychometric properties in data from different countries (Van de Vijver & Leung, 1997a). A majority of international research focuses on establishing construct

equivalence; in these studies, statistical procedures can be used to determine invariance of factor structures or equal correlations between variables. Construct equivalence exists if equal factor structures are obtained in different cultural populations. It can then be concluded that the underlying construct is identical in these cultures. Construct equivalence is assumed to have been established if there is no construct bias. It should be noted that construct equivalence does not imply that there is no method or item bias. For example, the same factor structure is maintained if all variables are multiplied by a constant or if a constant is added to each variable. Correlations (and consequently factor loadings) remain unaffected by these transformations. The two other types of equivalence, measurement unit, and scalar equivalence can be introduced to overcome this problem.

Measurement unit equivalence refers to a situation where the measurement unit is equal across populations, but where the origin of the measurement scale is different. An example from physics is the measurement of temperature, where degrees Celsius and degrees Kelvin are measured in the same units, but where the zero (offset) differs. The functional relationship between the two scales is clear in this case, but with psychological measurements this is often not the case. Usually, the results of a test can be validly compared between subjects within countries, but between countries some form of bias may lead to a difference in the level of scores. Such a difference amounts to a difference in scale origin (Van de Vijver & Leung, 1997a). The same kind of difference in the level of scores may hold for a questionnaire to measure purchase intention in marketing. In other words, does a '4' on a 5-point scale refer to the same intention to purchase the product in different countries? This is only the case if the origin of the scale is the same for the cultural populations concerned. Thus, measurement unit equivalence is assumed if there is no construct bias and no method bias.

In the case of measurement unit equivalence, differences in patterns of scores (more measurements on the same variable, or measurements under different conditions) can be compared within and across countries. However, scores on a single variable can only be compared within countries. Thus, a purchase intention score can be compared within individuals (intention to purchase product A or B) or across individuals within the same country.

Scalar equivalence or full-score equivalence exists if the scale also has an equal origin across countries. Then, purchase intention scores can also be compared directly between individuals from different countries. For example, it may be

concluded that 'a person i from country K with a score of '4' on the scale refers to the same probability of buying as person j from country L with a score of '4' on the scale'. Scalar equivalence is the highest level of equivalence. Scalar equivalence is established if there is no evidence of construct bias, method bias, and item bias[6].

2.5 Other equivalence frameworks

In Van de Vijver and Leung's equivalence framework (1997a), measurement issues are emphasized. Other frameworks used in cross-cultural psychology in the past, in marketing, and in international business studies, include other issues in addition to measurement.

Of these frameworks, Van de Vijver and Poortinga's framework (1982) is most closely related to that of Van de Vijver and Leung (1997b). The other frameworks, originating in marketing (e.g., Douglas & Craig, 1983; Craig & Douglas, 2000) and in business research (e.g., Sekaran, 1983), are more broadly defined than the Van de Vijver and Leung framework. They also include issues on the research topic, the research design, and/or sampling. To provide insight into the various approaches of equivalence in the different frameworks, they are described below.

2.5.1 Van de Vijver and Poortinga's framework (1982)

The focus of the Van de Vijver and Poortinga framework (1982) is on conditions for the universality of concepts and the validity of cross-cultural differences in psychological data. They define four categories of universals: conceptual universals, functionally equivalent universals, metrically equivalent universals, and scalar-equivalent universals. Conceptual universals refer to theoretical concepts at a high level of abstraction. It should be noted that they avoid the term equivalence with respect to the concept 'conceptual universals', stressing that 'universality of concepts cannot be refuted in (quasi-) experimental studies, since no empirical referents are supplied' (Van de Vijver & Poortinga, 1982, p. 389). They consider it impossible to empirically establish equivalence of concepts, unless 'it is indicated

[6] It should be noted here that this holds for uniform item and uniform method bias. If there is nonuniform method or item bias, the analysis of bias becomes more complicated (cf. Van de Vijver & Leung, 1997a).

which observable variables are considered relevant and how these should be measured' (Van de Vijver & Poortinga, 1982, p. 389). The category labeled 'conceptual universals' is missing in the Van de Vijver and Leung framework, because conceptualization is excluded. In their framework, Van de Vijver and Leung focus on equivalence of measurement of concepts that have already been operationalized.

Functionally equivalent universals (or weak universals) 'contain concepts for which empirical referents have been specified and for which construct validity has been demonstrated' (Van de Vijver & Poortinga, 1982, p. 389). Construct validity (Cronbach & Meehl, 1955) of psychological measures can be established by means of psychometric analysis techniques such as factor analysis and covariance structure models. Functional equivalence as defined here corresponds to construct equivalence in the Van de Vijver and Leung framework.

Metric equivalent universals (or strong universals) include concepts that are measured 'in the same metric' across countries, but the scales may have a different origin in different countries. Metric equivalence as it is defined here is the same as measurement unit equivalence in Van de Vijver and Leung's framework .

Scalar equivalent universals (or strict universals) have an equal metric and the same origin across all countries studied; hence this concept has the same meaning as in the Van de Vijver and Leung framework.

2.5.2 The framework for marketing by Douglas and Craig (1983)

Douglas and Craig's framework (1983) is more broadly defined than that of Van de Vijver and Leung. It incorporates measurement issues, but also topics studied, elements of the research process (e.g., translation), as well as the research environment. As such, Douglas and Craig's framework (1983) is a collection of issues that should be taken into account to obtain data that can be compared across countries. Douglas and Craig consider three forms of equivalence, construct, measurement, and sampling equivalence, which are essential to establishing comparability in various aspects of research design and implementation.

Douglas and Craig (1983) define three aspects of construct equivalence:

1) *Conceptual* equivalence is 'concerned with the interpretation that individuals place on objects, stimuli or behavior, and whether these exist or are expressed in similar ways in different countries and cultures' (Douglas & Craig, 1983, p. 138).

2) *Categorical* equivalence 'relates to the category in which objects or other stimuli are placed' (Douglas & Craig, 1983, p. 139). Categorical equivalence refers to comparability in product class definitions, and in background or socio-demographic classes that exist between countries.

3) *Functional* equivalence relates to whether the concepts, objects or behaviors studied have the same role or function in all countries studied. Douglas and Craig have taken this notion from Berry (1969).

Douglas and Craig consider examination of equivalence a two-step procedure: 'once construct equivalence has been examined, the next step is to consider measurement equivalence' (Douglas & Craig, 1983, p. 140). They consider construct and measurement equivalence as highly interrelated insofar as the measure is an operational definition of the construct. They distinguish three aspects of measurement equivalence:

1) *Translation* equivalence refers to the translation of the research instrument into another language so that it can be understood by respondents in different countries, and has the same meaning in each research context. They emphasize translation of verbal as well as nonverbal stimuli. Douglas and Craig (1983) consider translation equivalence 'a central issue in the establishment of construct validity, since this is the stage in the research design at which the construct is defined in operational terms' (Douglas & Craig, 1983, p.141).

2) *Calibration* equivalence refers to equivalence 'not only with regard to monetary units and measures of weight, distance and volume, but also other perceptual cues, such as color, shape, or form, which are used to interpret visual stimuli' (Douglas & Craig, 1983, p.140). Douglas and Craig, in discussing calibration equivalence, also refer to 'product grading, product quality and safety regulations' (Douglas & Craig, 1983, p. 140) that should be similar to attain this form of equivalence.

3) *Metric* equivalence refers to 'the scoring or scalar equivalence of the measure used' (Douglas & Craig, 1983, p. 142). They consider two aspects important in determining metric equivalence, (1) the specific scale or scoring procedure used to establish the measure and (2) the equivalence of responses to a given item or scale in different countries. Issues involved are familiarity with rating scales and scoring procedures used, and the meaning and interpretation of a score in another context. Douglas and Craig state that metric equivalence, in contrast to the other types of equivalence, can be assessed only after the data have been collected.

Douglas and Craig (1983) do not explicitly define sampling equivalence, but they mention two issues on which sampling equivalence should focus:

1) The *relevant respondent(s)* in the household or organization to be sampled. This 'relevant' respondent can be different regarding background characteristics. For example, the decision maker regarding a dessert that is bought can be the child in one country (US), and the parent in another (France).

2) *Sample representativity*. In their book, they mention the difficulties encountered in determining a proper sampling frame. Due to inaccurate or absent census data in developing countries it is (almost) impossible to obtain a representative sample. Apart from representativity, they mention the problem of generalizing results of one sample to other samples within the country, or to samples in other countries.

2.5.3 Frameworks in business research

In business studies, as in cross-cultural psychology and marketing, equivalence of research measures is considered an important topic. In various publications (Sekaran, 1983; Mullen, 1995; Singh 1995), equivalence issues are described extensively. According to Sekaran (1983), equivalence is linked to various stages in the research process. She mentions equivalence in function, instrumentation, data-collection methods, sampling design, and data-analysis. As in marketing, *functional equivalence* is associated with the behavior that people from different countries display, given comparable choices in situations. For example, "what

mode of transport do people choose if they have to travel a distance of about 100 kilometers?" Functional equivalence is often assumed in business research (Sekaran, 1983). *Instrumentation equivalence* includes equivalence in translation, reactions to scales, and existing bias, such as courtesy bias. Data collection not only comprises the method used, but also timing, interviewer status, and type of research (longitudinal or cross-sectional). All these aspects may lead to bias. Sampling equivalence covers issues such as representativity, matching of samples and whether inferences may be made to other cultural populations. The terminology used by Sekaran (1983) covers the whole research process, but it is not very precise and has no clearly defined meaning: construct and measurement unit equivalence are dealt with in several parts of her framework. For example, measurement unit equivalence is dealt with in instrumentation (e.g., courtesy bias) as well as in data collection (interviewer bias). Scalar equivalence in the sense of Van de Vijver and Leung (1997a) is not included in this framework.

In later articles on equivalence in business research (Mullen, 1995; Singh, 1995) other concepts to address equivalence are used. Mullen (1995) adopts the (overlapping) aspects of measurement equivalence introduced by Douglas and Craig (1983): translation, calibration, and metric equivalence. In his 1995 article, Singh construes another framework regarding measurement issues in international business research. One issue he extensively covers is construct equivalence. In his definition, construct equivalence involves three main aspects: (1) functional and conceptual equivalence, (2) instrument equivalence, and (3) measurement equivalence. In addition, he introduces (4) model equivalence. Thus, equivalence according to Singh (1995) comprises:

1) *Functional* and *conceptual* equivalence. They examine whether a given construct serves the same function and is expressed similarly (in terms of attitudes or behaviors) in different countries. Functional equivalence requires that 'the focal concept serves the same function in different nations' and conceptual equivalence that 'the concept or construct is expressed in similar attitudes or behaviors across nations' (Singh, 1995, p. 605).

2) *Instrument* equivalence explores whether the scale items and other stimuli (e.g., instructions) are interpreted similarly in different countries.

3) *Measurement* equivalence examines whether each scale item measures the underlying construct in the same way. Measurement equivalence is established if 'the factor loadings and error variances are identical for each scale item' (Singh, 1995, p. 605). These definitions imply that measurement equivalence corresponds to construct equivalence as defined by Van de Vijver and Leung (1997a).

4) *Model* equivalence is defined separately as a way to test for equivalence in cross-national data sets (Singh, 1995: p. 606). Specifically it incorporates: (1) estimation of significant differences between regression coefficients in the datasets, (2) adjustments of unequal reliability for all measurements involved, (3) multivariate and bivariate comparisons with some control over error rate, such as a multivariate test followed by bivariate testing, and (4) possibilities for testing construct equivalence by testing and comparing coefficients in path models, factor models and structural equation models.

As defined by Singh (1995), model equivalence is a combination of external validity and measurement unit equivalence (as defined by Van de Vijver & Leung, 1997a). It can be concluded that construct and measurement unit equivalence (Van de Vijver & Leung, 1997a) is included in the equivalence framework by Singh (1995). Scalar equivalence which allows for comparisons across individuals within as well as between countries is not included in this framework.

2.6 Comparisons between frameworks

The previous sections have shown that overlapping equivalence frameworks exist in cross-cultural psychology, marketing and business research. Although they deal with the same problem, that is, obtaining information that can be compared across countries or cultural groups, the terminology used differs as well as the central concern of each classification. Van de Vijver and Leung (1997b) and Van de Vijver and Poortinga (1982) focus on measurement issues; Douglas and Craig (1983) and Sekaran (1983) focus on the research process, and Singh (1995) focuses on construct and measurement. The meaning of terms differs despite the use of the same words. Functional equivalence in Van de Vijver and Poortinga (1982) comes fairly close to construct equivalence (referring to abstract constructs such as skills or traits); functional equivalence in the other frameworks refers to actual behaviors or objects having the same function or role (referring to concrete

actions). Instrument equivalence refers to vocabulary and concepts in Sekaran, while instrument equivalence in Singh refers to a similar interpretation of the scale items and the interview instructions. Another remarkable difference in terminology is found in metric equivalence. In the Douglas and Craig framework, it is very broad including, for example, reactions to the instrument, scale use, and meaning of scores. Metric equivalence in Van de Vijver and Poortinga refers to a property of the score variables as a measurement scale. It may be concluded that the various definitions used in literature are at least confusing and vague.

In Table 2-3, the frameworks are compared on the basis of three aspects, namely (1) topic of research, (2) measurement, and (3) sampling. In the next sections, these aspects will be described.

2.6.1 Topic of research

The topic of research refers to the concepts used, to issues of research design, and to research objects, such as products used, the product category they belong to, and the function these products have in a certain society. Especially in marketing and in international business studies, several forms of equivalence have been defined under this heading.

Conceptual equivalence is included in most frameworks. The term also refers to similar issues. *Functional equivalence* is included in the frameworks by Douglas and Craig (1983), Sekaran (1983), Van de Vijver and Poortinga (1982). However, the meaning of functional equivalence differs in each. In marketing and business research, functional equivalence refers to the function a product has in a certain country. For example, a bicycle can be used for transport or for recreation. In cross-cultural psychological research, functional equivalence refers to structural equivalence, for example, a measurement characteristic that can be tested across countries or cultures studied. *Categorical equivalence* is only mentioned in the equivalence framework by Douglas and Craig (1983). It refers to the specific product category a product belongs to. For example, pastry can be part of the dessert category, or it can belong to the snack category. Knowing which category a product belongs to is essential in applied marketing research, because construct equivalence is hard to obtain if people refer to different matters in answering survey questions.

Table 2-3 Equivalence in cross-national research

	Van de Vijver and Leung (1997)	Van de Vijver and Poortinga (1982)	Douglas and Craig (1983)	Sekaran (1983)	Singh (1995)
TOPIC OF RESEARCH		Conceptual *universals*	Conceptual Functional Categorical	Functional Instrument (e.g., vocabulary, conceptual, inferences) Data collection equivalence (e.g., response, timing)	Conceptual Functional Instrument
MEASUREMENT	**Construct equivalence**	Functional	Translation Calibration	Equivalence in scaling	Measurement
	Measurement unit equivalence	Metric	Metric	Equivalence in data analysis	Model
	Scalar equivalence	Scalar			
SAMPLING			Sampling	Sampling	

It should be emphasized that equivalence with respect to category, function, and calibration cannot be tested after the data have been collected, as is possible for construct, measurement unit, or scalar equivalence. For example, equivalence of the product category has to be investigated in qualitative studies before a quantitative study is initiated. Letting people assign products they consider alike to groups, or to ask them what products belong to the product category 'desserts' or 'vegetables' can achieve this. In this way, the appropriate category can be determined and non-equivalence can be investigated. If not explicitly tested in a study, it can only be assumed that products fall within the same product category.

2.6.2 Measurement

Equivalence in measurement includes translation, data administration, and data analysis in the frameworks in marketing and in international business studies. As such, measurement equivalence is a collection of aspects that should be taken into account before valid comparisons across countries can be made.

Translation equivalence is included in the frameworks by Douglas and Craig (1983) and Sekaran (1983). Douglas and Craig consider translation separately; Sekaran mentions translation in obtaining instrument equivalence. In the other equivalence frameworks, translation is not included as such. In Van de Vijver and Leung (1997b), translation is part of the research design and method. To them, translation is a prerequisite to obtaining equivalence in constructs measured, but it is not a formal aspect in their framework. Equivalence and translation are closely related: if a construct is properly operationalized and the items are properly translated, the probability of obtaining equivalent results increases.

Calibration in the Douglas and Craig framework refers to (1) aspects for which the same measurement unit can be obtained; examples are currencies, weights, and lengths, or (2) aspects that never are the same, but should be taken into account in the research design. One example is the color traditionally worn by women at weddings in different countries; in China red is used, whereas the traditional color for the wedding dress in Western countries is white. In the former case, calibration can be dealt with in the equivalence framework by converting measurements into the same units (e.g., all currencies in dollars). In this way, scalar equivalence in the sense of Van de Vijver and Leung can be established. In calibrating colors, different colors can be selected which have a similar meaning in the various countries. We can thus assign the same scores to formally different entities.

Obviously, it is desirable that researchers provide explicit evidence of the similarity of meaning in such cases.

The terminology used for equivalence in measurement or 'metric' is rather confusing. In Poortinga (1989), metric equivalence refers to measurement unit equivalence in the sense of Van de Vijver and Leung (1997a). In Douglas and Craig (1983), metric equivalence comprises structural equivalence as well as measurement unit equivalence. Aulakh and Kotabe (1993) state that 'metric equivalence can be taken care of by standardizing, ipsatizing, or normalizing the data' (p. 20). This suggests making the mean scores on items equal; so, it refers to measurement unit equivalence In later research in marketing (e.g., Dawar & Parker, 1994), metric equivalence is interpreted as 'structural equivalence' in the sense of Van de Vijver and Leung (1997a). In Singh (1995), measurement equivalence overlaps with 'construct equivalence' and the addition in his framework of 'model equivalence' addresses issues of external validity of the model. As such, this 'model equivalence' cannot be directly related to the Van de Vijver and Leung framework. In conclusion, the terminology used in the literature to address measurement equivalence does not lead to an unified approach to establish measurement equivalence.

2.6.3 Sampling

Sampling is included in equivalence frameworks by Douglas and Craig (1983) and Sekaran (1983); in the other frameworks, it is absent. In Van de Vijver and Leung (1997a), sampling is defined separately as part of the research process; it has consequences for inferences that can be made based on the data collected. In most research designs in cross-national research, samples are made comparable at the start of the study by, for example, matching respondents to background characteristics. Examples are student samples (Lee & Green, 1991; Durvasula, Andrews, Lysonski & Netemeyer, 1993; Dawar & Parker, 1994), samples of employees working for the same firm (Hofstede, 1980), or samples of people that have the same profession (teachers or students: Schwartz & Bilsky, 1987, 1990; Schwartz, 1992; automobile dealers: Kumar, Scheer & Steenkamp, 1995). By choosing samples that are alike, the interpretation of results is less equivocal, because the differences found are less likely to reflect sample characteristics. For example, if student samples are chosen, it is assumed that these subjects are similar in terms of age, income, education level, and social class.

2.7 A new framework in marketing

The terminology on equivalence is confusing; some researchers focus on the whole research process, while others limit themselves to analysis and interpretation of data. The confusion may have contributed to the fact that few researchers in international marketing address the issue (Aulakh & Kotabe, 1993). The state-of-the-art review in cross-cultural marketing research provided by Malhotra, Agarwal, and Peterson (1996) does not reduce this confusion; construct equivalence in their framework includes not only aspects such as conceptual equivalence, but includes measurement equivalence as well. To avoid confusion, it is proposed here to adopt the Van de Vijver and Leung (1997a) framework in marketing. In this framework, the term 'equivalence' is reserved for those issues that can be assessed through statistical analysis.

The framework focuses on measurement issues; issues of content or design of the research process are separated from their framework. The framework considers design aspects as procedures that, when properly addressed, can help reduce bias or enhance equivalence. A large body of knowledge is available in the literature that helps to reduce bias in several stages of the international research process. Guidelines for translation (Brislin, 1986; Van de Vijver & Hambleton, 1996) and guidelines for test administration, documentation, and score interpretation (Van de Vijver & Leung, 1997a) are available in the literature. These guidelines have been formulated for psychological and educational instruments, but they are equally applicable in marketing and business research. Van de Vijver and Leung (1997a) consider all aspects that may lower equivalence as sources of potential bias. Possible sources of bias exist at each stage of the research process: in item selection, data collection, sampling, et cetera, Van de Vijver and Leung suggest including variables that may explain bias in a questionnaire. These variables can be used later in the data analysis phase to help explain cross-cultural differences. For instance, in marketing research, these can be variables that describe the product category or the usage situation of a product, or demographic characteristics such as income or social class of respondents. All these variables can have a profound influence on reactions to questions on products.

Particularly in marketing, elements mentioned by Douglas and Craig (1983), like product category and product functionality, are essential in understanding consumer responses across countries. The category a product belongs to, or the situation in which it is used, explain product evaluation to a large extent (see for

example Eimers, Pieters & Verhallen, 1995, on product evaluation in Italy). In an international context, the category as well as the usage situation may differ and may explain differences in product evaluations, attitudes, and opinions. Variables at the country level may explain differences and similarities in responses given by people from different countries. Especially variables that need to be measured at the individual level should be included in the study. If relevant data are not part of the data collection, it is impossible to add them afterwards. Adding variables afterwards is only possible at an aggregate level; for example, 'GNP per capita' can be added as an explanatory variable afterwards. Adding context variables to a questionnaire enables us to determine sources of bias in cross-national marketing research.

Two characteristics make the Van de Vijver and Leung framework particularly valuable:

1) it is a clear framework that focuses on measurement issues only

2) it explicitly excludes sources of bias from the equivalence framework as such; sources of bias are considered separately as variables that can be used to explain findings of cross-cultural differences

The adoption of this framework in marketing eliminates the confusion in terminology regarding equivalence that has existed for so long. In this framework, equivalence is re-defined as that part of the cross-national research process that can be controlled by means of statistical analysis procedures. In this re-definition, issues in the design stage of a study, such as allocation of products to a product category, translation of questionnaires, and sampling do not belong to equivalence. At the design stage, measures can be taken to increase the probability that data are equivalent. However, evidence of equivalence can only be obtained by statistical analysis procedures after data have been collected.

In international marketing, like in cross-cultural psychology, equivalence in constructs and measures is important. In both scientific fields, scales with known psychometric properties are seldom available; scales thus have to be created and validated in a cross-national context. In marketing, the context people live in is also important in explaining consumer behavior. Also, various specific sources of bias may emerge, such as, for example, different interviewing practices, brands, and product categories. It should be noted that the use of different interviewing

methods, to mention one example, can be detrimental. If telephone interviewing is used in one country and personal interviewing is used in another, any observed difference between these countries can be caused by this method effect. Due to differences at a certain stage of the marketing research process, achieving equivalence may become impossible. Therefore, researchers should aim at an equal research design, an equal method of data collection, and equal instruments, if comparisons between countries are required.

2.8 Conclusion

Douglas and Craig's equivalence framework (1983) covers the whole marketing research process, including operationalization of constructs, measurement procedures, and sampling. Attention is also paid to the influence context (e.g., different product categories) may have on research on different cultural populations. The aspects covered by Douglas and Craig are essential in marketing. However, linking equivalence to various aspects, such as the product category, the concepts used, and the measurement scales, makes their framework confusing. It is not clear what can be assumed to be equivalent and what should be explicitly tested for equivalence.

In the Van de Vijver and Leung framework, equivalence is a quality of data. Differences in the international environment, such as different product categories or a different functionality of products, are excluded from their equivalence framework. In the framework by Van de Vijver and Leung, equivalence issues are described more systematically than in the framework that is more commonly used in marketing (Douglas and Craig). In Van de Vijver and Leung's framework, there are only three types of equivalence: construct, measurement unit, and scalar equivalence. These types are nested; this means that construct equivalence is a prerequisite for measurement unit equivalence and that both construct and measurement unit equivalence are prerequisites for scalar equivalence.

Equivalence can be established at each of the three aforementioned levels. Factors that may introduce bias in the data should be included in the research design if findings are to be validated and explained afterwards. Thus, in cross-national research, two steps need special attention. Firstly, every variable that may be a cause of bias in the research process should be included in the design, so that bias can be assessed afterwards. This implies detailed registration of several variables,

such as characteristics of the interviewer, the time of day, the interviewee, and his/her background variables (e.g., age in years; income in thousands of Euro). Secondly, all three forms of bias, construct, method and item bias, should be assessed. Valid comparisons can be made, or valid explanations can be given in a cross-national context, only if these two steps are taken. In most current studies, only construct bias, if at all, is assessed. The framework by Van de Vijver and Leung (1997a), placed within the marketing research process, as presented in this chapter, may help researchers establish the degree to which their results can be compared across cultural populations.

Chapter 3. Inferences in Culture Comparative Research

3.1 Introduction

On the basis of samples of subjects in several countries, marketing researchers make inferences about entire populations in those countries. Measures of product evaluation are used to determine consumers' perception of product characteristics, but also to forecast sales and profits associated with the product (Dillon, Madden & Firtle, 1994; Moore & Pessemier, 1993). Hence, marketing researchers generalize their findings to other subjects in the populations than those studied, and to broader behavioral domains.

In all culture comparative studies, such generalizations are prone to bias (Malpass & Poortinga, 1986). If comparisons between several cultural populations are to be made, equivalence issues are important (e.g., Mullen, 1995; Van de Vijver & Leung, 1997a,b; Craig & Douglas, 2000; Kumar, 2000; Chapter 2 this thesis). There are measures that can be considered comparable across cultural populations; for instance, the comparability of the body height of subjects in different populations is not open to debate. However, for most measures in marketing, it is not clear whether the meaning and implications are the same. Consequently, their comparability is uncertain. For example, it cannot be taken for granted that the way of olive oil is used for day-to-day cooking or as a luxury product is the same across countries in the European Union. Moreover, if abstract topics such as innovativeness are measured, establishing the level of equivalence is necessary before inferences on the similarity or difference of this topic across cultural populations can be made.

The extent to which valid inferences can be made depends on the cultural populations and the topics studied (Poortinga & Malpass, 1986; Berry, Poortinga, Segall & Dasen, 1992). In addition, it depends on the level of equivalence that can be established (see Chapter 2 this thesis). In this chapter, the inferences will be discussed that are allowed, given the level of equivalence of measurements. For this purpose, the notions population and universe (behavioral domain) are further described in the next section. Next, a classification in universes suggested in the literature (Poortinga & Malpass, 1986; Berry, Poortinga, Segall & Dasen, 1992) is extended to marketing. The crucial role of bias in making inferences in culture comparative research is emphasized.

3.2 Populations and universes

In almost all studies in a national as well as in a multi-country setting, data are either implicitly or explicitly gathered as a basis for making statements about other behaviors in addition to those observed, and to other persons than the respondents. However, the specific research design and sample determine whether inferences are possible about 1) other people and 2) broader domains of behavior. Cronbach, Gleser, Nanda, and Rajaratnam (1972) make this distinction by using the terms population and universe. They use the term population exclusively for people and the term universe exclusively for research domains and the items used to measure them. In this chapter, we adopt the same terminology.

3.2.1 A classification of populations

In culture comparative research, sampling is a two step procedure. The first step concerns the sampling of the cultural populations to be studied; the second step concerns the subjects to be sampled within these cultural populations. In general, three sampling procedures can be distinguished, namely convenience sampling, systematic sampling, and random sampling (Van de Vijver & Leung, 1997a). Each of these three procedures applies to the selection of cultural populations as well as subjects.

A majority of cultural comparative studies uses convenience sampling. Usually, countries are taken as proxies for various cultures. Specific countries are not chosen on the basis of theoretical considerations. An example is the kind of study where several countries in the world are compared using international students

enrolled in a particular university at the time a research project was conducted (e.g., Dawar & Parker, 1994). Other examples of convenience sampling are studies where countries are included because the researchers had collaborators in those countries (e.g., Lee & Green, 1991). If the choice of cultural populations is guided by theoretical principles, we speak about systematic sampling. Such studies, to the best of our knowledge, are absent in the marketing literature.

In a few studies, a large number of countries was chosen. This is called random sampling. Their main purpose is to test the universality of theories. This procedure is rare in culture comparative research due to practical constraints such as cost and lack of collaborators within countries. Nearest to random sampling are the studies by Hofstede (1980) and Schwartz (1992, 1994). Schwartz used samples from about 40 countries in his studies on the structure of human values, and Hofstede used 53 countries in his study on work-related values. As yet, in both the data sets collected by Hofstede and Schwartz, countries in Africa are rarely involved. The Hofstede study included countries in which subsidiaries of a large multinational corporation (IBM) were located (Hofstede, 1980). In the studies by Schwartz (1992, 1994) countries were included in which he could find collaborators.

After selecting countries, the second step is the selection of subjects within each country. A true random sample or a stratified random sample of each population can be drawn (Van de Vijver & Leung, 1997a). However, studies including a random sample from the whole population are scarce in academic comparative marketing research. Most comparative studies employ student samples (e.g., Ng et al., 1982; Hofstede & Bond, 1984; Parameswaran & Yaprak, 1987; Lee & Green 1991; Andrews, Durvasula & Netemeyer, 1994; Chen, Lee & Stevenson, 1995). A notable exception is a study by Ter Hofstede, Steenkamp, and Wedel (1999), who used a sample of European consumers from each of twelve countries in the European Union.

3.2.2 A classification of universes

The term universe is used for research domains. In this section, we will first describe research domains and then go into measurement of concepts in marketing.

3.2.2.1 Research domains

In cross-cultural psychology, a research domain can be defined as either identical or non-identical across cultural populations (Poortinga & Malpass, 1986; Berry, Poortinga, Segall & Dasen, 1992). They consider domains such as physical scales and shared educational domains (e.g., arithmetic) as identical. There is non-identity in the case of non-shared domains (e.g., local flora). We adopt this distinction. However, it must be emphasized that the distinction between identical and non-identical universes is a gradual one. This gradualness is connected to the extent to which a domain can be fully described. First, the extremes, the identical, and the non-identical universes are presented. Then, a middle category between those extremes is outlined.

In *identical universes,* measuresare mainly on physical dimensions, often consisting of one or a few items. One example is the audible sound level of pure tones (defined in Hz), where sound level is expressed on a scale in dB. This universe can be completely described as all measurable sound levels can be listed. This universe is identical across cultural populations. An example of a *non-identical universe* is the range of food products in Africa and Europe. The local food market in Africa and the local food market in the Western World are not similar. People from these continents do not share (too) many elements in this domain. This universe is not identical across cultural populations.

There is a middle category, between those extremes, for which the extent to which they are identical has to be assessed. Assessing identity is easier as it is possible to completely describe or list the elements in a universe.

In psychology, a universe that can be fully described is, for example, 'addition of one-digit numbers'. In (applied) marketing research, many universes can, in principle, be fully described. If a list can be made, the extent to which universes are identical across cultural populations can be investigated. For instance, if brand ownership of cars is the topic of a study, a list can be made of the cars for sale in each country. For this, a researcher may use secondary data on car sales in which brand market shares for all brands in a country are provided[7]. Thereafter, the researcher can determine which car brands are available in all countries studied.

---

[7] World Car Industry Forecast Report, a publication of the Global Automotive Group; Standard & Poor's DRI, Published by McGraw-Hill.

Other domains cannot be completely listed. For instance, domains like product evaluation (e.g., Moskowitz, 1983) refer to a large number of behaviors and opinions. Focus groups or qualitative interviews may help researchers to determine what characteristics determine the product's evaluation in several countries. Thus, it is possible to create a description of the domain and to make a set of items that can be considered to represent adequately this domain. This does not hold for any domain. There are domains that refer to abstract characteristics, such as values (e.g., Rokeach, 1973; Schwartz, 1992, 1994). In such domains, making a list of all relevant elements in the domain is impossible. Elements are selected that supposedly capture the essence of values or traits (Berry, Poortinga, Segall & Dasen, 1992). Hence, there is no single unequivocal operationalization of the construct that holds across all cultural populations studied.

3.2.2.2 Measurement of concepts

The extent to which universes can be described was addressed in the previous section. In this section, the focus is on measurement of concepts. For some concepts, measures are physical, such as measures of length, weight, or pitch. These concepts can be operationally defined. Such measures are an exception in marketing, but they exist. For instance, the amount of olive oil used over a time period can be measured in grams.

Most measures in marketing cannot be operationally defined. Instead, a set of items is used as an approximate representation of the domain of interest. Some items may be relatively concrete and refer to something like the texture of a pudding (Moskowitz, 1983). For instance, subjects are asked to evaluate the product on characteristics, employing items such as 'smooth' or 'watery'. An entire measurement instrument consists of some twenty items that together are a representative sample of the universe 'product evaluation'. Hypothetical constructs are less concrete than product characteristics. But they are very frequently used in marketing. Examples include personality variables (Kale, 1995b; Aaker, 1999) and human values (Kahle, 1983; Schwartz & Bilsky, 1987, 1990; Grunert & Scherhorn, 1990; Kamakura & Mazzon, 1991; Schwartz, 1992). Values and personality traits are abstract and usually operationalized in short sentences; each value is represented by one item. Other hypothetical constructs that are represented by multi-item measures include scales to measure ethnocentrism (Shimp & Sharma, 1987), or consumer innovativeness (Baumgartner & Steenkamp, 1996). In

short, to measure a hypothetical construct, a scale is developed that captures the essence of the construct.

3.3 Levels of inferences

3.3.1 Inferences about larger populations

In the selection of a population a fundamental dilemma exists between precision and generalizability (see also Blalock, 1982). Precision in culture comparative research can be attained if (1) homogeneous samples are used, (2) the social settings the subjects are placed in are simple and similar, and (3) the tasks given to them are simple and can be repeated so that reliable results can be obtained. A typical example of a precise measurement is a laboratory situation where all or most variables that might affect the measures are controlled. It is obvious that a laboratory situation is not representative for situations in real life. Thus, the more precise a measure becomes, the less generalizable it will be. Of course, even for simple variables measured in a laboratory situation, there will be bias in the measurements due to imperfections in the sampling design, the tester - testee interaction and recording of responses. However, such bias will be small compared to bias in field surveys. In culture comparative research in marketing, studies in a laboratory setting are, as far as we know, lacking. In field studies where several sources of bias may contaminate the data, attention for bias is crucial to prevent alternative explanations of results.

To minimize bias in the data a majority of cultural comparative studies uses homogeneous samples, often students. This makes an alternative explanation of results due to sample-specific differences less likely. However, students are by no means an adequate reflection of the population in their country. Thus, inferences about other subjects than students cannot be made. A sample that is representative of the population in the countries studied is seldom used in academic research; in applied research, such samples are more common. After all, for making inferences about the entire population in a country, a representative sample is necessary.

3.3.2 Inferences about universes

Researchers make inferences about a certain universe or behavioral domain, but usually the boundaries or limitations of these inferences are not clear. In general, the closer an inference is to the data, the more likely it is to be valid. An example from psychology where the interpretation is close to the data is a simple psychological test for children on adding pairs of one-digit numbers (addition of one-digit numbers is the universe). From results on the test in several countries, it may be inferred that children are skilled at this task or lack arithmetical skills. However, it may not be inferred that children in a country with low scores lack the ability to make additions or even, that they are less intelligent. This inference exceeds the reach of the measurement instrument. It is far more likely that the differences in test scores are caused by differences in schooling. Such differences in schooling introduce bias and thus lower the comparability of inferences about psychological traits of persons. In this example, an inference about skill is valid; an inference about ability is only possible if alternative explanations of results are ruled out.

Three levels of inference, from low to high, are distinguished in the psychological literature (Poortinga, 1989; Berry, Poortinga, Segall & Dasen, 1992). A low level inference remains close to the measurement domain, for example, 'aided brand awareness of cars' (a list of all brands for sale in countries can be made). The instrument is representative of the domain; the content validity is high (Carmines & Zeller, 1979), and the extent to which the measurement reflects the universe is high. A medium level refers to inferences about universes that refer to abstract unobservable psychological traits for which construct validity has been established. At this level, the extent to which the instrument reflects the universe has to be investigated. With high level inferences, the domain of generalization is virtually unrestricted. Poortinga (1989) refers to high level when differences are interpreted post hoc in terms of some cultural characteristic that was not included in the research design.

In Table 3-1, it is outlined whether inferences are possible, given (1) the extent to which universes can be completely described (see Section 3.2.2), (2) the way of operationalization of the concepts, and (3) bias that may emerge in measuring the concepts.

No inference. For operationally defined measures such as 'olive oil used in one week (in centiliters)', there is no inference. The measurement is equal to the universe. It is unlikely that construct bias will occur (this is indicated by the '-' in Table 3-1). In this case, there may be method bias, e.g., there might have been inaccuracy in reading the measuring glass. However, this type of method bias is, in principle, controllable (this is indicated by the '(+)' in Table 3-1). Item bias is non-existent as only one variable is measured, that is non-inferential. At this level, valid cross-cultural comparisons can be made. For instance, one might compare the amount of olive oil used in Greece and Italy and conclude that Greek subjects use twice the amount Italian subjects do.

Table 3-1 Bias and inference levels in cross-cultural comparisons

	INFERENCE LEVEL	*No*	*Low*	*Medium*	*High*
	Elements can be listed?	YES	YES	NO	NO
	CONCEPTS	operational	non-operational	non-operational	non-operational
BIAS	*Construct*	-	(+)	+	?
	Method	(+)	+	+	?
	Item	(+)	+	+	?

Low level inference. A low level inference can be made if the elements in a universe can be fully described or listed. The measurement instrument consists of items that more or less cover the whole domain. The instrument is a sample from all possible elements in a universe. For example, the elements can be listed in a study on 'aided brand awareness of cars'. Insight can be given into the elements in this universe (car brands that are available in the countries studied). It is possible to use secondary information about brand market shares in the countries studied (e.g., information from the Global Automotive Group)[8]. Measuring 'aided brand awareness is direct, one item per brand may suffice (Rossiter, 1999).

[8] It should be noted that in the description of a universe 'cars available in a country' there is no inference; the measurement is equal to the concept.

A researcher may ask whether subjects know a car brand, for example 'Do you know Fiat?'. However, even if the elements in an universe can be listed, such measures might be prone to all three types of bias (see also Chapter 2). At the same time, it is possible to reduce construct bias by using secondary information. Measurement may also be affected by item and method bias. Method bias may be caused by interviewers who help interviewees by providing additional information or repeating the item, whereas other interviewers do not. Item bias may emerge if interviewees in a particular cultural population do not understand a question.

Other examples of a universe at the low inference level are topics such as margarine use and the situations it is used in. In this case, as in the previous example, a list can be made of the margarine brands for sale in the countries studied. In addition, a qualitative pre-study can be held in which subjects are asked in which situations they use the product. Thus, there are possibilities to control for the influence of construct bias by performing a pre-study. However, method bias may occur. For instance, subjects may say they use a brand because it is a well known A-brand, even though they actually do not use it as they find it too expensive. However, it is possible to check answers. For example, a subject can be asked whether she uses a specific brand; on the other hand, the researcher may obtain an empirical referent of the subject's answer by doing a pantry check (see Verhage & Cunningham, 1989). Item bias may occur if respondents in a particular cultural population do not understand one of the questions. Absence of method and item bias have to be investigated (this is indicated by '+' in Table 3-1).

Medium level inference. If a domain cannot be fully described and if concepts cannot be operationally defined, this is a case of (at least) medium level inference. Examples include measures of, for example, abstract characteristics, such as 'product evaluation', 'innovativeness', and 'service quality'. In this situation, measures are more prone to bias than in situations of low level inference. At the medium level, the definition of the constructs as well as the measurement is prone to bias. No list is available of all possible elements that constitute the universe; such a list cannot be made. In some cases, researchers may achieve consensus on what elements constitute the domain. For instance, in product evaluation of food (e.g., Moskowitz, 1983), or in service quality assessment (e.g., Parasuraman, Zeithaml & Berry, 1985), elements that constitute the domains are presented. However, it cannot be assumed that such elements hold equally in all cultural populations. Scales that have been developed in one country cannot be assumed to be free from bias in other countries. All three types of bias (see Chapter 2) may

affect data in this type of research. This implies that researchers have to assess whether their constructs are equivalent across cultural populations. This can be investigated through statistical procedures, such as confirmatory factor analysis (e.g., Steenkamp & Baumgartner, 1998). A prerequisite is the use of multi-item scales to measure constructs; such scales are often used in academic research, but in applied research they are not, thereby limiting the possibility to establish construct equivalence. If multi-item scales are used, the presence of construct bias may be assessed. As at the other inference levels, method and item bias may invalidate the results. One or several items may be misunderstood (invoking item bias) or all items may be affected by response styles (invoking method bias). At this inference, level all types of bias have to be investigated to ascertain the validity of comparisons across cultural populations (this is indicated by '+' in Table 3-1).

High level inference. There is a high level inference if the universe cannot be properly defined in the same way in different cultural populations, as a result of unclarity about what is and what is not included. Examples are 'adaptability' in psychology, and 'market orientation' in marketing. The environment a subject or a company operates in has to be taken into account when a measure of 'adaptability' or 'market orientation' is developed. However, it is impossible to develop a measure that contains the same elements across cultural populations. At this inference level, defining the construct is hard if not impossible; as a consequence, any inference is likely to be invalid (this is indicated by '?' in Table 3-1).

3.4 Conclusion

In culture comparative research, bias is a major threat to the validity of the data. The type of universe and the way in which concepts are operationalized determine how serious this threat is. In general, the threat of bias is lowest if a concept can be operationalized directly. In such cases, the universe is identical. For example, in marketing, the measure for number of products used can be free from construct and item bias. Possible sources of method bias such as inaccuracy in reading a measuring-glass are usually negligible and valid cross-cultural comparisons can be made.

In this chapter, it was argued that two parameters determine the level of inference in culture comparative research. These parameters are (1) whether a universe is operationally defined or not, and (2) the extent to which a universe can be described in terms of observable items or variables. These parameters determine the extent to which there may be bias, and the extent to which this bias can be controlled by the researcher.

In most studies, there is some construct bias. There may be construct bias for measures that have not been operationally defined. If the topic is concrete (e.g., product evaluation), construct bias may be reduced by introducing controls. For instance, additional data (e.g., on product use, or usage situation) may be collected to avoid an alternative interpretation of the results. However, many concepts in marketing are abstract. Examples include general values (e.g., Rokeach, 1973; Kahle, 1983; Schwartz, 1992, 1994), or hypothetical constructs such as 'ethnocentrism' (e.g., Shimp & Sharma, 1987). For such concepts, construct bias is difficult to control.

In virtually any study, there may be method bias and/or item bias. These are types of bias that occur at all inference levels, for completely and incompletely described domains, and for operationally defined as well as for non-operational concepts. Both types of bias are hardly addressed in culture comparative marketing research although they may invalidate results to a large extent (see e.g., Greenleaf, 1992a).

Chapter 4. Response Style In Rating Scales In EU Countries

4.1 Introduction

In Chapter 2, it was explained that equivalence of research findings can be investigated at three levels, namely construct equivalence, measurement unit equivalence, and scalar equivalence (Van de Vijver & Leung, 1997a). If equivalence is addressed in cross-cultural marketing studies, the emphasis is on construct equivalence of scales or models with known psychometric properties, developed in one country (mostly the United States), for use in other countries. Examples of scales and models that were tested in several countries are the CETSCALE for measuring consumer ethnocentrism (Netemeyer, Durvasula & Lichtenstein, 1991), the Change Seeker Index (CSI, Steenkamp & Baumgartner, 1995), the Fishbein model (Lee & Green, 1991), and the Rokeach value structure (Kamakura & Mazzon, 1991). In studies that used these models, the data were analyzed at a structural level. The emphasis was on the question whether the structure of a particular construct was or was not the same across cultural populations.

However, even if construct equivalence is established, there is no guarantee that there are no other forms of bias in the scores (Tanzer, 1995; Van de Vijver & Leung, 1997a). Bias does not only arise from the definition of the construct, but also from the measuring-instrument that is employed. Method bias is a form of bias that affects scores at the level of the instrument. Method bias may arise from, for example, a difference in familiarity with the response procedures, or different conditions during interviewing. Some of these sources of method bias can be avoided by letting subjects practice before answering questions in a questionnaire

(see e.g., Nunnally, 1978). Physical conditions during interviews should be made as similar as possible across interviews (for example in classrooms), or if this is not possible, a variable assessing aspects of the interview situation can be added to the dataset. Such a variable can be employed afterwards in an analysis of covariance (ANCOVA) or regression analysis to assess the effect of the interview situation on the outcomes. This helps to rule out alternative explanations of results.

In addition to the testing environment, a major source of method bias is due to response styles[9] (e.g., Hui & Triandis, 1989). These are also described as one of the main sources of bias that invalidate cross-cultural marketing research (Usunier, 1996; 2000). Response style refers to a subject's tendency to respond to questionnaire items regardless of item content. An example is 'yeasaying', where subjects tend to agree with items in general, regardless of item content. Response styles are a concern in culture comparative research as they may invalidate the comparability of outcomes (e.g., Douglas & Craig, 1983; Greenleaf 1992a,b; Van de Vijver & Leung 1997a,b).

In culture comparative research, direct comparisons between scores are often made. Especially in such direct comparisons, response style may invalidate content-based comparisons of scores and may make comparisons across cultural populations at least questionable (e.g., Bachman & O'Malley, 1984; Poortinga, 1989). If scores are compared directly, they have to have the same meaning in a quantitative sense as well. For example, a score of '4' on a five-point rating scale that means 'moderately agree' has to reflect the same level of agreement across the cultural populations studied. This can only be assumed to be the case if the differential effects of response style can be ruled out. If scores on rating scales do not have the same meaning across cultural populations, comparisons between scores are erroneous or at least ambiguous.

Studies on response style are relevant for academic cross-national marketing research and cross-national consumer research as well as marketing practice. An example is the measurement of purchase intention on a rating scale, which is a

---

[9] 'Response style' in this thesis only refers to ways of giving answers to questions in questionnaires. It should be noted that the same term is sometimes used in the literature to refer to responding to requests, such as a request by a friend (e.g., Bresnahan, Ohashi, Liu, Nebashi, Liao, 1999) We do not use this sense.

common method in product evaluation research (Dillon, Madden & Firtle, 1994). If positive but invalid results emerge, the decision is made to go ahead with distribution and promotion, and resulting sales can be below expectations (see e.g., Hibbert, 1993). Another example is domain specific market segmentation (Van Raaij & Verhallen, 1994), where three types of items, namely general values, domain specific attitudes, and product specific attitudes are combined in one segmentation model. If this type of research is done in several countries, it is a prerequisite that effects of response style are the same in those countries across the different items (Van Herk, Verhallen & Barzilay, 1994).

Despite persistent cultural differences in consumer behavior, research on this topic is rather limited (Manrai & Manrai, 1996). To the best of our knowledge, no studies exist in which response style is assessed in a cross-national setting across various items that cover several behavioral domains. The purpose of this study is to explore whether response style effects can be identified in six countries in the European Union (EU), across three behavioral domains, for items at three levels of abstraction, and for two types of rating scales. The data are from surveys in the behavioral domains of cooking, washing, and personal care. Together, the surveys involve more than 6500 subjects.

The structure of this chapter is as follows. First, attention will be paid to response style in general. Then, response style is described in relation to background characteristics, response scale format, and item content. Next, the method is described, and the results are reported. Finally, the implications of the results and directions for further research are discussed.

4.2 Response Style

4.2.1 Terminology

Response style has been an issue in educational and psychological measurement since around 1950 (Messick, 1991). Before this time, it was usually taken for granted that the responses subjects gave were a true reflection of their knowledge or their opinion. In the 1940s, research was started on the problem of response style and Cronbach (1946, 1950) published two reviews on the issue. Cronbach (1950) used the term 'set' (not 'style') to refer to stylistic consistencies separately

from content in responses. The term 'set' refered to a temporary reaction or a situation demand on a subject, such as time pressure, or a specific item format. The term 'set' thus suggests that by employing another item format or by doing the study at another time, unbiased measures can be obtained. However, some response sets appeared to be relatively stable (Cronbach, 1946; Messick, 1968) and it was suggested that these response sets might reflect aspects of personality as well. To emphasize this point, Jackson and Messick (1958) proposed renaming 'response sets' 'response styles' as they considered the subject to display bias consistently across time and situations. The terms 'response set' and 'response style' are both still used in literature; some researchers use the term 'set' (e.g., Hui & Triandis, 1985b; the Chinese Culture Connection, 1987; Forsman, 1993), whereas others use the term 'style' (e.g., Bachman & O'Malley, 1984; Messick, 1991; Greenleaf, 1992a). The two terms are mixed up in the literature (see e.g., Greenleaf, 1992b) and for the sake of clarity, we will adopt the term 'response style' in this thesis from now on. The main reason for our choice is the use of the term response style in the marketing literature (e.g., Schaninger & Buss, 1986; Greenleaf, 1992a). More specifically, we define response style in this thesis as: response tendencies due to artifacts of measurement that are reliable and independent of the traits which are intended to be measured. Reliable in this definition means that there is consistency in response patterns across instruments; i.e. yeasaying response style on various item sets correlates substantially.

4.2.2 Typology

There are three prominent response styles that are mentioned in the psychological literature (Paulhus, 1991), namely acquiescence, extreme response bias, and socially desirable responding[10]. Acquiescence is the tendency to agree rather than

[10] It should be noted that Broen and Wirt (1958) mention eleven different types of response style (they call these sets). These styles are: (1) tendency to agree, (2) tendency to disagree, (3) evasiveness, (4) tendency to extreme agreement or disagreement, (5) carefulness, (6) speed, (7) tendency to guess, (8) tendency to check many items, (9) tendency to spread judgments on a rating scale, (10) tendency to rate to the left of center, and (11) tendency to rate to the right of center. Styles 1,2,4,9,10, and 11 are styles that fit into our definition of response style in rating scales. They can be considered forms of either yea-saying or extremity response style. The other styles (3, 5, 6, 7 and 8) are beyond the scope of this report. They apply to psychological testing, where subjects give right or wrong answers or check items (on a dichotomous scale) that apply to them.

disagree with items in general, regardless of item content. Subjects who tend to agree are called 'yea-sayers', as they are inclined to say 'yes' to items. The opposite of a yeasayer is a 'naysayer', a subject who tends to say 'no', regardless of item content. Extreme response bias (ERB) (Paulhus, 1991) is the tendency to use the extremes on a rating scale (e.g. the 1 and 5 on a five-point scale). In other publications, this tendency is also called Extreme Response Style (ERS) (e.g., Crandall, 1982; Greenleaf, 1992b). The third prominent response style is socially desirable responding (SDR). SDR is the tendency of subjects to make themselves look good in relation to current cultural norms when answering questionnaire items (Mick, 1996). Social desirability is especially important when measuring personality or self-reports of sensitive behavior. In personality and social psychological research, social desirability is the most frequently studied type of response style (Paulhus, 1991). In most marketing studies, the threat of social desirability is of lesser importance, as questions do not pertain to socially sensitive issues. However, in some cases, subjects may choose socially desirable responses, for instance, if questions are posed on topics such as materialism or compulsive buying (Mick, 1996).

Of the three aforementioned response styles, social desirability is the one for which most control measures have been developed. Two well known measures are the Edwards Social Desirability Scale (Edwards, 1957) and the Marlowe-Crowne Social Desirability Scale (Crowne & Marlowe, 1960). These scales are used to correct the scales one is interested in afterwards. Specific measures of acquiescence do not exist, but acquiescence can be countered by reversing half of the items used in a study. For example, in assessing a personality trait, the scale that is used usually has half of the items formulated positively and the other half negatively with respect to the trait concerned (see e.g., Knowles & Nathan, 1997; Ten Berge, 1999). Such a balanced item set does not eliminate acquiescence, but it distributes acquiescent responses evenly to positively and negatively formulated items. These balanced sets then provide the opportunity to develop separate measures for the trait and the acquiescence component. The trait score is based on the number of positive responses to the positively formulated items, plus the number of negative responses to the negatively formulated items. In the acquiescence measure, the number of positive answers to positively formulated items is added to the number of positive answers to negatively formulated items (Knowles & Nathan, 1997). As for acquiescence, there are no specific measures (such as a scale) for extreme response style (ERS). However, Greenleaf (1992b) has made a first step towards creating such a measure. He argued that if an ERS measure is defined as the

proportion of extreme responses, it will be more accurate if items are uncorrelated and have equal extreme response probabilities. In addition, he outlined a procedure to create and validate an ERS measure using items from a large US survey and showed that ERS is a valid, stable response style.

4.2.3 Response style: bias, truth or both?

For some years now, there has been a controversy among researchers on the issue of response style (see e.g., Messick, 1991). Some researchers believe response style provides important information on respondents (Jackson & Messick, 1958), while others believe that this is a myth (Rorer, 1965). Several publications have since indicated that response style is far from mythical. For example, there is evidence that there are differences between subjects in displaying social desirability response style (e.g., Middleton & Jones, 2000), extreme response style (e.g., Stening & Everett, 1984; Hui & Triandis, 1989; Greenleaf, 1992a,b), and yeasaying[11] response style (e.g., Moum, 1988; Greenleaf, 1992a).

It should be noted that differences in responses are not only attributable to response style, but also have a real attitude component (Jackson & Messick, 1958; Bentler, Jackson & Messick, 1971). In his review, O'Donovan (1965) indicated that normal subjects tend to have a higher tendency to extreme response if items are more meaningful to them. In a study by Forsman (1993), ERS depended on a categorical rejection of items with a negative content and not on acceptance of items with a positive content. He concluded that the results were attributable to response style as well as the content of the items. Similar conclusions were drawn by Greenleaf (1992a), in his study on consumer attitudes. It must be noted that in order to separate response style and content, more than a single item set has to be included in a study. With only one set of items, measuring one construct that is assessed differently by subjects, construct bias and method bias cannot be separated (see also Chapter 2). Consequently, the results might contain either, (a) only true information, (b) only response style (bias), or (c) both true information and bias. This affects the inferences that can be made based from the results. If the results contain only true information, cross-cultural differences are valid. If they contain

[11] The term yeasaying (e.g., Greenleaf, 1992a) is used in marketing. Therefore, we will use this term. In the literature, the term acquiescence (e.g., Martin, 1964; Ray, 1983; Watson, 1992) is used to refer to the same phenomenon.

response style or bias only, the results found are erroneous. If they contain both true information and bias, the outcome is equivocal, as we do not know which part of the variance is attributable to true information and which part to bias.

Several methods can be envisaged to assess response styles. One way is to measure the same construct (or trait) using various different methods (the so called multitrait-multimethod designs, e.g., Campbell & Fiske, 1959; Bagozzi & Yi, 1991; Marsh & Byrne, 1993). If various methods provide the same outcomes for each trait, then response bias is less likely. Another way to assess response style is to develop balanced scales (see Section 4.2.2). Balanced scales contain item pairs that are logical opposites (e.g., Ray, 1983). In marketing research, developing balanced scales may prove to be very difficult, as it is in sociology (Schuman & Presser, 1996). For some items, there is no logical opposite. In addition, questionnaires in (applied) marketing research usually cover various topics, and due to financial and time constraints, the number of questions that can be asked per topic is limited. Still another way to assess response style, and the method we employ here, is to determine the consistency of responses across various item sets with a heterogeneous content (e.g., Martin, 1964). In studies on response style employing this method, the use of items with a low inter-item correlation is preferred (Greenleaf, 1992b). For example, consider a set of items that measures only one construct, so that the items are highly correlated. Here, it is more likely that subjects will give similar answers to all those items. In such a situation, there are two explanations for the resulting answers: they can be either due to response bias or due to strong feelings about the construct (i.e. content of the items). If more constructs and items with a low inter-item correlation are employed, the confusion of content and bias is reduced as it is less likely that subjects will be positive about all the independent constructs. If similar response tendencies across various different item sets emerge, response bias is likely.

4.3 Factors affecting response style

4.3.1 Response style and cultural population

In the last few decades, a number of empirical studies were published in which differences in response styles were reported for several cultural populations. For instance, response style has been explored in different ethnic groups in the USA. It was explored in Afro-American and white American high school seniors (Bachman & O'Malley, 1984), and in groups of Hispanic and non-Hispanic whites (Hui & Triandis, 1989; Marín, Gamba & Marín, 1992). In those studies, a similar result was found: the non-white American subjects tended to display yeasaying more often. Studies focusing on response style in different countries are scarce. In a study by Chen, Lee, and Stevenson (1995), a difference in response style was found between subjects from the US, Canada, Japan, and Taiwan. In their student samples, American students had higher mean scores (i.e. they tended to display yeasaying) than Japanese and Taiwanese students and typically avoided the midpoint of the rating scales (i.e. they tended towards ERS). In other studies employing Korean and US student subjects (Chun, Campbell & Yoo, 1974; Lee & Green, 1991), a similar result was found. American students displayed more extreme scoring than Koreans did. Apart from these studies in the US and Asia, some research was done in Arab countries as well. Soueif (1968) compared extreme responses of students in Egypt, Syria, and Jordan. He found that Egyptian subjects gave more extreme responses than Jordanian and Syrian subjects did.

Studies on response style employing European samples are scarce. One study was carried out in which small samples (between 8 and 26 subjects) in Italy, France, and the USA were compared (Fioravanti, Gough & Frere, 1981). The authors concluded that overall social desirability ratings were similar across the three countries, but they found some significant differences for the desirability of specific items. In a descriptive study by Williams (1991), differences in rating scale based results were reported for German and Italian subjects. Williams found that the mean was usually higher in Italian samples. However, no empirical data were provided in his study. Differences in scale means have also been noticed between Greek, British and Belgian subjects (Steenkamp & Baumgartner, 1998). In Steenkamp and Baumgartner's study which covered only one construct (the

CETSCALE to measure ethnocentrism; Shimp & Sharma, 1987), Greek subjects had the highest mean, followed by British and Belgian subjects.

4.3.2 Response style and type of rating scale

In studies focusing on response style, several types of rating scales have been employed; examples include five-point (Bachman & O'Malley, 1984; Hui & Triandis, 1985b, 1989; Krosnick & Alwin, 1988), six-point (Greenleaf, 1992a, 1992b), seven-point (Barker & Kaciak, 1992; Chen, Lee & Stevenson, 1995), and ten-point rating scales (Hui & Triandis 1985b, 1989). From these studies, no consistent result has emerged, with respect to what format is least or most susceptible to response style effects. One major reason for this is that in most studies only one type of rating scale is employed. An exception is the study by Hui and Triandis (1989). In their study, involving male American Navy recruits aged 17 to 25, they found that Hispanics displayed more extreme response style (ERS) than Non-Hispanic recruits, when 5-point rating scales were used. However, when using 10-point rating scales the difference in ERS disappeared. Thus, it is possible that another response format affects response style. This would imply that response style is not a permanent characteristic of subjects and that it can be avoided by employing another scale format.

4.3.3 Response style and behavioral domain

In cross-cultural studies that assess response style, multi-dimensional constructs covering items at various levels of abstraction are preferred. Most studies that assess bias assume unidimensional structures or constructs with an a-priori known psychometric structure (Millsap & Everson, 1993; Öngel & Smith, 1994). If item bias is assessed, the unidimensionality of scales is a statistical prerequisite (see e.g., Shepard, 1982) and items have to be homogeneous with respect to the construct they measure. In assessing construct bias, the same preference for unidimensional constructs exists (see e.g., Steenkamp & Baumgartner, 1998).

Most studies that assess response style involve one type of item at a fairly high level of abstraction, such as personality (Couch & Keniston, 1960; Crandall, 1982; Forsman, 1993; Knowles & Nathan, 1997), work related attitudes (Hui & Triandis, 1989), attitudes towards current and future life (Bachman & O'Malley, 1984), or general attitudes, interests, and opinions (Greenleaf, 1992b). There are few publications in which response style is investigated using items at different

levels of abstraction. An exception is the study by Schaninger and Buss (1986), in which concrete attributes were employed to assess product characteristics and more abstract attributes were employed to assess energy consumption. Schaninger and Buss found that response style effects carried over from one type of attribute to another.

In a cross-national context, items at several levels of abstraction have only been employed simultaneously in studies that focus on the reliability of measures. Davis, Douglas, and Silk (1981) investigated the reliability of three types of measures, namely background characteristics, involvement in household decisions, and life-style in five countries. They found no systematic tendency to exhibit high or low reliability across five countries (the United States, the UK, France, Belgium and Canada). However, differences in reliability appeared depending on the type of variable. Hard variables such as demographic information are systematically measured more reliably than soft variables such as life-style information. A similar result was found by Parameswaran and Yaprak (1987) in their study in the United States and Turkey. They found that reliability coefficients were higher for hard product specific attributes when compared to more soft general country and general product attributes. In their study, differences in reliability between countries were also found: the US sample had more reliable responses than the Turkish sample. In these studies, no differences in the levels of scores were reported. However, reliabilities might be affected by yeasaying response style. If subjects give about the same (extreme) answers to items, the reliability of the scale based on these items increases.

4.4 Assessing response style

If a consistent way of responding to items is found independent of behavioral domain and rating scale response format, then it is likely that this is a response style (cf. Martin, 1964). On the other hand, if more positive answers are given to items referring to one behavioral domain than to another domain, then a real difference in attitude is likely. Also, if there is no consistent response pattern across different types of rating scales, then this points to an absence of response style. Thus, to explore whether there are differences in response style between countries, we formulate the following research question:

RQ: Are there consistencies in responses to items between EU countries, independent of rating scale type and behavioral domain?

4.5 Method

To explore response style, we used data from surveys in 6 countries in the EU, Greece, France, Spain, Italy, Germany, and the United Kingdom. All data were originally collected for commercial purposes and were made available to us by Unilever Research in Vlaardingen and Philips Domestic Appliances and Personal Care in Groningen, both in The Netherlands. The surveys covered three behavioral domains, namely cooking, washing, and personal care. We employed the surveys to explore response style across various types of rating scales and across items at various levels of abstraction, namely product specific attitudes, domain specific attitudes, and general values. In the following sections, samples and measures are described, followed by results, and the conclusion and discussion.

4.5.1 Samples and Measures

4.5.1.1 Survey I

4.5.1.1.1 Subjects

Survey I is from an international study on cooking behavior, which was carried out in the period 1988 through 1990 in three European countries, Greece, Italy, and France, for Unilever Research in the Netherlands. In each country, a national quota sample of homemakers was drawn that was representative with respect to age, region, and rural/urban residence. The sample sizes were 580, 598, and 532 in Greece, Italy, and France, respectively. Subjects in all three countries were female homemakers, aged 20 to 60. Mean ages were 38.1 in Greece, 40.1 in Italy, and 41.5 in France.

4.5.1.1.2 Measures

Using a face-to-face data collection procedure, information was gathered on domain specific attitudes, cooking behavior, product usage, product specific attitudes, and demographics. Bilinguals translated the questionnaires from English

into Greek, Italian, and French. As a second step, a discussion between project coordinators of the research agencies in the participating countries was held to check the equivalence of the translations. The items used in this study are 19 items on domain specific attitudes and 32 items on product specific attitudes. Examples of the domain specific items are 'I choose products that save me cooking time' and 'I consume more calories than I need'. The items were measured on five-point rating scales. The items have symmetrical response intervals ranging from 'strongly disagree' (= 1) to 'strongly agree' (= 5). All scale points were labeled.

The product specific attitudes are concrete; they refer directly to a specific product used in a specific application. In this study, attitudes were measured on two applications in the cooking domain. The first was spreading on bread or toast, and the second was making pastry. These specific applications, making pastry and spreading, were chosen since these applications can be considered functionally equivalent in an international context (Douglas & Craig, 1983). The products used in these applications were butter and margarine in all three countries. In total, eight items per application-product combination were used to measure product specific attitudes. Thus, in the questionnaire, it was clear for a subject to what specific application (e.g., making pastry) and to what specific product (e.g., butter) an item referred. An example of an item is 'it is a product that offers value for money'. The items were measured on nine-point rating scales. The items used had symmetrical response intervals ranging from 'totally disagree' (=1) to 'totally agree' (=9) with only the end points and the middle point labeled. The middle point was labeled 'neither agree nor disagree'. All items were formulated positively.

4.5.1.2 Survey II

4.5.1.2.1 Subjects

Survey II, carried out in 1993, concerned 'washing of clothes'. This survey covered six EU countries, namely Greece, Italy, France, Germany, the UK, and Spain. The sample sizes were 281, 299, 313, 295, 298, and 303 in Greece, Italy, France, Germany, the UK, and Spain respectively. Subjects in all samples were female homemakers, aged 20 to 60. Each quota sample was representative with respect to age and family size for the population being studied in the countries. Mean ages were 40.5 in Greece, 44.3 in Italy, 39.7 in France, 40.1 in Germany, 40.1 in the UK, and 41.8 in Spain.

4.5.1.2.2 Measures

Using a face-to-face data collection procedure, information was gathered on washing clothes. Items included domain specific attitudes and socio-demographic information. Bilinguals translated the questionnaires from English into Greek, Italian, French, Spanish, British, and German. As a second step, a discussion between the project coordinators of the research agencies in the participating countries was held to check the equivalence of the translations. The subjects assessed eleven items on washing. All items were formulated positively. An example of an item is 'Doing the laundry for your family gives you a lot of satisfaction'. These items were measured on a 5-point rating scale, where 1='don't agree at all' and 5='agree completely'. All scale points were labeled.

4.5.1.3 Survey III

4.5.1.3.1 Subjects

Survey III was held in 1996 in five EU countries, namely the UK, France, Germany, Italy, and Spain. Subjects in this survey were male members of a large European marketing research panel. In each country, the samples were representative with respect to age. To make the data sets comparable in age to the two previous surveys, only men aged 20 to 60 were included in the present analysis. Adjusted final sample sizes were 580, 550, 634, 806, and 453 in the UK, France, Germany, Italy, and Spain, respectively. Mean ages were 39.9 in Germany, 39.8 in the UK, 37.5 in Spain, 39.2 in Italy, and 37.9 in France.

4.5.1.3.2 Measures

In this survey, a mail questionnaire was employed to collect information on domain specific attitudes, general values, and socio-demographic variables. The study included 45 items on attitudes towards shaving. Each subject assessed the items with his own shaving method in mind, namely electric or blade shaving. All items were formulated positively. These 45 domain specific items were assessed on five-point rating scales, with the endpoints labeled 1 = 'disagree strongly' and 5 = 'agree strongly'. Examples of items are 'You feel fresh during shaving' and 'It is simple and easy to use'. The 45 items can be split into six separate sets of items that each cover a more or less coherent subdomain in shaving experience, such as the result and the convenience of the shaving method.

The general values are the List of Values (LOV) by Kahle (1983). The LOV is composed of nine items, which were assessed on nine-point rating scales. The scales were labeled 1='very important' to 9 = 'very unimportant'; middle points of the rating scales were not labeled. Subjects were given the LOV items and the instruction that it 'is a list of things that some people look for or want out of life'. They were asked to study the list and then rate each item on the nine-point scale. An example of a LOV item is 'Security'. Bilinguals translated the questionnaires from English into Italian, French, Spanish, and German. As a second step, a discussion between the project coordinators of the research agencies in the participating countries was held to check the equivalence of the translations.

4.5.2 Results

In assessing response style, a first step is to determine whether the items are actually measuring more than one concept. For each survey, principal components analyses were performed. For the domain specific attitudes in the survey on cooking, five components were found in each country. For the product specific attitudes two components were obtained when margarine was assessed and three when butter was assessed. For the washing survey, four components were found, and for the personal care survey, there were six components for the domain specific attitudes and two for the general values[12]. Employing the components, we defined item sets in such a way that each set contained a minimum of four items. If this condition was not satisfied, all items within a behavioral domain were taken together. In Table 4-1 and Table 4-2, the various item sets we employed are presented. In total, nineteen item sets were used. In the survey on cooking, we had five item sets; in the survey on washing 1 item set, and in the survey on personal care thirteen (six sets for electric shaving, six sets for blade shaving and one set for the LOV). It should be noted that in the personal care survey each subject answered either the items on electric or on blade shaving; thus, each subject assessed seven item sets.

---

[12] In order to determine whether there was factorial agreement, we performed a target rotation of the component structure in each country towards the component structure across all countries within the datasets. After this, Tucker's phi coefficient of agreement (Tucker, 1951) was calculated. A majority (75%) of the coefficients was higher than .90, a value below .90 points to incongruities (Van de Vijver & Leung, 1997a).

In addition, mean inter-item correlations within sets were calculated to assess whether each item set contains items with a different content. This is one criterion for defining a relevant item set for assessing response style (Greenleaf, 1992b). Mean inter-item correlations are not high (see Appendix 4.1), indicating measurement of various content. For the domain specific attitudes, the correlations range between .03 (cooking) and .56 (shaving behavior), and for product specific attitudes (cooking), the correlations range between .20 and .34. On average, the mean inter-item correlation is highest for the general values and the shaving attributes. The general values' inter-item correlations ranged from .29 in France to .57 in the UK. The mean inter-item correlation for the shaving attributes ranged from .20 in France for the attributes on irritation while shaving electrically to .56 in Italy on the items covering the shaving process while shaving electrically.

4.5.2.1 Yeasaying and extreme response indices

To obtain scores for yeasaying and extreme response behavior, each subject's response profile across the items was employed. The profile refers to the frequencies of the various response categories (1's, 2's, 3's et cetera which the subject used to indicate responses on the items). An example of a subject's profile while assessing nine items is twice a '1', three times a '2', once a '3', twice a '6' and once a '9'. Based on such profiles per subject, indices for yeasaying and an extreme response were calculated.

Following Bachman and O'Malley (1984), we first computed two indices, one for agreement and one for disagreement. The agreement index was the proportion of responses in the highest response category (e.g., 'totally agree')[13]. The value of the index varies between 0.00 and 1.00, where a 1.00 refers to a subject who totally agreed with all items. The disagreement index was the proportion of responses in the lowest response category ('totally disagree' or other most negative label on the

[13] All items were formulated positively, which implied that respondents endorsed the rating scale values on the positive side of the rating scale. On the 5-point rating scales, the values 1 and 5 are taken to calculate indices. On the 9-point rating scales only the values 1 and 9 were taken. The reason for not including the 2 and the 8 is that for the items on product specific attitudes (cooking) and general values (LOV, personal care) the extreme values (1 or 9) were endorsed in more than 40% of the items. This number would increase to about 55% if the values 2 and 8 would have been included in calculating the indices.

rating scale). Again, the value of the index varies between 0.00 and 1.00, where a 1.00 refers to a subject who totally disagreed with all items. The yeasaying index was calculated by subtracting the disagreement index from the agreement index. The extreme response index was calculated by adding the agreement and the disagreement index. Thus, the yeasaying and the extreme response index are, by definition, not totally independent. For example, if a subject only agrees with items (and thus the disagreement index = 0.00), then the yeasaying index and the extreme response index will be the same.

The yeasaying and the extreme response index were calculated separately for each item set and for each country. For example, for subjects in each country in survey I (cooking domain) five scores were obtained, that is, one for the domain specific attitudes and four for the product specific attitudes.

4.5.2.1.1 Yeasaying index

In Table 4-1 and Table 4-2, an overview is given of the yeasaying indices for the various item sets in each country. The yeasaying index differs across item sets and across countries; the scores on the index range from -.391 (Germany, electric shaving, emotional aspects) to .509 (Greece, cooking, product specific attitudes). The negative index means that subjects more disagree than agree with the items. The positive index of .509 means that subjects assess more than half of the items in the highest response category.

For each of the nineteen item sets, an Analysis of Variance was performed to determine whether indices differed across countries. Following this analysis, differences between countries were tested using the Tukey-HSD test. Results showed that there are differences in yeasaying indices between countries. In 18 out of 19 item sets, differences in the yeasaying index were found between countries (F-test, all $p < .001$).

The yeasaying indices in Table 4-1 indicate that Greek subjects tend to be more positive than Italian and French subjects. The Greek subjects have yeasaying indices of .372 and .356 respectively for the domain specific attitudes on cooking and washing. For Italian and French subjects, these indices are lower: in the washing domain, they are about zero and in the cooking domain the indices are .236 for Italian and .177 for French subjects.

Table 4-1 Yeasaying index Cooking and Washing

	Domain	Item content	Scale type	Number of items	Greece	Italy	France	Germany	The UK	Spain	F	df	p	η^2
	Cooking													
1	Domain specific	attitudes	5-point	19	$.372^{b,c}$	$.236^{c}{}_{a}$	$.177_{a,b}$				177.46	2,1688	<.001	.174
2	Product specific	spreading - butter	9-point	8	$.313^{b,c}$	$.151_{a,c}$	$.202^{b}{}_{a}$				11.32	2,713	<.001	.030
3		spreading - margarine	9-point	8	$.457^{b}$	$.206_{a,c}$	$.403^{b}$				11.47	2,556	<.001	.040
4		pastry - butter	9-point	8	$.407^{b,c}$	$.172^{c}{}_{a}$	$.115_{a,b}$				36.17	2,653	<.001	.100
5		pastry - margarine	9-point	8	$.509^{b,c}$	$.218_{a}$	$.233_{a}$				61.16	2,591	<.001	.171
	Washing													
6	Domain specific	attitudes	5-point	11	$.356^{b,c,d,e,f}$	$.063_{a}^{e}$	$.072_{a}^{e}$	$.045_{a}^{e}$	$-.087_{a,b,c,d,f}$	$.049_{a}^{e}$	129.94	5,1783	<.001	.267

Tukey HSD test $p < .05$; superscript indicates a higher score; subscript indicates a lower score

a: significantly different from Greece; b: significantly different from Italy; c: significantly different from France; d: significantly different from Germany e: significantly different from the UK; f: significantly different from Spain

Table 4-2 Yeasaying index Personal Care

	Domain	Item content	Scale type	Number of items	Greece	Italy	France	Germany	The UK	Spain	F	df	p	η^2
	Personal care													
	electric shaving													
7	Domain specific	shaving: emotional	5-point	12		$-.033^{c,d,e}$	$-.202_{b,f}{}^{d}$	$-.391_{b,c,e,f}$	$-.250_{b,f}{}^{d}$	$-.058^{c,d,e}$	34.76	4,923	<.001	.131
8	Domain specific	shaving: sensorial	5-point	10		$.239^{c,d,e}$	$.040_{b,f}$	$.023_{b,f}$	$.063_{b,f}$	$.223^{c,d,e}$	16.65	4,923	< .001	.067
9	Domain specific	shaving: result	5-point	8		$.181^{c,d,e}$	$.056_{b}$	$.019_{b}$	$-.010_{b,f}$	$.118^{e}$	8.79	4,923	< .001	.037
10	Domain specific	shaving: convenience	5-point	7		$.482^{d,e}$	$.395_{b}$	$.364_{b,f}$	$.369_{b}$	$.476^{d}$	4.81	4,923	< .001	.020
11	Domain specific	shaving: irritation	5-point	4		.219	$.237^{d}$	$.124_{c}$	.145	.106	3.32	4,923	< .05	.014
12	Domain specific	shaving: process	5-point	4		$-.056_{d}$	$-.076_{d}$	$.116^{b,c,e,f}$	$-.087_{d}$	$-.130_{d}$	14.96	4,923	< .001	.061
	blade shaving													
13	Domain specific	shaving: emotional	5-point	12		$.036^{c,d,e,f}$	$-.122_{b}{}^{d}$	$-.231_{b,c,f}$	$-.179_{b,f}$	$-.079_{b}{}^{d,e}$	37.94	4,2088	< .001	.068
14	Domain specific	shaving: sensorial	5-point	10		$.372^{c,d,e}$	$.156_{b,f}$	$.223_{b,f}$	$.187_{b,f}$	$.324^{c,d,e}$	33.75	4,2088	< .001	.061
15	Domain specific	shaving: result	5-point	8		$.359^{c,d,e,f}$	$.245^{e}{}_{b}$	$.256^{e}{}_{b}$	$.170_{b,c,d,f}$	$.273^{e}{}_{b}$	19.47	4,2088	< .001	.036
16	Domain specific	shaving: convenience	5-point	7		$.325^{c,d,e}$	$.235_{b,f}$	$.232_{b,f}$	$.192_{b,f}$	$.344^{c,d,e}$	16.59	4,2088	< .001	.031
17	Domain specific	shaving: irritation	5-point	4		$.128^{c,d,e,f}$	$.010_{b}$	$.045_{b}$	$-.036_{b}$	$.034_{b}$	12.51	4,2088	< .001	.023
18	Domain specific	shaving: process	5-point	4		$.102^{c,e,f}$	$-.001_{b,d}$	$.111^{c,e,f}$	$.037_{b,d}$	$-.042_{b,d}$	15.26	4,2088	< .001	.028
19	General	values (LOV)	9-point	9		$.403^{c,d,e}$	$.316_{b,f}$	$.288_{b,e,f}$	$.350^{c,d}{}_{b}$	$.402^{c,d}$	16.64	4,3018	< .001	.022

Tukey HSD test $p < .05$; superscript indicates a higher score; subscript indicates a lower score

a: significantly different from Greece; b: significantly different from Italy; c: significantly different from France; d: significantly different from Germany e: significantly different from the UK; f: significantly different from Spain

The Table shows substantial differences between Greece and Italy and between Greece and France on domain specific attitudes on cooking and washing. Six out of six times, the Greek yeasaying index is higher than the Italian index; this in itself already differs significantly from a chance distribution ($p < .05$) (Binomial probability, Hays, 1973). The indices for Greek and French subjects are different on five occasions; the Greek index is five times higher, which again differs from a chance distribution ($p < .05$). Differences in yeasaying indices are also found between French and Italian subjects. In fifteen of the nineteen item sets, the Italian and French indices were significantly different. In thirteen out of fifteen cases, the Italian index is higher ($p < .01$). The Spanish yeasaying index is higher than the French subjects' index in five out of five different item sets ($p < .05$). The number of times the yeasaying index in Germany and the UK is lower than in Italy and Spain and consistently differs from a chance distribution ($p < .05$). The yeasaying indices for German, French, and British subjects do not show consistent differences.

The differences in yeasaying indices are only meaningful if they account for a substantive proportion of the variance or if the effect size η^2 (Cohen, 1988) is meaningful relative to e.g., treatment effects. As can be seen in Table 4-1 and Table 4-2, the average η^2 in this study is .073 (range from .014 (item set on irritation in electric shaving) to .267 (item set on attitudes towards washing)). Cohen (1988) calls this a 'medium' to 'large' effect size and Peterson, Albaum, and Beltramini (1985) reported that in most studies of consumer behavior, η^2 attributable to treatments is below .090. Thus, the amount of variance accounted for by yeasaying is certainly non-negligible.

4.5.2.1.2 Extreme response index

Results for the extreme response index are presented in Table 4-3 and Table 4-4. As in yeasaying, for extreme response behavior, there are significant differences between countries (F-test, $p < .001$). Results for the Greek subjects are striking. In all six item sets in which Greek subjects participated, the extreme response index is higher for Greek than for Italian and French subjects (Tukey HSD, $p < .05$). Hence, Greek subjects display more extreme response behavior than Italian and French subjects do.

Table 4-3 Extreme response index Cooking and Washing

	Domain	Item content	Scale type	Number of items	Greece	Italy	France	Germany	UK	Spain	F	df	p	η^2
	Cooking													
1	Domain specific	attitudes	5-point	19	$.560^{b,c}$	$.479_{a}^{c}$	$.379_{a,b}$				169.49	2,1688	<.001	.167
2	Product specific	spreading - butter	9-point	8	$.492^{b,c}$	$.261_{a}$	$.300_{a}$				22.89	2,713	< .001	.060
3	Product specific	spreading -margarine	9-point	8	$.513^{b,c}$	$.239_{a,c}$	$.424_{a}^{b}$				16.38	2,556	< .001	.056
4	Product specific	pastry - butter	9-point	8	$.547^{b,c}$	$.255_{a}$	$.244_{a}$				34.32	2,653	< .001	.095
5	Product specific	pastry - margarine	9-point	8	$.547^{b,c}$	$.269_{a}$	$.270_{a}$				60.48	2,591	< .001	.170
	Washing													
6	Domain specific	attitudes	5-point	11	$.488^{b,c,d}$	$.345_{a,e,f}^{d}$	$.376_{a,e,f}^{d}$	$.246_{a,b,c,e,f}$	$.443^{b,c,d}$	$.447^{b,c,d}$	44.51	5,1783	<.001	.111

Tukey HSD test $p < .05$; superscript indicates a higher score; subscript indicates a lower score

a: significantly different from Greece; b: significantly different from Italy; c: significantly different from France; d: significantly different from Germany e: significantly different from The UK; f: significantly different from Spain

Table 4-4 Extreme response index Personal care

	Domain	Item content	Scale type	Number of items	Greece	Italy	France	Germany	UK	Spain	F	df	p	η^2
	Personal care													
	electric shaving													
7	Domain specific	shaving: emotional	5-point	12		.448	$.384_{d}$	$.476^{c}$	.409	.442	2.90	4,923	< .05	.012
8	Domain specific	shaving: sensorial	5-point	10		$.420^{c,d,e}$	$.281_{b}$	$.280_{b,f}$	$.267_{b,f}$	$.374^{d,e}$	9.74	4,923	< .001	.040
9	Domain specific	shaving: result	5-point	8		$.401^{c,d,e}$	$.236_{b,f}$	$.237_{b,f}$	$.249_{b}$	$.333^{c,d}$	11.89	4,923	< .001	.049
10	Domain specific	shaving: convenience	5-point	7		$.573^{c,d,e}$	$.421_{b,f}$	$.410_{b,f}$	$.425_{b,f}$	$.571^{c,d,e}$	11.48	4,923	< .001	.047
11	Domain specific	shaving: irritation	5-point	4		$.420^{d}$	.363	$.324_{b}$	.331	.346	2.49	4,923	< .05	.011
12	Domain specific	shaving: process	5-point	4		$.448^{c,d,e}$	$.273_{b,f}$	$.330_{b,f}$	$.328_{b,f}$	$.520^{c,d,e}$	14.00	4,923	< .001	.057
	blade shaving													
13	Domain specific	shaving: emotional	5-point	12		$.422^{c,e}$	$.341_{b,f}$	.398	$.374_{b}$	$.423^{c}$	6.83	4,2088	< .001	.013
14	Domain specific	shaving: sensorial	5-point	10		$.476^{c,d,e,f}$	$.274_{b,f}$	$.304_{b,f}$	$.278_{b,f}$	$.389_{b}^{c,d,e}$	39.85	4,2088	< .001	.071
15	Domain specific	shaving: result	5-point	8		$.471^{c,d,e,f}$	$.350_{b}$	$.340_{b,f}$	$.328_{b,f}$	$.410_{b}^{d,e}$	19.57	4,2088	< .001	.036
16	Domain specific	shaving: convenience	5-point	7		$.462^{c,d,e}$	$.308_{b,f}$	$.307_{b,f}$	$.308_{b,f}$	$.471^{c,d,e}$	33.82	4,2088	< .001	.061
17	Domain specific	shaving: irritation	5-point	4		$.354^{c,d,e}$	$.278_{b}$	$.241_{b,f}$	$.279_{b}$	$.334^{d}$	9.00	4,2088	< .001	.017
18	Domain specific	shaving: process	5-point	4		$.423^{c,e}$	$.307_{b,f}$	$.361_{f}$	$.335_{b,f}$	$.444^{c,d,e}$	13.44	4,2088	< .001	.025
19	General	values (LOV)	9-point	9		$.488^{c,d,e}$	$.394_{b,f}^{d}$	$.348_{b,c,e,f}$	$.411_{b,f}^{d}$	$.462^{c,d,e}$	25.61	4,3018	< .001	.033

Tukey HSD test $p < .05$; superscript indicates a higher score; subscript indicates a lower score

a: significantly different from Greece; b: significantly different from Italy; c: significantly different from France; d: significantly different from Germany e: significantly different from the UK; f: significantly different from Spain

In Italy, the extreme response index is higher than in France. In 12 out of 13 item sets where scores are different, the Italian index is higher ($p < .01$). A similar result emerges between Italy and Germany (11 out of 11; $p < .01$) and between Italy and the UK (11 out of 12; $p < .01$). Italian subjects consistently tend to have a higher extreme response index than subjects in the Western European countries. As in Italy, the extreme response index in Spain is consistently higher than in France, Germany, and the UK ($p < .05$). Between France, Germany, and the UK, the index is not consistently different. The indices are also similar in Italy and Spain ($p = .38$).

The index of effect size η^2 (Cohen, 1988) is reported in the last column of Table 4-3 and Table 4-4. The average effect size is .057, ranging from .011 (items on irritation in electric shaving emotional and irritation aspects) to .170 (items on product specific attitudes on margarine for making pastry). The effect size is about 'medium' (Cohen, 1988). Thus, it is likely that the extreme response style, to a certain extent, affects inferences based on rating scale based information.

4.5.2.2 Response style and type of rating scale

In the food domain and in the personal care domain, different types of rating scales were employed, namely five-point and nine-point rating scales. In the food domain, three countries are included: Greece, Italy, and France. In the personal care domain, five countries are included: France, Italy, Germany, the UK, and Spain. In describing results, we start with the food domain, followed by personal care.

In the food domain, the yeasaying index is highest in Greece independent of rating scale type. For France and Italy, results are equivocal. For five-point scales, the Italian indices are higher, but for nine point scales on product specific attitudes, the differences are not consistent in one direction. In the personal care domain, the indices based on 5- and 9-point scales are consistent in one direction. Here, the scores are consistently highest in Italy and Spain and lowest in the UK, France and Germany.

To determine whether response tendencies carry over from one scale to the other, Pearson correlations for samples per country were calculated for yeasaying and extreme indices on five- and nine-point scales in the food and personal care domain. The average correlation for the yeasaying indices in the cooking domain was .15, ranging from .10 in Italy to .23 in France. For extreme response indices,

the average correlation was .29, ranging from .27 in Greece to .30 in France. In the personal care domain, the average correlation for the yeasaying indices was .18, ranging from .13 in Spain to .21 in Italy; for the extreme response indices, the average was .29, ranging from .25 in France and Germany to .33 in Italy. All correlations were positive and significant ($p < .01$), indicating that response tendencies carry over from one rating scale type to the other.

4.5.2.3 Response style and behavioral domain

In our study, three behavioral domains were included, namely cooking, personal care, and washing. Some differences on the yeasaying and extreme response indices are notable. In the domains where Greek subjects were included, their indices were consistently higher, whatever the item set. In the personal care domain, the indices for France, Germany, and the UK were consistently lower than those of Italy and Spain. This held for both the yeasaying and the extreme response index. In the washing domain, results were different. British subjects had a lower yeasaying index than Italian, French, German, and Spanish subjects, and they had a higher extreme response index than Italian, French, and German subjects. The reason for this was that they tended to be very negative about the items on washing. The German subjects reacted in the opposite direction; they had a lower extreme index than subjects in the other countries.

As Italian and French subjects were included in all three domains, and assessed domain specific attitudes, product specific attitudes, and general values, we focused on them to further investigate the effect of behavioral domain on response style. Results for the yeasaying and the extreme response index in France and Italy are not consistently different for the domain specific attitudes. In the cooking and personal care domain, the indices are higher in Italy than in France, but in the washing domain the indices are equal. For product specific attitudes in the cooking domain, the four item sets do not show consistent differences. As can be seen in Table 4-1, the yeasaying index is higher for the item sets on spreading than for the item sets on making pastry. This may well reflect an actual difference in consumer behavior. According to a survey in the food domain[14], about 87% of the French population used some product for spreading, against only 48% in Italy. As a consequence, subjects in France could be more positive about using a product for

[14] Unilever Kitchen Studies 1990; Unilever Research, Vlaardingen, The Netherlands

spreading on bread or toast. Finally, for the general values, the yeasaying index and the extreme response index are higher in Italy than in France.

4.5.2.4 Response style and response profiles on items

In the previous sections, yeasaying and extreme response indices were compared based on sets of items in different behavioral domains. Results showed some consistent differences between countries. In this section, an additional step is made to rule out the possibility that individual items caused this result. For example, in Greece only two item sets in which 5-point scales were assessed were included in the study. A few items with a very high proportion[15] of the highest scores might have affected our results. To assess this possibility, we focus on the response distributions of the items.

If there is no yeasaying or extreme response style, items with the same mean should have about the same number of endorsed scores on a five-point rating scale. As a consequence, the mean number of scores endorsed in the middle of the scale (2, 3 or 4) can be employed to predict the number of 5s. If there is no response style, then items with a higher mean will have a higher number of highest scores (5) and this result will be independent of country.

Across the three surveys, 75 items were measured on 5-point rating scales. For these items, the mean of the scores on the middle values of the rating scales (2 to 4) was calculated per country. Subsequently, intervals of means were chosen in such a way that about 5 items in each country fell within an interval. Intervals that satisfied this criterion were 2.89 through 2.97, 3.02 through 3.09, 3.24 through 3.29 and 3.41 through 3.45.

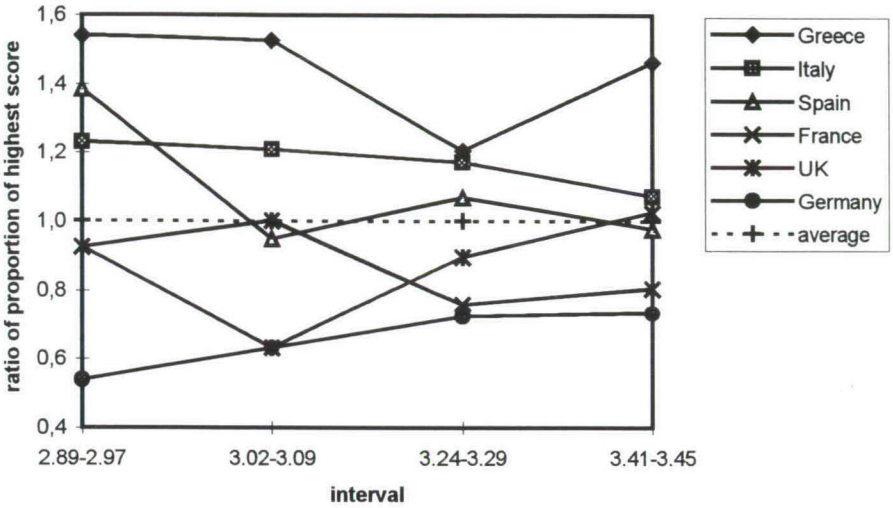
[15] Proportions were taken instead of numbers of observations, because this number differs across surveys. Taking proportions made the data comparable across surveys.

Table 4-5 ANOVA: Proportions of '5' in response patterns

	F	df	p
interval	97.02	3,165	< .001
country	18.03	5,165	< .001
interval by country	1.53	15,165	.103
R^2	.73		

The proportion of 5 was monotonically increasing with interval. This means that items with a lower mean have a lower proportion of 5 and items with a higher mean have a higher proportion of 5s. An Analysis of Variance (ANOVA) revealed that both interval ($F_{(3,165)} = 97.02$, $p < .001$) and country ($F_{(5,165)} = 18.03$, $p < .001$) are important main effects (see Table 4-5). The interaction was not significant. The total proportion of explained variance was high ($R^2 = .73$). The main effect of country indicates that there are differences in response style across countries. The proportion of 5 is significantly higher for the Greek subjects. The average proportion of 5 across the four intervals is 23%. For Greek subjects, this proportion is 36%, i.e. 13% higher. For German subjects, the proportion 5 is 8% lower than the average. For the other countries, these percentages are 6% higher for Italian subjects, 1% higher for Spanish, 4% lower for French, and 6% lower for British subjects. These results imply that there are some notable differences in response style across Europe.

Figure 4-1 Relative proportion of responses in the highest category (5)



In Figure 4-1, an overview is given for all countries of the ratio of 5 within each interval. The value 1.0 indicates the average across the six EU countries in the study; if a country is not different from the overall average, the ratio is 1.0. A ratio of 1.5 indicates that subjects within that country tend to endorse the 5 on the five-point scales 50% more often than average. Comparisons between countries indicate that Greek subjects consistently have the highest proportion of 5, namely about 40% higher across the four intervals. German subjects endorse 5 about 35% less often than average. The other countries score in between, Italy is second highest, followed by Spain, France, and the UK respectively.

In the items employing nine-point scales, only one group of items with a similar mean on the middle points could be identified. Despite the fact that it was only one group, results revealed that the proportion of 9 was highest in Greece. Based on these results, we infer that the higher proportion of extreme values reported in the previous sections cannot be ascribed to a few items only, but that there are systematic differences in response styles between EU countries.

4.6 Discussion and Implications

In this study, response style as reflected in a yeasaying and an extreme response index was investigated in six countries in the European Union. An important conclusion is that Greek subjects tended to have a higher yeasaying and a higher extreme response index than subjects in the other EU countries. The scores for Greek subjects were higher, independent of rating scale type and behavioral domain. This higher yeasaying tendency was also demonstrated when items with a similar mean in the middle range of rating scales were examined for the proportion of extreme positive ratings. The proportion of extreme rating points was consistently the highest in Greece. Thus, we conclude that Greek subjects have a higher tendency to yeasaying and extreme response than German, British, French, Italian, and Spanish subjects. This result is compatible with earlier findings in the literature where Greek subjects had a higher mean than British and Belgian subjects (Steenkamp & Baumgartner, 1998).

Spanish and Italian subjects tended to have about the same yeasaying and extreme response indices. The yeasaying and extreme response index in both countries was higher than for the British, German, and French subjects. German, British, and French subjects had about the same score both on the yeasaying and on the extreme response index. If anything, German subjects tended to be the most extreme response avoiding of all EU subjects, but differences within Western Europe were small.

Results on Italy and France were obtained from all surveys. In general, Italian subjects tended to have higher yeasaying and extreme response indices than French subjects. In the item sets on general values, domain specific attitudes concerning personal care, and attitudes concerning cooking the Italian score is higher. On the attitudes concerning washing, no consistent differences were found. Product specific attitudes concerning cooking showed equivocal results. The influence of the specific content of the items rather than response style differences appears to have determined the outcome. Thus, differences in scores are not independent of behavioral domain across all countries. This finding is not remarkable, as large differences in cooking behavior and product usage are known in the EU (Grunert, Baadsgaard, Larsen & Madsen, 1996).

In general, our findings indicate that there are tendencies to yeasaying and extreme response behavior in EU countries. Our results show that the differences and

similarities found between countries are consistent across two rating scale types and more often than not consistent across the three behavioral domains. This consistency in results means that there is method bias due to response styles in rating scale use in European countries. However, this consistent result does not imply that the differences in rating scale scores are entirely due to such bias effects. For example, it is unclear whether all differences in scores such as between Greece and the other countries are attributable to bias. But we have to conclude that a substantial part of the differences in scores is attributable to bias.

Ignoring differences in response styles may lead to unintended consequences. Take, for example, studies to predict sales. In these studies, it is common to employ the level of the scores or the 'top box' of an item as an indication of the purchase intention (Morwitz, Steckel & Gupta, 1997). Building on our results, researchers would infer that Greek subjects are much more likely to purchase products than, for example, German subjects are, if no correction for response styles are made. After correction for response styles, this inference might be the opposite.

In conclusion, researchers in theory as well as practice need to think about differences in response styles when assessing pan-European marketing research information. Response styles such as yeasaying and extreme response style are by no means a myth (cf. Rorer, 1965), or a negligible source of variance (cf. Nunnally, 1978). In our study, they proved to be a substantial source of variance.

Appendix 4.1. Mean inter-item correlations attitudes

Domain	Items		Number of items	Greece Mean (sd)	Italy Mean (sd)	France Mean (sd)	Germany Mean (sd)	The UK Mean (sd)	Spain Mean (sd)
Cooking	Domain specific		19	.05 (.10)	.04 (.13)	.03 (.16)			
	Product specific	spreading - butter	8	.20 (.13)	.22 (.12)	.26 (.18)			
	Product specific	spreading -margarine	8	.26 (.11)	.29 (.24)	.31 (.23)			
	Product specific	pastry - butter	8	.26 (.16)	.24 (.16)	.21 (.19)			
	Product specific	pastry - margarine	8	.33 (.16)	.27 (.18)	.34 (.19)			
Washing			11	.09 (.09)	.05 (.14)	.07 (.14)	.10 (.17)	.05 (.17)	.10 (.13)
Personal care									
	Domain specific electric	shaving: emotional	12		.36 (.11)	.37 (.09)	.49 (.13)	.40 (.13)	.39 (.12)
	Domain specific electric	shaving: sensorial	10		.47 (.14)	.37 (.10)	.43 (.11)	.42 (.12)	.46 (.13)
	Domain specific electric	shaving: result	8		.45 (.15)	.42 (.11)	.48 (.13)	.43 (.12)	.46 (.12)
	Domain specific electric	shaving: convenience	7		.32 (.11)	.25 (.11)	.36 (.07)	.46 (.11)	.31 (.13)
	Domain specific electric	shaving: irritation	4		.56 (.06)	.40 (.11)	.43 (.20)	.44 (.10)	.48 (.09)
	Domain specific electric	shaving: process	4		.29 (.08)	.20 (.18)	.28 (.10)	.26 (.13)	.32 (.22)
	Domain specific blade	shaving: emotional	12		.37 (.12)	.35 (.07)	.48 (.12)	.43 (.13)	.38 (.11)
	Domain specific blade	shaving: sensorial	10		.41 (.09)	.39 (.09)	.46 (.09)	.40 (.10)	.45 (.09)
	Domain specific blade	shaving: result	8		.37 (.11)	.35 (.12)	.42 (.12)	.39 (.15)	.38 (.10)
	Domain specific blade	shaving: convenience	7		.30 (.08)	.27 (.08)	.30 (.08)	.31 (.08)	.29 (.11)
	Domain specific blade	shaving: irritation	4		.47 (.08)	.40 (.08)	.46 (.18)	.42 (.14)	.47 (.07)
	Domain specific blade	shaving: process	4		.28 (.09)	.27 (.12)	.21 (.14)	.27 (.09)	.26 (.12)
	General values	LOV	9		.41 (.15)	.29 (.15)	.24 (.15)	.57 (.13)	.39 (.12)

Chapter 5. Antecedents of Response Style in Regions in the EU

5.1 Introduction

The globalization of businesses has confronted marketing researchers with the question whether research results from different countries can be compared. Differences in response style (see Chapter 4) appear to contaminate the comparability of outcomes, because subjects from one country assess attributes systematically more positively or more extremely than subjects from other countries, regardless of item content. Finding the antecedents of differences in response styles is an essential step obtaining data that are comparable across countries. Antecedents of differences in response style may be in the environment people live in or in the subjects themselves, such as in their socio-demographic characteristics. In this chapter, antecedents of response styles will be explored for some 55 regions in 5 countries within the European Union

5.2 Differences in response style between countries

Differences in response style between countries have been mainly investigated in student samples (e.g., Chen, Lee & Stevenson, 1995; Chun, Campbell & Yoo, 1974; Soueif, 1968). Other studies employed students as well as employees (Yeh, Kim, Chompreeda, Rimkeeree, Yau & Lundahl, 1998).

These studies were mainly done in the US and Asia (China, Korea, and Thailand). A study by Soueif (1968) is an exception; it was done in Egypt, Syria, and Jordan. In the EU, there are few studies that assess response style. One study (Williams,

1991) is anecdotal and reports that Italian subjects have a higher mean on rating scales than German subjects. In another study (Steenkamp & Baumgartner, 1998), Greek subjects were found to have a higher mean than British and Belgian subjects on the CETSCALE (Shimp & Sharma, 1987). However, in this study, response styles were not explicitly assessed. In this thesis (Chapter 4), evidence for differences in response style was found between Greek, Italian, Spanish, and Western-European subjects from France, Germany, and the UK. Greek subjects had the highest yeasaying score, followed by Italian and Spanish subjects; the lowest yeasaying scores were found in Western Europe.

Studies on response style employing representative samples of subjects of the total population in different countries are scarce. The first study assessing response styles in such samples is by Baumgartner and Steenkamp (1999). They reported differences in response styles between 11 EU countries. However, they did not mention which countries systematically scored higher or lower.

5.3 Regional differences within countries

Thus far, studies across countries have been considered. An implicit assumption in these studies is that response style is a homogeneous characteristic of all subjects within one country. However, many countries are not homogeneous; there are regional differences within countries (Douglas & Craig, 1997). In large countries such as the US and China, there are geographical subcultures and even for smaller countries in Europe such as the UK, Italy, and Spain, a North-South difference has been reported (Usunier, 1996). In general, differences in consumer behavior are expected to be smaller within countries than between countries, but differences in consumer behavior within countries may not be negligible. For example, a study on European food customs (Askegaard & Madsen, 1998) reveals that most countries have homogeneous food customs (e.g., the UK), but there are also countries with heterogeneous food customs (e.g., Germany). To the best of our knowledge, regional differences in response style have not been investigated in the EU. A purpose of the present study is to assess the extent to which regional differences in response style within the EU do occur and how far they may be explained by the country a region is located in. We also look at North-South differences.

5.4 Socio-demographic differences in response style

Differences in rating scale data have been reported for various demographic groups distinguished by age, household income, or education level (e.g., Hamilton, 1965; Krosnick & Alwin, 1988; Watson, 1992; Greenleaf, 1992a,b). In his review, Hamilton (1965) already reported accumulated evidence that children and elderly subjects gave more extreme responses than subjects aged 20-59. Later studies (e.g., Greenleaf, 1992a,b) support this finding. As extreme response style, yeasaying is positively related to age. Older subjects tend to use the positive side of the rating scale more often or display more acquiescence (Winkler, Kanouse & Ware, 1982; Moum, 1988; Greenleaf, 1992a,b).

Research indicates that education is negatively associated with yeasaying (Greenleaf, 1992a,b; Moum, 1988; Krosnick & Alwin, 1988; Winkler Kanouse & Ware, 1982; Watson, 1992; Narayan & Krosnick, 1996) and with extreme response style (Greenleaf, 1992a,b; Stening & Everett, 1984). Like education, household income is found to be negatively associated with yeasaying and extreme response style (Greenleaf, 1992a,b). Regarding gender, there is contradictory evidence. In Greenleaf (1992a), men are reported to have a higher yeasaying score, but in Moum (1988), women have a higher yeasaying score. In Watkins and Cheung (1995) and in Marín, Gamba and Marín (1992), no significant differences for gender were found. Also, no significant differences for gender were found in the study by Greenleaf (1992b) on extreme response style. Based on these findings, we expect that yeasaying and extreme response style are positively related to age if subjects are 20 years of age or older. In addition, we expect that both response styles are negatively related to education and household income.

5.5 Method

5.5.1 Subjects

Data in this study are from a survey on Personal Care[16] held in 1996 in five EU countries, namely the UK, France, Germany, Italy, and Spain. This data set is also used in Chapter 4; there, only men aged 20 to 60 were included in the analyses. In this chapter, no selection of subjects was made and all men aged 15 to 90 were included. Mean ages in the complete data set were 45.0 in Germany, 43.3 in the UK, 40.8 in Spain, 44.3 in Italy, and 43.4 in France.

5.5.2 Measures

In the survey, a mail questionnaire was employed to collect information on domain specific attitudes and socio-demographic variables. The study included 45 items on attitudes towards shaving. For further details on the items, we refer to Section 4.5.1.3.2.

The socio-demographic information included age, education level, and household income. Age was measured in years. Education level was measured on a 4-point rating scale (1 = 'left full time education at 15/16 years old', 2 = 'left full time education at 17/18 years old', 3 = 'obtained advanced level' and 4 = 'University studies'). For the purposes of this chapter, education level was recoded into two categories; '0' stands for low education (combining rating scale scores 1 and 2) and '1' stands for high education (combining rating scale scores 3 and 4). The scores were recoded, because the rating scale scores 3 and 4 did not have the same meaning in France as in the other countries[17]. Monthly (gross) household income was measured on a five-point rating scale in the local currencies. We re-calculated the income rating scale to the equivalence in Euro[18], so these scales were the same

[16] Philips DAP Groningen, The Netherlands; data collected by GfK/IHA Lausanne, Switzerland.

[17] In France, the fourth category was 'supérieur ou université'; this implied that subjects having no university degree also endorsed this category.

[18] The Euro was introduced on January 1st 1999; at the time of the data collection, the ECU (European Currency Unit) was used.

across countries. The rating scale categories were 1 = 'Less than Euro 750', 2 = 'Euro 751 - 1500', 3 = 'Euro 1501 - 2250', 4 = 'Euro 2251 - 3000', 5 = 'Over Euro 3000'. For the purposes of this chapter, monthly household income was recoded into two categories 0 and 1; '0' stands for low income (combining rating scale scores 1, 2 and 3) and '1' stands for high income (combining rating scale scores 4 and 5). Scores were recoded to avoid rating scale categories with few observations.

A variable indicating the 55 regions subjects live in was included in the dataset. An overview of the regions is given in Appendix 5.1. In addition, the latitude and longitude of each region was added to the dataset. Subjects with missing values either on the domain specific attitude items or on the socio-demographic information were excluded from analyses. Final sample sizes were 666, 729, 768, 1014, and 567 in the UK, France, Germany, Italy, and Spain, respectively.

5.6 Results

In the Personal Care survey in Chapter 4, six Principal Components for the group of blade shaving men and six Principal Components for the group of electric shaving men were found. We started with assessing whether the component structure for blade and electric shaving is the same for these two pan-European Principal Components Analysis (PCA) solutions. The components for blade shaving men as well as for electric shaving men were: 'emotional', 'sensorial', 'result', 'convenience', 'no irritation,' and 'process' (see also Chapter 4). To determine whether there is factorial agreement between the PCA solutions for blade and electric shaving, Procrustean rotation was done. This is one of the techniques for establishing construct equivalence (Van de Vijver & Leung, 1997a). Our results after Procrustean rotation show that five out of six identity coefficients (Zegers & Ten Berge, 1985) were higher than .90. This is a value below which incongruities are non-negligible (Van de Vijver & Leung, 1997a). For the first five components, the average identity coefficient is .96 (range .95 to .98). For the component 'process', the identity coefficient is .89. Based on this result, we decided to drop the latter component, and to combine the data for blade and electric shaving for the first five components into the remainder of this chapter.

In total, five indices for yeasaying (i.e. one for each component) and five indices for extreme response were available for the analyses. We assessed whether these

indices could be added to form a composite measure for yeasaying and extreme response respectively. PCA of the five yeasaying indices revealed one component with one dominant eigenvalue, accounting for 58.5% of the total variance. The factor loadings ranged from .87 for 'sensorial' to .69 for 'no irritation'. For extreme response, PCA on the five extreme response indices revealed one dominant eigenvalue, accounting for 63.9% of total variance. The factor loadings range from .88 for 'sensorial' to .70 for 'emotional'.

Based on these results, we decided to calculate one overall index for yeasaying, and one overall index for extreme response. The overall indices were the unweighted sum of the five indices divided by 5; this implies that the yeasaying index (as in Chapter 4) ranges from -1.0 to 1.0, and that the extreme response index ranges from 0.0 to 1.0. Cronbach's alpha was .82 and .86 for the composite of yeasaying and extreme response respectively; these reliabilities appear satisfactory (Nunnally, 1978). We call the composites 'YEA' and 'EXTREME' in the next sections.

5.6.1 Multi-level analysis

Our analysis of response styles involved two levels of analysis, subjects, and regions. Subjects and regions have a hierarchical relationship, because lower level observations (subjects) are nested within higher levels (regions). An appropriate technique to take the nested structure of data into account is multi-level modeling (e.g., Bryk & Raudenbush, 1992; Kreft & De Leeuw, 1998). In multi-level modeling, there is the notion that the relationships between variables at the lowest level are not necessarily the same for each higher level. For example, the relationship between education level and yeasaying may be different across regions. An interesting property of multi-level analysis is that variables can be used to explain variance at the subject level and that other variables can be used to explain variance at the region level. This provides insight into reasons why regions may differ. The latter is not possible in other forms of analysis, such as analysis of covariance (ANCOVA).

A (statistical) property of multi-level modeling is that it takes into account that subjects within one region may be more alike than subjects in different regions. A measure for the dependency within groups is called intra-class correlation (Kreft & De Leeuw, 1998). If subjects are more alike, observations are, from a statistical point of view, less independent. Multi-level modeling adjusts for dependencies within groups and thus prevents that we consider effects to be significant when

they are not. This is a non-negligible issue. For instance, if an intra-class correlation is .05 (in groups of about 100 subjects) the probability of wrongly rejecting the null hypothesis is inflated from the assumed 0.05 to 0.43 (Barcikowski, 1981).

For the subsequent analyses, we used the computer program Hierarchical Linear Modeling, HLM (Bryk & Raudenbush, 1992; Bryk, Raudenbush & Congdon, 1996). In the data, the intra-class correlation is $\rho = .22$ for YEA and $\rho = .19$ for EXTREME, respectively. This implies that subjects within one region show similarities, and that using a technique like ANCOVA would have resulted in a too high estimate of the amount of variance explained by differences between regions.

5.6.2 Individual and regional antecedents of response style

We started with the means-as-outcomes regression model (Bryk & Raudenbush, 1992). In this model, intercepts are random and the slopes of the regression lines are fixed. Later in this chapter, it will be tested whether making the slopes of the regression lines random improves the fit of the model. At the individual (subject) level, the following model was used:

$$(1)\quad RSINDEX_{ij} = \beta_{0j} + \beta_1\, AGE_{ij} + \beta_2\, EDUC_{ij} + \beta_3\, INC_{ij} + \beta_4\, METHOD_{ij} + r_{ij}$$

where $RSINDEX_{ij}$ is the response style index (YEA or EXTREME), AGE is age measured in years, EDUC is education (0 and 1, for low and high respectively), and INC is income (0 and 1, for low and high respectively). METHOD is shaving method (0 is blade shaving, 1 is electric shaving), β_1 to β_4 are regression coefficients, and r_{ij} is an error term. METHOD is included in the model, because differences in the level of the scores may still be present for subjects assessing blade or electric shaving respectively. A similar component structure does not mean that levels of scores are similar; structural equivalence does not imply measurement unit equivalence (see Chapter 2).

Differences in response style at the region level are predicted by the country to which they belong and by the location on the globe. Thus, at the region level the following two models were used:

$$(2a)\quad \beta_{0j} = \gamma_{00} + \gamma_{01}\, FRANCE_j + \gamma_{02}\, GERMANY_j + \gamma_{03}\, SPAIN_j + \gamma_{04}\, ITALY_j + u_{0j}$$

$$(2b)\quad \beta_{0j} = \gamma_{00} + \gamma_{01}\, LATITUDE_j + \gamma_{02}\, LONGITUDE_j + u_{0j}$$

In equation (2a), $FRANCE_j$, $GERMANY_j$, $SPAIN_j$, and $ITALY_j$ indicate whether or not a region is situated in these countries[19]. They are all binary variables (code 0 and 1, where 1 indicates that a region is located in the specific country). This coding implies that the outcomes at the regional level are all relative with respect to the UK. In equation (2b) $LATITUDE_j$ and $LONGITUDE_j$ indicate the position of each region on the globe[20].

In both equations (2a) and (2b), γ_{00} is the intercept. The γ_{01}, γ_{02}, γ_{03}, and γ_{04} are the effects of the respective variables on the response indices and u_{0j} is an error term. The coefficients are all unstandardized regression coefficients, consistent with the HLM approach (Bryk & Raudenbush, 1992).

5.6.2.1 Socio-demographics and country

The results for equations (1) and (2a), for both YEA and EXTREME, are given in Table 5-1. The effects of the demographic variables on YEA and EXTREME are as expected. Age has a positive effect on yeasaying ($\beta_1 = .003$, $p < .001$); education and income have a negative effect ($\beta_2 = -.050$, $p < .001$ and $\beta_3 = -.040$, $p < .001$). Age also has a positive effect on extreme response ($\beta_1 = .002$, $p < .001$); education and income have a negative effect ($\beta_2 = -.041$, $p < .001$ and $\beta_3 = -.018$, $p < .05$). Yeasaying is affected by shaving method ($\beta_4 = -.023$, $p < .05$). This significant coefficient means that subjects who shave electrically display less yeasaying than subjects who shave with a blade. Extreme responding is not affected by shaving method. The socio-demographic variables were effective in accounting for 2,9% and 1,2% of the individual variance in YEA and EXTREME respectively.

---

[19] The UK is not included. The dummy coding for the other countries implies that a region where $FRANCE_j$, $GERMAN_j$, $SPAIN_j$, and $ITALY_j$ are all 0 is part of the UK. Setting the UK to 0 is arbitrary.

[20] The global coordinates of each region were determined by taking the coordinates of the center of each region. In doing so, the coordinates as given in The New Oxford Atlas (Oxford University Press, 1978) were used.

Table 5-1 YEA and EXTREME: country & socio-demographics

	YEA			EXTREME		
	Coefficients	T-value	p	Coefficients	T-value	p
Intercept (γ_{00})	-.006	-.25	.81	.285	15.08	<.001
Main effects: Region level						
France (γ_{01})	.062	2.47	<.05	.012	.62	.54
Italy (γ_{02})	.152	6.52	<.001	.106	5.67	<.001
Germany (γ_{03})	.006	.24	.81	-.005	-.23	.82
Spain (γ_{04})	.095	3.57	<.01	.078	3.69	<.01
UK (by definition)	*0*	*0*		*0*	*0*	
Main effects: Individual level						
Age (β_1)	.003	11.36	<.001	.002	5.98	<.001
Education (β_2)	-.050	-5.27	<.001	-.041	-4.71	<.001
Income (β_3)	-.040	-4.26	<.001	-.018	-2.09	<.05
Method (β_4)	-.023	-2.35	<.05	.016	1.85	.07

Table 5-1 also shows that yeasaying is positively associated with a region in Italy ($\gamma_{02} = .152$, $p < .001$), Spain ($\gamma_{04} = .095$, $p < .01$), and France ($\gamma_{01} = .062$, $p < .05$). A region in Germany is not significantly different from a region in the UK. The results for extreme response show a positive association between being located in Italy ($\gamma_{02} = .106$, $p < .001$) and Spain ($\gamma_{02} = .078$, $p < .01$); being located in France or Germany is not significantly different from being located in the UK. Country was effective in accounting for 42.3% (YEA) and 42.9% (EXTREME) of variance in the response style indices across regions in the EU.

5.6.2.2 Socio-demographics and global coordinates

Other variables, such as whether a region is located more in the South or in the North of Europe, may also be associated with yeasaying or extreme response style. The results of the HLM model, including latitude and longitude, are presented in Table 5-2. The effects of the demographic variables on YEA and EXTREME are similar to the results of Table 5-1. This is expected, as the variables at the individual level were the same as in the previous analysis. Results at the region level show that the global coordinates of a region can affect YEA and EXTREME. If a region is located further in the South (= lower latitude), then more yeasaying and extreme response style has occurred. The East-West distinction in location of a region does not noticeably affect response behavior. Longitude and latitude of a region were effective in accounting for 25,7% and 23,0% of variance at the regional level in YEA and EXTREME respectively. This is less than the amount of variance accounted for when country was used to explain regional differences.

Of course, there is an overlap between both analyses. For example, all regions in Italy are located more to the South, than regions in Germany and the UK; all regions in the UK are located more to the West than regions in Germany. When equations (2a) and (2b) were combined, the amount of variance explained at the regional level was increased to 44,9% (YEA) and 44,5% (EXTREME). Thus, the gain in combining regional location and country is very limited, compared to country alone (42.3% and 42.9%, respectively).

Table 5-2 YEA and EXTREME: global coordinates & socio-demographics

	YEA			EXTREME		
	Coefficients	T-value	p	Coefficients	T-value	p
Intercept (γ_{00})	.058	3.50	< .01	.324	22.24	< .001
Main effects: Region level						
Longitude (γ_{01})	.003	1.89	.06	.002	1.36	.18
Latitude (γ_{02})	-.008	-4.21	< .001	-.007	-4.18	< .001
Main effects: Individual level						
Age (β_1)	.003	11.41	< .001	.002	6.01	< .001
Education (β_2)	-.051	-5.30	< .001	-.042	-4.81	< .001
Income (β_3)	-.038	-4.00	< .001	-.017	-1.99	< .05
Method (β_4)	-.025	-2.57	< .05	.015	1.67	.10

In this study, shaving method, referring to a subject's behavior, was included at the individual level. For YEA, it proved a significant addition to the model (see Table 5-1 and Table 5-2). Men who shave with an electric shaver engage in less yeasaying. It should be noted that this might also be a real difference in attitude: men who use a blade may be more positive about shaving than men who use an electric shaver. The latter proposition was further investigated by fitting a HLM model to separate response style indices for the components 'emotional', 'sensorial', 'result', 'convenience' and 'no irritation'. Results for yeasaying showed that the coefficient for method differed per component. For 'emotional', 'sensorial', and 'result', it is negative (-.067, -.138, -.189; all, $p < .001$), indicating a higher score if a man used a blade; for 'convenience' and 'no irritation', the coefficient is positive (.144 and .133; both, $p < .001$). These results point to a real

difference[21], as 'convenience' and 'no irritation' are the aspects that distinguish shaving with a blade from shaving with an electric shaver.

5.7 Conclusions and discussion

In this chapter, differences in response style indices were explained at the level of individual subjects and at the level of regions within countries in the European Union. At the individual level, response style is positively affected by age and negatively affected by household income and education. These effects are in the expected directions. In addition, the coefficients for age are about equal in size to coefficients found in other studies on response style (Greenleaf, 1992a; Baumgartner & Steenkamp, 1999). Measures for education and income are not collected in the same way as the measures used in other studies, so we cannot compare them directly in size.

Our results imply that an increase in age of 30 years increases the yeasaying index by .09 (= 4.5%) and the extreme response index by .06 (= 6%). If education and income are combined, the differences in the yeasaying and extreme response indices increase even further. For example, compare two subjects using the same shaving method on yeasaying. One subject is 30 years old with a high education and a high income, and the other subject (in the same of region) is 60 years old with a low income and a low education. The younger subject has an expected yeasaying index of about 0.00, the older subject has an expected yeasaying index of about .18 (= 9% higher). In using 5-point rating scales, this means that the younger subject has an average rating scale score of about .50 less than the older subject. These effects are non-negligible.

At the regional level, the variables indicating the country a region was located in were effective in accounting for about 42% of variation between regions. These results imply that the country is important in explaining differences in response styles. However, the percentage of explained variance also implies that regional differences should not be ignored. For instance, the global coordinates of a region, measured by latitude and longitude, explained some additional variance at the region level (2,5%). It is likely that there are additional variables that can be used

[21] Internet; Electric shaver page, 1999; http://iavbbs.com/gflinn/INDEX.HTM

to explain differences in response style between regions. For instance, there may be differences rural and urban regions or between richer and poorer regions within the respective countries (Edye & Lintner, 1996). The data used in this study do not have sufficient detail to investigate the effects of grade of urbanization or wealth on response style indices. Some regions in the dataset cover a large area, including a metropolitan as well as its rural surroundings. For example, in Italy Trentino-Alto Adige includes Turin and a rural area.

It should be noted that the amount of variance that was explained at the individual level is far less than the amount of variance explained at the region level. In addition, total variance was considerably higher at the individual level than at the region level for both YEA and EXTREME. Variance at the region level accounted for approximately 15% of total variance in the indices. Thus, further research should particularly focus on additional antecedents of response style at the individual level. These antecedents may be in a subject's personality (e.g., Hamilton, 1968; Knowles & Nathan, 1997), or in his/her behavior. In this study, shaving method, referring to a subject's behavior, was included at the individual level. For YEA, it proved a significant addition to the model. Men using an electric shaver displayed less yeasaying.

Appendix 5.1. Regions within the countries

France	Region Paris, North, East, Parisian Basin: East, Parisian Basin: West, South East, Mediterranean, West, South West
The UK	Scotland, Greater London, South East, South West, East Midlands and East, West Midlands, North West, Yorkshire & Humberside, North England, Wales
Italy	Piemonte-Valle d' Aosta, Lombardy, Liguria, Trentino - Alto Adige, Veneto-Firiuli VG, Emilia Romagna, Toscana, Marche-Umbria, Lazio, Abruzzi-Molise, Campania, Sicily, Sardinia
Spain	Barcelona Metropolitan, North East, East, South, Madrid Metropolitan, Center, North East, North West
Germany	Mecklenburg-Vorpommern, Brandenburg, Sachsen-Anhalt, Thüringen, Saxony, Berlin East, Berlin West, Schleswig-Holstein, Lower Saxony, North Rhein-Westphalia, Rhineland-Palatinate & Saarland, Hessen, Baden-Württemburg, Bavaria

Chapter 6. Rating Versus Ranking In Cross-National Research

6.1 Introduction

In Chapters 4 and 5, effects of response style in cross-cultural data sets were demonstrated. In spite of the fact that rating scales are susceptible to response styles, they are used frequently in international marketing research. They are used to measure constructs such as attitudes, preferences, and values; examples are measures of product category interest and familiarity (Dawar & Parker, 1994), attitudes toward advertising (Durvasula, Andrews, Lysonski & Netemeyer, 1993), fairness and relationship quality (Kumar, Scheer & Steenkamp, 1995), proficiency, commitment, integration, and differentiation in new product development (Song & Parry, 1997), and exploratory buying behavior (Baumgartner and Steenkamp, 1996; Steenkamp, Ter Hofstede & Wedel, 1999).

In Chapters 4 and 5, only rating scales were used. A main reason for using rating scales is their convenience. For instance, Munson and McIntyre (1979) mention that rating scales are used for measuring values and attributes instead of rankings, because they: (1) allow for ties between items; (2) are easy to administer; (3) are less time-consuming; and (4) are less difficult for respondents. However, ratings have two potential drawbacks (Alwin & Krosnick, 1985): (1) they may reduce respondents' willingness to make precise distinctions, and (2) they are susceptible to problems of response style. These drawbacks are non-negligible as the amount of variance due to response styles can be substantial (see e.g., Chapter 4).

Several procedures for correcting differences in response style have been proposed. For instance, it has been suggested that variance due to response styles be removed

through a correction procedure such as within-subjects standardization, also known as ipsatizing (e.g., Douglas & Craig, 1983; Helsen & Kotabe, 1998). Others suggest another measurement procedure: ranking (e.g., Kamakura & Mazzon, 1991). Subjects are asked to rank items in terms of relative importance, thereby avoiding rating scales entirely (Smith & Schwartz, 1997). In strict ranking, it is required that one item is considered the most important, another the second most important and so on, until the last item that is considered the least important. It is required that no two items are considered equally important; each item should be given its own unique score. Thus, a subject is forced to discriminate between items; no ties are allowed. Ranking is an alternative to rating if the items can be ordered by importance relative to one another. This holds, for example, in studies on brand preference (e.g., Dillon, Madden & Firtle, 1994), or studies on values (e.g., Rokeach, 1973; Kahle, 1983). However, the appropriateness of those procedures for correcting differences in response styles has hardly, or not at all, been addressed in previous cross-national research.

In the present chapter, we focus on the appropriateness of ranking given the existence of response style in rating scales. As the study of values has been a key theme in cross-national research in the last decades (e.g., Douglas & Craig, 1997), values studies are emphasized. The data we employ are scores on the List of Values (LOV, Kahle, 1983); this is a values scale for which it is explicitly stated that items can be either rated or ranked (e.g., Bearden, Netemeyer & Mobley, 1993). A sample of about 4500 male subjects from five countries in the European Union, namely Italy, Spain, France, Germany, and the UK (see also data set in Section 4.5.1.3) is used.

The purpose of this chapter is fourfold. First, we investigate whether response behavior in rating and ranking of values is the same across five countries in the European Union. Second, we investigate whether specific prototypical response patterns, i.e. patterns derived from certain response styles in ratings, occur and whether these can be explained by differences in socio-demographic characteristics. Third, we assess whether these response patterns and socio-demographic characteristics affect the correlations between ratings and rankings of the nine LOV items. Finally, we investigate whether or not the predictive validity of ratings or rankings of the LOV is the same and whether this result is related to the response patterns that subjects have displayed.

6.2 Values: rating or ranking?

Most researchers agree that the main features of values are: (1) values are beliefs, (2) values refer to desirable goals and to the modes of conduct that promote these goals, (3) values transcend specific actions and situations, (4) values serve as standards to guide the selection or evaluation of behavior, people, and events, and (5) values are ordered by importance relative to one another (cf. Smith & Schwartz, 1997, p. 80). Especially the latter feature emphasizes the possibility to measure value orderings with a ranking as well as with a rating procedure.

Rating scales and rankings are both used to measure values. In the Rokeach Value Survey (RVS, Rokeach, 1973), ranking is the most common procedure (e.g., Kamakura & Mazzon, 1991). In the Schwartz value survey, rating scales have been used (Schwartz, 1992), and with the List Of Values (LOV, Kahle, 1983), ranking (e.g., Kamakura & Novak, 1992) as well as rating (e.g., Homer & Kahle, 1988; Grunert, Grunert & Beatty, 1989; Grunert, Grunert & Kristensen, 1993; Goldsmith, Freiden & Kilsheimer, 1993) has been applied.

There is no consensus about what method is preferred for studying values in a cross-national context. Some argue that ranking is the most appropriate (Kamakura & Mazzon, 1991); others argue that rating should be preferred (Ng, 1982; Schwartz, 1992).

Both rating scales and rankings have advantages and disadvantages. Rating has an advantage if a subject considers two or more items equally important. On the other hand, using a rating scale may invoke response bias such as yeasaying (e.g., Greenleaf 1992a; Chapter 4 and 5 this thesis). Yeasaying may even result in a situation where a subject endorses all items as being '(very) important'. In such situations, a ranking in which a subject is forced to discriminate between items is preferred. Ranking has an advantage if a subject has a hierarchically ordered picture of the items in his/her mind. On the other hand, rankings do not provide valid results if a subject considers several items as equally important. In that case, assigning a forced unique score to each item increases measurement error in the data (see e.g., Barnard & Ehrenberg, 1990).

6.2.1 Studies comparing ratings and rankings

Several comparisons have been made between ratings and rankings. The majority of studies on rating and ranking are between-subject designs, in which a subject either rated or ranked the items (Barnard & Ehrenberg, 1990; Alwin & Krosnick, 1985; Krosnick & Alwin, 1988; Braithwaite & Law, 1985). In these studies, a direct within-subject comparison between ratings and rankings is impossible.

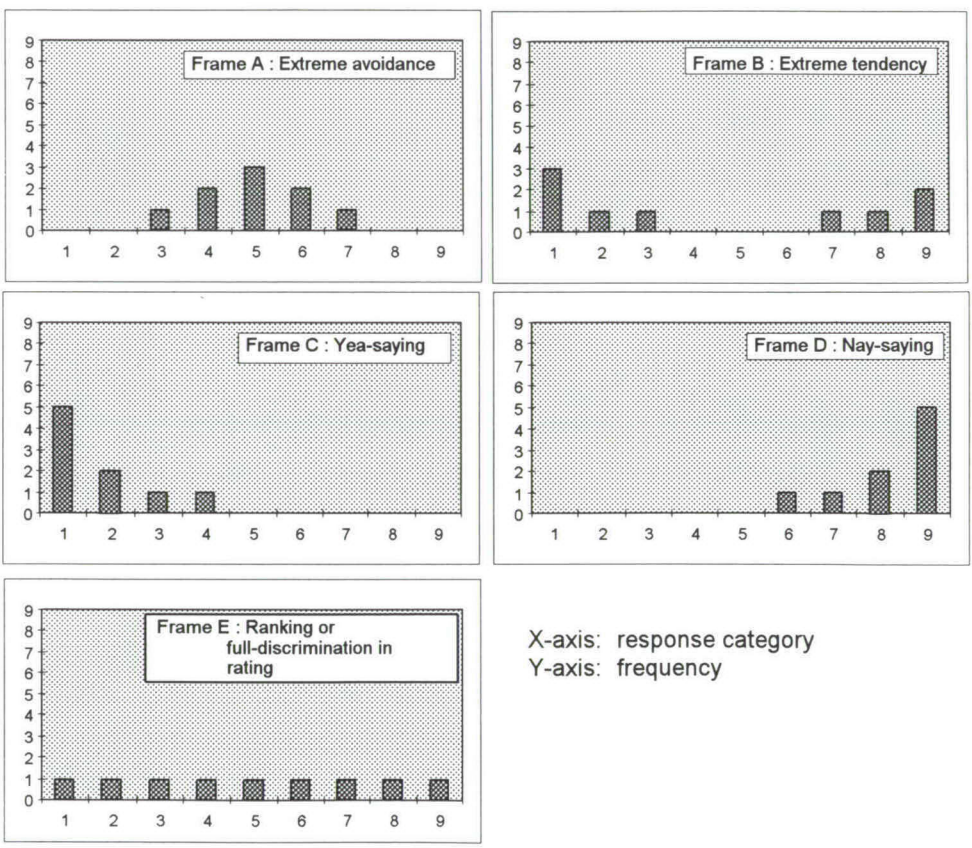
There are a few studies where the same subjects did ratings and rankings (Moore, 1975; Rankin & Grube, 1980; Russell & Gray, 1994; Maio, Roese, Seligman & Katz, 1996). Unfortunately, the results are not consistent regarding which method should be preferred. Moore (1975) sees advantages in ratings if data are to be compared after aggregation across individuals. However, if data are to be compared at the individual level, he has no preference. Rankin and Grube (1980) found that reliabilities for rankings were somewhat higher than for the ratings in a test - retest situation. Maio, Roese, Seligman and Katz (1996) prefer ratings to rankings in value research, because they found that subjects rated about 32% of their values equally. Thus, subjects apparently consider several values to be equally important and do not consider finer distinctions, which are imposed by ranking. Finally, Russell and Gray (1994) state that rankings have advantages over ratings if the items are highly discriminable. They found that results of rating and ranking became more alike if respondents were instructed to compare the items with one another while rating. All in all, the literature does not provide unambiguous results as to which measurement procedure should be preferred.

6.3 Prototypical response patterns in rating

In rating items, such as values, a respondent is free to assign a score to each item. If a respondent considers two items equally important, he can give the same score to both items. Thus, in rating it is possible that, in the most extreme case, each item is assigned the same score; for instance, a subject only endorses '5' on a five-point rating scale. Subjects who do not discriminate between items will have a highly peaked response pattern. Other subjects will have a normally distributed response pattern: some items are considered 'important', some 'unimportant', and most items are considered 'neither important nor unimportant'. Due to response style, typical response patterns in ratings may emerge (see e.g., Greenleaf, 1992a; Bijmolt, Wedel, Pieters & DeSarbo, 1998). Such response patterns can be

considered, at least in part, to be a consequence of the yeasaying and extreme response tendencies (see e.g., Bachman & O'Malley, 1984; Chapter 4 this thesis). To determine response patterns, answers given on rating scales are counted, without considering the content of the specific items. For example, on a nine-point rating scale, the frequencies of '1', '2', '3' et cetera, are counted to determine the response pattern for each subject.

Figure 6-1 Prototypical response patterns



In Figure 6-1, prototypical response patterns, based on nine-point rating scales, are presented. Frames A through E present prototypical patterns. In frame A, a prototypical pattern of a person avoiding extremes is presented. The score of '5', which is 'neither important nor unimportant', is most frequently chosen here, whereas the extreme scores '1' and '9' ('very important' and 'very unimportant')

are avoided. Frame B presents a prototypical pattern of a person who typically avoids the middle values of the rating scale and considers items either 'very important' or 'very unimportant'. This frame can be considered the counterpart of frame A. In frames C and D, prototypical response patterns of subjects showing yeasaying (C) and naysaying (D) are shown. Yea-sayers mainly use the positive side of the scale; they consider most or all items 'very important'. Frame D is the counterpart of frame C. Subjects with this pattern consider most or all items 'unimportant'. In Frame E, the prototypical response pattern of ranking is given. A similar pattern might emerge from a rating task if a subject discriminates maximally between the items, and gives each item another score. In this chapter, we explore the occurrence of these patterns in ratings of values.

6.4 Background characteristics

6.4.1 Response patterns

Response style in ratings is known to be different within as well as between countries. Within countries, differences in response style are known to exist between groups, which can be explained by demographic characteristics such as age, household income, and education level (e.g., Hamilton, 1965; Greenleaf, 1992a,b; Narayan & Krosnick, 1996; Chapter 5 this thesis).

As mentioned earlier, the two main response styles, extreme response style and yeasaying, are positively related to age (see e.g., Chapter 5). Older subjects tend to use the positive side of the rating scale more often or display more yeasaying. Education is negatively associated with yeasaying (e.g., Narayan & Krosnick, 1996; Chapter 5) and with extreme response style (e.g., Greenleaf, 1992a,b; Chapter 5). Like education, household income is found to be negatively associated with yeasaying and extreme response style (Greenleaf, 1992a,b; Chapter 5). Based on these findings, we expect that older subjects will be over-represented in the response patterns B (extreme tendency) and C (yeasaying), and under-represented in the other response patterns. In addition, we expect that subjects with a higher income and a higher education will be over-represented in the response patterns A (extreme avoidance) and E (full discrimination).

Response style differences between respondents in different countries have also been reported, such as between the US and Korea (Lee & Green, 1991), between the US, Canada, Japan, and Taiwan (Chen, Lee & Stevenson, 1995) and between Greece, Italy, France, Spain, the UK, and Germany (Chapter 4 this thesis). In the latter study, subjects in Greece tended to display more yeasaying and extreme response style than subjects in the other EU countries. Italian and Spanish subjects tended to display more yeasaying and extreme response style than subjects in the UK, France and Germany. Subjects in these three countries had about the same scores for yeasaying as for extreme response style. Based on these findings, we expect that there will be a relatively high proportion of subjects from Italy and Spain with response patterns B and C. We expect a higher proportion of subjects from the UK, France, and Germany with response pattern A (extreme avoidance).

6.4.2 Correlations between rating and ranking

In studies in which rating and ranking were compared, various kinds of samples have been used, such as students (Moore, 1975; Rankin & Grube, 1980; DeCasper & Tittle, 1988; Maio, Roese, Seligman & Katz, 1996), housewives (Barnard & Ehrenberg, 1990), and random samples of citizens throughout a country (Alwin & Krosnick, 1985; Krosnick & Alwin, 1988; Russell & Gray, 1994). In these studies, hardly any attempt has been made to explain differences in the correlations between rating and ranking in terms of these background characteristics. An exception is the study by Krosnick and Alwin (1988). They found that the results of rating and ranking became more similar if the subjects who rated all items the same were removed from the data. This response behavior occurred more often among respondents with less education. Based on these results, we expect that higher correlation between ratings and rankings will be found for subjects with more years of education. Higher correlations will generally result from response patterns in which subjects endorsed more different response categories. These are response patterns A, B and E. The lowest correlation is expected for subjects displaying either yeasaying (Frame C) or naysaying (Frame D) (see Figure 6-1).

For background characteristics such as age, household income and country of residence, there is no previous research proving direction for the magnitude of the correlation between ratings and rankings. Hence, no hypothesis could be formulated.

6.5 Predictive validity

In many research projects, values are employed as antecedents to other measures such as attitudes and behaviors (Homer & Kahle, 1988). In cross-national research, values have been related to measures such as domain-specific attitudes and product benefits (Kamakura, Novak, Steenkamp & Verhallen, 1994), self-reports of fashion leadership (Goldsmith, Freiden & Kilsheimer, 1993), product involvement scores (Grunert & Muller, 1996) and ethnocentrism and attitudes towards the past (Steenkamp, Ter Hofstede & Wedel, 1998). Hence, predictive validity of the measurement of values is important.

In the literature, few papers assess the predictive validity of ratings and rankings. An exception is the paper by Rankin and Grube (1980), which compared ratings and ranking of the Rokeach Value Survey (RVS, Rokeach, 1973). They found that ratings are a better predictor of attitudes than rankings; however, the differences were only slight. A study by Maio, Roese, Seligman, and Katz (1996) also indicated that ratings of the RVS tend to have a greater predictive validity than rankings. They found that ratings of values correlated higher with measures of attitude and with measures of judgments about the acceptability of various behaviors than rankings of those values.

There is also some evidence regarding the predictive validity of ratings and rankings in relation to response patterns. Maio, Roese, Seligman, and Katz (1996) found that for 'low and moderate differentiating' subjects, ratings have a higher predictive validity than rankings. For 'low differentiators' (i.e. subjects having many ties in their ratings), the forced discrimination induced by rankings is likely to provide no valid results. For subjects, who have few ties in the ordering of the items, the correlation between rating and ranking is expected to be higher (e.g., Krosnick & Alwin, 1988). Consequently, for those subjects the predictive validity of both measures is expected to be more similar. Based on these results, we expected that the predictive validity of ratings and rankings will be more similar for subjects with response patterns resembling patterns A, B or E (see Figure 6-1).

6.6 Method

Data in this study are from a commercial survey on Personal Care[22] held in 1996 in five EU countries, namely the UK, France, Germany, Italy, and Spain (see also Chapters 4 and 5). In this chapter, we used all subjects in the dataset, no a-priori selections were made.

6.6.1 Subjects

In the following analyses, subjects with missing values on either ranking or rating were excluded. Final sample sizes were 973, 1083, 943, 860, and 668 for France, Italy, Germany, the UK, and Spain, respectively.

6.6.2 Measures

The measures used in this chapter include the List of Values or LOV (Kahle, 1983), items on personal care, items on buying behavior, and demographic characteristics.

The LOV consists of nine values that can be either rated or ranked in order to determine their overall ranking of the least to the most important personal value (cf. Bearden, Netemeyer & Mobley, 1993). The fact that LOV can be either rated or ranked gives us the opportunity to investigate whether subjects react in the same way to ratings and rankings. The nine LOV statements were assessed using nine-point rating scales followed by a ranking of the same nine statements[23]. In the rating task, two scale points were labeled 1 = 'very important' to 9 = 'very unimportant'; the other values were not labeled. Each subject rated and subsequently ranked the nine value statements. Five variables were included that measured attitudes towards trying and buying personal care products. These variables were measured on two-point rating scales, labeled '1' = 'agree' and '2' = 'disagree'. In addition, demographic information on age, education and income was

[22] Philips DAP Groningen, The Netherlands; data collected by GfK/IHA Lausanne, Switzerland.

[23] The study on Personal care was a commercial survey. Only one version of the questionnaire was made. Therefore, the order of the rating and the ranking task was the same for each subject.

included. In this chapter, the same the demographic information is used as in Chapter 5.

6.7 Results

6.7.1 The LOV rated and ranked

This chapter emphasizes methodological aspects in rating and ranking. Therefore, we focus on how subjects rate and rank the List Of Values items. The ordering of the LOV items is quite different across countries (see Table 6-1). For example, the value 'sense of belonging' is the most important in Germany, whereas it is considered the least important in Italy.

From a psychometric point of view, there are significant differences between countries in their level of the ratings of the LOV. The mean rating across all LOV items is significantly different across countries ($F_{(4,4521)} = 13.67$, $p < .001$). Spanish and British subjects have the lowest average rating score with 2.7, and French and German subjects have the highest average with 3.0. This implies that if no response style effects are present, Spanish and British subjects consider these nine values more important than French and German subjects. The standard deviation across rating scores per subject is also significantly different across countries ($F_{(4,4521)} = 73.38$, $p < .001$). The highest standard deviation is 1.9 (France and Germany), and the lowest is 1.3 (the UK). French and German subjects differentiate more among the nine values than British subjects.

In general, the subjects in each country tie two or more values when rating the LOV. The mean score is not equal to or even close to 5.0 in any country. This would have been the case if subjects had a hierarchically ordered picture of the values in their minds (with equal distances between the values) and thus would have given each value its own unique score, as is done while ranking the items.

Table 6-1 Mean scores of ranking and rating on the LOV[24]

	RATING					RANKING				
	France	Italy	Germany	UK	Spain	France	Italy	Germany	UK	Spain
Sense of belonging	5.1	4.1	1.8	2.9	4.6	7.4	6.7	3.0	5.2	7.6
Excitement	4.6	4.0	5.2	4.0	3.6	6.8	6.5	7.3	7.2	6.5
Warm relationships with others	2.4	2.8	2.7	2.5	2.1	3.9	4.8	4.4	3.8	3.4
Self-fulfillment	3.9	2.2	3.5	2.6	2.4	6.5	3.8	5.6	5.2	4.6
Being well-respected	2.3	2.0	4.9	2.9	1.9	3.8	3.9	7.5	5.4	3.6
Fun and enjoyment in life	1.8	3.3	2.0	2.6	2.4	2.8	5.9	3.7	4.9	4.7
Security	2.3	2.1	2.0	2.4	2.0	4.4	4.4	3.8	4.2	3.9
Self-respect	1.9	1.9	2.0	2.1	2.2	3.5	4.0	4.2	3.7	4.4
A sense of accomplishment	3.1	2.4	2.5	2.5	2.8	5.8	5.0	5.4	5.4	6.3
Mean	3.0	2.8	3.0	2.7	2.7	5.0	5.0	5.0	5.0	5.0
Standard deviation	1.9	1.6	1.9	1.3	1.5	2.7	2.7	2.7	2.7	2.7

[24] 1= most important value; 9 = least important value

6.7.2 Response patterns

To determine the response pattern based on rating scale use,[25] the number of times a subject chose a '1', a '2' and so on was calculated. The resulting score profile per subject is called the[26] response pattern. To determine whether there were groups of subjects that displayed similar response behavior, a clustering was made based on each respondent's response pattern on the ratings of the LOV. We first performed a hierarchical clustering analysis (Ward method). Then we used the emerging cluster centers as input for a K-means clustering (Hair, Anderson & Tatham, 1987). Such two-stage clustering is a common procedure for clustering large data sets (e.g., Punj & Stewart, 1983). We examined solutions ranging from two to eight clusters and selected 5 clusters, as this yielded the best interpretation and the clusters differed significantly from one another with respect to each of the active variables. The five resulting clusters, called groups from this point on, had the following sizes: group 1 (N=651; 14%), group 2 (N=1446; 32%), group 3 (N=1351; 30%), group 4 (N=952; 21%), and group 5 (N=127; 3%).

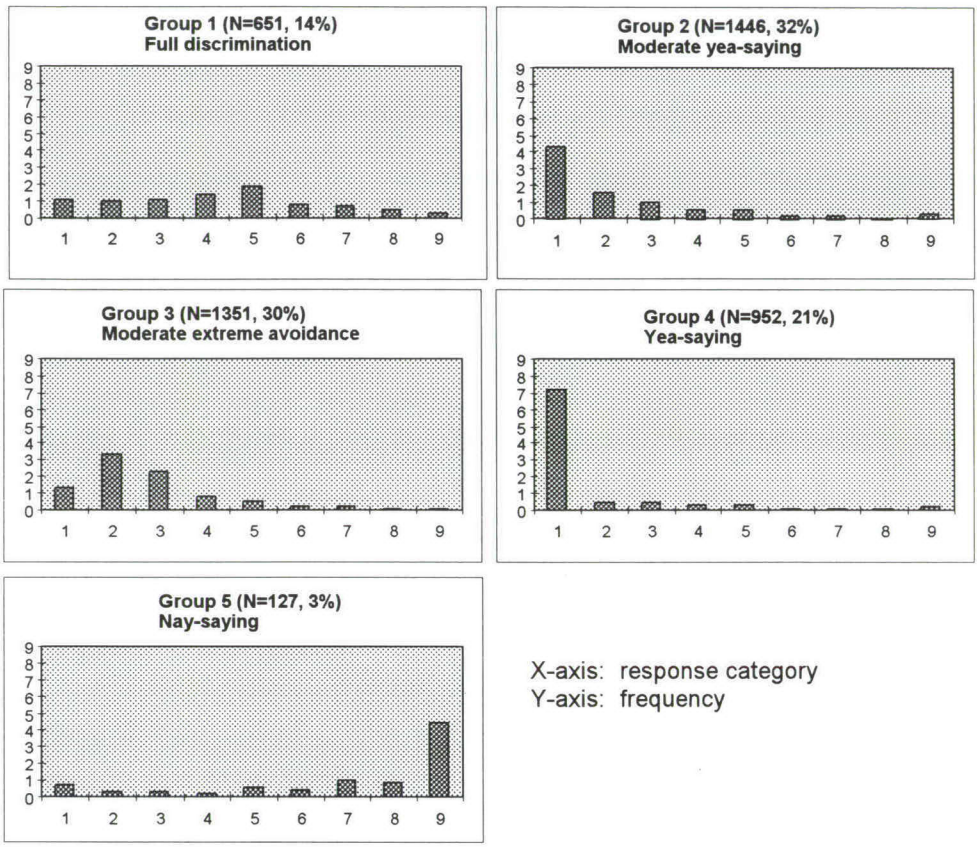
The average response pattern as observed in each group is presented in Figure 6-2. Subjects in group 1 spread their scores the most evenly across all numbers. Their rating pattern most closely resembled that of a ranking, where each number from '1' to '9' was used only once. Subjects in group 2 had a moderate yeasaying tendency. They considered approximately four out of nine values 'very important' and the remaining values were almost all given an unique score ranging from '2' through '9'.

---

[25] It should be noted that the response pattern based on ranking is the same for all subjects in the data set. Each subject has a pattern equal to the one in Frame E (Figure 6-1).

[26] There were only male subjects

Figure 6-2 Empirical response patterns based on LOV rating scores



Subjects in group 3 typically avoided yeasaying; the categories most frequently used by them were '2' and '3', indicating that they considered the underlying values 'important', but not 'very important'. Group 4 represents the extreme yea-

sayers.[27] About seven out of nine times, they used the score '1', indicating that they considered seven values 'very important.' Subjects in group 5 use category '9' (= 'not important at all') very frequently, thus qualifying them as nay-sayers. Hence, we observed yeasaying (groups 2 and 4, a total of 53 %), naysaying (group 5, 3%), moderate extreme avoidance (group 3, 30%), and fully discriminating (group 1, 14%) response patterns. However, the extreme tendency pattern, in which a subject uses both extremes of a scale, was absent.

6.7.3 Background characteristics and the effects on response patterns

Background characteristics may be related to the response patterns. For example, education level, income, and age could affect whether subjects use the extremes of a scale or not, or whether subjects discriminate between values or not. Our results revealed that the groups resulting from the cluster analysis differed in background characteristics (see Table 6-2).

Groups 1 (full discrimination) and 2 (moderate yeasaying) did not differ significantly from the grand total. This was not what was expected. For group 1, we expected a higher education than average and a higher income than average; in our data only income was slightly higher for this group. Group 3, the moderate extreme avoiders, included relatively more men from Germany, men who were young, and/or had a higher education, and/or a higher income. This result was as expected. Group 4, the yea-sayers, mainly included men from Italy and Spain, or men with a lower education or lower income. This was also in line with expectations. However, this group was also relatively young; it was expected that they would be older. Finally, men in group 5, the nay-sayers, were mainly from the UK and/or older.

---

[27] Subjects in group 4 considered values that which were opposites, such as 'excitement' and 'sense of belonging,' to both be 'very important'; they agreed with items independent of item content.

Table 6-2 Groups distinguished by LOV response patterns and background characteristics

	Total	Group 1	Group 2	Group 3	Group 4	Group 5
N (%)	4527 (100%)	651 (14%)	1446 (32%)	1351 (30%)	952 (21%)	127 (3%)
Country[28]						
France	22%	26%	24%	21%	14%	22%
Italy	24%	26%	23%	19%	31%	26%
Germany	21%	19%	21%	28%	13%	11%
The UK	19%	16%	17%	20%	21%	31%
Spain	15%	13%	14%	12%	21%	10%
Age[29]	43.1	44.4	43.8	42.0	42.0	47.3
Education[28]						
Left school 15/16	27%	27%	27%	25%	32%	27%
Left school 17/18	32%	31%	33%	28%	35%	25%
Advanced & University	41%	42%	40%	48%	33%	48%
Income[28]						
< 750 Euro	7%	7%	7%	6%	11%	10%
751 - 1500 Euro	18%	15%	19%	15%	21%	14%
1501 - 2250 Euro	29%	29%	30%	27%	28%	32%
2251 - 3000 Euro	23%	23%	24%	23%	23%	18%
> 3000 Euro	23%	26%	20%	29%	18%	26%

28 $\chi 2$ significant, $p < .001$

29 F-value, $p < .001$

6.7.4 Correlations between ratings and rankings

To calculate the correlations between ratings and rankings, the Spearman rank-order correlation statistic was used. Computations were done at the level of individual subjects, at which level the correlation between the ratings and the rankings was calculated for each subject; in Appendix 6.1, an overview of the correlations is given per country. Across all subjects, the average correlation is .66 (standard deviation is .38). There are some differences between countries, for example, the average correlation was higher in France and Germany than in the other EU countries.

The individual response patterns in ratings as well as background characteristics may have an effect on the magnitude of the correlation between ratings and rankings of this subject. Table 6-3 presents the results of an analysis of covariance, assessing the effect of response patterns and various socio-demographic variables on the correlation between ratings and rankings. The response pattern, country, education, income, and age all significantly affected the correlation between rating and ranking. The most important main effect was due to response pattern ($F_{(4,4214)} = 176.16$, $p < .001$). The correlation was lowest in group 5 (.04), the group with the naysaying pattern, and highest in groups 2 (.75) and 3 (.74), the groups with moderate yeasaying and moderate extreme avoidance, respectively (see also Table 6-2). This result was as expected for group 3; in group 1 the correlation was also high (.68). The correlation in group 4, the yeasaying group was .48. This correlation was lower than in groups 1 and 3, as was expected (see Section 6.4.1).

The next important main effect is country ($F_{(4,4214)} = 47.21$, $p < .001$). The correlation was highest in France and Germany (.74 and .76 respectively) and lowest in the UK (.59). Another significant main effect is education ($F_{(2,4214)} = 13.89$, $p < .001$), where advanced-level and university-educated subjects had an average correlation of about .69, whereas the less-educated subjects had an average correlation of about .63. Income also significantly affected the correlation between

ratings and rankings ($F_{(4,4214)} = 10.38$, $p < .001$). There was a monotonically increasing relationship; subjects with a higher income had a higher correlation between ratings and rankings.

Table 6-3 Explaining differences in correlations between ratings and rankings of LOV

	F-value	df	p
COVARIATE			
Age; b = 0.001	7.95	1	.005
FACTORS			
Income	10.38	4	<.001
Education	13.89	2	<.001
Country	47.21	4	<.001
Group	176.16	4	<.001
Group by country	11.56	16	<.001
Group by income	3.52	16	<.001
Group by education	1.39	8	.194
Income by country	1.46	16	.105
Income by education	.89	8	.527
Country by education	.84	8	.565
R^2_{adj}	.21		

Finally, age also affected the correlation: the regression coefficient (b) was .001. This means that a difference in age of 50 years increased the correlation by .05, which is a non-negligible effect.

Two significant interaction effects were present. The first is group by country ($F_{(16,4214)} = 11.56$, $p < .001$), which can be explained by pointing out that there were big differences in group 5. For German and French subjects, the correlations between ratings and rankings in group 5 were .61 and .50, respectively. This was below the average in each of these countries, but was expected. In the UK and Spain, the average correlations in group 5 were negative, -.51 and -.31, respectively. This implies that those subjects who stood for less than 5% of the samples in the UK and Spain reversed their rating and their ranking. The interaction between response pattern and income ($F_{(16,4214)} = 3.52$, $p < .001$) was also due to group 5. In the category 'Euro 2251-3000,' the average correlation is positive (.49), whereas it is about zero or even negative in the other income categories in group 5.

On the basis of these results, we concluded that rankings and ratings were more similar if subjects avoided extremes of the response scale, were moderate yeasayers, or exhibited full discrimination response patterns; were higher educated, and/or had a higher income and/or lived in Germany or France. However, on the basis of these results, it cannot be decided whether rating or ranking will provide better results. This can only be achieved if the predictive validity of both measures is assessed.

6.7.5 Predictive validity

In the previous sections, response patterns in ratings were distinguished. In addition to subjects' background characteristics, the effects these patterns had on the correlations between ratings and rankings were analyzed. In this section, the predictive validity of ratings and rankings is explored. For this purpose, five dichotomous attitude items were used on trying, choosing, and buying (personal care) products.[30] These five items together form a scale which was called 'Trybuy'; for this scale, Cronbachs' Alpha was .64, indicating that the reliability of the scale was adequate (Nunnally, 1978).

In the literature (Goldsmith, Freiden & Kilsheimer, 1993), a positive relationship was found between the value 'excitement' and 'being a fashion leader.' A fashion leader is a person who likes to buy new fashions and enjoys the process of buying them. We expect the scale on trying, choosing and buying products ('Trybuy') to correlate higher with the value 'excitement' than with the other values. Results of the correlations of the scale with all rated and ranked values are given in Table 6-4 and Table 6-5. The Spearman correlation of the value 'excitement' with the scale was higher than for the other values, for rating as well as for ranking. This was as expected.

It also is higher for rating than for rankings in all groups. Thus, the predictive validity for rating appears to be higher than for ranking. However, across all items, the correlation is higher for ratings than for rankings. This could be due to a degree

---

[30] These five variables are: (1) 'I like to buy and try new products', (2) 'I like sales assistants and demonstrators to invite me to try the products in the shop,' (3) 'I like to spend time choosing, trying and buying personal care products, toiletries and perfume,' (4) 'I like to be given samples when I buy cosmetics, personal care products or perfume,' and (5) 'I like to have a wide choice of different brands.'

of spurious correlation between the ratings of values and the ratings used for measuring the attitude items. This effect may exist, for example, because response styles carry over from one rating scale to the other (see e.g., Chapter 4 this thesis). For instance, the correlations between both rating scale measured data decreased after removing variance due to yeasaying response style from the data. In this case, the Spearman correlations between 'excitement' and the scale became .14 (total), .21 (group 1), .19 (group 2), .11 (group 3), .13 (group 4), and .26 (group 5), respectively. The correlation was least affected in group 1; this was expected as this group did not display evident yeasaying and/or extreme response style. Based on the results, it can be concluded that the predictive validity of ratings is (slightly) higher than that of rankings of values.

Table 6-4 Predictive validity of rating of the LOV with Trybuy

		GROUP				
	Total	1	2	3	4	5
Sense of belonging	.04**	-.02	-.01	-.00	.02	.11
Excitement	.22***	.20***	.23***	.15***	.15***	.12
Warm relationships	.12***	.00	.09***	.06*	.00	-.03
Self-fulfillment	.15***	.06	.09***	.12***	.07*	.07
Being well-respected	.18***	.02	.16***	.11***	.11**	.03
Fun and enjoyment	.14***	.03	.07**	.12***	.11**	-.16
Security	.08***	.11**	.01	.02	-.01	-.32***
Self-respect	.07***	.08*	.06*	.04	.00	-.23*
Sense of accomplishment	.10***	.00	.03	.06*	-.00	-.27**

*** $p < .001$; ** $p < .01$; * $p < .05$

Table 6-5 Predictive validity of ranking of the LOV with Trybuy

		GROUP				
	Total	1	2	3	4	5
Sense of belonging	-.07***	-.07	-.08**	-.06*	-.07*	.01
Excitement	.08***	.10**	.11***	.06*	.08*	.05
Warm relationships	.04**	.02	.04	.05	.03	.11
Self-fulfillment	.03	-.00	.01	.03	.06	.23**
Being well-respected	.06***	.03	.11***	.03	.00	.10
Fun and enjoyment	.01	.11**	.01	.04	.04	-.17
Security	.06***	-.11**	.03	-.06*	-.04	-.18*
Self-respect	.02	-.06	.01	-.02	-.02	.04
Sense of accomplishment	.06***	-.03	.10***	-.05	-.04	.16

6.8 Conclusion and discussion

To the best of our knowledge, the present study is the first in which representative samples, covering subjects from different countries, were included to investigate response behavior on ratings and ranking tasks. Other studies in this field have been based on homogeneous samples of students (Rankin & Grube, 1980; Maio, Roese, Seligman & Katz, 1996), or representative samples (Russell & Gray, 1994) in one country only. We have provided insight into response behavior across different countries and across different socio-demographic sections of the populations.

In the study, we identified groups of subjects, based on LOV response patterns of ratings in line with Bijmolt, Wedel, Pieters and DeSarbo (1998). Five groups were distinguished with the following patterns (see also Figure 6-2): a full-discrimination pattern (group 1); a moderate yeasaying pattern with several ties (group 2); an

extreme avoidance pattern (group 3); a yeasaying (group 4) and a naysaying pattern (group 5).

Our results revealed that the subjects' country of residence, household income, level of education, and age may be associated with differences between these response-pattern-based groups. Yeasaying (group 4), which is reflected here in considering all or most of the values to be very important, was found predominantly for subjects in Italy and Spain and/or subjects with lower education, lower household income, or younger age. Extreme avoidance (group 3), i.e., using rating scores in the middle of the scale more often, was more often found for higher-educated subjects and/or subjects living in Germany. These results were consistent with findings in the literature (Greenleaf, 1992a; Chapters 4 and 5, this thesis), which showed that higher-educated subjects tended to discriminate more among rating-scale points and that lower-educated subjects typically displayed yeasaying response behavior.

In general, ratings are affected by response styles such as yeasaying and extreme response style, whereas rankings are not. However, this does not mean that rankings are free of bias. Rankings can be used if a subject has a hierarchical ordering of items in mind; however, if this ordering contains ties, the ranking will be erroneous. If the subject's ordering approximates a hierarchy, the orderings of ratings and rankings can be expected to be more alike if the ratings are not affected by response styles. The results show that the correlations between ratings and rankings were more alike for subjects with a moderate yeasaying pattern with several ties (group 2) or an extreme avoidance response pattern (group 3). These tended to be subjects who were more educated, had a higher income, and/or lived in Germany or France. Rating and ranking show less similar results if subjects exhibited yeasaying or naysaying response behavior; those subjects were more frequently less educated and/or lived in Italy, Spain, and the UK (groups 4 and 5).

We acknowledge that, based on psychometric properties only, the Spearman rank-order correlation between rating and ranking was expected to be higher for subjects with a higher variance in their rating. However, our results revealed that the correlation was not the highest for the subjects with the highest variance in ratings (group 1). This shows that likewise subjects who assigned a different importance to the values tended to react differently towards the rating and the ranking task.

Our results imply that ratings and rankings of values can be considered to be quite similar alternatives if subjects do not display yeasaying or naysaying response styles and if their value ordering does not contain (many) ties. If subjects display yea- or naysaying behavior, rating will provide poor results, whereas if the true ordering contains numerous ties, ranking will provide poor results. In order to find out which method provides more valid results, the predictive validity of both procedures was explored. The predictive validity of ratings turned out to be slightly higher than that of rankings.

6.8.1 Implications

Our results for the patterns of ratings imply that the ordering of the LOV items for all subjects contained one or more ties in all five countries studied. This conclusion confirms the assumption by Ng (1982) that a fully-hierarchized value system can be too strong an assumption in cross-national research. Rating naturally allows for ties and may, therefore, be the preferred measurement procedure in cross-national research.

The results are relevant for academic research, as value research is currently done with both rating and ranking procedures, and there is no consensus about which method should be preferred. Our results for ratings showed a higher or at least equal predictive validity on some variables across all 5 EU countries and across the five groups of subjects that were distinguished on the basis of their response

pattern. For the samples most studied in academics, highly-educated younger subjects (students), the rating and ranking results were quite similar. As ratings are easier to administer and the predictive validity is certainly not lower, we are inclined to recommend the use of ratings in future value studies.

The results are also relevant for practice, as ratings are less cumbersome than rankings in data collection as well as in data analysis. Performing a ranking of more than 10 items is a difficult task, and subjects have to see the items to be able to perform the ordering. In telephone interviewing therefore, the ranking of items is not possible, while rating can be done quite easily. The results of this study are promising, as telephone interviewing is being used to an increasing degree these days (ESOMAR, 1997). In the analysis phase, furthermore, ratings possess certain advantages, as the scores are independent. Rankings pose problems in analyses as these scores are mutually dependent, and common techniques such as factor analysis cannot be performed. The use of ratings overcomes this drawback.

Appendix 6.1 Spearman rank-order correlations rating and ranking

		Total EU	France	Italy	Germany	UK	Spain
Total		.66	.74	.60	.76	.59	.61
Group	1	.68	.73	.65	.78	.59	.63
	2	.75	.79	.73	.79	.71	.75
	3	.74	.80	.64	.79	.73	.69
	4	.48	.58	.43	.58	.44	.46
	5	.04	.50	.21	.61	-.51	-.31
Age	15-24	.61	.69	.57	.67	.52	.60
	25-34	.68	.73	.63	.77	.61	.62
	35-44	.65	.75	.56	.77	.57	.60
	45-54	.68	.73	.63	.78	.62	.60
	55-64	.69	.80	.60	.79	.60	.67
	65 +	.65	.74	.60	.75	.54	.59
Income in Euros	< 750	.57	.60	.50	.72	.45	.58
	751 - 1500	.63	.73	.60	.74	.50	.62
	1501 - 2250	.67	.74	.62	.75	.54	.66
	2251 - 3000	.68	.75	.61	.78	.70	.68
	> 3000	.69	.74	.51	.78	.65	.57
Education	Left 15/16	.63	.73	.59	.74	.56	.62
	Left 17/18	.65	.73	.57	.76	.56	.59
	Advanced & University	.69	.75	.64	.77	.64	.63

Chapter 7. Conclusion, Implications and Future Research

7.1 Introduction

The objective of this thesis was twofold. The first objective was to provide a new and clear framework for addressing equivalence and bias in marketing research. The second objective was to provide insight into differences and similarities in response styles between countries in the European Union. Response styles, which are characteristic of rating scales, can be considered a main source of bias in culture-comparative research. However, in marketing, where rating scales are often used, response style is hardly ever addressed. In this thesis, the question is examined whether this is justified. For this purpose, we performed three empirical studies. Two empirical studies investigated the existence and some antecedents of response style in countries in the European Union. The third study investigated whether ranking, which is insensitive to two common types of response styles, can be considered an alternative to rating scales. In the empirical chapters, nationally representative samples of consumers from six countries in the European Union were used.

In this chapter, the main conclusions of the thesis are summarized and issues for further research are indicated.

7.2 Assessment of equivalence

In international marketing research, direct comparisons are often made between marketing research data from different countries. The equivalence of these data is thus a prerequisite. Equivalence refers to the extent to which results of a study are comparable across cultural populations. Yet, equivalence has barely been addressed in culture comparative studies. There are two possible reasons for this. The first is the confusing terminology that has been used to address equivalence. The second reason is a lack of uniformity, in dealing with the question of how equivalence could be examined. To make equivalence more open to researchers as well as to practitioners, a new and clear framework was evidently required. Existing frameworks drawn from the fields of marketing, psychology and business studies were compared (Chapter 2). It appeared that the equivalence framework as developed in psychology by Van de Vijver and Leung (1997a,b) was the most concise and clear. There are only three types of equivalence in this framework: construct, measurement unit, and scalar equivalence. These types of equivalence are nested; measurement unit equivalence presumes construct equivalence. Scalar equivalence, in turn, presumes the other levels of equivalence.

We have shown that this framework can be integrated in marketing research (Chapter 2). In doing so the equivalence framework developed by Craig and Douglas (2000), and which is currently used in marketing, was abandoned. In applying the new framework in marketing, equivalence is redefined as that part of the cross-national research process that can be controlled by means of statistical analysis procedures. In this redefinition, issues in the design stage of a study, such as allocation of products to a product category, translation of questionnaires, and sampling are considered possible sources of bias. Equivalence of those issues cannot be established a priori. In the design stage, measures can be taken to increase the probability that data are equivalent. In the proposed new framework for marketing, however, evidence of equivalence can only be obtained by statistical analysis procedures after the data have been collected.

Bias is a core issue in the new framework. Whether or not results can be considered equivalent is established at the end of the research process; in the course of the research process bias may emerge. In general, bias lowers the level of equivalence that can be reached. There may be bias in (1) the topics items refer to, (2) the measurement procedures or in the measurement scales, or (3) in certain items in the questionnaire. These biases are called construct, method and item bias, respectively. Each type of bias may be present in a given international marketing study. For instance, construct bias may emerge if the same product is used in different situations in various countries; method bias may emerge due to the data collection procedure or to differences in response styles; finally, item bias can be caused by a specific item not being applicable to one of the populations studied. Bias is not elusive; it can often be controlled. With this purpose in mind, the relationships between the stages in the research process and the sources of bias were discussed (Section 2.3). Bias can emerge in all stages of the research process. It was explained that attention to and, if possible, control of bias is necessary at all stages of the research process. In translation for example, bias may be reduced by following guidelines for translation (e.g., Van de Vijver & Hambleton, 1996); certain forms of bias in data collection can be checked afterwards by adding one or more variables to the data set.

The threat of bias is not the same for all types of data. The extent to which measurements are prone to bias is influenced by the topic of the research and the manner in which these measurements are interpreted. In general, the more concrete a topic becomes, the less prone to bias its measurements become. For instance, the measurement of a subject's amount of water consumption is less prone to bias than measurement of his or her value structure. This also influences the extent to which valid cross-cultural inferences can be made (Chapter 3). This depends on (1) the extent to which a research domain can be accurately described, and (2) whether concepts can be operationally defined or not. In marketing, several research domains can, in principle, be fully described. For instance, the car brands for sale

in a country can be listed. Such domains are relatively concrete and valid inferences about the domain can be made. The researcher has the possibility of controlling sources of bias in these data. If a research domain becomes more abstract, the listing of all of the elements may no longer be possible (see Section 3.2.2.1). Hypothetical constructs are created that are assumed to measure psychological traits (e.g., 'innovativeness'). There, the threat of bias is substantial. There may be construct, method, and item bias, and the researcher has limited possibilities of monitoring and controlling these in the study.

7.2.1 Implications

The studies in this thesis imply that bias may prove to be a problem if researchers wish to compare the results of studies in various countries. Taking into account the various sources of bias has implications for the design and the data analysis of a study. In designing a study, attention has to be paid to the identification of possible sources of bias. In international marketing research, possibilities for controlling sources of bias are available. Pre-studies can be done. For instance, a qualitative study may be linked to a quantitative study as is done in psychology (see, e.g., Hines, 1993). Moreover, it is argued that the use of qualitative research should be encouraged as a complement to quantitative research in future marketing studies (Day & Montgomery, 1999). The information from a qualitative study can be used to improve the quality of questions in a quantitative study. This may reduce bias. Qualitative studies, of course, need not be free from bias, but they help researchers in defining the research domain in different countries. In designing the main study, attention can then be paid to sources of bias that emerged in the pre-study.

Another implication of the new equivalence framework in marketing is that more attention should be paid to testing the equivalence of constructs and measures across cultural populations. However, this also extends to models. In marketing, conceptual models are used which incorporate variables at a different level of abstraction. Examples include models on food-related lifestyle (see e.g., Van Herk,

Verhallen & Barzilay, 1994; Brunsø & Grunert, 1998). Such models include different types of variables such as values and behavioral measures; those variables may be differently affected by bias (see Chapter 3); values are more prone to bias than operationally defined measures. The level of equivalence has to be investigated for each of those types. However, equivalence of scales does not imply that a model incorporating those scales can be assumed to be equal. The cross-cultural comparability of the model has to be assessed; ignoring this may lead to erroneous conclusions about differences or similarities between cultures.

7.3 Response style

In culture comparative research, instruments may be affected by method bias. A major potential cause of method bias in culture comparative research is response style. The effects of response style on the outcomes of studies have been shown in psychology, but in marketing there are hardly any publications on this issue. To add to the knowledge about response style in international marketing we performed three empirical studies.

Differences in response style were investigated in six countries in the EU, Greece, Italy, France, Spain, the UK and Germany. It was known that response style differed between the US and Asia, but differences in response style in countries within the European Union, had until recently never been compared. It was found (Chapter 4) that there are tendencies to yeasaying and extreme response behavior. The results showed that the differences and similarities found between countries are quite consistent across two rating scale types and to a lesser extent also consistent across three behavioral domains (i.e., cooking, washing, and personal care). In particular, it was found that Greek subjects tended to display more yeasaying and more extreme response behavior than subjects from Italy, France, Spain, the UK and Germany. These tendencies in response style were also more pronounced in Italy and in Spain than in France, the UK and Germany. The Germans tended to be

the least prone to extreme response style of all EU subjects. In Chapter 5, some antecedents of response style were investigated in 55 regions in the EU. Results indicated that there were differences between regions. It appeared that the country in which a region is located in and its global coordinates explained differences in response style at the regional level. At the individual-subject level, the socio-demographic characteristics of age, education, and income affected the extent to which subjects displayed yeasaying or extreme response style. Effects were significant and non-negligible.

The empirical research demonstrated the presence of differences in response style among countries in the EU. Variance due to such response styles may be removed by regressing items on response style indices (see Chapters 4 and 5). Subsequent analyses can then be carried out on the residualized scores. Another solution for handling response styles is to choose another response procedure such as ranking instead of rating. By making this choice, response styles like yeasaying and extreme response style are avoided. In Chapter 6, it was investigated whether rankings and ratings of the same items by the same subjects are similar across countries. In the ranking task, subjects were forced to assign a unique score, between 1 and 9, to each of nine items. In the study, five groups of subjects were distinguished with the following patterns (see Figure 6-2): a full-discrimination pattern; a moderate yeasaying pattern with several ties; an extreme avoidance pattern; a yeasaying; and a naysaying pattern. It appeared that ratings and rankings of the List of Values items (Kahle, 1983) were more alike for subjects with a 'moderate extreme avoidance' or 'moderate yeasaying' response pattern. These subjects were on average better-educated, had a higher income and/or lived in France or Germany. In the study, the predictive validity of ratings and rankings was explored; predictive validity for ratings was slightly higher for all groups of subjects.

7.3.1 Implications

Differences in response style were found in countries in the European Union. These differences were non-negligible. This implies that direct comparisons between levels of scores of subjects from different countries in the EU may lead to erroneous marketing decisions. Currently, many companies are active in several countries in the EU simultaneously. The use of surveys to determine performance has become routine. In such surveys, rating scales are often used. Based on the results from this thesis, it can be expected that customers in Greece will be more positive than customers from Germany. However, this is likely to reflect bias. Correction of the scores for response styles is required before valid decisions on customer service levels can be made.

Using another measurement procedure such as ranking to avoid response styles does not appear to be a solution. Our results suggest that ranking is an alternative to rating in studies in which students or other highly educated subjects are employed. For other subjects, however, rating still appears to be preferable to ranking. In the practice of marketing research in particular, a variety of respondents is used. Ranking would then not appear to be a good alternative.

7.4 Further research

7.4.1 Antecedents of response style

Several antecedents of response style were indicated in Chapter 5. At the individual level, a great deal of variance was left unexplained. Further research may focus on investigating other antecedents of response style at this level. In the literature, several antecedents have been indicated; personality (e.g., Hamilton, 1968; Knowles & Nathan, 1997) and meaningfulness of the item to the subject (e.g.,

O'Donovan, 1965) re but two examples. In our analyses, few categories were distinguished for education and income; in future research, measurement of these socio-demographic variables in amounts of money and years of education, respectively, would be preferable. The data sets in this thesis consisted of either men or women only. Future studies may include both men and women, despite the fact that most studies on response style report no differences for gender (e.g., Greenleaf 1992b; Watkins & Cheung, 1995).

At the regional or country level, as at the individual level, other antecedents may explain differences in response styles. At the regional level, this may include such characteristics as wealth. At the national level, characteristics come to mind such as Hofstede's dimensions of 'national culture' (Hofstede, 1980) or 'national character' (Clark, 1990). This merits further research.

7.4.2 Balanced scales

Another desirable future development is the construction of balanced scales in marketing. Balanced scales contain item pairs that are logically opposite (e.g., Ray, 1983). Balanced scales would provide an opportunity to calculate acquiescence (yeasaying), in particular in studying abstract hypothetical constructs measured with multi-item scales,. Variance due to acquiescence can be partialled from the scores afterwards (see e.g., Winkler, Kanouse & Ware, 1982; Ten Berge, 1999) and scores free from acquiescence bias can be used for making inferences.

7.4.3 Extent to which response styles are bias or also content

We have found that the size of the yeasaying and extreme response tendencies differed between domains and item sets with a different content (Chapter 4). In particular, it was found that the method of shaving was significantly related to the yeasaying tendency (Chapter 5). This shows that yeasaying is more than a mere error component. Ignoring the yeasaying component may lead to misrepresentation

of cross-cultural differences. Moreover, the interpretation of the yeasaying component deserves attention. Part of it may reflect a real difference in attitude between cultural populations.

We propose actual behavior may be used to assess the extent to which response style is bias or content. In a cross-national context in particular, measures of actual behavior provide a hard criterion for identifying differences and similarities in a behavioral domain in different countries. For instance, the amount of olive oil used in one week can be observed and weighed and this amount can be employed as a reference for establishing possible response bias in the rating of the item 'I use a great deal of olive oil'. An additional positive characteristic of actual behavior lies in the extent to which it can be the basis for comparisons across countries. Actual behavior can be operationally defined and is less prone to bias (Chapter 3). A person using 100 grams of olive oil in one country is using two times as much as a person using 50 grams independent of any demographic or cultural variable.

The validity of the outcomes can be determined by comparing the actual behavior in the domain studied to (1) the original scores and (2) the scores after correcting for bias. To relate attitudes to behavior a model like the Fishbein and Ajzen model (Fishbein & Ajzen, 1975) may be employed. This model has validity across some cultures (see Lee & Green, 1991). Moreover, it has been useful in food choice studies in the UK (Shepard, 1989). In the latter study, attitudes towards behavior were a good predictor of actual product use.

7.5 Epilogue

Based on marketing research information, important decisions are made by marketing managers. Such decisions increasingly involve several countries simultaneously. Thus, questions regarding comparability of research done in several countries simultaneously can be expected to grow. It is expected that the requirements concerning the quality of this research will increase in the coming years. Thus, active cooperation and collaboration between the academic study of marketing and its practice is needed (Day & Montgomery, 1999).

In this thesis, a framework was discussed for addressing equivalence in marketing research. It indicates how international marketing research can be raised to a higher level and it may help to increase the validity of results in international marketing research. Our framework was developed from a theoretical basis; in addition, some applications to commercial marketing research data were shown. However, more applications in the marketing field, in both academic investigation and in practice, are needed to test its usefulness.

References

A

Aaker, J.L. (1999), "The malleable self: The role of self-expression in persuasion," *Journal of Marketing Research*, 36(1), 45-57.

Albaum, G. and R.A. Peterson (1984), "Empirical research in international marketing," *Journal of international Business Studies*, (Spring/Summer), 161-173.

Alwin, D.F. and J.A. Krosnick (1985), "The measurement of values in surveys: A comparison of ratings and rankings," *Public Opinion Quarterly*, 49, 535-552.

Askegaard, S. and T.K. Madsen (1998), "The local and the global: Exploring traits of homogeneity and heterogeneity in European food cultures," *International Business Review*, 7, 549-568.

Aulakh, P.S. and M. Kotabe (1993), "An assessment of theoretical and methodological development in international marketing: 1980-1990," *Journal of International Marketing*, 1(2), 5-28.

B

Bachman, J.G. and P.M. O'Malley (1984), "Yea-saying, nay-saying, and going to extremes: Black-white differences in response styles," *Public Opinion Quarterly*, 48, 491-509.

Bagozzi, R.P and Y. Yi (1991), "Multitrait-multimethod matrices in consumer research," *Journal of Consumer Research*, 17(March), 426-439.

Barcikowski, R.S. (1981), "Statistical power with group mean as the unit of analysis," *Journal of Educational Statistics*, 6(3), 267-285.

Barker, A.T. and E. Kaciak (1992), "An examination of the equivalence of four measurement scales in cross-cultural research," *Journal of International Consumer Marketing*, 4(3), 77-93.

Barnard, N.R. and A.S.C. Ehrenberg (1990), "Robust measures of consumer brand beliefs," *Journal of Marketing Research*, 27(4), 477-484.

Baumgartner, H. and J.B.E.M. Steenkamp (1996), "Exploratory consumer buying behavior: Conceptualization and measurement," *International Journal of Research in Marketing*, 13(2), 121-137.

Baumgartner, H. and J.B.E.M. Steenkamp (1999), "Response styles in marketing research: A cross-national investigation," *Working paper*, Penn State University.

Bearden, W.O., R.G. Netemeyer and M.F. Mobley, Eds. (1993), *Handbook of Marketing Scales*, Newbury Park, CA: Sage.

Bentler, P.M., D.N. Jackson, and S. Messick (1971), "Identification of content and style: A two-dimensional interpretation of acquiescence," *Psychological Bulletin*, 76(3), 186-204.

Berry, J.W. (1969), "On cross-cultural comparability," *International Journal of Psychology*, 4, 119-128.

Berry, J.W., Y.H. Poortinga, M.H. Segall, and P.R. Dasen (1992), *Cross-cultural psychology. Research & applications.* Cambridge: Cambridge University Press.

Bijmolt, T.H.A., M. Wedel, R.G.M. Pieters, and W.S. DeSarbo (1998), "Judgements of brand similarity," *International Journal of Research in Marketing*, 15(3), 249-268.

Blalock, H.M. (1982), *Conceptualization and measurement in the social sciences.* Beverly Hills, CA: Sage.

Boote, A.S. (1982), "Psychographic segmentation in Europe," *Journal of Advertising Research*, 22(6), 19-25.

Botschen, G. and A. Hemetsberger (1998), "Diagnosing means-end structures to determine the degree of potential marketing program standardization," *Journal of Business Research*, 42, 151-159.

Braithwaite, V.A. and H.G. Law (1985), "Structure of human values: Testing the adequacy of the Rokeach Value Survey," *Journal of Personality and Social Psychology*, 49(1), 250-263.

Bresnahan, M.J., R. Ohashi, W.Y. Liu, R. Nebashi, and C-C Liao (1999), "A comparison of response styles in Singapore and Taiwan," *Journal of Cross-Cultural Psychology*, 30(3), 342-358.

Brislin, R.W. (1986), "The wording and translation of research instruments". In: *Field methods in cross-cultural research*, Lonner, W.J. and J.W. Berry (Eds.), 137-164.

Brislin, R.W., W.J. Lonner, and R. Thorndike (1973), *Cross-cultural research methods*. New York: Wiley.

Broen, W.E. and R.D. Wirt (1958), "Varieties in response sets," *Journal of Consulting Psychology*, 22(3), 237-240.

Brunsø, K. and K.G. Grunert (1998), "Cross-cultural similarities and differences in shopping for food," *Journal of Business Research*, 42, 145-150.

Bryk, A.S., and S.W. Raudenbush (1992), *Hierarchical Linear Models: applications and data analysis methods*. London: Sage.

Bryk, A.S., S.W. Raudenbush, and R.T. Congdon (1996), *HLM: Hierarchical linear and nonlinear modeling with the HLM/2L and HLM/3L Programs*. Chicago: SSI Scientific Software International.

C

Campbell, D.T. and D.W. Fiske (1959), "Convergent and discriminant validity by the multitrait - multimethod matrix," *Psychological Bulletin*, 56(2), 81-105.

Carmines, E.G. and R.A. Zeller (1979), *Reliability and validity assessment. Series: Quantitative Applications in the Social Sciences*, London: Sage.

Chee, H. and R. Harris (1998), *Global Marketing Strategy*. London: Financial Times, Pitman Publishing.

Chen, C., S-Y. Lee, and H.W. Stevenson (1995), "Response style and cross-cultural comparisons of rating scales among east Asian and north American students," *Psychological Science*, 6(3), 170-175.

Chinese Culture Connection (1987), "Chinese values and the search for culture-free dimensions of culture," *Journal of Cross-Cultural Psychology*, 18(2), 143-164.

Chun, K-T, J.B. Campbell and J.H. Yoo (1974), "Extreme response style in cross-cultural research. A reminder," *Journal of Cross-Cultural Psychology*, 5(4), 465-480.

Churchill, G.A. (1979), "A paradigm for developing better measures of marketing constructs," *Journal of Marketing Research*, 16(1), 64-73.

Churchill, G.A. (1991), *Marketing Research. Methodological Foundations*. Fifth Edition. Fort Worth: Harcourt Brace Jovanovich, Inc.,The Dryden Press.

Clark, T. (1990), "International marketing and national character: A review and proposal for an integrative theory," *Journal of Marketing*, October, 66-79.

Cohen, J. (1988), *Statistical power analysis for the behavioral sciences*. Hillsdale New Jersey: Lawrence Erlbaum Associates.

Couch, A. and K. Keniston (1960), "Yeasayers and naysayers: Agreeing response set as a personality variable," *Journal of Abnormal and Social Psychology*, 60(2), 151-174.

Craig, C.S. and S.P. Douglas (2000), *International Marketing Research*. Second Edition, New York: Wiley.

Crandall, J.E. (1982), "Social interest, extreme response style, and implications for adjustment," *Journal of Research in Personality*, 16, 82-89.

Cronbach, L.J. and P.E. Meehl (1955), "Construct validity in psychological tests," *Psychological Bulletin*, 52(4), 281-302.

Cronbach, L.J., G.C. Gleser, H. Nanda, and N. Rajaratnam (1972), *The dependability of behavioral measures. Theory of generalizability for scores and profiles*. New York: Wiley.

Cronbach, L.J. (1946), "Response sets and test validity," *Educational and Psychological Measurement*, 6, 475-494.

Cronbach, L.J. (1950) "Further evidence on response sets and test design," *Educational and Psychological Measurement*, 10, 3-31.

Crowne, D.P. and D. Marlowe, (1960), "A new scale for social desirability independent of psychopathology," *Journal of Consulting Psychology*, 24(4), 349-354.

D

Davis, H.L., S.P. Douglas, and A.J. Silk (1981), "Measure unreliability: A hidden threat to cross-national marketing research," *Journal of Marketing*, 45 (spring), 98-109.

Dawar, N. and P. Parker (1994), "Marketing universals: Consumer's use of brand name, price, physical appearance and retailer reputation as signals of product quality," *Journal of Marketing*, 58(2), 81-95.

Day, G.S. and D.B. Montgomery (1999), "Charting new directions for marketing," *Journal of Marketing*, 63 (Special Issue), 3-13.

DeCasper, H.S. and C.K. Tittle (1988), "Rankings and ratings of values: Counseling uses suggested by factoring two types of scales for female and male eleventh grade students," *Educational and Psychological Measurement*, 48, 375-384.

Dillon, W.R., T.J. Madden, and N.H. Firtle (1994), *Marketing research in a marketing environment.* Third Edition. Irwin.

Douglas, S.P. and C.S. Craig (1983), *International Marketing Research.* Englewood Cliffs, N.J.: Prentice Hall.

Douglas, S.P. and C.S. Craig (1992), "Advances in international marketing," *International Journal of Research in Marketing*, 9, 291-318.

Douglas, S.P. and C.S. Craig (1997), "The changing dynamic of consumer behavior: Implictions for cross-cultural research," *International Journal of Research in Marketing*, 14, 379-395.

Durvasula, S., Andrews, J.C., Lysonski, S., and R.G. Netemeyer (1993), "Assessing the cross-national applicability of consumer behavior models: A model of attitude toward advertising in general," *Journal of Consumer Research*, 19(4), 626-636.

E

Edye, D. and V. Lintner (1996), *Contemporary Europe. Economics. Politics and Society.* Hertfortshire, UK: Prentice Hall Europe.

Edwards, A.L. (1957), *The social desirability variable in personality assessment and research.* New York: Dryden Press.

Eimers, Y.M., R.G.M Pieters, and Th.M.M. Verhallen (1995), "Product evaluation as a function of consumer type and usage situation," In: *Integrating views on economic behavior.* Antonides, G., W.F. van Raaij, and S. Maital (Eds.). Deventer: Kluwer.

ESOMAR (1997), "ESOMAR Annual Study on the Market Research Industry 1996," ESOMAR, European Society for Opinion and Marketing Research, Amsterdam.

Ettenson, R. (1993), "Brand name and country of origin effects in the emerging market economies of Russia, Poland and Hungary," *International Marketing Review,* 10(5), 14-36.

F

Fioravanti, M., H.G. Gough, and L.J. Frere (1981). "English, French and Italian adjective check lists," *Journal of Cross-Cultural Psychology*, 12(4), 461-472.

Fishbein, M. and I. Ajzen (1975) *Belief, Attitude, Intention and Behavior: An introduction to Theory and Research.* Reading, UK: Addison Wesley.

Forsman, L. (1993), "Giving Extreme Responses to Items in Self-Esteem Scales: Response Set or Personality Trait?," *European Journal of Psychological Assessment*, 9(1), 33-40.

Frear, C.R., M.S. Alguire, and L.E. Metcalf (1995), "Country segmentation on the basis of international purchasing patterns," *Journal of Business & Industrial Marketing*, 10(2), 59-68.

G

Garlington, W.K. and H.E. Shimote (1964), "The Change Seeker Index: A measure of the need for variable stimulus input," *Psychological Reports*, 14, 919-924.

Gatignon, H., Eliasberg, J., and Robertson, T.E. (1989), "Modelling multinational diffusion patterns: An efficient methodology," *Marketing Science*, 8(3), 231-247.

Goldsmith, R.E., J.B. Freiden, and J.C. Kilsheimer (1993, September/October), "Social values and female fashion leadership: A cross-cultural study," *Psychology & Marketing*, 10(5), 399-412.

Graves, D. (1973), *Management research: A cross-cultural perspective.* San Francisco, CA: Jossey-Bass.

Greenleaf, E.A. (1992a), "Improving rating scale measures by detecting and correcting bias components in some response styles," *Journal of Marketing Research*, 29(2), 176-188.

Greenleaf, E.A. (1992b), "Measuring extreme response style," *Public Opinion Quarterly*, 56, 176-188.

Grunert, K.G., A. Baadsgaard, H.H. Larsen, and T.K. Madsen (1996), *Market orientation in Food and Agriculture*. Boston/Dordrecht/London: Kluwer Academic Publishers.

Grunert, K.G., Grunert, S.C., and S.E. Beatty (1989), "Cross-national research on consumer values," *Marketing and Research Today*, 1, 30-39.

Grunert, S.C. and G. Scherhorn (1990), "Consumer values in West Germany underlying dimensions and cross-cultural comparison with North America," *Journal of Business Research*, 20(2), 97-107.

Grunert, S.C. and Th.E. Muller (1996), "Measuring values in international settings: Are respondents thinking 'real' life or 'ideal' life?," *Journal of International Consumer Marketing*, 8(3 /4), 169-185.

Grunert, S.C., K.G. Grunert, and K. Kristensen (1993), "Une méthode d'estimation de la validité interculturelle des instruments de mesure: Le cas de la mesure des valeurs de consommateurs par la liste des valeurs LOV," (On a method for estimating the cross-cultural validity of measurement instruments: The case of measuring consumer values by the List of Values LOV), *Recherche et Applications en Marketing*, 8(4), 5-28.

H

Hair, J.F.Jr., Anderson, R.E., and R.L. Tatham (1987), *Multivariate Data Analysis. With readings.* Second Edition. New York: Macmillan Publishing Company.

Hamilton, D.L. (1968), "Personality attributes associated with extreme response style," *Psychological Bulletin*, 69(3) , 192-203.

Harman, H.H. (1976), *Modern Factor Analysis.* Third Edition Revised. Chicago: The University of Chicago Press.

Hays, W.L. (1973), *Statistics for the social sciences.* Second Edition. Holt, London UK: Rinehart and Winston.

Helsen, K., K. Jedidi and W.S. DeSarbo (1993), "A new approach to country segmentation utilizing multinational diffusion patterns," *Journal of Marketing*, 57(4), 60-71.

Hibbert, E. (1993), "Researching international markets - How can we ensure validity of results?," *Marketing and Research Today*, November, 222-228.

Hines, A.M. (1993), "Linking qualitative and quantitative methods in cross-cultural survey research: Techniques from cognitive science," *American Journal of Community Psychology*, 21(6), 729-746.

Hofstede, G. (1980), *Culture's Consequences: International differences in Work-Related Values.* Beverly Hills, CA: Sage.

Hofstee, W.K.B., J.M.F. Ten Berge, and A.A.J. Hendriks (1998), "How to score questionnaires," *Personality and Individual Differences*, 25, 897 - 909.

Holland, P.W. and D.T. Thayer (1988), "Differential item performance and the Mantel-Haenszel procedure," In: *Test Validity*, H. Wainer and H.I. Braun (Eds.) 129-145, Hillsdale NJ: Erlbaum.

Homer, P.M. and L.R. Kahle (1988), "A Structural Equation Test of the Value-Attitude-Behavior Hierarchy," *Journal of Personality and Social Psychology*, 54(4), 638-646.

Hui, C.H. and H.C. Triandis (1985a), "Measurement in cross-cultural psychology. A review and comparison of strategies," *Journal of Cross-Cultural Psychology*, 16(2), 131-152.

Hui, C.H. and H.C. Triandis (1985b), "The instability of response sets," *Public Opinion Quarterly*, 49, 253-260.

Hui, C.H. and H.C. Triandis (1989), "Effects of culture and response format on extreme response style," *Journal of Cross-Cultural Psychology*, 20(3), 296-309.

J

Jackson, D.N. and S. Messick (1958), "Content and style in personality assessment," *Psychological Bulletin*, 55, 243-252.

Jain, S.C. (1993), *Marketing Planning & Strategy*. Fourth Edition, Cincinnati, Ohio: South-Western Publishing Co.

Jeannet, J.P. and H.D. Hennessey (1992), *Global marketing strategies*. Second Edition, , Boston, London, Melbourne: Houghton Mifflin Company.

Johansson, J.K., S.P. Douglas, and I. Nonaka (1985), "Assessing the impact of country of origin on product evaluations: a new methodological perspective," *Journal of Marketing Research*, Vol. XXII, November, 388-396.

Jöreskog, K.G. (1971), "Simultaneous factor analysis in several populations," *Psychometrika*, 36(4), 409-426.

K

Kahle, L.R. (1983), *Social values and social change: Adaptation to life in America*. New York: Praeger Publishers.

Kale, S.H. (1995a), "Grouping Euroconsumers: A Culture-Based Clustering Approach," *Journal of International Marketing*, 3(3), 35-48.

Kale, S.H. (1995b), "Using Consumer Temperament as a Basis for Global Segmentation," In: *Marketing Today and for the 21st Century*, M. Bergadaà (Ed.), 557-568. Cery-Pontoise: ESSEC.

Kamakura, W.A., Novak T.P., Steenkamp J.B.E.M., and Th.M.M. Verhallen (1994), "Identification de segments de valeurs pan-europeens par un modele logit sur les rangs avec regroupements successifs." (Identifying Pan-European value segments with a clusterwise rank-logit model), *Recherche et Applications en Marketing*, 8(4), 29 - 55.

Kamakura, W.A. and J.A. Mazzon (1991), "Value Segmentation: A model for the measurement of values and value systems," *Journal of Consumer Research*, 18(2), 208-218.

Kamakura, W.A., and T.P Novak (1992), "Value-system segmentation: Exploring the meaning of LOV"," *Journal of Consumer Research*, 19(1), 119-132.

Kearney, M., M.J. Gibney, J.A. Martinez, M.D.V. de Almeida, D. Friebe, H.J.F. Zunft, K. Widhalm, and J.M. Kearney (1997), "Perceived need to alter eating habits among representative samples of adults from all member states of the European Union," *European Journal of Clinical Nutrition*, 51, Supplement 2, 30-35.

Kinnear, T.C. and J.R. Taylor (1996), *Marketing research: an applied orientation.* New York: McGraw-Hill.

Klein, P.R. (1991), "Global advertising: Does research find a cross-cultural effect?," *Applied Marketing Research, a journal for practitioners*, 31(1), 17-26.

Knowles, E.S. and K.T. Nathan (1997), "Acquiescent Responding in Self-Reports: Cognitive Style or Social Concern," *Journal of Research in Personality*, 31, 293-301.

Kotabe, M. and K. Helsen (1998), *Global marketing management.* New York: John Wiley & Sons, Inc.

Kotler, Ph. (2000), *Marketing Management.* The Millennium Edition. Upper Saddle River NJ: Prentice Hall, Inc.

Kreft, I., and J. De Leeuw (1998), *Introducing Multi-level modeling.* London: Sage.

Kroeber, A.L. and C. Kluckhohn (1952), "Culture: A critical review of concepts and definitions," *Papers of the Peabody Museum of American Archaeology and Ethnology*, 47 (1), 1-223.

Krosnick, J.A. and D.F. Alwin (1988) "A test of the form-resistant correlation hypothesis. Ratings, rankings and the measurement of values," *Public Opinion Quarterly*, 52, 526-538.

Kumar, N., L.K. Scheer, and J.B.E.M. Steenkamp (1995), "The effects of supplier fairness on vulnerable resellers," *Journal of Marketing Research*, 31(1), 54-65.

Kumar, V. (2000) *International marketing research*. Upper Saddle River NJ: Prentice-Hall Inc.

L

Lee, C. and R.T. Green (1991), "Cross-cultural examination of the Fishbein behavioral intentions model," *Journal of International Business Studies*, 2, 289-305.

Lenneräs, M., C. Fjellström, W. Becker, I. Giachetti, A. Schmitt, A.M. Remaut de Winter, and M. Kearney (1997), "Influences on food choice perceived to be important by nationally-representative samples of adults in the European Union," *European Journal of Clinical Nutrition*, 51, Supplement 2, 8-15.

Leung, K. and M.H. Bond (1989), "On the empirical identification of dimensions for cross-cultural comparisons," *Journal of Cross-Cultural Psychology*, 20(2), June, 133-151.

Levitt, T. (1983), "The globalization of markets," *Harvard Business Review*, May-June, 92-102.

Lord, F.M. (1980), *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

M

Maheswaran, D. (1994), "Country of origin as a stereotype: Effects of consumer expertise and attribute strength on product evaluations," *Journal of Consumer Research*, 21(2), September, 354-365.

Maio, G.R., N.J. Roese, C. Seligman, and A. Katz (1996), "Rankings, ratings, and the measurement of values: Evidence for the superior validity of ratings," *Basic and Applied Social Psychology*, 18(2), 171-181.

Malhotra, N.K. and D.F. Birks (1999), *Marketing research: An applied approach*. Pearson Education Ltd., Essex, England.

Malhotra, N.K., J. Agarwal and M. Peterson (1996), "Methodological issues in cross-cultural marketing research," *International Marketing Review*, 13(5), 7-43.

Malpass, R.S. and Y.H. Poortinga (1986), "Strategies for design and analysis," In: *Field methods in cross-cultural research*, Lonner, W.J. and J.W. Berry (Eds.), 47-83.

Manrai, L.A. and A.K. Manrai (1996), "Current issues in the cross-cultural and cross-national consumer research," *Journal of International Consumer Marketing*, 8(3/4), 9-22.

Marín, G., R.J. Gamba, and B.V. Marín (1992), "Extreme response style and acquiescence among Hispanics. The role of acculturation and education," *Journal of Cross-Cultural Psychology*, 23(4), December, 498-509.

Marsh, H.W. and B.M. Byrne (1993), "Confirmatory factor analysis of multi-group-multimethod self-concept data: Between group and within-group variance constraints," *Multivariate Behavioral Research*, 28, 313-349.

Martin, J. (1964), "Acquiescence—measurement and theory," *British Journal of Social and Clinical Psychology*, 3, 216-225.

Maucher H.O. and P. Brabeck-Lethmathe (1993), "The impact of the single European market on regional product and price differentiation - the example of the European food industry," In: *European Marketing. Readings and Cases.* Halliburton C. and R. Hünerberg (Eds.) 201- 217, Addison Wesley.

Messick, S. (1991), "Psychology and methodology of response styles," In: *Improving inquiry in social science: A volume in honor of Lee J. Cronbach.* Snow, R.E. and D.E. Wiley, 161-200.

Messick, S. (1968), "Response sets,". In: D.L. Sills (Ed.) International encyclopedia of the social sciences, 492-496. New York: Macmillan.

Mick, D.G. (1996), "Are Studies of Dark Side Variables Confounded by Socially Desirable Responding? The Case of Materialism," *Journal of Consumer Research*, 23(2), 106-119.

Middleton, K.L. and J.L. Jones (2000), "Socially Desirable Response Sets: The Impact of Country Culture," *Psychology & Marketing*, 17(2), 149-163.

Millsap, R.E. and H.T. Everson (1993), "Methodology Review: Statistical Approaches for Assessing Measurement Bias," *Applied Psychological Measurement*, 17(4), 279-334.

Moore, M. (1975), "Rating versus ranking in the Rokeach value survey: An Israeli comparison," *European Journal of Social Psychology*, 5(3), 405-408.

Moore, W.L. and E.A. Pessemier (1993), *Product Planning Management. Designing and Delivering Value*. McGraw-Hill International Editions.

Morwitz, V.G., J.H. Steckel, and A. Gupta (1997), "When do purchase intentions predict sales?," *Working paper*, Marketing Science Institute, Report No. 97-112, June.

Moskowitz, H.R. (1983), *Product testing and sensory evaluation of foods. Marketing and R&D approaches*. Westport, Connecticut USA: Food & Nutrition Press, Inc..

Moum, T. (1988), "Yea-saying and mood-of-the-day affects in self-reported quality of life," *Social Indicators Research*, 20, 117-139.

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

Munson, J.M. and S.H. McIntyre (1979), "Developing practical procedures for the measurement of personal values in cross-cultural marketing," *Journal of Marketing Research*, 16(1), 48-52.

N

Narayan, S. and J.A. Krosnick (1996), "Education moderates some response effects in attitude measurement," *Public Opinion Quarterly*, 60, 58-88.

Netemeyer, R.G., S. Durvasula, and D.R. Lichtenstein (1991), "A cross-national assessment of the reliability and validity of the CETSCALE," *Journal of Marketing Research*, vol. XXVIII, 320-327.

Ng, S.H. (1982), "Choosing between the ranking and rating procedures for the comparison of values across cultures," *European Journal of Social Psychology*, 12, 169-172.

Ng, S.H., A.B.M. Akthar-Hossain, P. Ball, M.H. Bond, K. Hayashi, S.P. Lim, M.P. O'Driscoll, D. Sinha, and K.S. Yang (1982), "Human values in nine countries," In: *Diversity and unity in cross-cultural psychology*. Rath, R., H.S. Asthana and J.B.H. Sinha (Eds.), 196-205. Lisse, the Netherlands: Swets & Zeitlinger.

Nunnally, J.C. (1978), *Psychometric Theory*. Second Edition. New York: McGraw Hill.

O

O'Donovan, D. (1965), "Rating extreme: Pathology or meaningfulness?," *Psychological Review*, 72(5), 358-372.

Ohmae, K. (1989), "Managing in a borderless world," *Harvard Business Review*, May-June, 152-161.

Öngel, U. and P.B. Smith (1994), "Who are we and where are we going to ?," *Journal of Cross-Cultural Psychology*, 25(1), 25-53.

P

Parameswaran, R. and A. Yaprak (1987), "A cross-national comparison of consumer research measures," *Journal of International Business Studies*, spring, 35-49.

Parasuraman, A., V.A. Zeithamland L.L. Berry (1985), "A conceptual model of service quality and its implications for future research," *Journal of Marketing*, 49(Fall), 41-50.

Paulhus, D.L. (1991), "Measurement and control of response bias," In: *Measures of Personality and Social Psychological Attitudes*, Volume 1. San Diego, CA: Academic Press Inc.

Peter, J.P. and J.C. Olson (1996), *Consumer behavior and marketing strategy*. Irwin.

Peterson, R.A., G. Albaum and R.F. Beltramini (1985), "A meta-analysis of effect sizes in consumer behavior experiments," *Journal of Consumer Research*, 12(June), 97-103.

Poortinga, Y.H. (1975), "Limitations on intercultural comparison of psychological data," *Nederlands Tijdschrift voor de Psychologie*, 30, 23-39.

Poortinga, Y.H. and R.S. Malpass (1986), "Making inferences from cross-cultural data," In: *Field methods in cross-cultural research*. Lonner, W.J. and J.W. Berry (Eds.), 17-46.

Poortinga, Y.H. and F.J.R. Van de Vijver (1982), "Cross-cultural generalization and universality", *Journal of Cross-Cultural Psychology*, December, 387 - 408.

Poortinga, Y.H. (1989), "Equivalence of cross-cultural data: An overview of basic issues," *International Journal of Psychology*, 24, 737-756.

Punj, G. and D.W Stewart (1983), "Cluster Analysis in Marketing Research: Review and Suggestions for Application," *Journal of Marketing Research*, 10(2) ,134-148.

R

Rankin, W.L. and J.W. Grube (1980), "A comparison of ranking and rating procedures for value system measurement," *European Journal of Social Psychology*, 10, 233-246.

Ray, J.J. (1983), "Reviving the problem of acquiescent response bias," *The Journal of Social Psychology*, 121, 81-96.

Rokeach, M.J. (1973), *The nature of human values*. New York: Free Press.

Rorer, L.G. (1965). "The great response-style myth," *Psychological Bulletin*, 63, 129-156.

Rossiter, J.R. (1999), "Construct Measurement in marketing," *Working Paper*, Department of Marketing, University of Wollongong, NSW, Australia, July 1999.

Russell, P.A. and C.D. Gray (1994), "Ranking or rating? Some data and their implications for the measurement of evaluative response," *British Journal of Psychology*, 85, 79-92.

S

Samiee, S. and I. Jeong (1994), "Cross-cultural research in advertising: An assessment of methodologies," *Journal of the Academy of Marketing Science*, 22(3), Summer, 205-217.

Schaninger C.M. and W.C. Buss (1986), "Removing Response-Style Effects in Attribute-Determinance Ratings to Identify Market Segments," *Journal of Business Research*, 14, 237-252.

Schöllhammer, H. (1973), "Stategies and methodologies in international business and comparative management research," *Management International Review*, 13(6), 17-31.

Schuman, H. and S. Presser (1996), *Questions and Answers in Attitude Surveys. Experiments on Question Form, Wording, and Context*. Thousand Oaks, London: Sage Publications.

Schwartz, S.H. and W. Bilsky (1987), "Toward a psychological structure of human values," *Journal of Personality and Social Psychology*, 53, 550-562.

Schwartz, S.H. and W. Bilsky (1990), "Toward a theory of the universal content and structure of values: Extensions and cross-cultural replications," *Journal of Personality and Social Psychology*, 58, 878-891.

Schwartz, S.H. (1992), "Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries," *Advances in Experimental Social Psychology*, 25, 1-65.

Schwartz, S.H. (1994), "Are there universal aspects in the structure and contents of human values?," *Journal of Social Issues,* 50(4), 19-45.

Sekaran, U. (1983), "Methodological and theoretical issues and advancements in cross-national research," *Journal of International Business Studies*, Fall, 61-73.

Serpell, R. (1979), "How specific are perceptual skills?," *British Journal of Psychology*, 70, 365-380.

Shepard, L.A. (1982), "Definitions of Bias," In: *Handbook of methods for detecting test bias*. R.A. Berk (Ed.) Baltimore: The John Hopkins University Press.

Shepard, R., (1989), "Factors Influencing Food Preferences and Choice," In: *Handbook of Psychophysiology of Human Eating*. Shepard R. (Ed.), 3-24. John Wiley & Sons, Ltd.

Shimp, T.A. and S. Sharma (1987), "Consumer Ethnocentrism: Construction and Validation of the CETSCALE," *Journal of Marketing Research*, 24(August), 280-289.

Singh, J. (1995), "Measurement issues in cross-national research," *Journal of International Business Studies*, Third Quarter, 597-619.

Smith, P.B. and S.H. Schwartz (1997), "Values," In: *Handbook of Cross-cultural Psychology, Second Edition, 3, Social Behavior and Applications*, Berry, J.W., M.H. Segall and C. Kagitçibasi (Eds.), 77-118. Needham Heights: Allyn & Bacon.

Smith, P.B., S. Dugan and F. Trompenaars (1996), "National culture and the values of organizational employees. A dimensional analysis across 43 nations," *Journal of Cross-Cultural Psychology*, 27(2), March, 231-264.

Song, X.M. and M.E. Parry (1997), "A Cross-National Comparative Study of New Product Development Processes: Japan and the United States," *Journal of Marketing*, 61(2), 1-18.

Soueif, M.I. (1968), "Extremeness, indifference and moderation response sets: a cross-cultural study," *Acta Psychologica*, 28, 63-75.

Sriram, V. and A.M. Forman (1993), "The relative importance of products' environmental attributes: A cross-cultural comparison," *International Marketing Review*, 10( ), 51-70.

Steenkamp, J.B.E.M. and H. Baumgartner (1995), "Development and cross-cultural validation of a short form of CSI as a measure of optimal stimulation level," *International Journal of Research in Marketing*, 12(2), 97-104.

Steenkamp, J.B.E.M. and H. Baumgartner (1998), "Assessing Measurement Invariance in Cross-National Consumer Research," *Journal of Consumer Research*, 25(June), 78-90.

Steenkamp, J.B.E.M., F. Ter Hofstede and M. Wedel (1999), "A Cross-National Investigation into Individual and National Cultural Antecedents of Consumer Innovativeness," *Journal of Marketing*, 63(2), 55-69.

Stening, B.W., and J.E. Everett (1984), "Response styles in a cross-cultural managerial study," *The Journal of Social Psychology*, 122, 151-156.

Szabo, S., J. Orley and S. Saxena (1997), "An approach to response scale development for cross-cultural questionnaires," *European Psychologist*, 2(3), 270-276.

T

Tanzer, N.K. (1995), "Cross-Cultural Bias in Likert-Type Inventories: Perfect Matching Factor Structures and Still Biased?," *European Journal of Psychological Assessment*, 11(3), 194-201.

Ten Berge, J.M.F. (1986), "Rotatie naar perfecte congruentie en de Multiple Groep Methode," *Nederlands Tijdschrift voor de Psychologie*, 41, 218-225.

Ten Berge, J.M.F. (1999), "A legitimate case of component analysis of ipsative measures, and partialling the mean as an alternative to ipsatization," *Multivariate Behavioral Research*, 34(1), 89-102.

Ter Hofstede, F., A. Audenaert, J.B.E.M. Steenkamp and M. Wedel (1998), "An investigation into the association pattern technique as a quantitative approach to measuring means-end chains". *International Journal of Research in Marketing*, 15(1), 37-50.

Ter Hofstede, F., J.B.E.M Steenkamp and M. Wedel (1999), "International Market Segmentation Based on Consumer-Product Relations," *Journal of Marketing Research*, 36(1), 1-17.

Terpstra, V. and R. Sarathy (1994), *International marketing*. Sixth Edition, The Dryden Press, Harcourt Brace College Publishers.

Tucker, L.R. (1951), "A method for synthesis of factor analysis studies," Personnel Research Section Report No. 984. Washington, DC: Department of the Army.

U

Usunier, J.C. (1996), *Marketing across cultures*. Second Edition. Prentice Hall Europe.

Usunier, J.C., (2000), *Marketing across cultures*. Third Edition. Pearson Education Limited.

V

Van de Vijver, F.J.R. and K. Leung (1997a), *Methods and data analysis for cross-cultural research*. Sage.

Van de Vijver, F.J.R. and K. Leung (1997b), "Methods and data analysis for cross-cultural research," In: *Handbook of Cross-Cultural Psychology, 2nd Ed., Volume I. Theory and Method*, Berry, J.W., Y.H. Poortinga and J. Pandey (Eds.), 257-300. Boston: Allyn and Bacon.

Van de Vijver, F.J.R. and Y.H. Poortinga (1982), "Cross-cultural generalization and universality," *Journal of Cross Cultural Psychology*, December, 387-408.

Van de Vijver, F.J.R. and Y.H. Poortinga (1997), "Towards an integrated analysis of bias in cross-cultural assessment," *European Journal of Psychological Assessment*, 13(1), 21-29.

Van de Vijver, F.J.R. and R.K. Hambleton (1996), "Translating tests: Some practical guidelines," *European Psychologist*, 1(2), 89-99.

Van der Kloot, W.A. (1997), *Meerdimensionale schaaltechnieken voor gelijkenis- en keuzedata. Ruimtelijke modellen voor psychologie, marktonderzoek en andere wetenschappen*. Utrecht: Lemma.

Van Herk, H., Th.M.M. Verhallen, and J. Barzilay (1994), "Methodological Issues in international segmentation," In: *Marketing: Its dynamics and challenges*. Bloemer, J., J. Lemmink and J. Kasper (Eds.), 1311–1314. Maastricht NL: EMAC.

Van Herk, H. and Th.M.M. Verhallen (1995a), "Prerequisites for international segmentation," In: *Marketing Today and for the 21st Century*, M. Bergadaà (Ed.), 2053-2057. Cery-Pontoise: ESSEC.

Van Herk, H. and Th.M.M. Verhallen (1995b), "Response effects in international research in the food area," In: *Proceedings of the Second International Conference on "The Cultural Dimension of International Marketing'*, 392-402. Odense, Denmark.

Van Raaij, W.F. and Th.M.M. Verhallen (1994), "Domain-specific Market Segmentation," *European Journal of Marketing*, 28(10), 49-66.

Verhage, B. and W.H. Cunningham (1989), *Grondslagen van de marketing*. Leiden: Stenfert Kroese Uitgevers.

W

Watkins, D., and S. Cheung (1995), "Culture, gender, and response bias. An analysis of responses to the self-description questionnaire," *Journal of Cross-cultural Psychology*, 26(5), 490-504.

Watson, D. (1992), "Correcting for Acquiescent Response Bias in the Absence of a Balanced Scale," *Sociological Methods & Research*, 21(1), 52-88.

Williams, J. (1991), "Constant questions or constant meanings? Assessing intercultural motivations in alcoholic drinks," ESOMAR International Research Ljubljana, February, 159-175.

Winkler, J.D., D.E. Kanouse and J.E. Ware Jr. (1982), "Controlling for Acquiescence Response Set in Scale Development," *Journal of Applied Psychology*, 67(5), 555-561.

Wyndham, J. and R. Goosey, 1997. 'It is time we started using statistics!' *Marketing & Research Today*, 25(4), November, 244-253.

Wyer, R.S. (1969), "The Effects of General Response Style on Measurement of Own Attitude and the Interpretation of Attitude-Relevant Messages," *British Journal of Social and Clinical Psychology*, 8, 104-115.

Y

Yeh, L.L., K.O. Kim, P. Chompreeda, H. Rimkeeree, N.J.N. Yau, and D.S. Lundahl (1998), "Comparison in use of the 9-point hedonic scale between Americans, Chinese, Koreans, and Thai," *Food Quality and Preference*, 9(6), 413-419.

Yip, G.S. (1995), *Total Global Strategy. Managing for Worldwide Competitive Advantage*. Englewood Cliffs, N.J.: Prentice Hall.

Z

Zegers, F.E. and J.M.F. Ten Berge (1985), "A family of association coefficients for metric scales," *Psychometrika*, 50, 17-24.

Author Index

H

J

K

Subject Index

Samenvatting

De globalisatie van markten voor goederen en diensten is, in de afgelopen decennia, voor bedrijven steeds belangrijker geworden. In de laatste 50 jaar is de wereldwijde handel zeventien keer zo groot geworden. Daar bedrijven meer en meer betrokken zijn bij globale handel is globale marketing van essentieel belang geworden. Dit laatste vereist kennis over de culturele, economische, wettelijke en geografische verschillen tussen de markt in het eigen land en de markten binnen andere landen. Om inzicht te krijgen in die complexe internationale omgeving hebben bedrijven behoefte aan marktinformatie. De vraag naar marktinformatie is derhalve ook sterk groeiend; in vergelijking met 1980 wordt er nu wereldwijd vijf keer zoveel besteed aan marktonderzoek. Bij dit marktonderzoek worden ook steeds vaker vergelijkingen gemaakt tussen verschillende landen.

Als er vergelijkingen tussen landen gemaakt worden is de vergelijkbaarheid of de 'equivalentie' van de marktinformatie een basisvoorwaarde. Equivalentie bestaat als informatie uit verschillende landen direct vergeleken kan worden. Bijvoorbeeld, het aantal inwoners in Duitsland en Nederland kan direct vergeleken worden. Echter, veel zaken kunnen niet direct vergeleken worden; bijvoorbeeld, 'innovatief gedrag' betekent niet per definitie hetzelfde in verschillende landen. Hieruit blijkt dat de bepaling van equivalentie van informatie van vitaal belang is; als informatie niet equivalent of gebiast is kunnen getrokken conclusies op z'n minst ambigu zijn of zelfs geheel fout. Ondanks het risico van foute conclusies is er in de praktijk van het marktonderzoek weinig bekend over equivalentie. Bovendien, ook in de academische literatuur is er weinig aandacht voor onderzoeksmethodologie bij internationaal vergelijkend onderzoek in het algemeen en voor equivalentie in het bijzonder.

In dit proefschrift worden de bestaande frameworks voor het benaderen van equivalentie naast elkaar gezet. Deze frameworks komen uit de marketing (e.g., Craig & Douglas, 2000), de cross-culturele psychologie (Van de Vijver & Leung, 1997a) en business studies (b.v., Singh, 1995). Uit de vergelijking tussen de verschillende frameworks blijkt dat Van de Vijver en Leung's framework het meest beknopt en duidelijk is. In dit framework zijn maar drie vormen van equivalentie, namelijk construct equivalentie, equivalentie van meeteenheid en scalaire equivalentie. Deze drie typen van equivalentie zijn genest; er kan alleen equivalentie van meeteenheid zijn als er construct equivalentie is. Om scalaire equivalentie te krijgen dienen de andere twee vormen van equivalentie te bestaan.

In hoofdstuk 2 wordt beschreven hoe Van de Vijver en Leung's framework geïntegreerd kan worden in marketing. Met de keuze voor dit framework wordt het framework dat gebruikelijk is in marketing, het framework van Craig en Douglas (2000), verlaten. Door het nieuwe framework te gebruiken wordt equivalentie opnieuw gedefinieerd als dat deel van het cross-culturele onderzoeksproces dat gecontroleerd kan worden met behulp van statistische analyse procedures. In deze nieuwe definitie worden zaken in de design fase van een studie, zoals allocatie van produkten aan produktcategorieën, vertaling van vragenlijsten en streekproef trekkingen beschouwd als mogelijke bronnen van bias. Equivalentie van deze zaken kan niet apriori vastgesteld worden. Wel kunnen in de design fase maatregelen genomen worden om de kans te vergroten dat gegevens equivalent zijn. Echter, in het voorgestelde nieuwe framework voor marketing, kan enig *bewijs* van equivalentie alleen worden verkregen door middel van statistische analyse procedures nadat alle gegevens verzameld zijn.

Bias is een kernbegrip in het nieuwe framework. Bias kan ontstaan in de loop van het onderzoeksproces; of resultaten als equivalent beschouwd kunnen worden wordt vastgesteld aan het einde van het onderzoeksproces. In het algemeen is het zo dat bias het niveau van equivalentie dat bereikt kan worden verlaagt. Er kan bias zijn in (1) de onderwerpen waarnaar items refereren, (2) de meetprocedures en de

gebruikte meetschalen, of (3) in bepaalde items in de vragenlijst. Deze vormen van bias worden respectievelijk construct, methode en item bias genoemd. Elk van deze vormen van bias kan voorkomen in internationale marketing studies. Bijvoorbeeld, construct bias kan ontstaan als hetzelfde produkt in andere situaties wordt gebruikt in verschillende landen; methode bias kan ontstaan door de methode van dataverzameling of door verschillen in response stijlen bij respondenten; tenslotte, item bias kan ontstaan als een specifiek item niet van toepassing is in een van de bestudeerde populaties. Bias is niet ongrijpbaar; vaak kan het onder controle gehouden worden. Voor dit doel zijn in dit proefschrift de relaties tussen de fasen in het onderzoeksproces en de bronnen van bias besproken (zie Hoofdstuk 2). Bias kan optreden in alle fasen van het onderzoeksproces. Er is uitgelegd dat aandacht voor bias en, wanneer mogelijk, controle voor bias noodzakelijk is in alle fasen van het onderzoeksproces. Bijvoorbeeld, bij vertalen van vragenlijsten kan bias gereduceerd worden door de richtlijnen voor vertalingen te volgen (bijv., Van de Vijver & Hambleton, 1996); bepaalde vormen van bias in de dataverzameling kunnen naderhand gecontroleerd worden door extra variabelen aan de data set toe te voegen. Hierbij kan, bijvoorbeeld, gedacht worden aan het opnemen van een variabele die aangeeft welke interviewer het vraaggesprek heeft uitgevoerd.

De dreiging van bias is niet gelijk voor alle typen gegevens. De mate waarin metingen gevoelig zijn voor bias wordt beïnvloed door het onderwerp van de studie en de interpretatie van de resultaten. In het algemeen geldt dat naarmate een onderwerp meer concreet wordt het minder gevoelig is voor bias. Bijvoorbeeld, de meting van iemands water consumptie is minder gevoelig voor bias dan meting van zijn/haar waarden structuur. Dit beïnvloedt ook de mate waarin valide cross-culturele inferenties mogelijk zijn (zie Hoofdstuk 3). Dit hangt af van (1) de mate waarin het onderzoeksdomein accuraat beschreven kan worden, en (2) of concepten operationeel gedefinieerd kunnen worden of niet. In marketing kunnen verscheidene domeinen, in principe, volledig worden beschreven. Bijvoorbeeld, er kan een lijst gemaakt worden van alle automerken die te koop zijn in een bepaald land. Zulke domeinen zijn relatief concreet en valide inferenties met betrekking tot het domein

kunnen gemaakt worden. De onderzoeker heeft de mogelijkheid om bias in deze gegevens onder controle te houden. Als een onderzoeksdomein meer abstract wordt dan is het maken van een lijst van alle elementen niet mogelijk. Men creëert hypothetische constructen waarvan verondersteld wordt dat zij bepaalde psychologische traits (bijv. 'innovativiteit') meten. In dit laatste geval is de dreiging van bias substantieel. Er kan construct, methode en item bias zijn en de onderzoeker heeft beperkte mogelijkheden om deze te beheersen in de studie. Als er aandacht is voor bias in cross-culturele studies, dan is dat veelal voor construct bias en in mindere mate voor item bias. Aan methode bias wordt doorgaans geen aandacht besteed.

In cultuur vergelijkend onderzoek kan het (gehele) onderzoeksinstrument beïnvloed worden door methode bias. Een zeer belangrijke potentiële oorzaak van methode bias in cultuur vergelijkend onderzoek zijn response stijlen. De effecten van response stijlen op de uitkomsten van studies zijn aangetoond in de psychologie, maar in marketing zijn er nauwelijks publikaties over dit onderwerp. Om bij te dragen aan de bestaande kennis over response stijlen in internationale marketing zijn er in dit proefschrift drie empirische studies uitgevoerd (zie Hoofdstukken 4, 5 en 6).

Er is onderzocht of er verschillen in response stijlen bestaan in zes landen in de Europese Unie (EU) te weten, Griekenland, Italië, Frankrijk, Spanje, Duitsland en het Verenigd Koninkrijk. Het was reeds bekend dat er verschillen in response stijlen bestaan tussen de USA en Azië, maar verschillen in response stijlen tussen landen binnen de EU waren, tot nu toe, nog nooit vergeleken. Wij hebben gevonden dat er tendenties zijn naar 'yeasaying' en 'extreem response gedrag'. De resultaten tonen aan dat de verschillen en overeenkomsten, die gevonden zijn tussen landen redelijk consistent zijn over twee rating scale typen en, in mindere mate, over drie gedragsdomeinen (koken, wassen en persoonlijke verzorging). Er is gevonden dat Griekse respondenten meer geneigd zijn om 'yeasaying' en 'extreem response gedrag' te vertonen dan respondenten in Italië, Frankrijk, Spanje, Duitsland en het

Verenigd Koninkrijk. De neiging om deze response stijlen te vertonen was ook hoger in Italië en Spanje dan in Frankrijk, Duitsland en het Verenigd Koninkrijk. Van alle onderzochte nationaliteiten waren de Duitsers het minst geneigd om extreem te antwoorden. In hoofdstuk 5 is onderzocht wat de antecedenten van de response stijlen zijn. Hiervoor zijn 55 regio's in de EU gebruikt. Resultaten geven aan dat er verschillen zijn tussen regio's. Het bleek dat het land waarin een regio ligt en de globale coördinaten van de regio verschillen in response stijlen op het niveau van de regio bepalen. Land bleek vooral het verschil in response stijl te verklaren. Op het individueel niveau beïnvloeden socio-demografische kenmerken zoals leeftijd, opleiding en inkomen de mate waarin respondenten 'yeasaying' of 'extreem response gedrag' vertonen. Deze effecten waren significant en zeker niet verwaarloosbaar.

Het empirische onderzoek toont aan dat er verschillen in response stijlen bestaan tussen landen in de EU. Variantie veroorzaakt door verschillen in response stijlen kan verwijderd worden met behulp van regressie; items kunnen onafhankelijk gemaakt worden van de response stijl indices (zie hoofdstukken 4 en 5). Volgende analyses kunnen dan uitgevoerd worden op de residuen van deze regressie. Een andere oplossing om met response stijlen om te gaan is het gebruiken van een andere methode van dataverzameling zoals rangordeningen in plaats van rating. Door een dergelijke methode worden antwoordstijlen zoals 'yeasaying' of 'extreem response gedrag' vermeden. In hoofdstuk 6 is onderzocht of beoordeling op rating schalen en rangordeningen van dezelfde items door dezelfde respondenten gelijke resultaten opleveren over landen heen. Bij rangordenen worden respondenten gedwongen om een unieke score toe te kennen aan items. Bijvoorbeeld, het toekennen van een getal tussen 1 en 9 aan negen items. In de studie werden vijf groepen respondenten onderscheiden op basis van hun antwoordpatroon op de rating schalen. Er was een (a) 'full-discrimination' patroon, een (b) 'gematigd yeasaying patroon met verscheidende ties'; een (c) 'extremen vermijdend' patroon; een (d) 'yeasaying' en een (e) 'naysaying' patroon. Het bleek dat respondenten met patronen (b) en (c) meer overeenkomsten hadden tussen ratings en rangordeningen

dan de andere respondenten. Deze respondenten waren over het algemeen hoger opgeleid, hadden een hoger inkomen en/of woonden in Frankrijk of Duitsland. In de studie werd verder de predictieve validiteit van de twee meetmethoden onderzocht. Het bleek dat de predictieve validiteit voor ratings licht hoger was voor alle groepen respondenten. Ratings lijken dus de voorkeur te hebben boven rangordeningen in cross-cultureel onderzoek.

De uitkomsten van deze studies impliceren dat aandacht voor bias belangrijk is in studies, waarbij marktonderzoek informatie uit verschillende landen vergeleken wordt. Bij internationaal marktonderzoek zijn er diverse mogelijkheden om bias te controleren. Bijvoorbeeld, inzetten van kwalitatief (voor)onderzoek is een ideale manier om construct bias te beperken en na het uitvoeren van een kwantitatieve studie kunnen gegevens achteraf gecorrigeerd worden voor methode bias. Door aandacht voor bias in alle fasen van het marktonderzoek proces zal de kwaliteit van internationaal vergelijkend marktonderzoek toenemen. Bovendien zal correctie voor bias kunnen helpen om de validiteit van de uitkomsten van internationaal vergelijkend onderzoek te vergroten.

Curriculum vitae

Hester van Herk (1961) studied psychology at Leyden University with a major in research methodology and minors in social and organizational psychology and in mathematics. Before graduation in 1986 she started working at the Netherlands Central Bureau of Statistics. There she was a scientific researcher at the Department of Statistical Methods. After this, she worked at Solvay Duphar as a statistical software consultant. In 1989 she started working in marketing research. First at ABN Amro, later at the marketing research agency MarketResponse. Since 1993 she has been employed part-time at the department of Marketing at Tilburg University. First as a scientific researcher, since 1997 as an assistant professor. In addition to her work at the university, Hester is the director/owner of H. van Herk Marktonderzoek Advies, a marketing research and consultancy firm.

Hester has broad interests. For instance, since 1990 she has been an active member of the Dutch Society for Marketing Research (NVMI). At present she occupies the chair of its International Marketing Research group. Her research interests are in cross-cultural research, international marketing and research methodology. About these subjects she has already published articles in various journals.

Stellingen

behorende bij het proefschrift

Equivalence in a Cross-national Context: Methodological & Empirical Issues in Marketing Research

van

Hester van Herk

28 juni 2000

1. Het combineren van kennis uit de economie en de psychologie is noodzakelijk voor het verklaren van cross-culturele verschillen in consumentengedrag.

2. Verschillen in antwoorden op vragen door consumenten uit landen in de EU worden in belangrijke mate verklaard door verschillen in antwoordstijlen (dit proefschrift hoofdstuk 4).

3. De inconsistenties in het framework van Douglas en Craig (1983) voor equivalentie hebben ertoe geleid, dat dit framework zelden gebruikt wordt in marktonderzoek (dit proefschrift hoofdstuk 2).

4. Het gebruik van multi-level analyse in cultuur vergelijkend onderzoek verhoogt het begrip van de complexe relaties die binnen en tussen culturen bestaan (dit proefschrift hoofdstuk 5).

5. Ranking van items in plaats van rating is geen oplossing voor het antwoordstijl probleem (dit proefschrift hoofdstuk 6).

6. In Zuid-Europa vertoont men meer instemmend en extreem antwoordgedrag dan in Noord-Europa (dit proefschrift hoofdstukken 4 en 5).

7. Ongelijkheden in gedragsrepertoire, die het gevolg zijn van verschillen in de materiële omgeving, kunnen gemakkelijk overeenkomsten in psychisch functioneren verhullen.

8. Het inzetten van studenten als respondenten is eerder efficiënt dan effectief voor het verkrijgen van inzicht in culturele verschillen.

9. De toename van de werkdruk in Nederland leidt tot een daling in de animo voor vrijwilligerswerk, zoals bestuurstaken.

10. De Europese richtlijnen voor Dierentransport (1999) hebben nog niet geleid tot een aanwijsbare vermindering van het dierenleed bij dergelijke transporten.

11. Het aanspreken door buitenstaanders van een groep Rotarians, bestaande uit mannen en vrouwen, met 'heren van de Rotary' geeft aan dat men niet alleen binnen maar ook buiten Rotary nog moet wennen aan gemengde clubs.

Uitnodiging

Op woensdag 28 juni 2000
verdedig ik mijn proefschrift:

Equivalence in a
cross-national context:
Methodological &
empirical issues in
marketing research

De promotie vindt plaats
in de aula van de
Katholieke Universiteit
Brabant,
Warandelaan 2 in Tilburg.

Aanvang 14.00 uur precies.

Aansluitend is er een receptie
in de Goossenszaal
naast de aula.

Hester van Herk
Muiderweg 49
3891 DA Zeewolde
Tel. 036-5223920