2005

Faculteit
Toegepaste Economische Wetenschappen

Measurement Invariance Issues in International Management Research

Proefschrift voorgelegd tot het behalen van de graad van
Doctor in de Toegepaste Economische Wetenschappen
te verdedigen door

Alain DE BEUCKELAER

Promotors : Prof. dr. G.K. Janssens
Prof. dr. G. Swinnen

LIMBURGS UNIVERSITAIR CENTRUM
PARTNER IN DE UNIVERSITEIT LIMBURG

According to the guidelines of the Limburgs Universitair Centrum, a copy of this publication has been filed in the Royal Library Albert I, Brussels,as publication D/2005/2451/1

Promoters:

Professor dr Gerrit K. Janssens;
and Professor dr Gilbert Swinnen

(both at Limburgs Universitair Centrum [LUC], Belgium)

Other members of the Ph.D. committee:

Professor dr Mieke Van Haegendoren,
Vice-chancellor of LUC and president of the Ph.D. committee;

Professor dr Malaika Brengman (LUC, Belgium);
Professor dr Steffen Kühnel (Georg-August-Universität, Göttingen, Germany);
Professor dr Albert Satorra (Universitat Pompeu Fabra, Barcelona, Spain);
Professor dr Patrick Van Kenhove (Universiteit Gent, Belgium).

Printed in Belgium

Acknowledgements

I am indebted to the many people who have contributed to this Ph.D. project in one way or another. First and foremost, I would like to express my deep gratitude to the promoters of this Ph.D. dissertation, Professor dr Gilbert Swinnen and Professor dr Gerrit K. Janssens. I am forever indebted to them for their continual and extensive guidance and support. At the end of 2001, they offered me the opportunity to start a Ph.D. project at Limburgs Universitair Centrum (LUC) in Belgium. Since then, both promoters have provided me with many valuable suggestions and comments on parts of the manuscript in various stages of its development.

I would like to thank Professor dr Steffen Kühnel whose 'Advanced Course in Structural Equation Modelling' I attended at the 28th Essex Summer School in Social Science Data Analysis and Collection (University of Essex, Colchester, U.K., 1995). Professor Kühnel introduced me to some more specialised topics in Structural Equation Modelling (e.g. advanced estimation procedures, goodness-of-fit testing etc.). My prior experience with Structural Equation Modelling was a valuable asset when working on this Ph.D. project. I have benefited a lot from Professor Kühnel's lectures, as well as his ideas and comments on some (draft) chapters of this manuscript.

I am also indebted to Professor dr Albert Satorra, a well-respected scholar in the field of multivariate analysis methods for his stimulating thoughts (e.g. when we were both attending the 2004 SMABS conference in Jena, Germany). Next, I would like to thank dr Philip Barbonis, Professor dr Peter Bentler, Professor dr Malaika Brengman, Professor dr Jan Broeckmans, dr Conor Dolan, dr Hugo Duivenvoorden, Professor dr Patrick Groenen, Professor dr Patrick van Kenhove, and Professor dr Hans van Trijp for their helpful advice and suggestions. They all provided me with some useful remarks on one or more chapters of this manuscript.

Many thanks to Sunila Supavadeeprasit for all the time she has spent making all necessary language and style corrections to the text of this dissertation.

A special word of thanks goes to Professor dr Jacques Tacq. Since I have known him from my studies at the Catholic University of Brussels (M.Sc. programme in Quantitative Analysis in the Social Sciences), I realise that working with multivariate statistics can be extremely fascinating. I was very privileged to become his 'course assistant' a long time ago … and I must admit: I still enjoy making my (yearly) contribution to his course on multivariate statistics at The Catholic University of Brussels. Professor Tacq has shown me what multivariate statistics is about, how it can be applied, how one can inspire students to learn more about it, and whose courses I should attend to get acquainted with the

more advanced topics in the field of research methodology and statistics. Professor Tacq, for all of this, thanks a lot!

In this dissertation, data from existing research projects were analysed. I was fortunate to get exceptionally valuable data from Unilever's Global Human Resource Centre of Expertise (see Chapter 5 of this dissertation). I would like to thank Saskia Trienen and dr Brigitte Tantawy-Monsou for their willingness to provide me with the data, as well as some crucial background information on the study. I am also indebted to the Marketing Science Centre of Research International Ltd. (London), a large market research agency. The Marketing Science Centre gave me permission to work with data from two international consumer studies. Both studies are presented in Chapter 6 of this dissertation.

Further, I wish to express my deepest appreciation to my wife, Guo Ming, who encouraged me to keep on working on my dissertation till late evening hours, during weekends, and holidays. Ming, thanks for your support! During the first six years of Tom's life [our son!], Tom had a father who has spent more time playing 'with his computer' than with him. Tom, I hope I will be able to catch up with you soon.

Last but not least, I would like to thank my parents for their moral and financial support when doing my undergraduate studies. Without their support, I would never had the opportunity to start working on a Ph.D. project.

Alain DE BEUCKELAER

Berchem (Antwerp), Belgium
January 2005

"Observing without evaluating is the highest form of human intelligence"

Jiddu Krishnamurti (1895-1986),
Indian philosopher

For *Tom*, *Ming*, *Anny* and *Francois*

CONTENTS

Chapter 1. Introduction

"Factorial invariance is one of the most elegant conceptions that quantitative psychology has produced to date."

J.R. Nesselroade

1.1. Introduction

Are there any differences between people living in different parts of the world? What keeps them busy the whole day? What values in life guide their behaviour? Are they different or very similar? What about their roles in society? Are they children, parents, colleagues, or consumers? Do they exhibit similar behaviours in all of these roles? How are they seen by others, and how do they view other people?

Such questions are not just 'food for thought' for scientists who spend most of their time investigating human behaviour (e.g. anthropologists, psychologists, sociologists, political scientists, and economists). People who take influential decisions in today's 'global environment' are likely to be more successful if they know the answers to such questions. World-leaders, such as the president of the United States, is more likely to get the U.S. foreign (military) policy approved by other leading politicians if he can show that the measures taken (under that policy) are necessary to create a peaceful world. The reason for the approval is obvious. 'Feeling safe' (i.e. a peaceful environment) is a need which characterises mankind, and is, therefore, a reasonable motivation. To defend the U.S. foreign policy on the basis of economical considerations would be a formula for failure in international politics.

The same principles apply equally well in the field of international business. Marketers, for instance, are likely to be more successful in a global business environment if they can convince consumers that consumption of their products (or services) will help them achieve certain values in life which are most important to them (e.g. staying healthy). One may reasonably expect that products / services which are instrumental in achieving 'universal' values in life have more sales potential than products / services which help achieve values in life which are 'culture-specific'. Mobile phones, for example, are now sold all over the world as mobile communication may serve as a means to an end. In this context, the end may be that anyone can feel part of a group at any time. Having a sense of belonging may be considered to be a universal value in life. Unlike mobile phones, headscarves are a 'culture-specific' (or religion-specific) product. Wearing a headscarve

does not lead to the achievement of a universal value in life. Most non-Islamic women will never wear headscarves!

To launch products (or services) which are purchased by consumers all around world remains one of the biggest challenges for multinational companies. Many successful new products are designed such that they meet the needs of the 'global consumer'. To meet the needs of the global consumer, international comparative analysis is required to assess common and region-specific consumer values, consumer needs, and consumer attitudes (e.g. towards genetically modified foods). If the new product taps into consumer needs which are common across countries, then the product has an increased chance of becoming successful in all countries in the global market.

To make sure that meaningful conclusions can be drawn from international comparative research, the data on consumers is required to be comparable across countries (e.g. Kumar, 2000). There are many conditions that need to be fulfilled to guarantee data comparability across countries. These conditions are briefly explained later in this introductory chapter. Here and now, the focus is now on the validity and reliability of the measurement scales used in international management research. A wide variety of (complex) scales (or 'measurement instruments') have been developed to measure consumer values, consumer needs, or consumer attitudes. An example of a measurement instrument is a multi-item battery of statements measuring consumers' loyalty towards a brand. Many other examples can be found in a book by Bruner and Hensel (1997) and the 'Handbook of Marketing Scales' by Bearden and Netemeyer (1999). It is of crucial importance that the measurement instruments used in international research are meaningful from a cross-country perspective. In technical jargon, the measurement instrument is required to exhibit *measurement invariance* across countries

In this dissertation, an investigation will be made to assess the extent that the assumption of measurement invariance across countries is realistic in actual research practice. Actual research examples (i.e. case studies) from the field of international management research will be presented to the reader. It will be investigated whether or not the measurement instruments, which were used, satisfy the condition of measurement invariance across countries. If these measurement instruments turn out to be non-invariant across countries, it will be assessed[1] whether or not cross-country comparisons based on these non-invariant measurement instruments are truly unreliable. In this dissertation, the reliability of cross-country comparisons will also be investigated by means of a simulation study. More

[1] This assessment will be made only if it is technically possible (i.e. when the measurement instrument 'comes close to' satisfying the condition of measurement invariance across countries / cultures). More clarification will be given in chapters 5 and 6 of this dissertation.

details on the simulation approach will be given at the end of this first introductory chapter. In addition, the main research question of this Ph.D. research will be specified in a more formal way.

The next section provides a brief introduction to the field of international management research.

1.2. International management research

1.2.1. Introduction

Martinez and Toyne (2000) looked at the meaning of the word 'management' when used in combination with other words. They focused on the use of the word 'management' by both the Academy of Management and Kroontz (1980). On the basis of their study they conclude that 'management' has at least five different meanings. First of all, it is used as a qualifier when preceding or following such words as education, theories, and research. Secondly, it is also used to encompass the traditional managerial functions when following such words as 'conflict', 'human resource', 'operations', 'technology', and 'innovation'. In the context of the Social Issues in Management division of the Academy of Management, the word 'management' is narrowly interpreted as referring to corporate social responsibility and performance, and to business ethics. Fourthly, the definition of 'management' is broadened to include the activities and organisations serving a social need (i.e. activities or organisations that are not necessarily economic in orientation or focus). In the International Management division of the Academy of Management, the word (international) management is used to encompass the Academy's interpretation of the word's universal meaning, but with an international or cross-cultural dimension. In this Ph.D. dissertation, the focus is on the international dimension of management, and management research in particular.

The formation of large multinational companies and the continuous rise in cross-border trade has led to a significant increase in international business and management research (Hui, 1990; Boddewyn and Iyer, 1999). International research is commonly aimed at investigating differences and similarities between selected countries. A major problem is the complexity of international research. One of the main reasons why international research is very complex is because of the influence of 'culture' on the behaviour of individuals and organisations. The concept of culture can be defined[2] in many different ways. Culture can be viewed as *"a shared system of representations and meaning"* (Goodenough, 1971), *"a system of meaning, ideas and patterns of thought"* (Goffman, 1974), *"basic assumptions or value orientations on the nature of man's relationships to nature and to other human beings"* (Kluckhohn and Strodtbeck, 1961). Most definitions of culture centre on 'human values' occuring frequently in a particular society (CIM, 1999, p. 371). How culture influences human behaviour is well-formulated by Harris and Moran (1987):

[2] A review of many definitions of culture is given in Usunier (1996).

"Culture gives people a sense of who they are, of belonging, of how they should behave, and what they should be doing. It provides a learned, shared, and interrelated set of symbols, codes, and values that directly justify human behavior." (Harris and Moran, 1987).

Baligh (1994) views culture as a set of 'components'. Specific components that have an impact on international management research are (see Usunier, 1998): relational patterns (e.g. dominant family and kinship patterns), language and communication, institutional and legal systems, values and value systems, (behavioural) norms, time orientations (e.g. punctuality), mindsets (i.e. mental maps and structures which correspond to a certain type of world view, linked in particular to the language structure).

According to Martinez and Toyne (2000) a distinction should be made between 'internationalised' management (research), and (truly) 'international' management (research). In internationalised management (research) the focus is on identifying those environmental factors (e.g. cultural, legal, political, and social factors) that may have a significant influence on the management of an organisation's operations when extended to include a foreign location or when comparing two or more countries. The environmental factors (including culture) are not necessarily seen as having an effect on management theories. Internationalised management (research) is, therefore, culture-bound. In (truly) international management (research) environmental factors (i.e. including culture) are taken into account when building new management theories. As a consequence, international research in management may add either distinctive or unique knowledge to the body of management knowledge (Martinez & Toyne, 2000).

The goal of international research in management is often to develop theories and models which help identify and explain cross-cultural practices at a national, organisational, and managerial level. A deeper understanding obtained from such theories and models should ultimately result in some form of competitive advantage for the multinational company. Seeking competitive advantages is a necessity for every multinational company as the global arena in which they operate is becoming increasingly competitive. In particular, competitive advantages which are more difficult to obtain by competitors (e.g. due to the high complexity of copying the source of the competitive advantage) are more likely to result in unique benefits for the multinational company for a relatively long period of time (Porter, 1980).

Since competitive advantages exist in many areas and can be achieved, the field of international business or management research (in short: international management research) can be subdivided into many different subdisciplines or research areas. Two important research areas (for multinational companies) are: global human resource (HR) management and global marketing. At its most fundamental level, the purpose of global HR management is to establish patterns of HR practices across different cultures. Similarly, the purpose of global marketing is to establish patterns of consumer behaviour across different cultures.

In international research, the term 'culture' is often related to a particular 'country' (or a group of countries). Obviously, this is not correct since a country usually comprises diverse cultures. Despite the actual difference between both terms, the terms 'culture' and 'country' are often used interchangeably, also in this dissertation. If there is only one dominant culture in a particular country, then cross-countryl comparisons are expected to reveal differences between (dominant) cultures reasonably well. In case there are at least two dominant cultures in a particular country, one may always collect sufficient data on all (dominant) cultures to also allow for comparisons between these different cultures.

In the next section, a more detailed discussion on global HR management and global market and consumer research is provided. In addition, reference is made to other areas in international management research.

1.2.2. Global research in Human Resource (HR) management

Several authors have argued that management of HR constitutes one of the more innovative sources compared to the traditional ones such as capital, technology, and location (Bartlett and Ghoshal, 1991; Sparrow et al., 1994; Schuler and Rogovsky, 1998). More and more business executives recognise the importance of effective people management for both short and long-term competitiveness of firms. The ability to attract, develop, and motivate people is even more crucial when companies globalise and set up subsidiaries overseas (Schuler and Jackson, 1996; Taylor et al., 1996).

Even though global HR practices may create many competitive advantages for a multinational company, strategically co-ordinating different organisational units across national barriers is difficult to achieve (Torrington, 1994). An even more difficult issue is to find the right balance between two extremes, namely: constraining HR practices to be identical across national borders, and establishing global HR practices which are unique to each individual nation (or country). 'Thinking global, and acting local' may be the best strategy to follow. This implies that some local adaptations of the global HR policy should always be considered.

Some useful research instruments for monitoring (and improving) global HR policies are: common performance appraisal systems (Borsman, 1991; Pulakos, 1997), and global employee opinion surveys (Ryan et al., 1999). These (global) research instruments enable HR professionals to take policy measures which are meant to lead to a higher performance of the employees, and higher job satisfaction levels. As a consequence, employees may stay longer with company. In this dissertation, the measurement quality of a survey instrument used in a global employee opinion survey will be investigated.

According to Tung and Punnett (1993), the area of international HR management is only slowly developing as a field of academic study and much remains to be done in this field.

1.2.3. Global market and consumer research

Global market and consumer research is another important research area in international management research. According to Wang (1996), one of the most important roles of global market and consumer research is to facilitate strategic decisions regarding international marketing segmentation and marketing-mix based on consumers' responses to global marketing efforts. A global marketing strategy assumes that a common marketing-mix approach across national borders (or, alternatively, a common marketing-mix approach with only minor adaptations per country). Such a global strategy may generate a lot of benefits (or competitive advantages) to the multinational company. Cost reductions through economies of scale (e.g. costs of production, advertising, and distribution), improved quality of products, and increased bargaining and competitive power are examples of such benefits (Levitt, 1983; Yip, 1995). Mitra & Golder (2002) mentioned that several multinational companies have a business interest in at least thirty countries (e.g. L' Oréal, Procter & Gamble, Unilever, Coca Cola, McDonald's, etc.).

Clearly, there are many factors that make it difficult to establish a common (or global) marketing strategy. There is a high degree of variability between nations in terms of environmental factors. The environmental factors are indicated by the mnemonic 'SLEPT', which refers to the Social and cultural, Legal, Economic, Political, and Technological factors. Despite of the fact that all of these factors make a global marketing strategy more difficult, multinational companies recognise that certain groups of consumers in different countries often have more in common with one another than with other consumers in the same country. Hence, multinationals choose to serve segments that transcend national borders (Hassan and Katsanis, 1994). A good example of a global consumer segment is the 'Teenager Segment' (see Hassan & Katsanis, 1994).

In many industries, national borders are becoming less and less important as an organising principle for international activities. As a consequence, multi-domestic strategies have become less relevant over time (Yip, 1995). This trend is accelerated by several developments in the area of international business. These developments include: regional unification, shifts to open economies, global investment, manufacturing, and production strategies, expansion of world travel, rapid increase in education, literacy levels, and urbanisation among developing countries, convergence of purchasing power, lifestyles and tastes, advances in information and communication technologies, the emergence of global media, and the increasing flow of information, labour, money, and technology across borders (Alden et al., 1999; Gielens and Dekimpe, 2001; Hassan and Katsanis, 1994; Hassan and Kaynak, 1994; Mahajan and Muller, 1994; Parker and Tavassoli, 2000; Yip, 1995).

Many global companies such as Coca-Cola, McDonald's, Sony, British Airways, Ikea, Toyota, and Levi-Strauss have had success in integrating their international strategies. International market segmentation tools (e.g. ter Hofstede et al., 1999; Steenkamp and ter Hofstede, 2002) can be used as a means to identify global consumer segments. Consumers belonging to a global consumer segment may come from different nations, but they may have very similar consumer needs. Therefore, such consumers may be targeted with a global marketing strategy, regardless of the country to which they belong. As such, targeting global consumer segments combines the benefits of standardisation (e.g. lower costs, better product quality) with the benefits of adaptation (e.g. meeting the specific needs of consumers) (Steenkamp and ter Hofstede, 2002).

As shown in a literature study in the early 1990s (Aulakh and Kotabe, 1993), international marketing segmentation had received only very limited attention in international marketing. Based on a review of over hundreds of published articles on international marketing in the period 1980-1990, Auklah and Kotabe (1993) concluded that only 1% dealt directly with international market segmentation (Aulakh and Kotabe, 1993). Douglas and Craig (1992) also claimed that, unlike in domestic marketing where segmentation is a central issue, segmentation received only little attention in international marketing (Douglas and Craig, 1992). More recently, however, international marketing segmentation has gained much more popularity (Wang, 1996).[3]

[3] The reader who is interested in a historical assessment of the literature on international consumer research between the 1960s and the mid 1990s is encouraged to consult Wang (1996).

1.2.4. Other areas in international management research

Apart from research in global HR management and global marketing research there are more areas in international management research. Usunier (1998) wrote in his book entitled 'International and Cross-Cultural Management Research':

"International business issues have progressively expanded from the theory of internationalization of firms and foreign direct investment to, inter alia, export management (…), the relationships with host countries, and international business negotiations. More specialized topics have also been developed, some of which are quite typical of the international arena such as countertrade. The international dimension of functional areas has developed following the globalization of businesses and markets. … Marketing has [also] strongly developed an international dimension, with typical topics such as the issue of whether to standardize the marketing mix worldwide or to customize for local markets, or the influence of the country of origin on product and brand images. Areas such as organization studies have been compelled to internationalize their research by the increased need to understand headquarters-subsidiary relationships and organizational issues in the multinational corporation." (Usunier, 1998, pp.3-4)

In cross-cultural research methodology, there are two major 'schools of thought', namely emic and etic. The next section of the introductory chapter discusses how both schools address the critical issue of comparability of data across cultures. An alternative view, in particular: Berry's 'derived etic approach', is presented to researchers in international management. Further sections discuss more specific issues such as different types of equivalence which are necessary to ensure comparability of data across cultures. One type of equivalence, in particular: 'construct equivalence', is discussed in more detail. A wide variety of procedures are presented that allow researchers to test for construct equivalence across cultures. The last section of the introduction addresses the main research questions, the scope, and the outline of the dissertation. The 'funnel' shown in Figure 1.1 graphically depicts the logical structure of the introduction.

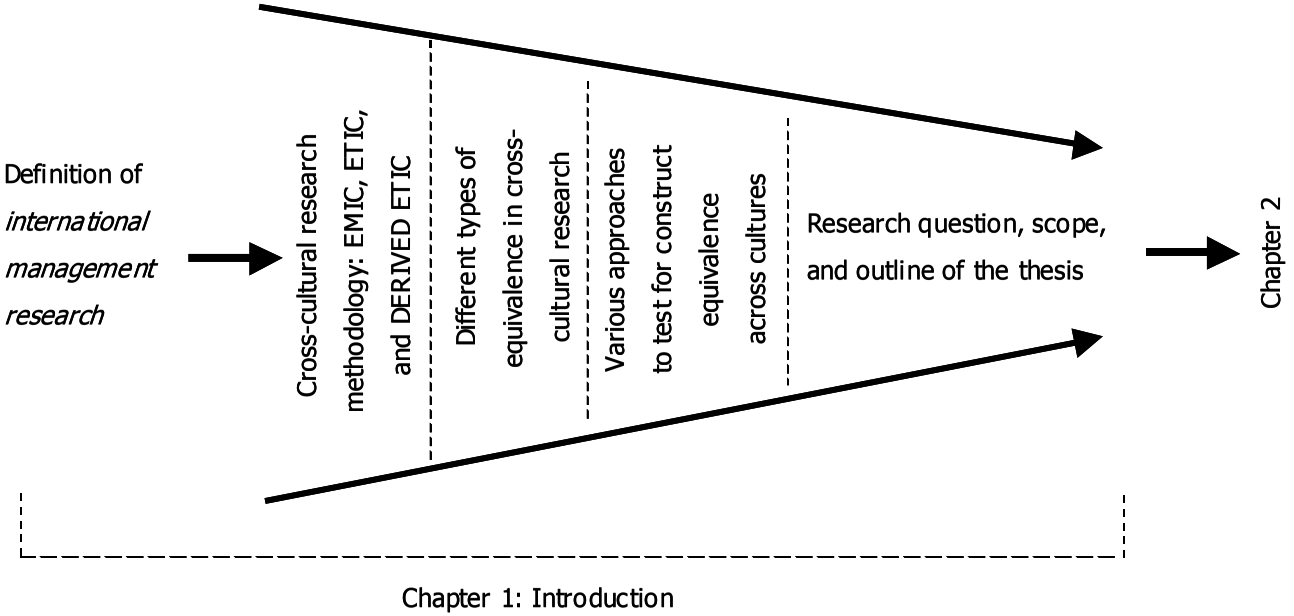


Figure 1.1.
Structure of the introduction (i.e. Chapter 1)

1.3. The issue of comparability of data

1.3.1. EMIC versus ETIC

As any type of cross-cultural research, international management research faces the problem of comparability of data across cultures (or countries). It is common that each culture is unique (at least to some extent), and people show high variations in terms of the 'cultural components'. Recall that the cultural components were specified in the beginning of the introductory chapter. Cross-cultural diversity between (and also within) countries may form a serious threat to making cross-national comparisons both in terms of chosen research methodologies as well as in terms of data obtained from multicountry studies. The specific characteristics of each individual culture (or country) may require different research methodologies, which may limit the comparability of data across countries (Kumar, 2000).

There are two major 'schools of thought' when it comes to cross-cultural research methodology (Hulin, 1987; Triandis and Marin, 1983). The first school, referred to as 'emic', believes in the uniqueness of each culture and emphasises the importance of studying the peculiarities of each culture, identifying and understanding its uniqueness. The study is typically culture-specific and inferences are made about cross-cultural similarities and differences in a subjective manner. The emic school attempts to reconstruct the experiential world of the individual through his/her reports and explanations. A proponent of the emic school of thought is H.C. Triandis (Triandis et al., 1980, 1981, 1985) along with many cultural anthropologists.

The other school, named 'etic', is primarily concerned with identifying similarities in terms of cultural components, and aims at developing pan-cultural or 'culture-fair' measures. According to this school, the measurement structure derived in one culture is expected to be universal and is applied to all cultures. If this assumption is legitimate, such measures make comparisons across cultures feasible and objective. Proponents of this school of thought are G. Hofstede, M. Rokeach, L.R. Kahle, S.H. Schwartz and many psychologists and marketing professionals (see Craig and Douglas, 2000; Kumar, 2000).

The terms 'etic' and 'emic' were introduced into anthropology in the 1960s by the linguist Kenneth Pike (1954, 1971). They were extrapolated from the distinction in linguistics between phonetic and phonemic.[4] Pike (1971) argued that the emic and the etic approach should not be perceived as opposite approaches. According to Pike, they describe the problem of cross-cultural

[4] The study of phonemics involves the examination of the sounds used in a particular language, while phonetics attempts to generalise from phonemic studies in individual languages to a universal science covering all languages. By analogy, emics apply only in a particular society, while etics are culture-free or universal aspects of the world (Berry, 1969).

comparability from two different standpoints, which lead to results, which shade into one another.

The major strengths and shortcomings of both the emic and the etic approach are summarised in Table 1.1.

Table 1.1.
Major strengths and shortcomings of the emic and etic approach

Emic approach	
Strengths	Weaknesses
It permits an understanding of the way in which a specific culture is constructed. It helps one to understand how individuals behave, and why exactly they behave the way they do (e.g. what the impact is of cultural influences). According to some proponents of the emic approach (e.g. Pike, 1971) only the emic approach provides a basis upon which a predictive science of behaviour can be expected to make progress, since even statistical predictive studies will in many instances prove invalid (see Pike, 1971).	Emic research is subject to *systematic bias*. Systematic bias occurs when individuals represent or misinterpret their own behaviour (Helfrich, 1999). Emic research is subject to *arbitrariness*. Arbitrariness refers to the subjective status of scientific knowledge (Helfrich, 1999).
Etic approach	
Strengths	Weaknesses
It provides a broad perspective about different events around the world, so that differences and similarities (in terms of the cultural components) can be recognised. Techniques for recording differing phenomena can be acquired. The etic approach is the only point of entry, since there is no other way to begin an analysis than by starting with a rough, tentative etic description of it (Pike, 1971). An etic comparison of selected cultures may allow the researcher to meet practical demands, such as financial or time limitations.	It is easy to overlook the differential aspects of cultural impact. It is easy to overlook that culture does not represent an independent variable in the usual sense* (Helfrich, 1999). The definition of the phenomena being studied (e.g. variables) may itself be culture-bound.

Note: *Culture is not an independent variable in the sense of an experimentally controlled variable. The assignment of individuals to different groups can, at best, be based on a selection according to their natural membership in that group (i.e. a 'quasi-experimental' research design) (Helfrich, 1999).

Several authors have suggested to combine the emic and the etic approach (Przeworski and Teune, 1970; Triandis, 1972; Davidson et al., 1976; Triandis and Marin, 1983; De Vera, 1985). Triandis (1972), for example, claimed that, in general, etic measures are needed to compare cultures and emic measures to fully understand them.

In the 1980s and 1990s a couple of authors (e.g. Berry, 1989; Helfrich, 1999) proposed alternative approaches which build on the strengths of both the etic and emic approach while minimising their weaknesses. In Helfrich's 'principle of triarchic resonance' (Helfrich, 1999), observed phenomena are the result of an interaction between three elements (the individual, the task, and the culture). The process of responding to situational demands may be universal (i.e. an etic point of view), but the situation can generate alternative behaviours depending on the particular culture and the particular individual (i.e. an emic point of view). As indicated by Helfrich (1999), the principle of triarchic resonance does not match with the idea of comparing groups/cultures based on construct scores (i.e. making cross-cultural comparisons). The reason for this is obvious. For the purpose of cross-cultural measurement etic measures are needed to measure a construct. Emic measures have to be removed as they form a threat to the cross-cultural applicability of the measurement instrument (Helfrich, 1999).

Berry (1989) proposed a five-step process that may provide a basis for an integrated approach to studying cultural differences. The steps in the process are:

> Step 1: Examine a research problem in one's own culture (emic A) and develop a conceptual framework and a set of relevant instruments.
>
> Step 2: Transport this conceptualisation and measurement to examine the same issues in a similar manner in another culture (i.e. 'imposed etic').
>
> Step 3: Enrich the imposed etic framework with unique aspects of the second culture (emic B).
>
> Step 4: Examine the two sets of findings for comparability.
>
> Step 5: If these findings are not comparable, the two conceptualisations will be considered to be independent. But, if they are comparable, then the common set, the 'derived etic', will form the basis of a unified etic framework.

Berry's approach is referred to as Berry's 'derived etic' (1989). The approach is graphically depicted in Figure 1.2.

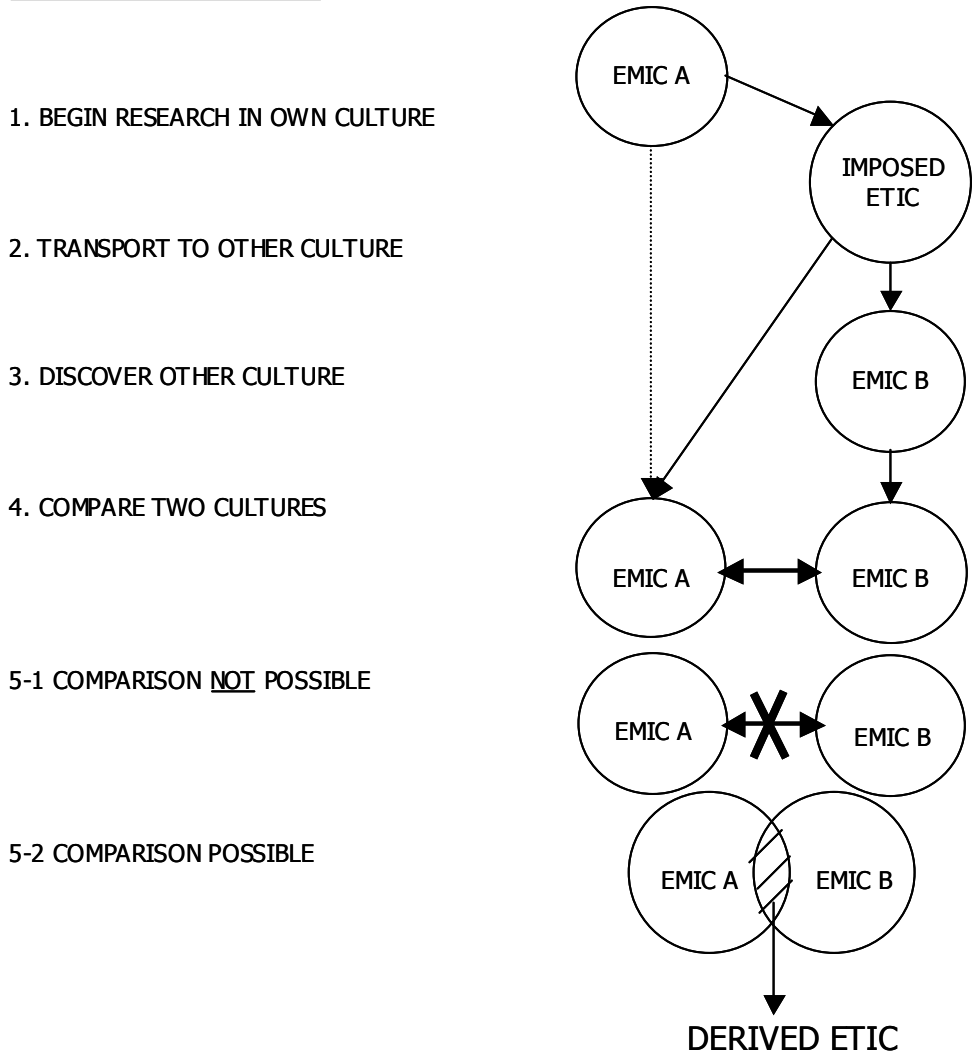


Figure 1.2.
Berry's five-step process (Berry, 1989)

Berry's five-step process provides a guideline for cross-cultural research at the operational level (Helfrich, 1999). It offers, at least in principle, an attractive alternative to researchers in the field of international management. The fact that the conceptualisation and measurement can be interchanged from one culture to another (i.e. 'imposed etic'), makes that the researcher can, at the very least, start studying certain phenomena in other cultures. By repeating Berry's five-step process in new cultures, a universal framework can be developed to explain the phenomena under study (Maheswaran and Shavitt, 2000). Several researchers in the field of international management have adopted an 'imposed etic' approach (e.g. Ryan et al., 1999; Ployhart et al., 2003). An emic approach would not be feasible as the research process would become too complex. Different (culture-specific) variables would need to be collected in each culture and then separate validation studies using different criteria would also need to be done (Ployhart et al., 2003).

Most conceptualisations (i.e. the phenomena under study) in international research are defined and operationalised in a Western country (typically the United States). In some cases it may well be that the conceptual domain and/or the measurement may not be totally transferable to other countries (Yaprak, 2003). Nevis (1983), for instance, has shown that Maslow's hierarchy of needs does not apply in Chinese societies. Taking this into account, it is clear that the enrichment of the imposed etic framework (i.e. step 3 in Berry's five-step process) is a crucial step in Berry's process. In international management research, it is very likely that this particular step in the process will not be executed. Time- and budget constraints often lie at the basis of such a decision. Such practical limitations may form a serious threat to the validity of comparisons made between countries (cultures).

1.3.2. Different types of equivalence

To ensure comparability of data it is necessary to establish 'equivalence' across cultures. Johnson (1998) listed over fifty types of equivalence which have been discussed in the literature. The necessity to establish equivalence across cultures runs throughout all stages of the research process.

Van Herk (2000, 2005) presented an overview of all relevant types of equivalence in all stages of the research process (i.e. problem definition, construct operationalisation, method of data collection, sampling, data collection, data preparation, testing for and establishing measurement invariance, and data analysis). A modified version of Van Herk's overview is presented in Table 1.2. This overview is excellent in that it clearly shows all important types of equivalence that are relevant in cross-cultural research. Some authors have introduced a process framework for creating sound cross-cultural research methodology designs (e.g. Malhotra at al., 1996; Cavusgil and Das, 1997). Cavusgil and Das' framework, for instance, is included in Appendix 1.1.

Table 1.2.
Overview of relevant types of equivalence in the research process

(partially based on the works by: Van Herk, 2000, 2005; Bauer, 1989; Salzberger et al., 1999; Kumar, 2000)

Stage in the research process	**Type of equivalence**	**Source of bias**	**Main type of bias**
1. Problem definition (research topics)	Functional equivalence (FUE)	Violation of FUE: The product, object / stimulus or behaviour does not serve the same *purpose* in different cultures	CONSTRUCT
	Conceptual equivalence (COE)	Violation of COE: The *interpretation* (i.e. the very real meaning) of objects and stimuli differs across cultures.	CONSTRUCT
	Category equivalence (C ATE)	Violation of CATE: The *categories* in which relevant objects or other stimuli are placed differ across cultures.	CONSTRUCT
2. Construct operationali-sation[#]	Equivalence in terms of operationalisa-tion (EOPE)	Violation of EOPE: The type of *study* or the *questions* differ across studies in different cultures.	CONSTRUCT (questions)
	Equivalence of instruments (or item equivalence) (EINS)	Violation of EINS: *Items* and/or *response formats* are not neutral across cultures.	METHOD / ITEM / ARS and ERS* ARS: Acquiescence response style ERS: Extreme response style (i.e. extreme response categories)
	Translation equivalence (TRE)	Violation of TRE: Questions / items do not have equivalent *meaning* across cultures.	ITEM

/ ...

Table 1.2. (continued)
Overview of relevant types of equivalence in the research process

Stage in the research process	**Type of equivalence**	**Source of bias**	**Main type of bias**
3. Method of data collection	Equivalence of data collection methods (EDCM)	Violation of EDCM: Data collection *methods* (face-to-face, telephone, e-surveys) and/or *stimuli* used differ across studies in different cultures.	METHOD
4. Sampling	Sampling equivalence (SE)	Violation of SE: The *target group* and/or *sampling frame* differ across studies in different cultures.	METHOD / SAMPLE
5. Data collection	Equivalence of research administration (ERA)	Violation of ERA: *Data collection procedures*, *interviewer selection processes*, and/or *the time frame* used differ across studies in different cultures.	METHOD
6. Data preparation / …	Equivalence of data handling (EDH)	Violation of EDH: *Data editing* and/or *data coding* procedures are dissimilar across studies in different cultures.	ITEM

Table 1.2. (continued)
Overview of relevant types of equivalence in the research process

Stage in the research process	**Type of equivalence**	**Source of bias**	**Main type of bias**
7. Testing for and establishing measurement equivalence#	Calibration invariance (CALE)	Violation of CALE: The *measurement units* used differ (or have a different meaning) across studies in different cultures.	CONSTRUCT / ITEM
	Configural invariance (CONE)	Violation of CONE: The *rough factor structure* of items across studies are not identical (i.e. pattern of zero and non-zero factor loadings differs across studies) in studies in different cultures.	CONSTRUCT
	Scalar invariance (SCAE).	Violation of SCAE: The *factor loadings* and/or *indicator intercepts* differ across studies in different cultures.	CONSTRUCT / ITEM
8. Data analysis	Equivalence of statistical methods used (ESTM)	Violation of ESTM: *Statistical methods* used to analyse the data are different across studies in different cultures.	-

Notes:

(1) *Van Herk (2000, 2005) referred to this block as 'research design';
(2) #Van Herk (2000, 2005) did not include this block.

Establishing 'construct equivalence' is crucial in cross-cultural research. A 'construct' can be generally defined as: *"a conceptual term used to describe a phenomenon of theoretical interest"* (Cronbach and Meehl, 1955; Nunnally, 1978; Schwab, 1980). The notions of 'constructs' and 'concepts' (i.e. conceptual terms) are similar, but they are not the same. Kerlinger (1986, p. 26) defines a concept as *"an abstraction formed by generalization from particulars"*. 'Violent acts', for instance, can be seen as a concept because people in our society are aware of certain behaviours of individuals (i.e. the particulars) which may be classified as 'violent acts' (i.e. the generalisation). A 'construct' is defined as a *"concept with added meaning"* (Kerlinger, 1986, p. 26). According to Kerlinger meaning is added because a deliberate and conscious attempt has been made to define, specify, and operationalise the concept for the purpose of scientific study. A construct makes it possible for the researcher to judge whether a particular instance is or is not a member of the category. The notion of 'violent acts' can be considered a construct once it is defined as "*intentional physical harm caused to a person"*. Other examples of constructs are general intelligence (in psychological research), national identity (in political research), employee satisfaction (in research in HR management), and consumer innovation adoption (in consumer research). These constructs are typically operationalised by means of a set of 'variables'. A variable is *"a construct that has been defined so that instances of it can be assigned value and counted"* (Kerlinger, 1986). Variables are expected to change either from one time to another or from one person (or unit) to another.

Hui and Triandis (1985) define four important types of (construct-related) equivalence. Their definitions are as follows:

> Conceptual / functional equivalence. A construct that can be meaningfully discussed in the cultures concerned is said to have cross-cultural conceptual equivalence. Conceptual equivalence is closely tied with functional equivalence, which in psychological research pertains to the similarity between the goals (or purposes) of the two behaviours. The concept of a bicycle may not be functionally equivalent across cultures as it may be (primarily) a means of transportation in one culture, and a means of recreation in another culture.
>
> When looking at gift-giving behaviour, for instance, different purposes exist between the U.S. and the Japanese consumers (Green and Alden, 1988). Some beliefs, such as secularism or traditionalism, and values, such as parochialism or cosmopolitanism, are likely to vary across different societies (Inglehart, 1997; Inglehart and Baker, 2000).
>
> Equivalence in terms of construct operationalisation. In order to be equivalent in terms of construct operationalisation, the construct should be operationalised using the same procedure. Operationalising

'aggression' in terms of verbal insults would lack equivalence between the mute population and the non-mute population.

Item equivalence. Item equivalence assumes that the construct is measured by the same instrument, and each item means the same thing to subjects from culture A as it does to those from culture B. The reader should realise that literal translations may not be appropriate as they may not be functionally equivalent to people from different cultures. Functional equivalence requires equivalence in terms of the connotations that people from different cultures have when interpreting words or expressions.

In order to establish item equivalence the two previous types of equivalence are presupposed.

Scalar equivalence. An instrument has scalar equivalence in two cultures if the construct is measured on the same metric. Inb order to establish scalar equivalence all previous types of equivalence are pressuposed.

Conceptual and functional equivalence is of particular importance in the first stage of the research process, which is referred to as 'problem definition' in Table 1.2. Equivalence in terms of construct operationalisation and item equivalence are considered in a later step in the research process, namely the research design (see Table 1.2.).

A couple of authors use the term 'scalar equivalence' (Van de Vijver and Poortinga, 1982; Hui and Triandis, 1983, 1985; Van de Vijver and Leung, 1997), while others use the term 'calibration equivalence' (Mullen, 1995) or '*measurement* equivalence' (Drasgow, 1984, 1987) to indicate that the construct is measured in all cultures using the same metric. In this dissertation, the term 'measurement invariance' is used. The term 'equivalence' and 'invariance' are used interchangeably. As indicated by Meredith (1993), it is only possible to compare two or more populations on the basis of their construct mean scores if measurement invariance is established across these populations.

In cross-cultural comparative research, many factors may form a threat to the necessity of measurement invariance (across cultures). Examples are: inaccurate translations (Orley, 1993; Temple, 1997; Voss et al., 1996), differences in response styles across cultures (Baumgartner and Steenkamp, 2001; Billiet and McClendon, 2000; Chen et al., 1995; Cheung and Rensvold, 2000; Greenleaf, 1992a, 1992b; Johnson et al., 1997; Smith, 2004; van Herk et al. 2004, Welkenhuysen-Gybels et al., 2003), differences in construct measurement over

time, in particular: α- (alpha-), β- (beta-), and γ- (gamma-) change[5] (Schmitt, 1982; Bartunek and Franzak, 1988; Millsap and Hartog, 1988; Schaubroeck and Green, 1989; Vandenberg and Self, 1993), and heterogeneity of populations (Oort, 1994). These are just a subset of those factors that may threaten measurement invariance.

Fortunately, a number of statistical procedures have been developed to assess construct equivalence (including measurement invariance) across cultures (or groups). Many of these statistical procedures[6] are based on the assumption that the construct is unidimensional rather than multidimensional. A brief discussion of some of these methods is provided in the next paragraphs.

[5] α - change is a change in the level of the trait (construct) in time; β - change is a change in the item response scale resulting from re-calibration (i.e. a redefinition of the measurement scale); γ - change is a change in item content resulting from a re-definition of the conceptual domain of the trait (construct).

[6] The IRT (DIF) approach and the multigroup MACS approach, which are explained later on in this introductory chapter, assume that the construct (i.e. the underlying latent variable) is unidimensional.

1.4. Testing for construct equivalence across groups

A number of statistical procedures have been proposed in the literature to test for construct equivalence across groups. Some of these statistical procedures test for measurement invariance, a particular type of construct equivalence. A distinction can be made between explorative and confirmatory methods. In the pursuing paragraphs, a couple of such methods are briefly described.[7]

1.4.1. Explorative methods

One approach to assess construct equivalence across groups is to examine (i.e. visually inspect) configural similarity of construct-related variables in multiple groups. A high degree of similarity across groups would indicate that constructs are very much equivalent across groups. Statistical methods such as principal component analysis / exploratory factor analysis (Kiers and ten Berge, 1989; Kiers, 1990; Katigbak et al., 1996), and multidimensional scaling[8] (Schwartz, 1992; Schwartz and Sagiv, 1995; Braun and Scott, 1998; Braun, 2000) have been used to obtain a configural representation of construct-related variables in two (or more) populations. These statistical methods have the advantage that they do not assume a unidimensional construct.

Some alternative approaches have been proposed to quantify the degree of similarity between the spatial configurations from two different populations. Provided that a Euclidean distance is used as a distance measure in both configurations, then the following transformations are admissable (Groenen, 2002):

(1) translation (i.e. shifting the origin)
(2) rotation
(3) reflection (i.e. multiplication of one or more axes by -1)
(4) dilation (i.e. multiplication of all distances by a constant).

There are at least two options in order to quantify the degree of similarity between two spatial configurations.

The first option is to calculate the well-known *coefficient of congruence* (Burt, 1948; Tucker, 1951; Wrigley and Neuhaus, 1955). The coefficient of congruence is a one-number index expressing the degree of similarity between two spatial

[7] A wider range of methods are discussed in Millsap & Everson (1993) (e.g. loglinear models, Mantel-Haenszel statistic, standardisation method, logistic regression method, logistic discriminant function approach).
[8] What multidimensional scaling does is represent the intercorrelations of the items in a multidimensional space.

configurations from separate multidimensional scaling (MDS) or factor analyses. This coefficient takes care of all four admissable transformations for spatial configurations based on a Euclidean distance (see higher) (Groenen, 2002). Leutner and Borg (1983) and Borg and Leutner (1985) have proposed statistical norms for the coefficient of congruence. Some alternative one-digit indices are: the s-statistic (Cattell, 1949; Cattell and Baggaley, 1960; Cattell et al., 1969), and coefficient kappa (Cohen, 1960). The simulation study by Guadagnoli and Velicer (1991) has shown that there are little differences in the accuracy of these three alternative measures of spatial similarity (i.e. coefficient of congruence, s-statistic, and coefficient kappa) in a wide range of experimental conditions.

The second option is to use a *Procrustean similarity transformation* (see Borg & Groenen, 1997, pp. 344-346) to bring the group-specific configural representations to an optimal point-by-point match (i.e. an 'optimal common space'). A Procrustean similarity transformation allows for all four admissable transformations for spatial configurations based on a Euclidean distance (Groenen, 2002). The product-moment correlation over all corresponding point coordinates in this 'optimal space' offers a measure of similarity between the two configural representations (Borg and Leutner, 1985). Statistical norms for such correlations have been presented by Langeheine (1980, 1982). Instead of computing the product-moment correlation, the coefficient of congruence may also be computed using all corresponding coordinates in the optimal common space. Paunonen (1997) has proposed statistical norms for the use of the coefficient of congruence in this particular situation. These statistical norms take into account the inflating effect of the Procrustean similarity transformation on the calculated coefficient of congruence (Cliff, 1966; Korth and Tucker, 1976; Brokken, 1983).

Instead of using a Procrustean similarity transformation to derive an optimal space, a (classical) Procrustean rotation (Hurley and Cattell, 1962) can also be considered. This approach is not recommended because a Procrustean rotation does not allow for translations and dilations (Groenen, 2002). As mentioned before, translations and dilations are two of the four admissable transformations with spatial configurations based on a Euclidean distance. Unlike this last option, the first two options do provide an adequate measure of the degree of similarity between the spatial configurations of two distinct populations.

Another explorative approach was proposed by Mullen (1995). Mullen's procedure permits assessment of measurement invariance across groups. As mentioned before, measurement invariance across groups is a more specific form of construct equivalence across groups. Mullen proposed an alternating least-squares optimal scaling approach (based on the PRINCIPALS analysis implemented in the software 'SAS') which assumes ordinal-level rather than interval-level data. Measurement invariance across groups was assessed by plotting the raw scores on the x-axis against the 'optimal scores' on the y-axis

for each group being studied. Parallel lines indicated that measurement invariance across groups was established. Non-parallel lines indicated the non-existence of measurement invariance across groups. Mullen's procedure is not that advantageous as it does not offer a statistical test to check whether two (or more) lines may be non-parallel just by chance.

The absence of a hypothesis-testing framework is the major shortcoming of all explorative methods proposed to test for construct equivalence. Confirmatory methods have the advantage that they do offer a hypothesis-testing framework. Proponents of explorative approaches try to overcome this limitation by adopting nonparametric approaches to statistical inference (e.g. a bootstrap procedure). Interested readers can refer to the literature that is dedicated to this topic (e.g. Mooney and Duval, 1993; Efron and Tibshirani, 1993).

1.4.2. Confirmatory methods

Two confirmatory methods are commonly used to test for measurement invariance (or the absence of measurement invariance) across groups: (1) models based on Item Response Theory (IRT), and (2) multiple-group Mean- and Covariance- Structure (MACS) models. A third confirmatory approach which is based on multigroup latent class analysis may complement the IRT-based and MACS- approach. The next paragraphs provide a non-technical explanation of these three confirmatory approaches.

1.4.2.1. An approach based on Item Response Theory

The first confirmatory approach is based on Item Response Theory (IRT) (Lord, 1980; Hambleton and Swaminathan, 1985; Embretson & Reise, 2000; Du Toit, 2003). As indicated by Singh (2004), Item Response Theory is relatively unknown in the marketing literature (and the management literature).

Some authors have used an IRT-based approach to test for the cross-cultural applicability of measurement instruments (e.g. Hulin et al., 1982; Hui and Triandis, 1983; Hulin & Mayer, 1985; Candell and Hulin, 1986; Hulin, 1987; Ellis et al., 1989; Ellis et al., 1993; Reise et al., 1993; Hambleton & Kanjee, 1995; Maurer et al., 1998; Robert et al., 2000; Schmit et al., 2000; Tomás et al., 2000; Cooke et al., 2001; Raju et al., 2002).

The Item Characteristic Curve (ICC), which is modelled in IRT, describes the conditional relationship between the probability of a particular item response and a respondent's position on the underlying latent variable (possibly a latent trait). The latent variable, expressed as theta (θ), is a continuous unidimensional construct that explains the covariance among item responses (Steinberg and Thissen, 1995).

Different nonlinear Item Characteristic Curves have been proposed to model the conditional relationship between a respondent's response to an (dichotomously scored) item and the latent variable (or construct):

(1) a 2-parameter ogive (i.e. probit) model (Lord, 1952),

(2) a 1-, 2-, and 3-Parameter Logistic (PL) model (Rasch, 1960 [1-PL], Birnbaum, 1957 [2-PL]; Birnbaum, 1968 [3-PL]).

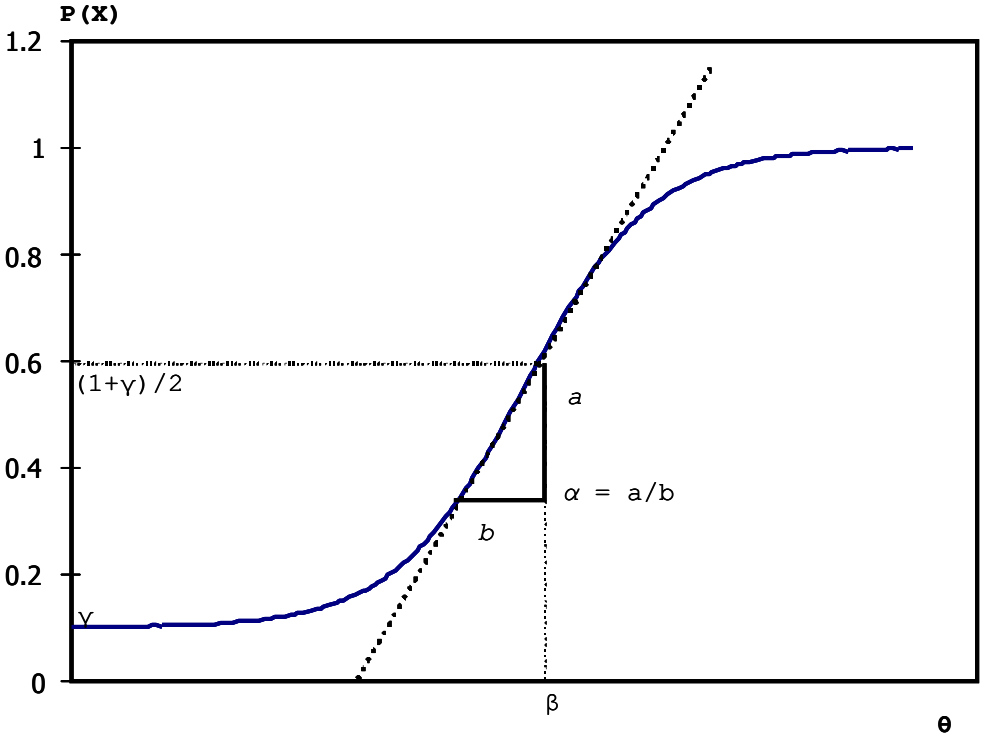


Figure 1.3.
A 3-Parameter Logistic model (1 item)

Notes:

(1) $P(X)$ denotes the probability of a particular item response (often a 'favourable' response [e.g. answering yes to a question or agreeing with a statement]);
(2) θ denotes the respondent's position on the underlying latent variable (latent trait).

Figure 1.3 shows an example of an ICC following a 3-Parameter Logistic model. The three parameters shown in Figure 1.3 are:

(1) the *location* parameter (β), which refers to the level of difficulty of an item or the 'positiveness' of an item. (included in the 1-, 2- and 3-PL model),

(2) the *discrimination* parameter (α), which indicates how well the ICC discriminates between people with adjacent positions on the underlying latent variable (or trait) (included in the 2- and 3-PL model). This parameter is defined only at the point of inflection (i.e. point $(1+\gamma)/2$ on the y-axis),

(3) the *pseudo-chance level* (γ), a lower bound. Even with a very low value on the underlying latent variable (or trait), the probability for a positive answer is at least γ (only in the 3-PL model).

The location parameter (β) is the most important model parameter. Its value indicates how strong the item 'loads' on the underlying latent variable/trait. Assume, for instance, one is looking for items to measure the construct 'intolerance towards minorities in the society'. One item may be: 'If you enter a shop which is owned by a person belonging to a minority group (e.g. having the Turkish nationality), would you refuse to buy something in that shop?'. Another item may be: 'If you are introduced to a person who belongs to a minority group (e.g. having the Turkish nationality), would you slap that person in the face?'. It is clear that, as far as intolerance is concerned, the value of the location parameter of the latter item is higher than the value of the location parameter of the former item.

In order to define the scale for the underlying latent variable / trait (i.e. θ), it is common practice to set the mean of the item parameters (or one item parameter) equal to one.

Three (crucial) assumptions are made in IRT.

Assumption 1: IRT assumes that item responses are *locally independent*.

> Local independence denotes that if the score on the latent variable (or trait) is held constant, there should be a correlation between the item responses which is not significantly different from zero (Thissen and Steinberg, 1988). This (weaker) definition of local independence makes reference only to the linear relationship (as measured by a correlation coefficient) between the item responses. The stronger definition of local independence makes reference to any dependence between the item

scores (given a constant score for the latent variable), not just linear dependence [9].

Both the weaker and stronger definitions of local independence have implications for the inter-dependencies between items and between items and the latent variable(s) (Bollen, 2002): (a) errors of measurement are independent (or uncorrelated), (b) items have no direct or indirect effects on each other, (c) there are at least two items to measure one latent variable, (d) each latent variable must have direct effects on one or more items, and (e) the items do not directly affect the latent variable. Taking these implications into account, it is obvious that the assumption of local independence (as specified in Item Response Theory) restricts the type of measurement structure of a latent variable (e.g. the latent variable cannot be considered to be 'a consequence' of its indicator variables [i.e. items], but should be seen as 'their cause'). A more thorough discussion on this is provided in Chapter 2.

Assumption 2: IRT assumes that item responses are *unidimensional*.[10]

Unidimensionality implies that the set of items assesses a single underlying latent variable (or trait) dimension (Reise et al., 1993). Many IRT models also have a supplementary assumption, namely that the (underlying) latent variable is normally distributed within the population.

Assumption 3: In IRT, *appropriate dimensionality* is assumed (Embretson & Reise, 2000).

Appropriate dimensionality is achieved when the IRT model contains the right number of latent variable/trait estimates per person for the data, and that the responses vary only according to the scale measuring the latent variable/trait.

---

[9] Formally, the (stronger) definition of local independence states that:
$P[Y_1, Y_2, \ldots, Y_k] = P[Y_1|\theta]\ P[Y_2|\theta]\ \ldots\ P[Y_k|\theta]$,
where $Y_1, Y_2, \ldots, Y_k$ are k random item responses, θ is a vector of latent variables, $P[Y_1, Y_2, \ldots, Y_k]$ is the joint probability of the item responses, and $P[Y_1|\theta]\ P[Y_2|\theta]\ \ldots\ P[Y_k|\theta]$ are the conditional probabilities.

[10] Some researchers have argued that unidimensionality is an unrealistic assumption (e.g. Hulin et al., 1982). Multidimensional IRT models have also been developed (see Glas and Verhelst, 1993, pp. 226-237). Salzberger et al. (1999) argued that one seldom tries to cover several dimensions within one item. To model complex phenomena, one may conduct separate unidimensional analyses rather than one multidimensional analysis.

The models developed by Lord (1952), Rasch (1960), and Birnbaum (1957, 1968) have mainly been designed for modelling ICC's, describing responses to items with only two answer categories (i.e. dichotomous responses). An overview of these models is shown in Appendix 1.2. According to Rasch (1960), (unidimensional) measurement, should be in line with a specific measurement paradigm referred to as 'specific objectivity'. Specific objectivity states that person parameters (i.e. θ's) have to be independent of specific items and vice versa (i.e. measurement should be sample-independent). Only the Rasch model (with only one location parameter) complies with this principle (see Salzberger et al., 1999). Additional parameters, such as the item discrimination parameter subdues this principle. Consequently, Rasch's model represents a very restrictive (and often unrealistic) measurement model. Birnbaum, for instance, considered IRT models which are more realistic in terms of the properties of the data.

From the late nineteen-sixties onwards, more advanced IRT models have been proposed to deal with more than two (i.e. 'polytomous') ordered responses. The most well-known examples are: Samejima's Graded Response Model (Samejima, 1969), Masters' Partial Credit Model (Masters, 1982), Muraki's Generalised Partial Credit Model (Muraki, 1992), and Andrich's Rating Scale Model (Andrich, 1978a/b). These models are based on the principle that k ordered categories can be modelled by defining: (1) k-1 boundaries between the adjacent response categories (i.e. adjacent-category models) or (2) k-1 boundaries between cumulative parts (cumulative-probability models). When boundaries are created between adjacent response categories, each boundary provides the probability of responding in category k instead of category k-1. Alternatively, when boundaries are created between cumulative parts, each boundary provides the probability of responding in category k or higher (Mellenbergh, 1995). These boundaries are referred to as Boundary Response Functions (BRFs).

These more advanced IRT models differ from one another in terms of the IRT model that is used to describe the Boundary Response Functions BETWEEN and WITHIN the items, respectively. Table 1.3 shows a comparison between these four IRT models for items with more than two ordered responses (i.e. 'polytomous ordered [or graded] responses).

Table 1.3.
A comparison between four IRT models for items with more than two ordered responses.

	Andrich RSM*	Masters' PCM	Muraki's GPCM	Samejima's GRM
IRT model used to model the BRFs BETWEEN the items	2-PL model (β, α)	1-PL model (β)	2-PL model (β, α)	2-PL model (β, α)
IRT model used to model the BRFs WITHIN the items	2-PL model (β, α)	1-PL model (β)	1-PL model (β, α)	1-PL model (β)
Adjacent-category model (ADJ) or cumulative probability model (CUM)?	ADJ	ADJ	ADJ	CUM

Notes:
(1) x-PL model stands for x (1- or 2-) Parameter Logistic model;
(2) *Andrich RSM assumes equal distances between adjacent categories across all items.

Consider the example of Samejima's Graded Response Model (GRM). In Samejima's GRM, each item i will have only one discrimination parameter (α_i), and k-1 location parameters β_{ik} (k being the number of [graded] responses, k=1,2,…,m_i). Samejima's Graded Response is based on the logistic function providing the probability that an item response will be observed in category k or higher (i.e. a cumulative probability):

$$P^*_{ik}(\theta) = P^*_{ik}(RC >= k|\theta) = \frac{1}{1 + \exp[-\alpha_i (\theta - \beta_{ik})]},$$

where RC refers to the chosen response category (Raju et al., 2002).

The (m_i-1) probabilities ($P^*_{i1}(\theta)$, $P^*_{i2}(\theta)$, . . . , $P^*_{i(m_i-1)}(\theta)$) represent the (m_i-1) Boundary Response Functions of item i.

For ordered responses $u_i = k$ (k = 1,2,3, …,m_i) where response m_i reflects the highest θ value, a Category Response Function (CRF) is determined by a difference between two probabilities:

$$P_i(u_i = k|\theta) = \frac{1}{1 + \exp[-\alpha_i (\theta - \beta_{ik})]} - \frac{1}{1 + \exp[-\alpha_i (\theta - \beta_{ik+1})]}$$

Note that the BRFs and CRFs for an item i depend on θ and the item parameters of Samejima's GRM (i.e. α and β parameters).

Item Response Models allow for testing of measurement invariance across two groups (i.e. the 'focal' and the 'reference' group). In IRT-terminology, measurement non-invariance of an item across groups is referred to as 'Differential Item Functioning' (DIF).

"DIF is said to occur whenever the conditional probability, $P(X|\theta)$, of a correct response or the agreement with an item for the same level of the latent variable differs for two groups." (Camilli and Shepard, 1994).

The key decision in a DIF analysis is the selection of the appropriate IRT model (Camilli and Shepard, 1994).

The graded (i.e. ordered) nature of the Likert-type response scales makes Samejima's Graded Response Model an obvious choice to test for the presence of DIF whenever Likert-type items are used (see Cooke et al., 2001; Maurer et al., 1998; Mellenbergh, 1994; Raju et al., 2002; Reise et al., 1993; Tomas et al., 2000). Van Zessen and De Beuckelaer (2000) used a rating scale version[11] of both Samejima's Graded Response Model and Muraki's Generalised Partial Credit Model to test for the presence of DIF in items scored on Likert-type scales.
A detailed presentation of the other IRT models presented in Table 1.3 (e.g. Andrich's Rating Scale Model, Masters' Partial Credit Model, Muraki's Generalised Partial Credit Model) is beyond the scope of this introductory chapter. The conceptual differences between Samejima's Graded Response Model and the alternative models (in terms of the adjacent or cumulative nature of the model and the specific parameters in the within-and between-item model) are clear from the information provided in Table 1.3. Readers who would like to study these alternative models thoroughly are encouraged to consult some of the following textbooks: Lord (1980), Hambleton and Swaminathan (1985), Embretson and Reise (2000), and/or Du Toit (2003).

The assessment of DIF can be done either at the item parameter level or at the Item Response Function (IRF) level. Assessing DIF at the item parameter level implies that one tests whether the item parameters are invariant across the focal and the reference group. Different approaches have been proposed to test for DIF at the item parameter level (Cohen et al., 1993; Thissen et al., 1988). All of these approaches are adequate when dealing with items with more than two ordered responses (Raju et al., 2002). When adopting the Differential Functioning of Items and Tests (DFIT)-approach proposed by Raju et al. (1995), it is evaluated whether item-level true scores are invariant for subjects with

[11] In a rating scale version the location parameter is 'splitted' in two parts: a threshold value for the answer category (is fixed across questions measured on the same scale), and an item-specific location parameter.

identical θs. This approach assesses the invariance of Item Response Functions across groups, instead of the invariance of item parameters.

Alternative IRT models (e.g. with or without DIF) can be mutually compared using a model-comparison approach (Thissen et al., 1986), provided that they are nested[12] within one another. The statistical comparison between alternative models is based on the G^2 statistic (i.e. –2 times the log of the likelihood function).

Some commercial software packages such as MULTILOG (Thissen, 1991) and PARSCALE (Muraki and Bock, 1997) offer the possibility to test for DIF when the observed variables have more than two ordered responses.

[12] Two models are 'nested' within one another if some parameters in one model are constrained to be equal across groups, whereas the same parameters in the other model are freely estimated across groups.

1.4.2.2. An approach using multigroup MACS models

The second confirmatory approach is based on a particular extension of Jöreskog's (1971) multigroup approach to Confirmatory Factor Analysis. This extension, which is also known as the multigroup Mean- And Covariance-Structure (MACS) model, has been introduced by Sörbom (1974, 1978). The multigroup MACS model is rooted in Classical Test Theory (CTT) (Lord and Novick, 1968; Nunnally, 1978; Crocker and Algina, 1986; Nunnally and Bernstein, 1994).

Initially, the MACS approach was used in the psychological and sociological literature to test for the cross-cultural applicability of measurement instruments (e.g. Drasgow and Kanfer, 1985; Miller et al., 1985; Watkins, 1989; Devins et al., 1997).

In this section, some key concepts of CTT are briefly discussed first. Next, the multigroup MACS model is briefly discussed. A more thorough explanation of the multigroup MACS model is provided in Chapter 3.

According to CTT, observed variable (i.e. indicator) scores, which measure the same underlying construct, are a function of a true score component and a measurement error score component:

$$X_{ri} = T_r + E_{ri}$$

where:

- X_{ri} represents the i^{th} observed variable score of respondent r (or, alternatively, examinee r);
- T_r represents the (unknown) true score of respondent r on the construct;
- E_{ri} represents a (nonsystematic) measurement error when using observed variable i to measure respondent's r score on the construct.

In CTT it is assumed that E_{ri} follows a Normal distribution with zero mean and variance σ^2_E. In addition, CTT[13] assumes that E_{ri} is distributed equally across all score levels. T_r is defined as the expected value across (repeated) realisations of X_{ri} (i.e. $E[X_{ri}]=T_r$). It is further assumed that the true score component and the measurement error score component are uncorrelated ($Corr(T_r, E_{ri})=0$ for all k observed variables). If this assumption is justified, it is possible to decompose the variance of the observed variable into a true score (variance)

[13] This assumption is not made in IRT modelling (Embretson, 1996).

component, and an error score (variance) component: $\sigma_X^2 = \sigma_T^2 + \sigma_E^2$ or $\sigma_T^2 = \sigma_X^2 - \sigma_E^2$. Based on this decomposition, the reliability of the scale composed of i observed variables can be further defined.
Definition of scale reliability (i.e. ρ_X^2) as the true score variance divided by the observed score variance leads to the following derivation:
$\rho_X^2 = \sigma_T^2 / \sigma_X^2$ (by definition) $= 1 - (\sigma_E^2 / \sigma_X^2)$. CTT has had such a big influence on scale construction and theory testing as it offers a suitable framework to assess the reliability of scales (Traub, 1994). The assumption of a zero correlation between the true score component and the measurement error score component is also a fundamental assumption in MACS modelling.

In a multigroup MACS approach, data on observed variables' mean scores (in every group) are analysed in addition to the variance-covariance information (in every group) (Sörbom, 1982). Originally, only observed variables which have an interval-(or ratio-) measurement level were used in multigroup MACS analysis. Recent methodological developments have made it possible to also deal with categorical variables in multigroup MACS models (e.g. Muthén, 1984; Jöreskog and Moustaki, 2001).

The fundamental hypothesis in the multigroup MACS model states that the variance-covariance matrix of the observed variables (in every group) and the vector of observed mean scores (in every group) is a function of a set of model parameters (in the same group).

In a multigroup MACS model, latent variables (or traits) are (at least in principle[14]) conceived as being unidimensional. The observed variables are seen as causal 'consequences' of one or more underlying latent variables. It is further assumed that the variance that is shared across observed variables can only be attributed to their common cause (i.e. one or more latent variables) and not to any other extraneous factor. As a result, the error terms of observed variables are assumed to be mutually uncorrelated. Their expected value is thus zero. The error term comprises two parts: a random part of true measurement error, and a random part that is specific to each indicator of a particular construct (e.g. Bagozzi, 1991).

The IRT (DIF) framework and the multigroup MACS framework are similar in that they both offer a hypothesis-testing framework. In both frameworks, it is possible to mutually compare an unrestricted model (e.g. not assuming measurement invariance across groups) with a restricted model (e.g. a model assuming measurement invariance across groups). The multigroup MACS

[14] One may question the unidimensionality in some specific cases; the second-order factors in a second-order factor analysis model are – by definition - not unidimensional (i.e. due to the first-order factor structure) .

approach is implemented in many commercial software packages (e.g. LISREL, EQS, MPLUS etc.).

There are, however, also substantial differences between the multigroup MACS approach and the IRT (DIF) approach. In both approaches, the relationship between an individual's response to an item (or question) and the person's position on the underlying latent variable (or trait) is represented by a functional relationship. The difference between both approaches lies in the form of that relationship. In the multigroup MACS approach, a linear relationship is used whereas IRT uses a nonlinear (S-shaped) relationship (e.g. a logistic curve). Secondly, in IRT modelling, one typically needs to compile a large number of items having substantially different location parameters (see Johnson, 1998). In the MACS approach, only a limited set of items (e.g. three to five) is needed to represent a single (unidimensional) construct.

The need to work with many items per construct makes the IRT approach a cost-ineffective option in international management research. Examples of the multigroup MACS approach to measurement invariance testing can be found in the following: Bagozzi and Edwards, 1998; Broderick, 1999; Durvasula et al., 1993, 2001; Grunert et al., 1993, 1994; Judge et al., 1998; Lastovicka, 1982; Mavondo et al., 2003; Myers et al., 2000; Riordan and Vandenberg, 1994; Ryan et al., 1999; Scholderer, 2004; Singh, 1995; Steenkamp and Baumgartner, 1995, 1998; Ueltschy et al., 2004; Vandenberghe et al., 2001; Wasti et al., 2000; Yoo, 2002; Yoo and Donthu, 2001, 2002.
Based on a review of 210 journal articles in eight management journals, Schaffer and Riordan (2003) concluded that 17% of all studies adopted a multigroup MACS modelling to test for the invariance of measurement scales across countries (cultures). In only 2% of the studies an IRT (DIF) approach was adopted. In 6% of the studies another approach to measurement invariance testing was used. From this, the reader will understand that in 75% (!) of all studies no attempt at all was made to test for measurement invariance of scales across countries (cultures).

Despite the many differences between the multigroup MACS approach and the IRT (DIF) approach, some methodological developments have also narrowed the gap between both approaches. Muthén (1984), for example, introduced IRT-like threshold parameters into structural equation models. For this purpose, he introduced the CVM estimator.[15] Muthén's measurement model is similar to a 2-parameter IRT model (with α- and β-parameter). This model is implemented in the Mplus software. In some special cases, IRT and factor analysis may provide similar (or identical) results. Takane and de Leeuw (1987), for example, have

[15] A very readable introduction to the theory behind Muthén's CVM estimator (i.e. an estimator for CFA models including nominal-, ordinal-, and/or interval-scaled variables) is provided by Hollis and Muthén (1987).

shown that IRT and factor analysis are identical provided that dichotomous variables are used.

1.4.2.3. An approach using multigroup Latent Class Analysis

When observed variables are of a (truly) categorical nature (for example: unordered categorical variables), a multigroup Latent Class (LC) analysis approach to measurement invariance testing across groups is possible. The paper by Clogg & Goodman (1985) provides a good introduction to the application of LC analysis in multiple groups.

In a LC analysis, it is assumed that the frequencies of the response patterns of some categorical indicator variables can be explained by a limited number of latent types (i.e. 'classes'). In other words, in a LC analysis, the identification of one or more underlying latent types provides knowledge of the latent class to which an individual belongs. Based on an object (or person's) latent class membership one can predict the responses on the categorical indicator variables.

In a LC framework, the notion of measurement invariance across groups implies the equivalence in the class-specific conditional response probabilities across groups. If measurement invariance across groups is established, it makes sense to test whether class sizes are equivalent across groups. If they do not differ across groups, the groups are said to be homogeneous with respect to the 'typological structure' of the items (Eid et al., 2003).

Eid et al. (2003) used the software PANMARK (van de Pol et al., 1996) to statistically compare an unrestricted model (e.g. with no cross-group constraints) with a more restricted model (e.g. assuming measurement invariance across groups and possibly equal class sizes across groups). The difference in the likelihood-ratio values for a given difference in degrees of freedom between both models determined the reject or acceptance of the more restricted model. If the likelihood-ratio difference is larger than the critical value of a Chi-squared distribution (with a given difference in degrees of freedom), then the more restricted model is to be rejected in favour of the unrestricted model (Eid. et al, 2003).

1.4.3. A complementary approach: An approach based on Generalizability Theory

Recently, Sharma and Wheathers (2003) proposed an approach based on 'Generalizability Theory' (i.e. 'G-Theory', see Brennan, 2001) to assess the generalisability of scales in cross-cultural (marketing) research.[16] G-Theory is based on variance-decomposition principles on which experimental designs and the analysis of variance (ANOVA) are based.

In the approach based on G-Theory, a crucial decision to make is whether the 'culture-factor' (or 'country-factor') should be viewed as a fixed or a random factor. When the culture-factor is considered to be fixed, the researcher is only interested in the cultures participating in the research. However, when culture is considered to be random, then the researcher would like to extrapolate results from this research to other cultures which are not included in this research.

G-Theory is advantageous in that it offers guidelines regarding the number of items and subjects needed to obtain a pre-defined level of generalisability for future studies. The approach is limited in that G-Theory requires sample sizes across cultures to be (nearly) equal.

Unfortunately, G-Theory does not provide statistical tests to compare the likelihood of alternative models. As a consequence, the approach based on G-Theory can not be considered to be a confirmatory approach to testing (different forms of) measurement equivalence across cultures. The approach based on G-Theory also fails to provide some diagnostic information as to which items do not exhibit measurement invariance across cultures.

Sharma and Wheathers (2003) positioned the approach based on G-Theory as a complementary method to the multigroup MACS approach. They did not claim that the approach based on G-Theory would offer a better alternative than the multigroup MACS approach when one aims to test for measurement invariance across groups.

[16] Katerberg et al. (1977) and Van de Vijver and Poortinga (1982) have also applied G-Theory to test for the equivalence of translated instruments.

1.5. Research questions, scope, and outline of the dissertation

The main objective of this dissertation is to investigate to which extent violations of the principle of measurement invariance across groups (or cultures) lead to wrong conclusions regarding construct (or factor) mean comparisons across groups. As Steenkamp and Baumgartner (1998) pointed out:

"If evidence supporting a measure's invariance is lacking, conclusions based on that scale [or measure] *are at best ambiguous and at worst erroneous".* (Steenkamp and Baumgartner, 1998)

To date, the (methodological) literature is not conclusive as to the extent to which measurement invariance should hold across groups. In this dissertation, it will be examined which measurement parameters need to be identical across groups (i.e. what level of measurement invariance should be established across groups).

In this dissertation, it is assumed that the data is metric, at least from an analysis point of view.[17] This implies that differences between the response categories (coded as: 1,2,…,K for a K-category scale) are expected to have substantial meaning.

The evaluation as to how threatening violations of the measurement invariance principle are, is based on a simulation study (i.e. Chapter 4), and two case studies in international management research (i.e. Chapter 5 and Chapter 6). The first case study deals with an international employee survey (i.e. international HR management), whereas the second concerns two international consumer studies. In all case studies, it is evaluated to which extent the principle of measurement invariance across groups is violated, and what the consequences are for the reliability and validity of factor mean comparisons across groups.

In this introductory chapter, a number of alternative statistical methods, which were designed to (formally) test the assumption of measurement invariance across cultures (groups), have been discussed. Only the multigroup MACS approach to measurement invariance testing will be further examined in this dissertation. As mentioned before, this approach is currently the most popular one in the field of international management research. The multigroup MACS approach is discussed in more detail in Chapter 3.

[17] However, in the simulation study (i.e. chapter 4) one will also deal with ordinal data, namely ordinal variables measured on a 5-point scale.

In Chapter 2, different types of measurement models for an underlying construct (or factor) are presented. As will be explained in Chapter 2, the use of the (multigroup) MACS approach is only legitimate with a specific type of measurement model for the underlying construct. The outline of this dissertation is graphically depicted in Figure 1.4.

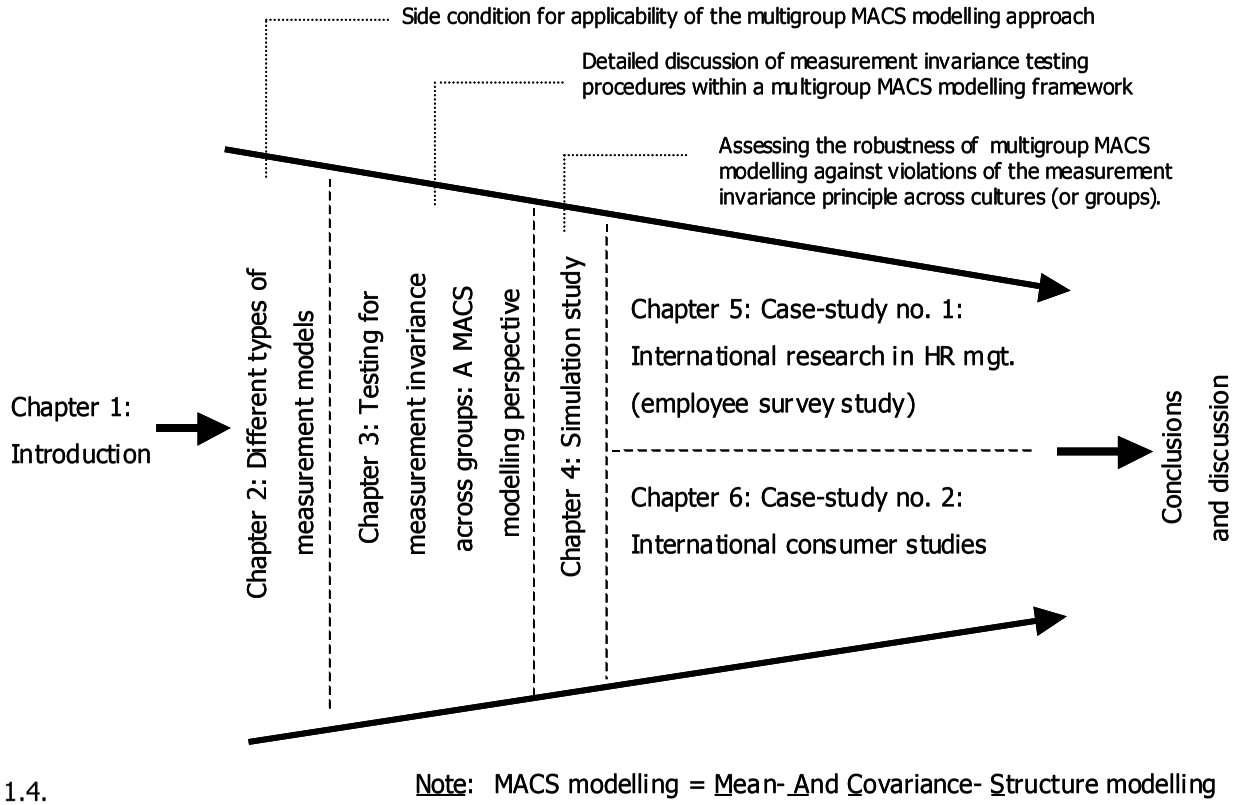


Figure 1.4.
Outline of this dissertation

Chapter 2. Different types of measurement models[18]

"The identification of a construct […] is more a matter of art than statistics."

D.J. Hand & C.C. Taylor

2.1. Constructs, observed variables, and causal theory

In Chapter 1, the notion of a 'construct' has been introduced. A construct was defined as: *"a conceptual term used to describe a phenomenon of theoretical interest"* (Cronbach and Meehl, 1955; Nunnally, 1978; Schwab, 1980). *Self-concept* is an example of a construct (e.g. Marsh and Hocevar, 1985; Byrne et al., 1989; Marsh, 1994; Marsh and Grayson, 1994). Self-concept is broadly defined as *"a person's self-perceptions, formed through experience with and interpretations of one's environment"* (Shavelson et al., 1976).

Edwards and Bagozzi (2000) view the construct itself as not real, but hypothetical. There exist, however, real phenomena to which researchers apply the construct. As such, the construct (general) self-concept is hypothetical. The five self-report items in the Self Description Questionnaire II[19] are real observed variables, which can be considered to be measures of (general) self-concept (see Marsh, 1990). Other scholars state that constructs exist in the mind of people (Loevinger, 1957, pp. 642) or in their imagination (Nunnally, 1978, pp. 96).

Kerlinger (1986, ch. 2) defined a 'construct' as *"a conceptual term with added meaning"*. Meaning is added to the conceptual term as a deliberate and conscious attempt has been made to define, specify, and operationalise the concept for the purpose of scientific study (Kerlinger, 1986). In this dissertation, it is assumed that constructs are operationalised by means of a number of 'variables' (i.e. phenomena which have been observed and measured). A variable has been defined in Chapter 1 (see Section 1.3.2) as *"a construct that has been defined so that instances of it can be assigned value and can be counted"*.

---

[18] Part of this chapter was published as a book chapter [De Beuckelaer, A. (2002). Comparison of Construct Mean Scores Across Populations: A Conceptual Framework (pp. 175-182). In S. Nishisato; Y. Baba, H. Bozdogan, and K. Kanefuji (Eds.), Measurement and Multivariate Analysis, Tokyo, Japan: Springer-Verlag].

[19] The Self Description Questionnaire II is a 102 item self-report inventory which measures self-concept in many areas ('academic', 'verbal', 'general school', 'physical abilities', 'physical appearance', 'same sex peer relations', 'opposite sex peer relations', 'parent relations', 'emotional stability', 'honesty' / 'thrustworthiness', 'total academic', 'general self'). The instrument may be ordered from the Self Research Centre (http://www.self.uws.edu.au).

With two (or more) constructs, it is possible to form a theory. A theory is defined as a *"set of interrelated constructs, definitions, and propositions that present a systematic overview of phenomena specifying relations among variables, with the purpose of explaining and predicting the phenomena* [under study]*"* (Kerlinger, 1986, p. 9). A theory may be considered a 'causal' theory if hypotheses are made about 'causes' and 'consequences'. A very simple causal theory (i.e. referred to as 'theory K') may be that construct A is expected to exert a (causal) influence on construct B. A causal diagram may be drawn to depict such a theory:

construct A -> construct B (theory K).

Alternative theories may assume a reverse causal relationship (i.e. theory L), a symmetric relationship (i.e. theory M), or no relationship at all (i.e. theory N):

construct B -> construct A (theory L)
construct B <-> construct A (theory M)
construct B construct A (theory N).

In the literature on causal modelling, the term 'construct' is often replaced by another term: 'latent variable'. MacCallum and Austin (2000), for example, stated that *"latent variables are hypothetical constructs that cannot be directly measured"*. Their definition of a latent variable explicitly stated that:

(1) latent variables are (hypothetical) constructs,

(2) latent variables cannot be measured directly.

A number of remarks have to be made with respect to MacCallum and Austin's definition of a latent variable:

(1) Even though a latent variable cannot be measured (or observed) directly, a latent variable can be measured indirectly through indicator variables,

(2) Indicator variables, shortly referred to as 'indicators', represent observed scores gathered through self-reports, interviews, observations, or some other means (Lord and Novick, 1968; DeVellis, 1991; Messick, 1995). The indicators 'capture' the real phenomena to which the term latent variable is applied. The five self-report items in the Self Description Questionnaire II, for instance, are supposed to be indicators of the latent variable '(general) self-concept',

(3) No assumptions are made about the dimensionality of the latent variable.

In this dissertation, a distinction will be made between a 'latent construct' and an 'emergent construct'. Whereas a 'latent construct' is assumed to be unidimensional, an emergent construct may be multidimensional. Both types of constructs will be explained in more detail in this chapter.

A scientific theory can be divided into two parts: a structural model and a measurement model (Jöreskog, 1973; Anderson and Gerbin, 1988). The structural model specifies (causal[20]) relationships between constructs. The measurement model describes (causal) relationships between constructs and indicators (Costner, 1969; Bagozzi and Phillips, 1982; Edwards and Bagozzi, 2002). The latter type of relationship is also referred to as an 'epistemic' relationship (i.e. a relationship describing the link between theory and data) (Cronbach and Meehl, 1955, and Fornell, 1982).

The (causal) connections between constructs on one hand, and between constructs and observed variables on the other hand, can be jointly represented in a causal diagram. An example is provided in Figure 2.1.

[20] According to principles of causality from the philosophy of science (e.g. Popper, 1959; Suppes, 1970), there are four conditions for establishing causality: (1) cause and effect are distinct entities, (2) association (i.e. often probabilistic association) is required, (3) temporal precedence (i.e. cause occurs before the effects) is required, and (4) rival explanations for the presumed relationship between the cause and the effect should be eliminated (Edwards and Bagozzi, 2000).

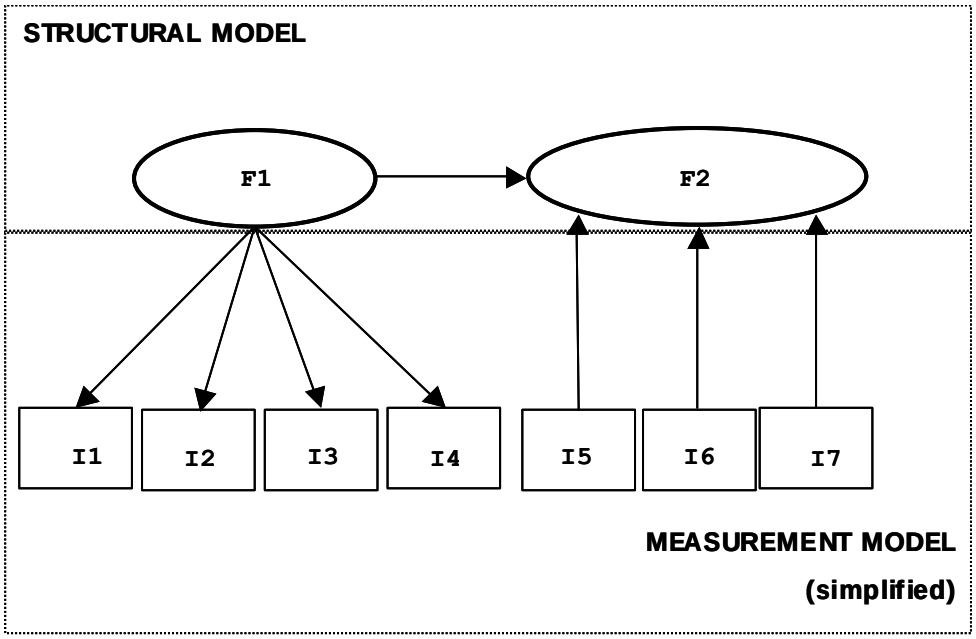


Figure 2.1.
A causal diagram to indicate hypothesised causal relationships

Note: F1 and F2 are constructs. I1 to I4 are indicators (i.e. observed variables) of F1. I5 to I7 are indicators of F2.

The structural model provides a degree of abstraction that permits researchers to generalise about relationships between the theoretical constructs, rather than making concrete statements restricted to the relationship between more specific indicators (Bollen, 2002). The nature and direction of relationships between constructs and indicators are of crucial importance because they constitute an auxiliary theory that bridges the gap between abstract theorectical concepts (i.e. constructs) and measurable empirical phenomena (i.e. indicators) (Edwards and Bagozzi, 2000).

In the next section, different types of constructs are discussed. Subsequent sections deal with multivariate statistical methods to test for measurement invariance across groups (i.e. Section 2.3.), and the identification of the true nature of the construct (i.e. Section 2.4).

2.2. Latent constructs versus emergent constructs

2.2.1. Different types of constructs

Several authors make a distinction between two types of constructs: (truly) 'latent' constructs and 'emergent' constructs (Blalock, 1964, pp. 162-169; Bollen, 1984; Cohen et al., 1990; Bollen and Lennox, 1991; Cole et al., 1993; MacCallum and Browne, 1993; Edwards and Bagozzi, 2000; Diamantopoulos and Winklhofer, 2001).

2.2.1.1. Latent constructs

If theory suggests that indicators are merely observable reflections (rather than determinants) of the construct, it is most plausible that the construct influences its indicators (and not the other way around). The construct is then called a 'latent construct', and the indicators are called 'reflective' indicators. The latent construct is the common cause of its indicators. It is a unidimensional construct.

An example of a latent construct is shown in Figure 2.2. Examples of latent constructs are listed in Exhibit 2.1 (see further). Additional examples from the management/marketing literature are provided in Appendix 2.1.

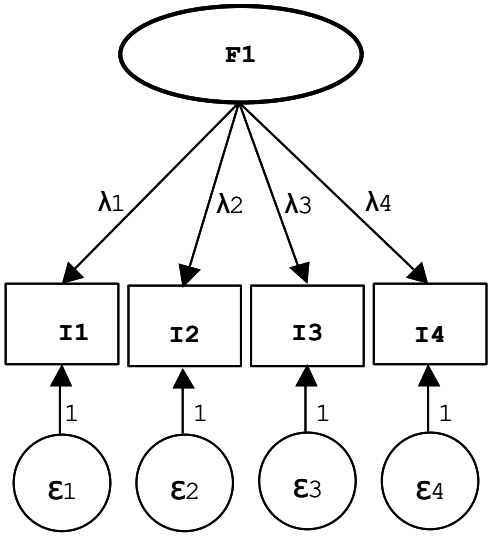


Figure 2.2.
Example of a latent construct

Notes: For reasons of simplicity the variance of F1 (i.e. Var(F1)), and the variances of the measurement error terms (i.e. $Var(\varepsilon_1), Var(\varepsilon_2), Var(\varepsilon_3), Var(\varepsilon_4)$) are not shown in this Figure. The indicators are indicated by I_i (with i referring to the number of the indicator), the construct as F1, the factor loading of indicator i as λ_i, the measurement error of indicator i as ε_i, and the variance of F1 as Φ_1 (the last parameter is not shown in the figure).

The following equations represent the measurement model for the latent construct F1:

$$I_1 = \lambda_1 F1 + \varepsilon_1$$
$$I_2 = \lambda_2 F1 + \varepsilon_2$$
$$I_3 = \lambda_3 F1 + \varepsilon_3$$
$$I_4 = \lambda_4 F1 + \varepsilon_4$$

The covariance between indicators I_i and I_j of the same construct (i.e. F1) equals $\lambda_i \lambda_j \Phi$ (Bollen, 1989; Bollen and Ting, 2000) .

Exhibit 2.1.
Examples of latent constructs

'General intelligence' (Cohen et al., 1990)
(with subsets of an IQ test as indicators)

Different types of *personality traits* (e.g. 'big 5' personality model discussed by Goldberg, 1990)
(with self or other reports of behavioural tendencies and preferences as indicators).

'Quantitative reasoning' (Bollen, 2002)
(with test scores on several tests of quantitative reasoning as indicators)

'Self-esteem' (Bollen, 2002)
(with degree of agreement with questions about self-worth as indicators)

'Depression' (Cohen et al., 1990)
(with self-reports of feeling states as indicators)

2.2.1.2. Emergent constructs

If one believes that the construct is the result (i.e. the combined effect) of its indicators, the construct and its indicators are referred to as an emergent construct and formative indicators, respectively. An emergent construct may be perceived as some kind of (e.g. a linear) composite measure of its indicators (MacCallum and Browne, 1993).

Figure 2.3 shows an example of an emergent construct. More examples of emergent constructs are shown in Exhibit 2.2 (see further). Additional examples taken from the management/marketing literature are listed in Appendix 2.2.

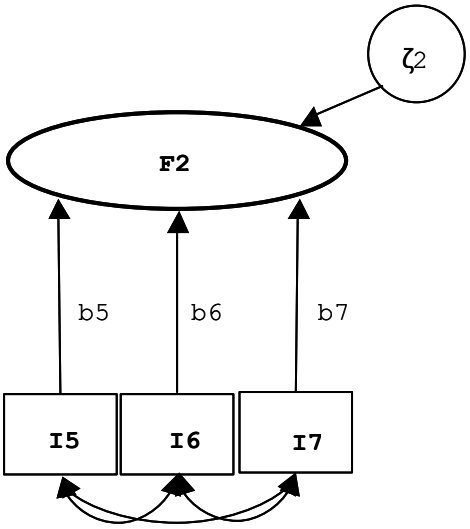


Figure 2.3.
Example of an emergent construct

Note: ζ_2 may be assumed to be zero.

The following equation represents the measurement model for the emergent construct F2:

$$F2 = b_5 I_5 + b_6 I_6 + b_7 I_7 + \zeta_2$$

The indicators are indicated by I_i (with i referring to the number of the indicator), the construct as $F2$, the path coefficient of indicator i as b_i. The disturbance term (i.e. ζ_2) consists of all of the other variables that influence $F2$ that are not included in the measurement model. The disturbance term is uncorrelated with the individual indicators (i.e. $Covar(I_i, \zeta_2)=0$) and $F2$. Obviously, this measurement model is, in isolation, (statistically) underidentified. Some authors (e.g. Bagozzi and Fornell, 1982; Diamantopoulos and Winklhofer) omit the disturbance term (ζ_2) in the equation above (i.e. they assume that the expected value of ζ_2 is zero).

Exhibit 2.2.
Examples of emergent constructs from the literature

'Socioeconomic status' (Hauser and Goldberger, 1971; Heise, 1972; Marsden, 1982; Bollen and Lennox, 1991)
(with indicators: occupational prestige; education; and income)

'Exposure to discrimination' (Bollen and Lennox, 1991)
(with indicators: race; age; sex)

'Social support' (MacCallum and Browne, 1993)
(with indicators: duration of caregiving; hours per day caregiving; days off from caregiving)

'Family socialisation' (Heise, 1972)
(with indicators: mother's liberalism; father's liberalism)

'Relationship Closeness' (Berscheid et al., 1989; see also Bollen and Lennox, 1991)
(with indicators: 38 activities)

'Exposure to media violence' (Bollen, 2002)
(with indicators: time spent watching violent television programmes; time spent watching violent movies; and time spent playing violent video games)

'A mother's availability to interact with and monitor any given child' (Cohen et al., 1990)
(with indicators: number of children in the family; illness of the mother; hours of maternal employment)

'Time spent in social interaction' (Bollen, 2002)
(with indicators: time spent with friends; time spent with family; and time spent with coworkers)

'Quality of life' (Bollen and Ting, 2000)
(with indicators: self-reported health; hapiness; and economic status)

'Vulnerability to heart attack' (Cohen et al., 1990)
(with indicators: high blood cholesterol; high blood pressure; smoking; family history of heart disease)

'Chronic health' (in the elderly)' (Liang, 1986; Cole et al., 1993)
(with indicators: cancer; arthritis; and indicators representing some nervous disorders)

'Accuracy of memory' (Bollen and Ting, 2000)
(with indicators: a number of details correctly recalled)

Formative measurement underlies the use of so-called 'block variables'. Block variables represent a summary of the effect of several (observed) variables (Igra, 1979; Marsden, 1982). A block variable yields a single summary estimate

of the effects of observed variables in the block on some outcome variable. The observed variables in the block are seen as distinct causes of the outcome variable. In management research, constructs are often treated as emergent when their indicators describe different facets of a general concept (Blau et al., 2001; Westphal & Zajac, 2001). In principle, indicators should only be viewed as formative indicators when they are 'causes' of the construct (i.e. when they cause variation in the construct).

Bagozzi and Edwards (1998) distinguished between four types of measurement models for constructs depending on the extent to which blocking was used to summarise item-(or component-) level data. The four types of models are:

(A) Total disaggregation model

(B) Partial disaggregation model

(C) Partial aggregation model

(D) Total aggregation model.

The models are graphically depicted in Appendix 2.3. In the total disaggregation model (i.e. model A), each item is used as an independent indicator for a construct, whereas in the total aggregation model (i.e. model D) all items are summed (or averaged) to form one composite of the construct. The partial disaggregation model (i.e. model B) and the partial aggregation model (i.e. model C) are a compromise between both extremes (i.e. the total disaggregation model and the total aggregation model). Yoo (2002) uses a partial disaggregation model to model the (17-item) consumer ethnocentrism (CETSCALE), which has been developed by Shimp and Sharma (1987).

Obviously, in all models where items are blocked (i.e. all models except for the total disaggregation model) it is implicitely assumed that the blocking variable represents an emergent construct. Further, in a multi-group setting (e.g. a multi-country study), measurement invariance across groups is implicitly assumed as far as the blocking variables are concerned. This assumption is problematical as it is not formally testable.

Diamantopoulos and Winklhofer (2001) discuss four issues that are critical to create sound formative constructs (i.e. content specification, indicator specification, indicator collinearity, and external validity). Interested readers can refer to their paper for an extensive discussion.

2.2.1.3. Mixed constructs (and MIMIC models)

Sometimes constructs are presented both as a latent and an emergent construct. Cohen et al. (1990) provided an example of 'physical ill health', which can either be measured by its causes (i.e. cerebrovascular and cardiovascular disease, muscular-skeletal disease, cancer, and immune system related problems) or by its consequences (i.e. pain severity and persistence, energy level, fatigue-proneness, and activity limitation).

A construct can possibly have both formative and effect indicators. The corresponding measurement models are referred to as MIMIC (i.e. Multiple Indicators, MultIple Causes) models (Hauser and Goldberger, 1971; Jöreskog and Goldberger, 1975). Bollen and Lennox (1991) stated that the Center for Epidemiological Studies Depression Scale (CES-D; Radloff, 1977) has some effect indicators (e.g. 'I felt depressed' and 'I felt sad'), as well as some causal indicators (e.g. 'I felt lonely'). As explained in Appendix 2.5, some authors consider the MIMIC model to be an alternative to the multigroup MACS model when one aims at testing for measurement invariance of scales across groups (e.g. nations or cultures). Because of some crucial shortcomings of the MIMIC approach to measurement invariance testing (e.g. the inability to identify differences in factor loadings across groups), the MIMIC approach is not considered in this dissertation. The shortcomings of the MIMIC approach are explained in more detail in Appendix 2.5.

2.2.1.4. Other types of constructs

Fornell (1982) stated that, apart from reflective and formative indicators, indicators can also be symmetric. If indicators are symmetric, then it makes no sense to determine a causal direction between the indicators and the construct. Instead, indicators are mapped into 'an abstract space'. An example of such an abstract space may be a two-dimensional representation of the structure of consumer values in a specific country. Exploratory statistical methods (e.g. principal components analysis, multidimensional scaling, etc.) are most appropriate to determine such a structure.

2.2.2. Implications for construct measurement

A lot of debate exist as to which type of construct (i.e. latent or emergent) is most frequently encountered in the social sciences. Horn and McArdle (1992), for example, claimed that almost all concepts in the behavioral sciences represent emergent constructs. This claim is in contrast with the literature on psychological concepts which suggests that for personality traits and attitudes, in particular, latent constructs are more appropriate (Sörbom, 1981; Fornell, 1982; Fornell and Bookstein, 1982; Bollen, 2002; Borsboom et al., 2003). The examples of constructs mentioned earlier in this chapter make it plausible that both types of constructs do occur simultaneously in the social sciences (and in international management).

The knowledge of whether observed variables are either reflective or formative indicators is crucial for several reasons:

(1) reflective indicators have other measurement properties than formative indicators (Bollen, 1984; Bollen and Lennox, 1991; Bollen and Ting, 2000);

(2) treating variables as effect indicators while they are formative indicators leads to model specification error. Inconsistent parameter estimates and misleading conclusions (also between constructs) are the likely consequences of model misspecification (Bollen and Lennox, 1991; MacCallum and Browne, 1993; Bollen, 2002). Furthermore, the researcher's understanding of the causal effects in the variable system[21] may be seriously distorted as a consequence of assuming wrong causal connections (Bollen and Ting, 2000).

As mentioned above, reflective indicators are different from formative indicators as far as their measurement properties are concerned. They are different in the following three aspects:

Aspect 1: With respect to the mutual *correlations* between the indicators measuring the construct:

> Reflective indicators should be highly correlated, as they measure essentially the same thing, namely a unidimensional (latent) construct.
>
> The correlations among formative indicators are not explained by the measurement model. Nothing is known about neither the strength (i.e.

[21] Huberty and Morris (1989) define a 'variable system' as: "a collection of conceptually interrelated variables that, at least potentially, determine one or more meaningful underlying variates (or constructs)".

high / low) nor the direction (positive / negative) of their mutual correlations, unless sufficient guidance is provided on substantial grounds (e.g. a priori knowledge). As a consequence, 'internal consistency measures' are useless as an estimate of scale reliability whenever the indicators are formative indicators (Diamantopoulos and Winklhofer, 2001). An exception is Bentler's (1968) maximal internal consistency measure[22], which is a maximal internal reliability measure for a composite measure (i.e. an emergent construct). This measure has only become popular recently (Hancock and Mueller, 2001).

Aspect 2: With respect to the *interchangeability* of indicators measuring the construct:

> It is crucial to understand that the omission of a (necessary) indicator may lead to invalid measurements if the construct is emergent, as one crucial dimension (or facet) is not taken into account. If the construct is latent, however, no effects on the adequacy of measurement of the construct are to be expected when an indicator is omitted (apart from a potential loss of reliability due to the smaller number of indicators). Obviously, if the most reliable indicator is removed from the whole set of indicators, there may be a substantial loss in reliability of measurement.
>
> Emergent constructs may require a more exhaustive list of indicators.

Aspect 3: With respect to corrections for *measurement error*:

> From a modelling perspective, measurement error can be taken into account if the construct is latent, but not if the construct is emergent (Bollen, 1984; Bollen and Lennox, 1991). With latent indicators, the variance in true scores is lower than the variance in indicator scores (i.e. due to the correction for measurement error); with formative indicators, the opposite is true (see Fornell et al. 1991; Diamantopoulos and Winklhofer, 2001). The assumption that no measurement error would affect the indicator scores is unrealistic in most situations.

In a recent paper, Smith and Reynolds (2001) discussed the cross-cultural applicability of scales used to measure (perceived) service quality. Service quality is usually measured by means of the 'SERVQUAL scale'. This scale has been introduced by Parasuraman, Berry, and Zeithaml (1988), and has been modified in a later stage by the same authors (Parasuraman, Berry, and Zeithaml, 1994). Smith & Reynolds (2001) discuss some major threats to the cross-cultural/cross-national applicability of the SERVQUAL scale (e.g. non-

[22] This measure is implemented in the software EQS.

equivalencies across nations due to cultural differences in quality expectation and/or extreme response styles). It is remarkable that Smith and Reynolds do not make any statement regarding the nature of the service quality construct (i.e. latent or emergent). They do point out that there is a lack of consensus relating to the dimensionality of the service quality construct (Smith & Reynolds, 2001, p. 461), but they do not elaborate on the latent or emergent nature of possible subdimensions of the construct. Meanwhile, the reader will have understood that it is impossible to test for measurement invariance of the SERVQUAL scale without making assumptions regarding the nature of (subdimensions) of the service quality construct. In another study, Ueltschy et al. (2004) use an alternative scale to measure service quality (i.e. the SERVPREF scale). According to Ueltschy et al. (2004), service quality construct may be perceived as a one-dimensional latent construct (see Ueltschy et al., 2004, figure p. 905). More details can be found in their paper (Ueltschy et al., 2004).

In the next section, some multivariate analyses techniques used to model latent or emergent constructs will be evaluated. In addition, the extent to which these techniques allow for testing measurement invariance (in a multiple-group context) will be discussed. The last section of this chapter will provide some guidelines as to how the true nature of a construct can be identified.

2.3. Multivariate statistical methods to test for measurement invariance (across groups)

2.3.1. Latent constructs

Reflective measurement underlies classical test theory (Lord and Novick, 1968), reliability estimation (Nunnally, 1978), item response theory (Lord, 1980; Hambleton and Swaminathan, 1985; Du Toit, 2003), factor analysis (Kim and Mueller, 1978), and latent class analysis (Lazarsfeld and Henry, 1968; Goodman, 1974; McCutcheon, 1987). Each theory treats an observed variable as a function of a construct plus error. As a consequence, confirmatory approaches to test for measurement invariance across groups as discussed in Chapter 1 (i.e. multiple-group MACS model, DIF models based on IRT, and the multigroup latent class model) are justified whenever constructs are latent.

2.3.2. Emergent constructs

The idea of formative measurement seems to match with the 'canonical model', which forms the basis of many classical multivariate techniques (Fornell, 1982). In the canonical model, canonical discriminant functions are created using observed variables as input variables. To be able to create such discriminant functions, a grouping variable is needed (i.e. a variable representing group-membership). In Multivariate Analysis of Variance (MANOVA), for example, linear (canonical) discriminant functions are computed so that these groups show maximal differentiation in terms of their scores on the discriminant functions. The discriminant functions provide composite measures (i.e. a weighted sum of indicators), which can be interpreted as emergent constructs (Cole et al., 1993) provided that all formative indicators needed to define the construct have been taken into account. In the same way, principal components (in a principal components analysis, see, for example: Kim and Mueller, 1978) and canonical variates (in a canonical correlation analysis, see Thompson, 1984) can be interpreted as emergent constructs.

Discriminant functions (or composite measures) are a cause of concern as the numerical indicator weights are 'optimal' from a statistical point of view only (e.g. [in MANOVA]: providing optimally discriminating discriminant functions between groups). The indicator weights which determine the construct may not be relevant from a theoretical point of view (Cole et al., 1993).

Researchers may prefer to work with an optimal weighting procedure because measurement models with formative indicators are often underidentified (Bollen, 1989; MacCallum and Browne, 1993). Using an optimal weighting procedure

may be the only option to circumvent the identification problem (unless the causal relationships between the construct and other consequences of that construct [other constructs or indicators] are known and included in a more elaborated model[23]). A possibility exists whereby some observed variables, which are wrongly assumed to represent formative indicators of the construct, get high indicator weights not because they are important indicators of the construct, but because they differentiate well between the groups (Cole and Maxwell, 1985).

In sum, the fact that there is no theory on the basis of which a measurement model can be built on, makes it very hard to define emergent constructs in a reliable and valid way. This is, however, not the only problem with emergent constructs.

When dealing with emergent construct in a multiple-group situation, there is no way to formally test for measurement invariance across groups. It is (implicitly) assumed that the discriminant function(s), which represent the emergent construct(s), can be meaningfully applied to all groups. This assumption is referred to as the 'homogeneity of regression' assumption (Marsh and Grayson, 1990). The homogeneity of regression assumption represents an extreme form of (assumed) measurement invariance. The main problem here is that this assumption is not formally testable.

A further problem relates to the stringent modelling assumptions. The combination of the homogeneity of regression assumption and the assumption of equal variance/covariance matrices across populations makes MANOVA a too stringent multivariate technique to be used in practice (Kühnel, 1980; Stelzl and Schnabel, 1992). An alternative multivariate technique, namely Partial Least Squares[24] [PLS] (see Wold, 1982,1985; Fornell and Cha, 1994; Chin, 1998), poses less stringent assumptions, albeit that PLS also fails to provide a means of testing for measurement invariance across groups. The PLS approach is often used in one-country studies. Examples within the applied economics and management literature can be found in: Jagpal (1981), Fornell and Bookstein (1982), Steenkamp and van Trijp (1996), Sirohi et al. (1998), Hulland (1999), and Rodgers (1999).

For reasons mentioned above, the conclusion seems justified that none of the available multivariate techniques are optimal for making cross-group comparisons based on an emergent construct. This may explain why some

[23] Simultaneous estimation of both the measurement part of the model and the structural part of the model may lead to an identified (or overidentified) model.
[24] Partial Least Squares (PLS) can be used to model both latent and emergent constructs (see Fornell and Bookstein, 1982). PLS is often considered to be an alternative to Confirmatory Factor Analysis, for example in: Fornell (1982); Fornell and Bookstein (1982), Chin (1995, 1998), Steenkamp and van Trijp (1996), Retzer and Fusso (1999).

researchers have used multivariate techniques that are appropriate for latent constructs (e.g. factor-analytic techniques), even when dealing with constructs which are emergent (Cole et al., 1993). Cohen et al. (1990) provided empirical examples of constructs which were treated as if they were latent constructs, even though evidence existed that these constructs were emergent. These constructs are listed in Appendix 2.4.

2.4. Identifying the true nature of constructs

One of the main challenges in choosing a measurement model, is to determine the direction of causation between constructs and their indicators (i.e. identifying the true nature of the constructs). As mentioned before, assuming a wrong type of construct may invalidate the results obtained from empirical research. In practice, the true nature of the construct is hard to identify.

A fortunate researcher knows about the true nature of the construct (i.e. latent or emergent). Substantive reasoning (i.e. reasoning based on theoretical knowledge) combined with evidence from prior empirical research (e.g. from fitting certain statistical models which assume a latent measurement structure) may provide strong indications as to whether the construct is latent or emergent. If the researcher is not that fortunate, then he has to find a way to identify the true nature of the construct. Several possibilities exist. They are discussed in the next sections.

2.4.1. Designed experiments

In exceptional cases, it is possible to design experiments that help to test whether variables are causal or effect indicators. In Bollen (1982), four indicators were proposed to measure (perceived) air quality: the colour, the clarity, the odour, and the overall quality of the air. All four indicators were hypothesised to be effect indicators of (perceived) air quality (i.e. hypothesis no. 1 [H1]). The alternative hypothesis suggested that the overall measure is a reflective indicator of (perceived) air quality and that the other indicators are causal indicators (i.e. hypothesis no. 2 [H2]). Cermak (1983) designed an experiment to test the plausibility of these alternative hypotheses. His reasoning was that if H2 is true, then the time it takes for an individual to respond should be greater for the overall measure of air quality than for the other measures (indicators). Alternatively, if H1 is true, then the time to respond should be essentially the same for all four measures. Bollen (1989, pp. 67) mentioned that some theories might lead to different predictions of response time. So the experiment is conditional on Cermak's (1983) assumptions about people's perceptions (of air quality). In most practical situations, such experiments are simply not feasible to execute (Bollen and Ting, 2000).

2.4.2. Mental experiments

Another possibility is to perform 'mental experiments', in which a researcher assumes a shift in the construct and then judges whether a simultaneous shift in all indicators is likely. If so, then this is consistent with an effect indicator specification. Alternatively, if the researcher assumes a shift in an indicator as leading to a shift in the construct even if there is no change in the other indicators, then this is consistent with a formative measurement model (Bollen, 1989, p. 65-67; Bollen and Ting, 2000). Heise (1972), for example, argues that the construct 'socioeconomic status' is caused by measures of education, income, and educational prestige on the basis of the premise that changes in these socioeconomic variables lead to changes in socioeconomic status, but not the reverse.

Theoretically, simultaneous reciprocal causation may exist between an indicator variable and the construct. An example (given by Bollen, 1989, pp. 66) is 'financial health' of a company measured by the stock price of the company's shares. Greater financial health can cause a higher stock price, and a higher stock price can increase financial health.

2.4.3. Confirmatory TETRAD analysis

Another approach is to model so-called (vanishing) 'tetrads'. Tetrads refer to the difference between the product of a pair of covariances and the product of another pair among four random variables (indicators). For a foursome of variables one can arrange the six covariances into three tetrads (Bollen and Ting, 2000):

$$T_{1234} = \sigma_{12}\,\sigma_{34} - \sigma_{13}\,\sigma_{24}$$
$$T_{1342} = \sigma_{13}\,\sigma_{42} - \sigma_{14}\,\sigma_{32}$$
$$T_{1423} = \sigma_{14}\,\sigma_{23} - \sigma_{12}\,\sigma_{43}$$

where the symbol T_{ijkl} is used to indicate a tetrad of the four variables $i, j, k,$ and l, and the symbol σ_{ij} refers to the population covariance between two indicators. A vanishing tetrad is a tetrad which has expectation zero.

In the next paragraphs, the assumption is made that there are exactly four indicators per construct. Cases in which there are less than four indicators per construct are discussed further on in the text.

Vanishing tetrads with four effect indicators for the construct

In case the four indicators are effect indicators it can be shown that the three tetrads mentioned above are equal to zero (i.e. 'vanishing tetrads'):

$$T_{1234} = \sigma_{12}\,\sigma_{34} - \sigma_{13}\,\sigma_{24} = (\lambda_1\lambda_2\Phi)\,(\lambda_3\lambda_4\Phi) - (\lambda_1\lambda_3\Phi)\,(\lambda_2\lambda_4\Phi) = \Phi^2\,(\lambda_1\lambda_2\lambda_3\lambda_4 - \lambda_1\lambda_2\lambda_3\lambda_4) = 0$$

$$T_{1342} = \sigma_{13}\,\sigma_{42} - \sigma_{14}\,\sigma_{32} = (\lambda_1\lambda_3\Phi)\,(\lambda_4\lambda_2\Phi) - (\lambda_1\lambda_4\Phi)\,(\lambda_3\lambda_2\Phi) = \Phi^2\,(\lambda_1\lambda_2\lambda_3\lambda_4 - \lambda_1\lambda_2\lambda_3\lambda_4) = 0$$

$$T_{1423} = \sigma_{14}\,\sigma_{23} - \sigma_{12}\,\sigma_{43} = (\lambda_1\lambda_4\Phi)\,(\lambda_2\lambda_3\Phi) - (\lambda_1\lambda_2\Phi)\,(\lambda_4\lambda_3\Phi) = \Phi^2\,(\lambda_1\lambda_2\lambda_3\lambda_4 - \lambda_1\lambda_2\lambda_3\lambda_4) = 0$$

where Φ is the variance of the latent construct.

The three vanishing tetrads are determined by the latent measurement structure, not by the parameters of the model (e.g. λ_i [i=1,2,3,4] or Φ). This approach where vanishing tetrads are modelled based on expectations about the causal structure within the data is referred to as 'confirmatory tetrad analysis'. The technique was discussed in an earlier paper by Bollen and Ting (1993). The methodology was applied to test for a latent and an emergent construct in Bollen and Ting (2000).

Vanishing tetrads with four formative indicators for the construct

Because all observed variables are exogeneous, there are no constraints on the covariances among the formative indicators. Except for the unlikely circumstance that the values of the two pairs of covariances (included in the tetrad) cancel each other out exactly, none of the tetrads are expected to be vanishing tetrads.

If, however, the indicators are not linearly related, then their covariances would tend towards zero. If both sides of the tetrad difference have covariances equal to zero, then the possibility that the tetrad vanishes still exists. For this reason, Bollen and Ting (2000) suggested some additional tests:

(1) a statistical significance test to test the null-hypothesis that each covariance appearing in the tetrad is zero

(2) to verify what measurement parameter values (e.g. factor loadings and the variance of the construct) are obtained if the (corresponding) effect indicator model would be estimated. If one or more of the factor loadings or the variance of the construct is not

significantly different from zero, then the causal indicator model may be more plausible than the effect indicator model.

Vanishing tetrads with less than four indicators per construct

Bollen and Ting (2000) discussed in an appendix on how vanishing tetrads can be used when less than four indicators are used to measure a latent (or an emergent) construct. Their idea was to include indicators of other latent constructs in the vanishing tetrads. They specified all vanishing tetrads for a variety of measurement models which differ in terms of the direction of the causal relationship between the constructs and their indicators. The interested reader is encouraged to refer to the original paper by Bollen and Ting (2000). Computational algorithms from Glymour et al. (1987) and Spirtes et al. (1994) can be used to derive the vanishing tetrads.[25]

[25] To date, this approach can only be used with recursive linear Structural Equation Models (i.e. models without 'feedback loops'). Extensions to non-recursive linear Structural Equation Models are currently under development.

2.5. Conclusions

In this chapter, the importance of specifying the correct causal direction between constructs and their indicators was shown. Substantive reasoning and / or theory (e.g. checking the temporal priorities between constructs and their indicators; mental experiments), and empirical checks (e.g. designed experiments and/or confirmatory tetrad analysis) may help the researcher determine whether a construct is latent or emergent, given the set of indicators.

After reviewing the literature on different measurement models, it is apparent that none of these approaches are fool-proof. The results of mental experiments are, to a large extent, subjective as they are largely based on the current beliefs of the researcher. Designed experiments are often not feasible, and –if they are– they require a test strategy which may be based on wrong a priori assumptions (e.g. about response times in Cermak's experiment). Confirmatory tetrad analysis only leads to tentative conclusions, because imperfect reflective indicators may yield covariances that deviate from the pattern expected for such indicators, and formative indicators may exhibit covariances that happen to follow the pattern expected for reflective indicators (Edwards and Bagozzi, 2000).

It is the author's conviction that none of these approaches can substitute substantive reasoning (i.e. reasoning based on theoretical knowledge). At best, they can provide 'additional checks' when trying to identify the true nature of the construct.

Chapter 3. Testing for measurement invariance across groups: A Mean- And Covariance- Structure (MACS) modelling perspective

"Statistics is a method for panning precious order from a sand of complexity."

I. Stewart

3.1. Introducing Mean- And Covariance- Structure (MACS) Modelling

In the 1970's, Dag Sörbom proposed a general method for studying differences in factor means[26] and factor structure between groups (Sörbom, 1974, 1978). This model is known as the Mean- And Covariance- Structure (MACS) model. The MACS model is an extension of Jöreskog's factor-analytic model in multiple groups (Jöreskog, 1971). Like Jöreskog's model, the MACS model is based on asymptotic theory and the principle of Normal Theory Maximum Likelihood (NML). Since its introduction, many psychometricians and cross-cultural psychologists have used the MACS model. The main reason for its popularity among researchers may be attributed to the fact that the MACS model provides an excellent statistical tool to answer two very relevant research questions (RQ):

Research question no. 1 (RQ1):
Does the factor model (with multiple indicators) lead to 'valid' factor quantifications in every group of interest?

Research question no. 2 (RQ2):
Are the (estimated) factor means significantly different across groups?

These research questions are adequately addressed by evaluating a series of competing MACS models. A hypothesis-testing framework can be used for this purpose (see for example: Steenkamp and Baumgartner, 1998; Vandenberg and Lance, 2000; Vandenberg, 2002; De Beuckelaer, 2002). The sequence of statistical model tests, which represent alternative hypotheses about the factor and mean-structure across groups, is described further on in this chapter. The term 'MACS approach' is used throughout this chapter to indicate that one or more MACS models are used to test alternative hypotheses about the factor- and mean-structure across groups.

[26] The scale and origin set for the factor (i.e. the latent variable) are fixed in an arbitrary way.

(partial) Measurement invariance of indicators (across groups)

A direct comparison of factor means (i.e. RQ2) is only meaningful and valid from a substantive point of view under certain conditions. The indicators, which are used to measure the factor under study, are required to exhibit 'measurement invariance' across groups. Measurement invariance of indicators exists when the numerical values across groups are on the same measurement scale (Drasgow, 1984, 1987). It is implied that all[27] measurement parameters should, at least in principle, be identical across groups. Measurement parameters which relate to indicators of a one-factor model are: factor loadings, indicator intercepts, and unique variances.

Indicators that do not satisfy the condition of measurement invariance across groups may show different numerical values when the (underlying) factor score is identical, but group membership is different. Such indicators are 'biased' and lead to wrong statistical conclusions in terms of the true factor differences across groups. Potential sources of indicator bias (i.e. differences in measurement parameters across groups) should be investigated (i.e. RQ1) before cross-group differences in factor means are tested (i.e. RQ2). Sometimes, only a subset of all measurement parameters (for example: factor loadings, but not indicator intercepts) satisfies the condition of measurement invariance across groups. This phenomenon is referred to as 'partial measurement invariance' of indicators across groups. Some authors (Muthén and Christofferson, 1981; Marsh and Hocevar, 1985; Byrne, 1989; Byrne et al., 1989; Reise et al., 1993; Byrne and Watkins, 2003) claim that when dealing with partial measurement invariance, factor means across groups can be compared. Others (Rock et al., 1978; Labouvie, 1980; Meredith, 1993; Marsh and Grayson, 1994; Little, 1997) argue that cross-group invariance of indicator intercepts and factor loadings is needed in order to meaningfully compare factor means across groups. The simulation study presented in Chapter 4 will provide the answer as to whether the cross-group invariance of factor loadings and indicator intercepts is an absolute necessity.

The next sections provide some statistical background concerning the MACS approach (Section 2), and explain a procedure that is typically used (within the MACS approach) to test for measurement invariance of indicators across groups (Section 3). In the last section (Section 4), a sequence of hierarchically nested MACS models is proposed for usage in international (cross-cultural) research.

[27] Exceptions to this rule are the unique variances. As will be explained further on in this chapter, the requirement of equality of unique variances (or indicator reliabilities as they are a direct function of unique variances and the factor variance) is an overly restrictive condition.

3.2. Statistical Background

3.2.1. Model specification

Consider a linear factor model with k (common[28]) factors and p indicators (in total):

$$x_g = \upsilon_g + \Lambda_g \xi_g + \delta_g \quad (1)$$

where:

subscript g denotes group membership (g = 1,2,…,G);
G represents the total number of groups considered;
x_g is a (p x 1)- vector of indicator scores;
υ_g is a (p x 1)- vector of indicator intercepts;
ξ_g is a (k x 1)- vector with k common factors;
Λ_g is a (p x k)- matrix of factor loadings (or regression weights);
δ_g is a random (p x 1)- vector of residuals (i.e., including random and systematic error).

In this chapter, all p indicators in x_g are assumed to be reflective indicators of the factor (construct). In addition, all indicators are at their lowest level of aggregation. As such, the 'total disaggregation model' (as shown in Appendix 2.3) describes the relationship between the (common) factors and their indicators. The p indicators in x_g are independent and identically distributed according to a multivariate normal distribution with mean vector μ_g , and variance-covariance matrix Σ_g. Furthermore, the following (standard) modelling assumptions are made:

$$E(\delta_g) = 0 \text{ and } Corr(\delta_g, \xi_g) = 0$$

Under these assumptions the group-specific mean- and variance-covariance-structures of x_g can be expressed as:

$$\mu_g(\theta_g) = \upsilon_g + \Lambda_g \kappa_g \quad \ldots (2)$$

$$\Sigma_g(\theta_g) = \Lambda_g \Phi_g \Lambda_g^{'} + \Theta_g \quad (3)$$

[28] 'Common factors' are factors that are thought to exert an influence on multiple indicators.

where:

μ_g represents a mean-structure in group g;
Σ_g represents a variance-covariance- structure in group g;
θ_g represents the complete set of (model) parameters to be estimated in group g;
κ_g is a (k x 1)- vector of factor means;
Θ_g is a (p x p)- matrix of unique variances;
Φ_g the (k x k)- matrix of factor variances and covariances.

In a (confirmatory) factor analysis model, it is commonly assumed that the matrix of unique variances in every group (i.e. Θ_g) is diagonal so that indicators' unique variances are uncorrelated.

It is also assumed that the parameters in equations (2) and (3) are identified (Jöreskog, 1969, Sörbom, 1974). Because only factor mean differences are defined, the factor mean in the first group (i.e. the 'reference group') is set equal to zero:

$$\kappa_1=0$$

The factor means in the other (G-1) groups are expressed as deviations from κ_1. This additional constraint is necessary because of an indeterminancy of the parameters υ_g and κ_g (in the special case where $\Lambda_1 = \Lambda_2 = \ldots = \Lambda_G$) (Sörbom, 1974).

3.2.2. Model estimation

When x_g follows a multivariate normal distribution with mean vector μ_g , and variance-covariance matrix Σ_g, the parameters in equations 2 and 3 can be estimated using a Normal Theory Maximum Likelihood (NML) estimation procedure. If NML estimation is used, the following discrepancy function is minimised (Browne and Arminger[29], p. 188):

$$F_{NML}(\theta) = \sum_{g=1}^{G} \left(\frac{N_g}{N} \right) f(\theta_g) \quad (4)$$

where:

N_g indicates the total number of observations in group g ($n=1,2,\ldots N_g$);
N indicates the total number of observations across all groups.

In Formula 4 (Browne and Arminger, p. 188):

$$f(\theta_g) = F_g + \left\{ \left(\bar{x}_g - \hat{\mu}_g \right)' \hat{\Sigma}_g^{-1} \left(\bar{x}_g - \hat{\mu}_g \right) \right\} \quad (5)$$

where:

$$F_g = \left\{ \log\left|\hat{\Sigma}_g\right| + tr\left[T_g \hat{\Sigma}_g^{-1} \right] - \log\left|T_g\right| - p \right\} \quad (6)$$

with:

$$T_g = \left(\frac{1}{N_g - 1} \right) \sum_{n=1}^{N_g} (x_n - \bar{x}_g)(x_n - \bar{x}_g)' \quad (7)$$

(T_g represents the sample variance-covariance matrix of the indicators)

The first term in Formula (5) (i.e. F_g) represents the standard Normal Theory ML function which is minimised in (classical) covariance- structure modelling.[30] The additional term [indicated between square brackets in Formula (5)] adds the weighted sum of squares resulting from the discrepancy between the vector with sample indicator means (i.e. $\bar{x}_g$) in group g, and the estimated means at population level in group g (i.e. $\hat{\mu}_g$). The distribution of the estimated means and covariances are assumed to be independent. This assumption may be violated in case of non-normal data.

[29] Browne and Arminger (1995) use a different notation when specifying this discrepancy function.
[30] The function $f(\theta_g)$ is different if the General Least Squares (GLS) estimator is used.

The statistic $N\,\hat{F}_{NML}(\Theta)$ with $\hat{F}_{NML}(\Theta)$ at its minimum represents the usual (loglikelihood ratio) test statistic (Browne and Arminger, 1995). The loglikelihood ratio statistic statistically compares the current model with the saturated model (i.e. a model with zero degrees of freedom). The loglikelihood ratio statistic as well as the NML based chi-square statistic is printed by all standard SEM modelling software.

For further details concerning the estimation procedure (e.g. derivatives of $F(\Theta)$), the reader is referred to Sörbom, 1974). With discrete, but ordinal data (e.g. Likert-types of scales), the assumption of multivariate normality of x_g is not legitimate. This violation does not affect the 'consistency'[31] of the ML estimator, but it turns the ML estimator into a(n asymptotically) less 'efficient'[32] estimator. The usefulness of the (standard) Maximum Likelihood estimator may, therefore, be questioned on conceptual grounds. A (conceptually) more elegant alternative would be the Weighted Least Squares (WLS) estimator, also known as the Asymptotically Distribution Free Estimator (ADF). This estimator, which has been introduced by Browne (1982, 1984), specifies an optimal weight matrix[33] leading to asymptotically unbiased parameter estimates (West et al., 1995). Though conceptually superior, the WLS does not perform well in practice. Simulation research has shown that WLS only produces trustworthy parameter estimates when (very) large samples are used (Hu et al., 1992; Chou and Bentler, 1995; Olsson et al., 1995; Curran et al., 1996). The same simulation studies have also demonstrated that the NML estimator and, in particular, Satorra and Bentler's (1994[34]) corrections for the NML test statistic provide a better alternative even when discrete, but ordinal indicators are used. Satorra and Bentler's (1994) scaled test statistic corrects the goodness-of-fit test (i.e. Chi-square) to better approximate chi-square under non-normality. Their scaled test does not adjust the parameter estimates, but robust standard errors are available. The parameter estimates are robust against violations of the assumption of multivariate normality of the data. Because of its better performance in these simulation studies, the scaled test statistic by Satorra and Bentler's (1994) will be used in the simulation study presented in Chapter 4.

[31] Consistent estimators are estimators that converge in probability (with increasing sample size) to the true parameter value (in the population).
[32] Asymptotic efficient estimators produce correct estimations of the mean squared error of parameter estimates. Asymptotic inefficient estimators fail to do so (Boomsma, 2003, p. 7-3).
[33] The optimal weight matrix consists of a combination of second- and fourth- order (central) product-moment terms.
[34] In addition to this paper, one may also consult earlier papers, such as Satorra & Bentler (1988) and Satorra (1992).

3.3. Testing for measurement (and factor mean) invariance across groups

3.3.1. Hypothesis testing

The MACS approach is especially attractive as it offers a hypothesis-testing framework. The hypothesis-testing framework allows the researcher to test a number of substantive questions. In the first phase of a hypothesis-testing procedure, the tenability of the assumption of measurement invariance across groups should be evaluated. A number of MACS models are evaluated to identify those measurement parameters that do (or do not) satisfy the condition of measurement invariance (of indicators) across groups. Statistical tests on the equality of factor means (and factor variances) across groups may be appropriate in the second phase of the hypothesis-testing procedure. As mentioned before, a comparison of factor means across groups is meaningful only if sufficient evidence is found to support the assumption of measurement invariance of indicators across groups.

3.3.1.1. A preliminary test

Several researchers recommend conducting a preliminary statistical test on the hypothesis of equality of variance-covariance matrices across groups (Jöreskog, 1971; Rock et al., 1978; Alwin and Jackson, 1981; Cole and Maxwell, 1985; Byrne et al., 1989; Horn and McArdle, 1992; Bagozzi and Edwards, 1998; Vandenberg and Lance, 2000; Vandenberg, 2002). More formally, this hypothesis is expressed as:

$$H_0 : \Sigma_1(\Theta) = \Sigma_2(\Theta) = \ldots = \Sigma_g(\Theta) = \ldots = \Sigma_G(\Theta)$$

where $\Sigma_g(\Theta)$ represents the variance-covariance matrix in group g (g=1,2,…,G).

The researchers' recommendation is based on the idea that, when covariances and variances are equal across groups, the assumption of measurement invariance of indicators across groups is legitimate. This implies that one should not proceed with any further tests on the equality of measurement parameters across groups (Jöreskog, 1971; Mulaik, 1975; Alwin and Jackson, 1981; Cole and Maxwell, 1985; Horn and McArdle, 1992; Bagozzi and Edwards, 1998; Vandenberg and Lance, 2000; Vandenberg, 2002). Unfortunately, the equality of variance-covariance matrices across groups does not provide a sufficient basis to assume cross-group measurement invariance of indicators. If the only parameter which is different across groups is the indicator intercept (e.g. temperature data expressed as Degrees Celsius [in one group] and Kelvin [in another group]), then identical variance-covariance matrices are obtained for all groups but, obviously, the condition of measurement invariance across groups is

not fulfilled. As this test on the equality of variance-covariance- structures across groups is not 'fool-proof', specific hypotheses in which specific measurement parameters (in particular: indicator intercepts) are constrained to be equal across groups should always be tested.

In the next paragraphs, some MACS models are presented which contain different assumptions regarding the degree of invariance (across groups). These MACS models may be tested in two successive phases (Byrne et al., 1989). In the first phase, MACS models are tested for various aspects of measurement invariance of indicators across groups. In the second phase, issues of 'structural invariance' rather than 'measurement invariance' are considered. Structural invariance issues, for example, relate to the equality of factor means, factor variances, and factor covariances across groups. The subdivision in two phases is in line with Anderson and Gerbing's (1988) argument that one should first understand what is being measured, before one investigates the mutual relationships among what is measured (see also Vandenberg and Lance, 2000).

3.3.1.2. Phase 1: Testing for measurement invariance of indicators across groups

'Congeneric factor invariance model'

The first MACS model states that the a priori pattern of fixed (i.e. nonsalient) and freed (i.e. salient) factor loadings is equivalent across groups (Horn and McArdle, 1992). The fixed factor loadings are typically constrained to zero. This model (H_0) is referred to as the *congeneric factor model*.

H_0: *Equivalent pattern of freed and fixed factor loadings across groups*

The congeneric factor model (H_0) is also referred to as the 'model of configural invariance' (Vandenberg and Lance, 2000). The model implies that all observed variables load on the same underlying factors[35], but the magnitude of the (nonsalient) factor loadings may differ across groups. The congeneric factor model has no cross-group constraints on estimated parameters (Marsh, 1994). Provided that this model is identifiable, it can be used as a 'baseline model' (i.e. a 'reference' model) for further models with more restrictions on the measurement parameters (e.g. Reise et al., 1993; Marsh, 1994; Bagozzi and Edwards, 1998; Vandenberg and Lance, 2000). If the congeneric factor model

[35] In the congeneric factor invariance model it is (at least in principle) possible that some indicators load on multiple (underlying) factors. From a 'measurement perspective', however, such indicators should not be selected. It is, therefore, assumed that 'double-loadings' (of indicators on factors) are not specified in the congeneric factor invariance model.

does not fit the data, then the conclusion must be that the basic meaning of the factor[s] (or construct[s]) differ across groups (De Beuckelaer, 2002).

'Metric invariance model'

The next test concerns the equality of factor loadings across groups. More formally, the restriction is:

$$H_1 : \Lambda_1 = \Lambda_2 = \ldots = \Lambda_g = \ldots = \Lambda_G$$

The measurement invariance condition under H_1 is referred to as 'metric invariance' of indicators across groups. The goodness-of-fit statistics of model H_1 (e.g. Chi-squared values with their degrees of freedom) can be mutually compared with the goodness-of-fit statistic of model H_0. If the goodness-of-fit statistic of model H_1 is not substantially worse than the goodness-of-fit statistic of model H_0 (i.e. given the difference in degrees of freedom of both models), it is concluded that equality of factor loadings across groups can be assumed. This metric invariance model can be statistically compared with the congeneric invariance model because they form a 'nested structure' (i.e. one model has a subset of the restrictions of the other model). The most commonly used goodness-of-fit statistic to compare nested models is the (ordinary) Chi-square statistic. Some researchers use goodness-of-fit statistics other than Chi-square for comparing nested statistical models, for example: Tucker and Lewis's nonnormed index (e.g. Little, 1997).

'Tau-invariance model'

The next step would be to test the hypothesis:

$$H_2 : \Lambda_1 = \Lambda_2 = \ldots = \Lambda_g = \ldots = \Lambda_G \quad \text{and}$$
$$\upsilon_1 = \upsilon_2 = \ldots = \upsilon_g = \ldots = \upsilon_G$$

Indicators that satisfy the invariance condition specified under H_2 are said to be 'tau-invariant'[36] across groups. Once cross-group equality of factor loadings and indicator intercepts is demonstrated (i.e. tau-invariance is established), factor means across groups (i.e. phase 2 of the hypothesis-testing procedure) can be compared. The equality of factor loadings and indicator intercepts across groups

[36] The term 'tau-invariance' (of the same indicator) across groups should not be confused with the term 'tau-equivalence' which is frequently used to indicate that two alternative measurements (indicators) have identical factor loadings with respect to the true factor (or construct) score (Traub, 1994).

provides sufficient evidence to conclude that the measurement scale used to score the indicators is identical across groups (Drasgow, 1984, 1987). The literature survey by Vandenberg and Lance (2000) showed that of all empirical research papers which dealt with some form of measurement invariance, only 12% tested for the condition of tau-invariance of indicators across groups.

'Parallel invariance model'

If a researcher has an interest in the extent to which the indicators are equally reliable across groups, he/she may consider conducting one more hypothesis. The additional hypothesis states that:

$$
\begin{aligned}
H_3 : \Lambda_1 &= \Lambda_2 = \ldots = \Lambda_g = \ldots = \Lambda_G \text{ and} \\
\upsilon_1 &= \upsilon_2 = \ldots = \upsilon_g = \ldots = \upsilon_G \text{ and} \\
\Theta_1 &= \Theta_2 = \ldots = \Theta_g = \ldots = \Theta_G \text{ and} \\
D_1(\Phi_1) &= D_2(\Phi_2) = \ldots = D_g(\Phi_g) = \ldots = D_G(\Phi_G)
\end{aligned}
$$

where $D_g(\Phi_g)$ refers to the elements on the main diagonal of the (k x k)-matrix of variances and covariances of ξ_g (g=1,2,…,G). As the elements on the main diagonal are variances, $D_g(\Phi_g)$ refers to the variances of ξ_g (g=1,2,…,G).

In model H_3, the equality of indicator reliabilities across groups is imposed by putting cross-group constraints on the unique variances and factor variances (Rock et al., 1978; Cole and Maxwell, 1985; Vandenberg and Lance, 2000). Strictly speaking, this model is too restrictive. Parallel invariance of indicators may also be tested by adjusting the indicators' reliabilities for group differences in factor variances. Consequently, hypothesis H_3' may be tested instead of hypothesis H_3.

$$
\begin{aligned}
H_3' : \Lambda_1 &= \Lambda_2 = \ldots = \Lambda_g = \ldots = \Lambda_G \text{ and} \\
\upsilon_1 &= \upsilon_2 = \ldots = \upsilon_g = \ldots = \upsilon_G \text{ and} \\
\left(\frac{D_1(\Phi_1)}{D_1(\Phi_1)+D_1(\Theta_{1,p})}\right) &= \left(\frac{D_2(\Phi_2)}{D_2(\Phi_2)+D_2(\Theta_{2,p})}\right) = \ldots = \left(\frac{D_g(\Phi_g)}{D_g(\Phi_g)+D_g(\Theta_{g,p})}\right) = \ldots = \left(\frac{D_G(\Phi_G)}{D_G(\Phi_G)+D_G(\Theta_{G,p})}\right)
\end{aligned}
$$

for all p indicators.

The MACS model specified under H_3' represent a less restrictive MACS model than the one specified under H_3. The latter model is, however, adequate to test for parallel invariance of indicators.

Parallel invariance of indicators is a rather extreme form of measurement invariance in which factor loadings, indicator intercepts, and indicator reliabilities

are assumed to be identical across groups. Alwin and Jackson (1981) advocate testing for parallel invariance prior to testing for differences in factor means across groups.

Conducting a test for parallel invariance is, however, not generally recommended because of the following reasons:

(1) cases in which the researcher has a substantive interest in the (invariance of) indicator reliabilities across groups are not so often encountered,

(2) the probability of establishing such an extreme form of measurement invariance would be very unlikely given that unique variances consist largely of random error (Hittner, 1995),

(3) equality of indicator reliabilities across groups is an overly restrictive condition when one aims at comparing factor means across groups.

Once hypothesis H2 (or H3) has not been rejected, it makes sense to proceed with the second phase of the hypothesis-testing procedure.

3.3.1.3. Phase 2: Testing for factor mean invariance across groups

'Equal factor means'

The next hypothesis to be tested, states that factor means are identical across groups:

$$H_4 : \kappa_1 = \kappa_2 = \ldots = \kappa_g = \ldots = \kappa_G$$

The model under H_4 can be statistically compared with the congeneric factor model (H_0), the metric invariance model (H_1), the tau-invariance model (H_2), and the parallel invariance model (H_3).

'Equal factor (co-)variances'

An additional restriction can be included to further hypothesise that the factor variances and factor covariances are identical across groups:

$$H_5 : \kappa_1 = \kappa_2 = \ldots = \kappa_g = \ldots = \kappa_G \text{ and}$$
$$\Phi_1 = \Phi_2 = \ldots = \Phi_g = \ldots = \Phi_G$$

Obviously, not all elements or Φ_g (g=1,2,…,G) have to be fixed across groups. For example, cross-group constraints on all factor variances (i.e. $V\Phi_g$) may be imposed, while freely estimating all factor covariances in all groups (i.e. $C\Phi_g$). Such a multigroup MACS model would take a position in the hierarchical sequence of MACS models in between the model specified under H_4 and the model specified under H_5. Equivalent factor variances across groups imply that the range that the factor/construct uses to respond to its indicators is equivalent across groups (Vandenberg and Lance, 2000). Equivalent factor covariances imply that the basic factor structure (i.e. the 'conceptual domain' of the factors) is invariant across groups (Vandenberg and Lance, 2000).

The next section looks at the exact sequence in which the (hierarchical) MACS models are to be tested.

3.4. Recommended sequence of MACS model tests

Vandenberg and Lance (2000) have shown in their literature study that different authors have used / proposed a different sequence of (hierarchically nested) multigroup MACS models (see Jöreskog, 1971; Rock et al., 1978; Alwin and Jackson, 1981; Cole and Maxwell, 1985; Drasgow and Kanfer, 1985; Schaie and Hertzog, 1985; Byrne et al., 1989; Horn and McArdle, 1992; Reise et al., 1993; Marsh, 1994; Nesselroade and Thompson, 1995; Bagozzi and Edwards, 1998; Chan, 1998; Steenkamp and Baumgartner, 1998; Taris et al., 1998; Vandenberg and Lance, 2000). This finding does not (necessarily) imply that some of the researchers have adopted a wrong sequence of statistical model tests.

As argued by Vandenberg and Lance (2000), the considerable difference in focus between these papers may justify somewhat different (although partly overlapping) sequences of hierarchical multigroup MACS model tests. Cole and Maxwell (1985), for instance, dealt with multigroup MACS model tests in the context of multitrait-multimethod analyses (i.e. situations in which a number of constructs/factors are measured using *multiple* measurement instruments). Chan (1998) proposed another sequence of MACS model tests to deal with latent *growth* modelling in multiple groups. Latent growth modelling aims at investigating the evolution of factor means over time. Steenkamp and Baumgartner (1998) introduced a sequence of hierarchical MACS model tests which they recommend for international (i.e. cross-cultural) research. The sequence of MACS model tests proposed by Steenkamp and Baumgartner (1998) and Vandenberg and Lance (2000) have especially stimulated the author's view on the 'recommended sequence' to test for the various MACS models in international (cross-cultural) research. The sequence of tests that is proposed in this chapter is also in line with recommendations made by Meredith (1993).

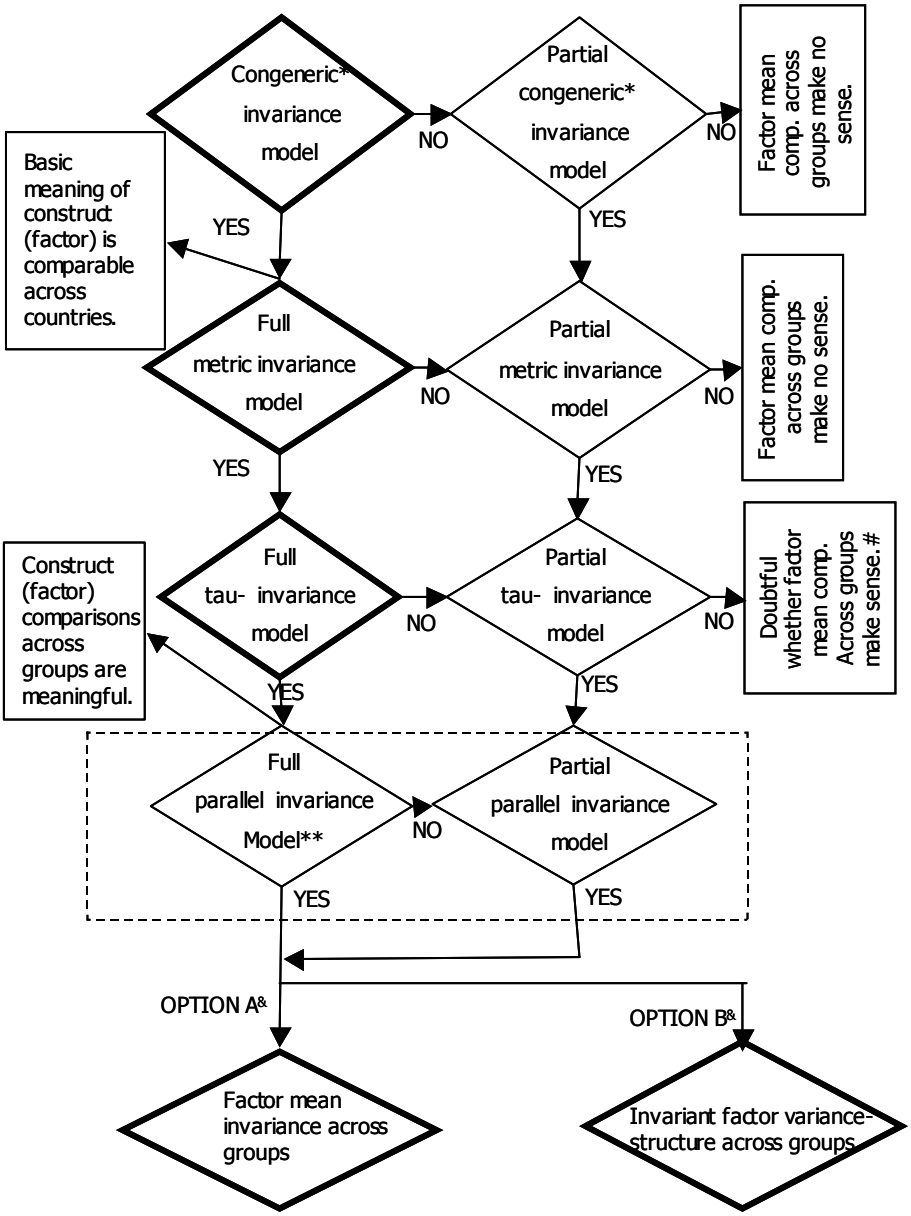
A second reason why the sequence of hierarchical MACS model tests may be altered lies in the particular characteristics of the data sampled from various groups (see Vandenberg and Lance, 2000). For example, when location parameters (i.e. indicator intercepts) are sample-specific, the model of tau-invariance of indicators across groups (i.e. model specified under H_2) is not relevant from theory. The reader should, however, recall that cross-group comparisons based on factor means would not be meaningful if the invariance of factor loadings and indicator intercepts is not established across groups (i.e. the requirement of tau-invariance across groups). Hence, cross-group factor mean comparisons would not be possible using such a dataset. However, in the (unlikely) case where an appropriate test-equating procedure (Engelen and Eggen, 1993) is available, comparability of factor means across groups may still be legitimate.

Some scepticism with respect to some recommendations made in the methodological literature concerning a recommended sequence of MACS model tests seems justified.

As explained before, the practice of conducting an initial hypothesis test on the equality of variance-covariance matrices across groups cannot be recommended for use in international (cross-cultural) research. An exception would be a situation in which the researcher has no substantial interest in verifying the hypothesis of equal location parameters (i.e. indicator intercepts) across groups. For instance, indicator intercepts may be expected to differ across groups as they reflect (predictable) 'response threshold differences' rather than a source of bias (see, for instance, Vandenberg and Lance, 2000). Response threshold differences across groups occur when respondents from one group (culture) are less inclined to respond positively (e.g. agreeing with a statement) than respondents from another group (culture). Some dedicated software (e.g. Mplus) allow for the specification group-specific threshold models so that these known response effects can be incorporated in the MACS model.

Furthermore, it is uncertain whether partially invariant indicators may be accepted in the (common) factor model. The simulation study in Chapter 4 will investigate the consequences of dealing with partially invariant indicators as far as the reliability of factor mean comparisons across groups is concerned. So, as long as the results of the simulation are not known, a reservation is made concerning the inclusion of partially invariant indicators in the (common) factor model.

To conclude, a procedure is recommended to assess measurement invariance of factor indicators in international (cross-cultural) research. The procedure is graphically displayed in Figure 3.1 (part 1) and Figure 3.2 (part 2). The procedure is partly based on the sequence of hierarchical MACS models proposed by Steenkamp and Baumgartner (1998) and Vandenberg and Lance (2002).



(Figure continues on the next page)

Figure 3.1.
Proposed hierarchical sequence of MACS model tests

(Figure starts on the previous page)

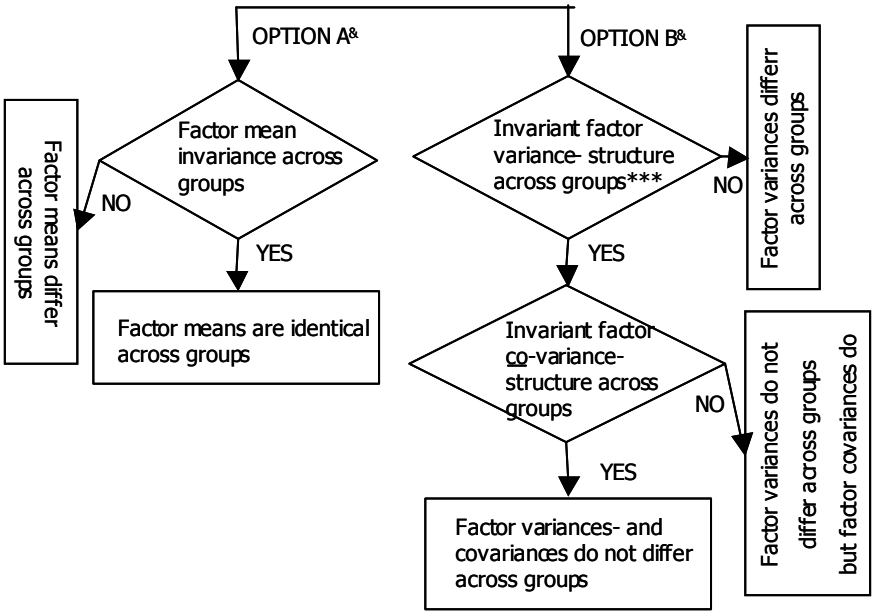


Figure 3.2. (continuation of Figure 3.1.)
Proposed hierarchical sequence of MACS model tests

Notes (with Figure 3.1.):

A dashed box indicates an optional hypothesis test (conduct this test if it is relevant from theory)

*Also referred to as 'configural invariance model';

**Two different versions of the parallel invariance test have been presented in the text (one being more restrictive than the other).

#Whether or not factor mean comparisons across groups make sense in this situation will be investigated in the simulation study in Chapter 4.

Notes (with Figure 3.2.):

&After choosing option A one may always choose option B or the other way around;

***One may immediately test for invariant factor variances- and covariances across groups. This implies that two subsequent steps represented in this flowchart are combined into one single hypothesis test.

In the previous sections, it has been argued that factor means can only be meaningfully compared across groups if all factor loadings and all indicator intercepts are invariant across groups (i.e. when tau-invariance across groups is established for all indicators). The attentive reader will have noticed that Figure 3.1 and 3.2. suggest that a comparison of factor means across groups can be made only if a subset of the indicators representing a factor exhibit tau-invariance across groups. In the simulation study which will be presented in Chapter 4, the correctness of statistical conclusions with regard to the factor mean difference test for various degrees of non-invariance of an indicator (with respect to factor loadings and/or indicator intercepts) is investigated. Based on the outcomes of this piece of research, modifications will be made to the proposed sequence of hierarchical model tests presented in Figure 3.1 and 3.2. .

The hypothesis-testing framework shown in Figure 3.1 and 3.2. allows the researcher to detect non-invariant measurement parameters across groups. As many applied researchers do not formally test for non-invariance of measurement parameters across groups, (false) assumptions are often made regarding the comparability of factor means across groups. This is why simulation research as presented in Chapter 4 is crucial for research practice.

Chapter 4. The robustness of factor mean comparisons against violations of the measurement invariance assumption across groups

"An ounce of replication is worth a ton of inferential statistics."

J.H. Steiger

4.1. Introduction

In Chapter 3 it was stated that the equality of factor loadings and indicator intercepts across groups (i.e. tau-invariance across groups) is a prerequisite to make cross-group comparisons at factor-level (e.g. comparison of factor means, factor variances, and factor covariances across groups). To date, there is no consensus in the methodological literature as to what level of measurement invariance across groups is required before such cross-group comparisons at factor-level are meaningful. Meredith (1993), Little (1997), and Steenkamp and Baumgartner (1998) firmly stated that tau-invariance across groups is required. Others believed that metric invariance across groups is required (e.g. Alwin and Jackson, 1981; Reise et al., 1993). Still others claimed that only a subset of all factor loadings (i.e. partial metric invariance across groups) would be sufficient to make meaningful cross-group comparisons at factor level (Muthén and Christofferson, 1981; Marsh and Hocevar, 1985; Byrne, 1989; Byrne et al., 1989; Reise et al., 1993). Less stringent invariance conditions, such as partial metric invariance, are proposed because of the growing belief that measurement instruments can hardly ever be totally equivalent across groups (e.g. Horn et al., 1983; Byrne and Watkins, 2003).

In this chapter, a simulation approach[37] is used to investigate the extent to which non-invariance of factor loadings and indicator intercepts may lead to false statistical conclusions in terms of (the reported significance of) factor mean differences across groups. This research question is especially relevant to a researcher who would not conduct any tests to identify possible sources of non-invariance of measurement parameters across groups. As argued by Cheung & Rensvold (1999), Vandenberg & Lance (2000) and Williams et al. (2003), researchers often do not formally test for possible sources of non-invariance of

[37] One may question whether an analytical approach (e.g. using 'power analysis') can be used for the purpose of this study. The problem with an analytical approach is that it is not accurate because of the categorical nature of the data (as used in this study). Categorical data are generated from Normally distributed (underlying) variates (or scores). As a consequence, the covariance structure calculated on the basis of raw (categorical) scores differs from the "implied" covariance structure (which is based on the parameter values specified for the various experimental conditions [e.g. factor loadings and indicator intercepts], and the 'true' scores as determined by the underlying normal distributions). It is because of this distortion of the covariance structure that the analytical approach is inaccurate.

measurement parameters across groups. Research on the effects of non- (or partial) invariance on group comparisons may be considered to be a welcome addition to the methodological literature. In a recent article Vandenberg (2002) stated:

> *"… the current article addresses some of the shortcomings in our understanding of the analytical procedures [to test for measurement invariance]. In particular, it points out the need to address (a) the sensitivity of the analytical procedures, (b) the susceptibility of the procedures to contextual influences, (c) how partial [measurement] invariance affects the tests of substantive interest, and (d) the triggers or causes for not supporting [measurement] invariance. In the hopes of stimulating further research on these topics, ideas are presented as to how this research may be undertaken."* (Vandenberg, 2002)

The current research focuses in particular on points (a) and (c) in the above citation.

The next sections deal with the method of research and the analysis plan (Section 2), the detailed results (Section 3), and the general conclusions from the research (Section 4).

4.2. Method

As mentioned in Chapter 3, Satorra and Bentler's scaled Chi-square statistic[38] was used in this simulation study because of its superior performance in earlier simulation studies. The design of this study was similar to the design of an earlier simulation study by Kaplan and George (1995). There were, however, some important differences. These differences will be explained in the next paragraphs.

4.2.1. Experimental design

4.2.1.1. Experimental conditions

The following design factors were used in the simulation study: number of indicators for the factor (factor 0); type of distribution of the indicators (factor 1); sample size in the different groups (factor 2); factor mean difference between groups at population level (factor 3); non-invariance of factor loadings and indicator intercepts (factor 4 and factor 5). One can distinguish between two groups of factors: 'side factors' and 'measurement non-invariance factors'. Side factors include all factors listed above except for factor 4 (i.e. non-invariant factor loading) and factor 5 (i.e. non-invariant indicator intercept). Factor 4 and factor 5 create non-invariance conditions as they determine the degree of measurement non-invariance of the non-invariant indicator. The simulation was set-up using a full-factorial[39] (experimental) design.

Side conditions

- Number of groups -

In the context of this simulation study, it was decided to work with two groups. More than two groups would have led to an experimental design which is far too complex to handle.

- Number of indicators (F0) –

The simulation study consisted of two separate simulation studies. In the first study, three indicators were specified for the factor to be measured (see Appendix 4.1, factor 0). In the second study, four indicators were specified. In

[38] More precisely, it is the mean- and variance adjusted Chi square statistic with robust standard errors which is used in this simulation study.

[39] Alternatively, one could also have used a 'fractional design'. A fractional design is more efficient in terms of the number of experimental conditions, but it does not allow to test for higher-order interactions (between the design factors).

actual market or public opinion research, one seldom includes more than four indicators for one factor. The reason is that a large number of indicators per factor would lead to a substantial increase in fieldwork costs. This explains why, in this simulation experiment, no conditions were included with more than 4 indicators for the factor.

- Distribution of indicators (F1) -

Three different distributions were specified for the indicators (see Appendix 4.1, design factor no. 1). In the first condition, the (standard) normal distribution was used (see Figure 4.1.). In the other two conditions, non-normal distributions arising from indicators with five response categories were specified. Five-point (Likert-type of) scales are very popular in consumer research and public opinion research to measure respondents' degree of agreement or disagreement with specific statements (items)[40]. The non-normal distributions are: (1) the uni-modal left-skewed distribution (see Figure 4.2.), and (2) the symmetric bi-modal distribution (see Figure 4.3.). Two different threshold models were specified to convert simulated z-scores (i.e. scores under the standard normal distribution) into five response categories in accordance with the proportions specified in Figures 4.2 and 4.3. (and in Appendix 4.1). These particular distributions were chosen because they are frequently encountered when working with five-point (disagree-agree) scales. Unlike this study, the earlier simulation study by Kaplan and George (1995) made only use of data which follows the (standard) normal distribution.

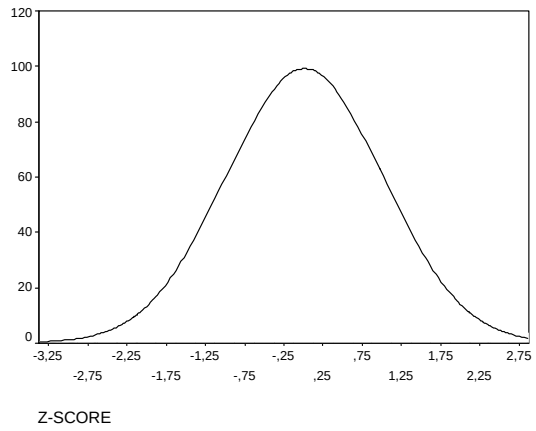


Figure 4.1.
(standard) Normal distribution

[40] If one would use only 4 category points per item, it would be better to work with (estrimated) 'polychoric correlations' rather than treating the data as if they were metric (see Wallentin, 2004).

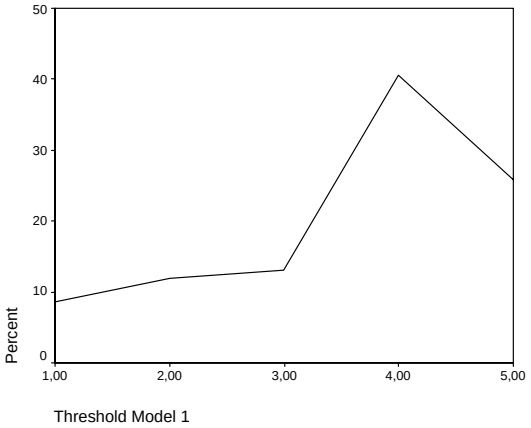


Figure 4.2.
Uni-modal left-skewed distribution (with five category points)

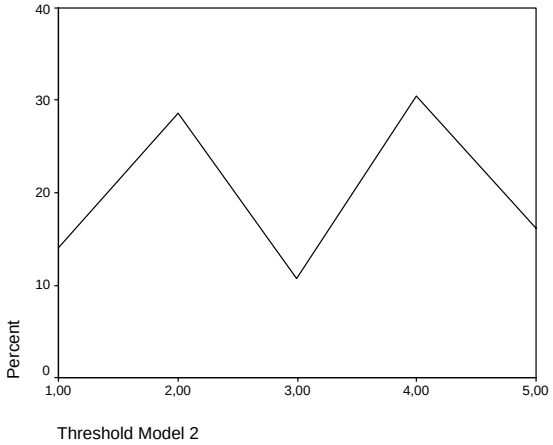


Figure 4.3.
Symmetric bi-modal distribution (with five category points)

- Sample sizes per group (F2) -

Sample sizes were held equal across groups (except for one setting) and varied between 200 and 750. More details are provided in Appendix 4.1 (see design factor no. 2).

Sample sizes which are substantially smaller than 200 observations per country are rather uncommon in international (commercial) market research (at least, according to the author's personal experience). The reason is that, in each country, a substantial number of observations should be collected on different types of consumer groups (e.g. product users versus non-product users, males versus females, families with children versus families without children, etc.).

- Factor mean differences at population level (F3) -

Factor mean differences at population level vary across experimental conditions. Three basic conditions apply:

(1) the factor mean difference at population level is zero. For estimation purposes the factor mean in group 1 is always fixed to zero. To ensure that the factor mean difference is zero at population level the factor mean in group 2 is also set equal to zero,
(2) the factor mean difference at population level is small (i.e. 0.15),
(3) the factor mean difference at population level is large (i.e. 0.30).

Further on in this chapter, it will be explained why a factor mean difference of 0.15 can be considered to be small, whereas a factor mean difference of 0.30 can be considered to be large.

Conditions 2 and 3 are represented by two experimental conditions: positive and negative discrepancy cases. In positive discrepancy cases, the factor mean in the second group is larger than the factor mean in the first group. In negative discrepancy cases the reverse condition applies (i.e. factor mean in group 2 is smaller than in the first group). So, the design factor 'factor mean difference' (between both groups) comprises five conditions in total (see also Appendix 4.1, design factor F3: factor mean difference). The variance of the factor is fixed to one in both groups. It will be explained in Section 4.2.1.2 why the simulation study considers negative discrepancy cases in addition to positive discrepancy cases.

Non-invariance conditions

- Non-invariance of factor loadings and indicator intercepts (F4 and F5) -

In this simulation study, measurement non-invariance was caused by one indicator only (out of three / four). This indicator is always the second one. This indicator may or may not have shown measurement non-invariance across the two groups. The non-invariant indicator has one or two measurement parameters (in particular: factor loadings and indicator intercepts) that differed across groups. These measurement parameters are referred to as 'λ_2' and 'Int_2', respectively. To refer to the value of these parameters in the second group, the suffix (G2) is added.

The settings for factor loadings resemble the settings specified by Kaplan and George (1995). Indicator reliabilities ranged between 0.24 (with factor loading equal to 0.4) and 0.56 (with factor loading equal to 0.8).[41] Differences in the indicator intercepts varied between 0.00 and 0.45. The latter value of the indicator intercept represents a distance of nearly one tenth of the 'length' of the total scale (i.e. five category points). The details are provided in Appendix 4.1 (design factors numbers 4 and 5). One may expect that differences in indicator intercepts across groups are more harmful than differences in factor loadings when (estimated) factor mean scores are to be compared across groups. Differences in indicator intercepts will bias estimated factor mean scores equally for each observation (or person), whereas the bias resulting from differences in factor loadings really depend on the observation's (or person's) score on the underlying construct.

In addition to non-invariance conditions, the corresponding invariance conditions were also included in the study. As will be explained in the analysis section, the invariance conditions provide a 'natural benchmark' against which the (statistical) performance of the factor mean difference test may be evaluated in non-invariance conditions. The study by Kaplan and George (1995) differs from this study in that the researchers did not include the possibility of unequal indicator intercepts in addition to unequal factor loadings (across groups).

[41] Indicator reliabilities are calculated as follows: 1 – (error variance / [lambda^2 + error variance]). The error variance is always fixed to 0.51 in the simulation study

4.2.1.2. Asymmetrical structure of the design

The specific non-invariance conditions (see Appendix 4.1, design factors number 4 and 5) showed that the experimental design has a structure which is asymmetrical. The asymmetric structure was a result of the experimental settings specified for the intercept of the (non-invariant) indicator. The indicator intercept of the non-invariant indicator in the second group could have been larger than (or equal to) the corresponding indicator intercept in group one. The factor loading of the non-invariant indicator in the second group could have been smaller than, equal to, or larger than the corresponding factor loading in group one. The asymmetry was thus caused by the experimental settings for the non-invariant indicator intercept, and not by the experimental settings for the non-invariant factor loading.

Due to the asymmetry, the effect of unequal indicator intercepts across groups on the (estimated) size of (absolute) difference in factor means across groups were different for positive and negative discrepancy cases. In positive discrepancy cases, unequal indicator intercepts increases the estimated discrepancy between factor means. In negative discrepancy cases, the estimated discrepancy between factor means decreases due to the inequality of indicator intercepts across groups. Therefore, the inclusion of negative indicator intercepts in the simulation design (in addition to positive indicator intercepts) would only lead to duplicate information as some conditions with a positive discrepancy between factor means would be identical to some other conditions with a negative discrepancy between factor means.

4.2.2. Simulation process

Multiple data files (i.e. 50[42]) were generated for each experimental condition. Several software programs were written to run the simulations. The programs took care of the data preparation and data extraction tasks. The actual parameter estimations were provided by a (dedicated) software program, namely Mplus (Muthén and Muthén, 1999). An overview of these programs is provided in Appendix 4.2. Examples of Mplus input files are provided in Appendix 4.3.

4.2.3. Analysis strategy

The results from the simulation study were analysed in two consecutive steps. These two steps are explained below.

Step1:
Correct and incorrect statistical conclusions

Using the simulated data files (i.e. one for every replication of an experimental condition), factor means were estimated for both groups. The estimation was carried out under the (possibly false) assumption that measurement invariance holds across groups (i.e. equality of factor loadings and indicator intercepts across groups). The model specified when estimating the (measurement) model parameters is the tau-invariance model. This model (with 3 factor indicators) is shown in Figure 4.4.

[42] Fifty replications per experimental condition is sufficient given that the simulation experiment involves such a large number of experimental conditions.

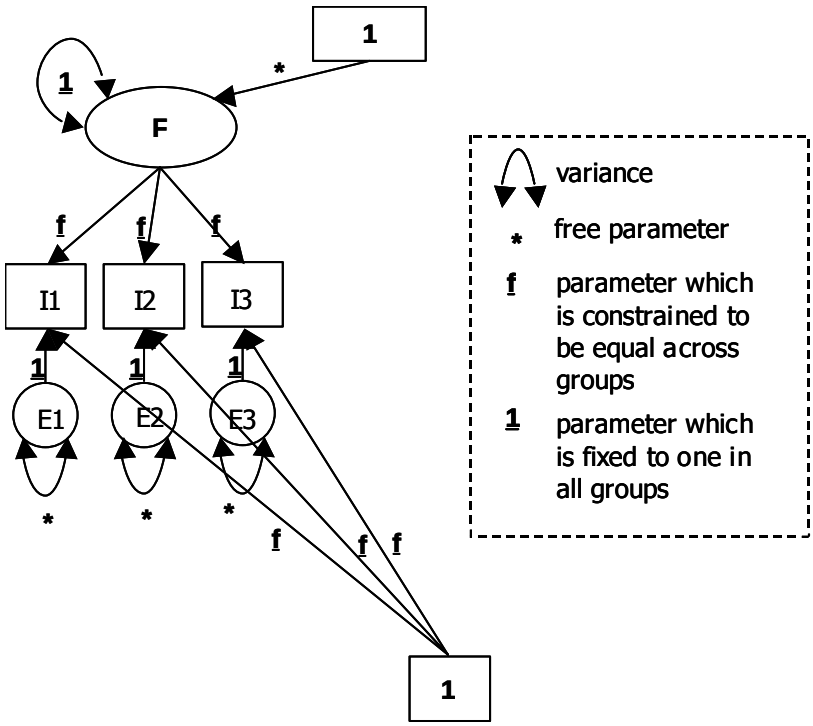


Figure 4.4.
The tau-invariance model (with 3 indicators)

Notes: F represents the factor; I1,I2, and I3 represent the 3 factor indicators; E1,E2, and E3 represent the unique variances of the 3 factor indicators; the mean value for F is fixed to zero in group 1 (for identification purposes), whereas it is freely estimated in group 2.

The robust ML procedure as implemented in the software MPlus (Muthén and Muthén, 1999) was used to estimate the model parameters. Testing whether the estimated factor mean in the second group differed significantly from the factor mean in the first group was straightforward. Because the factor mean in the first group was set equal to zero (for identification purposes), it was adequate to test whether the estimated factor mean in group two differs significantly from zero. A simple z-statistic (i.e. the estimated factor mean in the second group divided by its standard error) was used for this purpose. Provided that the estimated factor mean in group two is zero (i.e. the null-hypothesis holds), the z-statistic

follows a standard normal distribution, asymptotically. Since the true difference between factor means (if not zero) is known for simulated data, a conclusion can be made whether the statistical conclusion is correct or incorrect. Consequently, the correctness of the statistical conclusion concerning the difference in factor means across both groups, is known for every replication (of an experimental condition). The correctness of the statistical conclusion is flagged by a 'not correct' [0] / 'correct' [1] – indicator.

Next, the influence of the individual design parameters[43] on the *correctness* of statistical conclusions regarding the factor mean difference test across groups was assessed. Previous research (i.e. Kaplan and George, 1995) has shown that the effect of the difference between factor means at population level (i.e. factor 3) is dominant when compared to other effects. This is quite obvious as the probability of finding a significant difference between factor means in two independent samples is directly related to the size of the difference between factor means at population level. This effect is, however, not relevant for the research problem at hand. The main research question is to evaluate the extent to which measurement non-invariance conditions (i.e. factor 4 and 5) and certain side conditions (such as number of indicators, distribution of indicators and sample sizes) 'bias' factor mean comparisons across groups. Therefore, the factor mean difference at population level may be regarded as an 'extraneous factor'. Consequently, the effects of all other design parameters were assessed separately for various levels of the factor mean difference at population level.

The design parameters were indicated by means of binary variables (i.e. 0/1 variables). The following notation was used: Fi_Dj with i representing the number identifying the design factor and j indicating the number corresponding to the level within that factor. One level of each design factor was used as a 'reference' to quantify the effect of all other levels of that particular factor. As a consequence, k-1 binary variables are sufficient to represent all k levels of the design factor.

Two types of multivariate analysis techniques were used to assess the effects of the design parameters on the correctness of the statistical conclusion based on the difference test between factor means.

The first technique used was the *Classification And Regression Tree technique* by Breiman et al. (1984). The abbreviation C&RT is used in this chapter to refer to this technique. The results of a C&RT analysis are typically presented in a tree-based structure. The tree splits the whole sample containing all replications in two subsamples (i.e. 'binary splits') so that each subsample is maximally

[43] The design parameters are: the distribution (i.e. type of threshold model), the number of observations per group, the difference in factor means at population level, the degree of inequality of the non-invariant factor loading, the degree of inequality of the non-invariant indicator intercept (see Appendix 4.1.).

different from the other subsample when it comes to the percentage of correct statistical conclusions. Provided that the convergence criteria are not met, a subsequent (new) split is made for every subsample resulting from the previous split (see, for example, Appendix 4.5 and 4.6, first page). All design parameters (i.e. levels of a factor) which have not been used higher in the tree 'compete' with one another to split the current sample in two parts. A C&RTree analysis is in fact nothing more than a stepwise regression-type of analysis. Those design parameters that are most important in distinguishing groups of replications with a relatively high and low percentage of correct statistical conclusions appear in the upper level of the tree. Design parameters which are somewhat less important may pop up in the lower level of the tree (or may not lead to any sample split). The importance of the individual design parameters on which sample splits have been made, are reflected by the sequence in which these sample splits are made.[44] In addition to C&RT analyses, one could also perform CHAID (i.e. Chi-Square Automatic Interaction Detection) analyses.[45] Unlike the C&RT technique, CHAID does not make binary sample splits, but it splits at all factor levels which is selected as the 'splitting factor'.

[44] An alternative criterion would be the number of correctly classified cases. This (alternative) criterion is not used in this study.

[45] Such CHAID analyses have been conducted as well. The tree-based diagrams based on CHAID showed substantially more similarity across experimental parameter settings (e.g. for different levels of F3). The tree-based diagrams based on the C&RT analyses were considered to be more useful (and specific) than the tree-based diagrams based on CHAID analyses. For this reason it was decided to (only) include tree-based diagrams based on C&RT analysis in this chapter.

The second multivariate analysis technique that was used is *logistic regression analysis* (Hosmer and Lemeshow, 1989). The probability to be correct (i.e. p) is given by the logistic function:

$$p = Pr(Y_i = 1|x_i) = \frac{e^{z_i}}{1 + e^{z_i}}$$

where:

y_i is the outcome variable (e.g. a correct or incorrect statistical conclusion)
x_i is a vector of values for the i^{th} observation, and
$z_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i + \ldots + \beta_k X_k$.

The k design parameters X_1 to X_k may represent interval-scaled design parameters and/or binary design parameters. As explained before, all design effects in this study are represented by means of a series of binary parameters (i.e. $k-1$ binary variables for a design factor with k levels).

Using some elementary algebra, it follows from the expression of the logistic function that the ratio between the probability to be correct (p) and the probability to be incorrect ($1-p$) is expressed as:

$$\frac{p}{1-p} = \frac{Pr(Y_i = 1|x_i)}{Pr(Y_i = 0|x_i)} = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i + \ldots + \beta_k X_k} = e^{\beta_0} e^{\beta_1 X_1} \ldots e^{\beta_i X_i} \ldots e^{\beta_k X_k}$$

The ratio ($p/(1-p)$) is generally known as the odds ratio.

In a logistic regression, the (natural) logarithm of the odds ratio is used as the dependent variable:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i + \ldots + \beta_k X_k$$

The natural logarithm of the odds ratio is referred to as the logit. The regression coefficients (i.e. β_i's) indicate the expected change in the logit when the corresponding design parameter is increased by one unit, provided that all other design parameters are kept constant.

An easier interpretation of the effect of the design parameters is given by the coefficient e^{β_i} (i.e. a simple exponential function). The coefficient e^{β_i} can be interpreted as the expected change in the odds ratio when increasing the corresponding design parameter (i.e. X_i) by one unit (see Long, 1997, p. 80). The 'ceteris paribus' principle (i.e. all other design parameters are kept constant) is still assumed. In case the coefficients e^{β_i} are smaller than one, their

reciprocal values $1 / e^{\beta_i}$ may be considered to obtain only coefficients with a value greater than one. One should keep in mind that the latter (transformed) coefficients determine the reciprocal of the odds ratio and not the odds ratio. In other words, values higher than one for $1 / e^{\beta_i}$ imply a relatively higher chance to be incorrect.

In the logistic regression analyses presented in this study, the effects of all k design parameters were estimated simultaneously. Because of the simultaneous estimation of all design parameters, it may be expected that some of the design parameters, which did not appear in the C&RTree, were found to be statistically significant in the logistic regression model. The logistic regression model, in which all parameters are estimated simultaneously, provides thus a more detailed picture of all influential effects on the correctness of the statistical conclusion. In this study, the C&RTrees' are particularly useful because they clearly indicate the most important determinants of the correctness of the statistical conclusion based on the factor mean difference test across groups[46].

In a logistic regression model, it is possible to include interaction effects between the design parameters in the model. To include such interaction effects the right-hand side of the logistic regression equation should be extended with multiplicative terms, such as: $\beta_i X_i * \beta_j X_j$. Interaction effects between factor 4 (non-invariance of factor loadings) and factor 5 (non-invariance of indicator intercepts) were included in all logistic regression analyses.

Step2:
Robust and non-robust conditions

So far, the unit of analysis has been a replication of an experimental condition. As explained in Section 2, there were 50 replications per experimental condition. To assess the robustness of the experimental conditions against violations of the measurement invariance assumption across groups (i.e. step 2), aggregated data are needed. In particular, the data of all replications need to be aggregated for every experimental condition.

The analysis strategy is to use the total number of correct statistical conclusions of invariance conditions as a reference against which to evaluate the robustness

[46] It would also be possible to conduct a stepwise logistic regression analysis instead of a regression analysis in which all explanatory variables are estimated simultaneously. Due to the stepwise selection of explanatory variables in the equation, one may expect that the results obtained by means of a stepwise regression model would match somewhat more closely with the results obtained by means of C&RTrees.

of all (related[47]) non-invariance conditions. Based on the binomial distribution, a 99% confidence interval[48] is specified around the number of correct statistical conclusions of invariance conditions. If the number of correct statistical conclusions of a related non-invariance condition falls within this interval, the non-invariance condition is considered to be 'robust' against violations of the measurement invariance assumption across groups. Otherwise, it is not considered to be robust. Based on such an analysis, all non-invariance conditions are flagged with a 'not-robust' [0] / 'robust' [1]-indicator. In sum, the idea is to examine the decrease (or increase) in the number of correct statistical conclusions of non-invariance conditions using the number of correct statistical conclusions of invariance conditions as a benchmark (or reference). In other words, a test based on statistical inference is used. The tolerance region is determined by the 99% confidence interval based on the binomial distribution (i.e. the type-I error rate equals 1-0.99=0.01).

Similar to the first step C&RTree analyses and logistic regression models are used to determine the influence of the individual design parameters on the robustness of the experimental condition against violations of the measurement invariance principle across groups. As the unit of analysis is an experimental condition rather than a replication (within an experimental condition), there are only a limited (i.e. small) number of observations available for these analyses.

[47] Related non-invariance conditions are characterised by an identical factor mean difference between both populations (groups) and a non-invariant indicator having an unequal factor loading and/or indicator intercept across groups.

[48] The specification of a confidence interval (CI) is somewhat arbitrarily. Changing from a 99% CI to a 95% CI would not have a substantial impact on the decisions regarding robustness / non-robustness of the non-invariance condition. The mean difference* in the 'tolerance region' as specified by both confidence intervals is about 0.80 (standard deviation is about 0.60) for both 3- and 4- indicator conditions (*across all non-invariance conditions). This difference is very small (i.e. an average difference of [less than] 1 replication out of 50 replications per non-invariance condition). In three out of four cases the upper value within the tolerance region as specified by the 99% CI does not exceed 47 out of 50 replications (i.e. 47 is the value of the 3rd quartile).

4.3. Results

4.3.1. Correct and incorrect statistical conclusions

4.3.1.1. Descriptive results

The percentage of correct conclusions regarding the factor mean difference test varies around 66% across all simulated conditions, regardless of the number of indicators used in the factor model (i.e. 3 or 4 indicators). This is shown in Table 4.1. Table 4.1 also shows percentages of correct statistical conclusions tabulated for different levels of the factor mean difference at population level (i.e. the different levels for F3).

Table 4.1.
Percentage of CORRECT conclusions regarding the factor mean difference test

% of correct statistical conclusions	Factor mean in group 2				
	= -0.30	= -0.15	= 0.00	= 0.15	= 0.30
3 indicators Overall: 65.2%	62.5	32.9	55.5	79.4	95.5
4 indicators Overall: 67.3%	69.5 [+]	28.7 [-]	67.3 [+]	76.1 [-]	95.0

Note: A plus or minus sign between square brackets indicates the direction of significant increases [+] or decreases [-] in terms of the percentage of correct statistical conclusions (when comparing 4-indicator conditions versus 3-indicator conditions).

Table 4.1 shows that the percentages of correct statistical conclusions differ substantially across different levels for factor 3. It is not surprising that larger differences between factor means at population level result in a higher percentage of correct statistical conclusions (i.e. compare conditions F3=1 with F3=2 and F3=5 with F3=4). This finding is obvious, given the asymmetrical structure of the experimental design. The design is asymmetrical as a non-invariant indicator intercept is always larger in group 2. Conditions in which the non-invariant indicator intercept is smaller in group 2 (when compared to group 1) are not included in the simulation study. Consequently, there is an upward bias on the estimated factor mean in group 2 which is due to a higher indicator intercept in group 2. In positive discrepancy cases, the bias works in favour of a rejection of the hypothesis of equal factor means across groups. In negative discrepancy cases, the bias works in favour of a non-rejectance of the hypothesis of equal factor means across groups.

When mutually comparing conditions with 4 indicators for the factor versus conditions with 3 indicators, significantly different percentages of correct statistical conclusions were obtained. This is shown in Table 4.1 (see plus or minus signs indicated between square brackets). Taking into account the large sample sizes per cell (N=10800) in Table 4.1, it is not surprising that the

percentages differ significantly across 4- and 3- indicator conditions. A further inspection shows that none of both factor models (i.e. with 4 or 3 indicators) can be considered to be a 'winner'. The 4-indicator conditions report higher percentages of correct statistical conclusions for F3=1 and F3=3, whereas the 3-indicator conditions report higher percentages for F3=2 and F3=4.

A more detailed picture is provided in Tables 4.2 and 4.3. These tables show aggregated results for non-invariance conditions, partial invariance conditions (i.e. with a non-invariant factor loading or indicator intercept, but not both), and full invariance conditions (i.e. conditions with invariant factor loadings and invariant indicator intercepts) for both 3- and 4-indicator conditions.

Table 4.2.
Correct statistical conclusions (3-indicator cases)

NON-INVARIANCE CONDITIONS					
	Factor mean in group 2				
% of correct statistical conclusions	=-0.30	=-0.15	= 0.00	= 0.15	= 0.30
$\Delta\ \lambda_2$ = -0.20 (F4=1)	40.9	29.9	55.2	71.6	90.7
$\Delta\ \lambda_2$ = +0.20 (F4=3)	81.1	37.9	56.7	86.6	99.1
$\Delta\ int_2$ = 0.15 (F5=2)	71.8	25.8	73.7	79.3	96.2
$\Delta\ int_2$ = 0.30 (F5=3)	52.1	19.1	41.2	91.5	98.7
$\Delta\ int_2$ = 0.45 (F5=4)	39.4	35.0	19.2	96.9	98.9
C1: F4=1 & F5=2	46.3	16.4	71.9	69.0	92.0
C2: F4=1 & F5=3	23.2	22.3	41.9	86.6	97.1
C3: F4=1 & F5=4	22.3	45.8	21.2	94.1	97.9
C4: F4=3 & F5=2	92.6	37.0	76.4	88.0	99.6
C5: F4=3 & F5=3	79.4	20.3	42.4	95.1	100.0
C6: F4=3 & F5=4	62.7	25.2	18.3	98.8	99.6
(partial) INVARIANCE CONDITIONS					
	Factor mean in group 2				
% of correct statistical conclusions	=-0.30	=-0.15	= 0.00	= 0.15	= 0.30
$\Delta\ \lambda_2$= 0.00 (F4=2)*	63.4	31.0	54.6	80.1	96.6
$\Delta\ int_2$ = 0.00 (F5=1)*	86.8	51.7	87.7	50.0	88.0
C7: F4=2 & F5=2	76.4	24.0	72.9	80.8	97.1
C8: F4=2 & F5=3	53.6	14.7	39.2	92.8	99.0
C9: F4=2 & F5=4	33.1	34.0	18.1	97.9	99.2
C10: F4=1 & F5=1	71.6	35.1	85.7	36.8	75.8
C11: F4=3 & F5=1	97.9	68.9	89.4	64.6	97.1
FULL INVARIANCE CONDITION ('reference / control condition')					
	Factor mean in group 2				
% of correct statistical conclusions	=-0.30	=-0.15	= 0.00	= 0.15	= 0.30
C12: F4=2 & F5=1	91.0	51.2	88.1*	48.9	91.1

Notes: (1) The symbol 'Δ' refers to the difference in the value of the parameter between both groups;
(2) Ci with i=1,2,…,12 indicates the i^{th} combination of factor 4 and factor 5;
(3) *When only normally distributed indicators are used (F1=1) and the number of observations per group is equal or higher than 500 (F2=4 or F2=5) the percentage of correct statistical conclusions equals 96.0% (i.e. close to 95%, which is one minus the nominal type I-error rate).

Table 4.3.
Correct statistical conclusions (4-indicator cases)

NON-INVARIANCE CONDITIONS					
	Factor mean in group 2				
% of correct statistical conclusions	= -0.30	= -0.15	= 0.00	= 0.15	0.30
$\Delta\ \lambda_2$= -0.20 (F4=1)	45.1 [+]	21.0 [-]	69.3 [+]	65.0 [+]	89.1
$\Delta\ \lambda_2$= +0.20 (F4=3)	90.6 [+]	37.5	66.0 [+]	85.3	99.2
$\Delta\ int_2$ = 0.15 (F5=2)	76.9 [+]	30.5 [+]	80.7 [+]	72.9 [-]	94.5
$\Delta\ int_2$ = 0.30 (F5=3)	62.2 [+]	16.5	58.6 [+]	86.8 [-]	98.1
$\Delta\ int_2$ = 0.45 (F5=4)	50.3 [+]	15.8 [-]	39.7 [+]	93.6 [-]	99.0
C1: F4=1 & F5=2	51.9	17.6	80.1 [+]	59.4 [-]	86.9 [-]
C2: F4=1 & F5=3	33.2 [+]	12.1 [-]	61.1 [+]	79.2 [-]	95.8
C3: F4=1 & F5=4	21.9	19.8 [-]	46.2 [+]	87.0 [-]	97.2
C4: F4=3 & F5=2	95.4	45.0 [+]	82.7 [+]	83.9	99.4
C5: F4=3 & F5=3	89.3 [+]	25.2	55.0 [+]	92.8	99.8
C6: F4=3 & F5=4	78.7 [+]	11.7 [-]	36.0 [+]	98.1	99.9
(partial) INVARIANCE CONDITIONS					
	Factor mean in group 2				
% of correct statistical conclusions	= -0.30	= -0.15	= 0.00	= 0.15	0.30
$\Delta\ \lambda_2$= 0.00 (F4=2)*	72.6 [+]	27.5	66.4 [+]	77.9	94.7 [-]
$\Delta\ int_2$ = 0.00 (F5=1)*	88.4	51.9	90.0	51.1	88.3
C7: F4=2 & F5=2	83.3 [+]	28.9	79.2	75.2	97.1
C8: F4=2 & F5=3	64.1 [+]	12.1	59.7 [+]	88.4	98.9
C9: F4=2 & F5=4	50.4 [+]	15.9 [-]	36.9 [+]	95.7	99.8
C10: F4=1 & F5=1	73.3	34.6	89.9	34.2	76.3
C11: F4=3 & F5=1	89.1 [-]	68.2	90.4	66.6	97.7

Note: A plus or minus sign between square brackets indicates the direction of significant increases [+] or decreases in terms of the percentage of correct statistical conclusions (when comparing conditions with 4 indicators versus conditions with 3 indicators for the factor). A 99% confidence interval is used.

Table 4.3. (continued)
Correct statistical conclusions (4-indicator cases)

FULL INVARIANCE CONDITION ('reference / control condition')					
	Factor mean in group 2				
% of correct statistical conclusions	= -0.30	= -0.15	= 0.00	= 0.15	0.30
C12: F4=2 & F5=1	92.7**	53.0**	89.8*	52.4**	90.9**
SUMMARY STATISTICS					
Count of [.] (. = + or -)	13 /19	8 /19	14 /19	7 /19	2 /19
Count of [+]	12	2	14	1	0
Count of [-]	1	6	0	6	2

Notes:

(1) A plus or minus sign between square brackets indicates the direction of significant increases [+] or decreases in terms of the percentage of correct statistical conclusions (when comparing conditions with 4 indicators versus conditions with 3 indicators for the factor). A 99% confidence interval is used;
(2) *When only normally distributed indicators are used (F1=1) and the number of observations per group is equal or greater than 500 (F2=4 or F2=5) the percentage of correct statistical conclusions equals 95.0% (i.e. exactly one minus the nominal type I-error rate);
(3) **These percentages represent the (average) 'power' (i.e. one minus the type II-error rate) across side conditions (i.e. F1, F2).

Consider the full invariance condition in Tables 4.2 and 4.3. In the 'no difference' cases (i.e. F3=3), the percentage of correct statistical conclusions is about 88% in both tables. From statistical theory, 95 per cent correct statistical conclusions should be reported for these conditions. Ninety-five per cent is obtained by taking 100 (%) minus the nominal type I-error rate (5%) which is specified for testing the significance of the (estimated) cross-group difference between factor means. There are several reasons why the actual percentage (about 88%) differs from the expected percentage: (1) indicators may not be normally distributed (i.e. F1=1 or F1=2), and (2) sample sizes may be relatively small (e.g. N=400 or a smaller N). As indicated in the notes accompanying Tables 4.2 and 4.3, the expected rate of 95% correct statistical conclusions was nearly obtained in cases in which the indicators follow a standard normal distribution (i.e. F1=1) and sample sizes are sufficiently large (i.e. sample sizes are at least 500 per group). The percentages of correct statistical conclusions reported for positive discrepancy cases (i.e. F3=3 and F3=4) and negative discrepancy cases (i.e. F3=1 and F3=2) in the full invariance condition represent (average) 'power levels' when indicators do exhibit full measurement invariance across groups. The power of the test indicates the probability of detecting true factor mean differences at population level by means of the factor mean difference test.

The results with respect to the full invariance condition as reported in Tables 4.2 and 4.3 indicate that a factor mean difference of 0.15 may be considered to be 'small', whereas a factor mean difference of 0.30 may be considered to be 'large'. Why these labels are used is obvious from the percentage of correct statistical conclusions regarding the factor mean difference test. The percentage of correct statistical conclusions is rather 'small' for a factor mean difference of minus 0.15 (51.2 and 53.0% for 3- and 4-indicator cases, respectively), and a factor mean difference of (plus) 0.15 (48.9% and 52.4% for 3- and 4-indicator cases, respectively). In contrast, the percentage of correct statistical conclusions is very high for a factor mean difference of minus 0.30 (i.e. 91.0% and 92.7% for 3- and 4-indicator cases, respectively), and a factor mean difference of (plus) 0.30 (i.e. 91.1% and 90.9% in 3- and 4-indicator conditions, respectively).

In Table 4.3 a comparison was made between the percentages of correct statistical conclusions of 4- and 3-indicator conditions, respectively. A comparison was made with respect to every individual cell in the table. The summary with counts of significant differences in percentages is shown at the bottom of Table 4.3. Most significant differences were obtained for large negative discrepancy cases (i.e. F3=1) and 'no difference' cases (i.e. F3=3). In these cases, the percentage of correct statistical conclusions was consistently higher in 4-indicator conditions than in 3-indicator conditions. When differences in factor means are small (i.e. F3=2 and F3=4), a different picture emerges. In these cases, the percentage of correct statistical conclusions was consistently higher in 3- indicator conditions than in 4-indicator conditions. In sum, none of

the both factor models (with 3- or 4- indicators) outperformed the other in terms of a percentage of correct statistical conclusions which was consistently higher across all levels of factor 3.

4.3.1.2. Influence of design factors on the correctness of the factor mean difference test

As expected, the factor mean difference at population level (i.e. factor 3) turned out to be the most influential factor determining the correctness of the statistical conclusion regarding the factor mean difference test. This is shown in the first two C&RTrees (i.e. T1 and T2) presented in Appendix 4.5. C&RTrees T1 and T2 are based on 3- and 4- factor conditions, respectively. In both trees, factor 3 pops up as the first design factor to split the sample with all simulated replications in two subsamples (F3=2 versus F3<>2). Further down the trees, more sample splits are made using other levels of factor 3 as splitting variables. In addition, the first two logistic regression analyses shown in Appendix 4.4 indicated that the factor mean difference at population level was the strongest determinant of the correctness of the statistical conclusion regarding the factor mean difference test. The high values reported for e^B (B representing the unstandardised regression coefficient) support this finding. This finding is consistent with earlier results presented by Kaplan and George's (1995). For reasons explained in the analysis section, the effects of design parameters on the correctness of the statistical conclusion regarding the factor mean difference will be analysed separately for all levels of factor 3.

The results of the C&RTree analyses will be presented first. C&RTree analyses are particularly useful to assess the relative importance of the individual design parameters in terms of predicting the correctness of the factor mean difference test. Further on, the results of the logistic regression models will be presented. Logistic regression analysis are beneficial as they provide a more detailed picture of the individual effects (and the significance) of each design parameter on the correctness of the factor mean difference test.

C&RTree analyses

F3=1

In large negative discrepancy cases (i.e. F3=1), the non-invariant factor loading (i.e. factor 4) was selected as the first factor to split all (simulated) replications in two subsamples (A and B). This is shown in the C&RTrees T3 and T4 in Appendix 4.5. These C&RTrees represent 3- and 4-indicator conditions, respectively. This sample split indicated that a factor loading of 0.4 for the non-invariant indicator in group 2 (versus 0.6 in group 1) substantially lowered the probability of making correct conclusions regarding the factor mean difference

test. Further sample splits (in both subsamples) were based on the non-invariant indicator intercept (i.e. factor 5). The very first sample split distinguished between conditions with the largest (simulated) difference in the non-invariant indicator intercept (i.e. F5_D4) and conditions with a smaller (or zero) difference in the indicator intercept. Subsequent sample splits were made using other levels of factor 5 (e.g. F5_D3) as splitting variables. The larger the discrepancy in the non-invariant indicator intercept, the lower the probability of drawing the correct statistical conclusion based on the difference in factor means. This is logical as, in negative discrepancy cases, the direction of the difference in the non-invariant indicator intercept is opposite to the direction of the difference between factor means at population level. In sum, the results showed that, in large negative discrepancy cases, both a non-invariant factor loading and a non-invariant indicator intercept were factors that had a strong influence on the correctness of the statistical conclusion regarding the factor mean difference test.

F3=2

In small negative discrepancy cases (i.e. F3=2), the sample of simulated replications was first split according to differences in the non-invariant indicator intercept (i.e. factor 5). This is shown in the C&RTrees T5 and T6 in Appendix 4.5. In C&RTree T5, the first sample split was made using the third level of factor 5 (i.e. F5_D3) as the variable to split on. In C&RTree T6, the highest level of factor 5 (i.e. F5_D4) was used as the variable to split on. Further down in both trees, more splits were made using other levels of factor 5 as splitting variables. The implication is (once again) that a larger non-invariant indicator intercept has a strong negative impact on the percentage of correct statistical conclusions in negative discrepancy cases. Further inspection of C&RTrees T5 and T6 revealed that further sample splits were made using the degree of non-invariance of the factor loading as a variable to split on (e.g. F4_D1 and F4_D3). These findings support the conclusion that non-invariance conditions (i.e. factor 4 and factor 5) have a strong impact on the percentage of correct statistical conclusions.

F3=3

In the 'no difference' cases (i.e. F3=3), successive splits were made using various levels of factor 5 as splitting variables. This is shown in C&RTrees T7 and T8. The smaller the difference in the non-invariant indicator intercept, the higher the probability of drawing the right statistical conclusion with respect to the difference in factor means across populations. Non-invariant indicator intercepts enlarge the (estimated) difference between the factor means in both populations. As a consequence, the probability of rejecting the hypothesis of equal factor means at population level (i.e. the correct statistical conclusion here!) decreases.

F3=4

In small positive discrepancy cases (i.e. F3=4), the difference in the non-invariant indicator intercept was successively used as the design factor on which sample splits were made (see C&RTrees T9 and T10 in Appendix 4.5). Larger differences in the non-invariant indicator intercept enlarge the (estimated) difference between factor means at population level. The conclusion is that the probability of drawing the correct statistical conclusion (namely, a difference between the factor means in both populations) increased because of the bias introduced by the non-invariant indicator intercept.

F3=5

In large positive discrepancy cases (i.e. F3=5), the percentage of correct statistical conclusions regarding the factor mean difference test turned out to be very high (i.e. around 95% in both 3- and 4-indicator conditions). The C&RTrees T11 and T12 showed that the sample was first split using the first level of factor 4 (i.e. a non-invariant factor loading of 0.4 in group 2 versus a factor loading of 0.6 in group 1) as the variable to split on. The small difference in the percentage of correct statistical conclusions reported for both subsamples (as well as the size of the calculated measure of improvement) showed that this sample split was only marginally relevant. In conclusion, the difference in factor means at population level (+0.30) was large enough to ensure a very high proportion of correct statistical conclusions (i.e. close to 95%). Obviously, the bias caused by a non-invariant indicator intercept (as present in many simulated conditions) was, to a large extent, responsible for this high percentage in correct statistical conclusions.

Overall

Overall, the C&RT analyses showed that measurement non-invariance (as operationalised in this study) exerted a strong influence on the percentage of correct statistical conclusions regarding the factor mean difference test. A non-invariant indicator intercept, in particular, had a strong effect on the correctness of the statistical conclusion regarding the factor mean difference test. This effect could either be positive (in positive discrepancy cases [i.e. F3=4 and F3=5]) or negative (in negative discrepancy cases [i.e. F3=1 and F3=2] and the 'no difference' cases [F3=3]). The stronger impact of differences in the indicator intercept (as opposed to differences in the factor loading) was in line with the author's expectations for reasons explained earlier in this chapter (see Section 4.2.1.1, description of 'non-invariance conditions').

Logistic regression analyses

Tables 4.4 and 4.5 (see next pages) present five different logistic regression models for the 3- and 4-indicator conditions, respectively. One logistic regression model is presented for each level of factor 3. For all models presented, the percentage of correct classifications is high (i.e. 75.6%, 71.6%, 76.0%, 82.9%, 95.4% in Table 4.4; 78.4%, 76.2%, 74.8%, 80.1%, 94.8% in Table 4.5). The high percentage of correct classifications may be interpreted as an indication that one may have confidence in the interpretation of the regression coefficients as presented in these regression models.

Table 4.4.
Logistic regression models predicting the CORRECTNESS of the statistical difference test between factor means at population level (one-factor model with 3 indicators)

3 indicators		Factor mean in group 2														
		= -0.30 -2LL=10472.12 N=10800 CCR: 75.6 % RCDS: 62.5 %			= -0.15 -2LL=12344.37 N=10800 CCR: 71.6 % RCDS: 32.9 %			= 0.00 -2LL=10970.06 N=10800 CCR: 76.0 % RCDS : 55.5 %			= 0.15 -2LL=7827.01 N=10800 CCR: 82.9 % RCDS: 79.4 %			= 0.30 -2LL=2816.08 N=10800 CCR: 95.4 % RCDS: 95.5 %		
BINARY VARIABLE	SETTING	P value	B	e^B	P value	B	e^B	P value	B	e^B	P value	B	e^B	P value	B	e^B
Constant	Constant = 1	.00	3.24	N.R.	.00	.46	N.R.	.00	1.31	N.R.	.00	1.25	N.R.	.00	4.84	N.R.
Reference F1	Standard normal distribution	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
F1_D2	Uni-modal distr.	.38	-	-	.00	.17	1.18	.14	-	-	.04	-.15	.86	.11	-	-
F1_D3	Bi-modal distr.	.68	-	-	.31	-	-	.86	-	-	.48	-	-	.45	-	-
F2_D1	N=200/200	.00	-1.51	.22	.00	-.85	.43	.00	1.43	4.16	.00	-2.30	.10	.00	-3.43	.03
F2_D2	N=300/300	.00	-1.02	.36	.00	-.62	.54	.00	1.03	2.79	.00	-1.60	.20	.00	-2.42	.09
F2_D3	N=400/400	.00	-.59	.55	.00	-.42	.66	.00	.69	1.99	.00	-1.04	.35	.00	-1.61	.20
F2_D4	N=500/500	.00	-.39	.68	.00	-.25	.78	.00	.30	1.35	.00	-.65	.52	.00	-1.29	.28
Reference F2	N=750/750	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
F2_D6	N=200/400	.00	-1.23	.29	.00	-.76	.47	.00	1.12	3.07	.00	-1.78	.17	.00	-3.00	.05
F4_D1	λ_2 (G2) = 0.4	.00	-1.44	.24	.00	-.68	.51	.12	-	-	.00	-.57	.57	.00	-1.33	.27
Reference F4#	λ_2 (G2) = 0.6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
F4_D3	λ_2 (G2) = 0.8	.00	1.54	4.66	.00	.76	2.14	.36	-	-	.00	.72	2.06	.00	1.24	3.45
Reference F5#	Int_2 (G2) = .00	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
F5_D2	Int_2 (G2) = .15	.00	-1.17	.31	.00	-1.23	.29	.00	-1.05	.35	.00	1.65	5.20	.00	1.24	3.45
F5_D3	Int_2 (G2) = .30	.00	-2.27	.10	.00	-1.84	.16	.00	-2.56	.08	.00	2.83	16.87	.00	2.33	10.31
F5_D4	Int_2 (G2) = .45	.00	-3.17	.04	.00	-.73	.48	.00	-3.68	.03	.00	4.13	62.28	.00	2.59	13.31
F4_D1*F5_D2	Interaction effect	.83	-	-	.20	-	-	.34	-	-	.41	-	-	.45	-	-
F4_D1*F5_D3	idem	.91	-	-	.00	1.20	3.32	.05	.34	1.40	.43	-	-	.58	-	-
F4_D1*F5_D4	idem	.00	.87	2.38	.00	1.18	3.26	.02	.43	1.53	.08	-.51	.60	.52	-	-
F4_D3*F5_D2	idem	.61	-	-	.37	-	-	.75	-	-	.44	-	-	.25	-	-
F4_D3*F5_D3	idem	.35	-	-	.02	-.36	.70	.98	-	-	.19	-	-	.41	-	-
F4_D3*F5_D4	idem	.39	-	-	.00	-1.19	.30	.53	-	-	.65	-	-	.32	-	-

Notes: (1) Negative coefficients are underlined (or $e^B <= 0$); non-significant coefficients are not printed (except for the constant term); N.R. means 'not relevant';
(2) #Cross-group invariance condition (with respect to this measurement parameter);
(3) e^B is the impact on the odds ratio;
(4) -2LL means 'minus 2 times loglikelihood'; CCR means 'correct classification rate'; RCDS means 'rate of correct [statistical] decisions in sample' .

Table 4.5.
Logistic regression models predicting the CORRECTNESS of the statistical difference test between factor means at population level (one-factor model with 4 indicators)

4 indicators		Factor mean in group 2														
		= -0.30 -2LL=9504.62 N=10800 CCR: 78.4 % RCDS: 69.5 %			= -0.15 -2LL=11132.70 N=10800 CCR: 76.2 % RCDS: 28.7 %			= 0.00 -2LL=11296.120 N=10800 CCR: 74.8 % RCDS: 67.3 %			= 0.15 -2LL=8942.97 N=10800 CCR: 80.1 % RCDS: 76.1 %			= 0.30 -2LL=3044.35 N=10800 CCR: 94.8 % RCDS: 95.0 %		
BINARY VARIABLE	SETTING	P value	B	e^B	P value	B	e^B	P value	B	e^B	P value	B	e^B	P value	B	e^B
Constant	Constant = 1	.00	3.65	N.R.	.00	.61	N.R.	.00	1.61	N.R.	.12	-.10	N.R.	.00	4.83	N.R.
Reference F1	Standard normal distribution	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
F1_D2	Uni-modal distr.	.69	-	-	.02	.13	1.14	.73	-	-	.12	-	-	.23	-	-
F1_D3	Bi-modal distr.	.71	-	-	.53	-		.05	-.11	.90	.28	-	-	.09	-.20	.82
F2_D1	N=200/200	.00	-1.83	.16	.00	-1.09	.33	.00	1.29	3.63	.00	-2.40	.09	.00	-3.50	.03
F2_D2	N=300/300	.00	-1.26	.28	.00	-.60	.55	.00	.77	2.16	.00	-1.76	.17	.00	-2.41	.09
F2_D3	N=400/400	.00	-.85	.43	.00	-.46	.63	.00	.56	1.75	.00	-1.20	.30	.00	-1.62	.20
F2_D4	N=500/500	.00	-.51	.60	.00	-.30	.74	.00	.32	1.38	.00	-.78	.46	.00	-1.17	.31
Reference F2	N=750/750	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
F2_D6	N=200/400	.00	-1.44	.24	.00	-.82	.44	.00	1.07	2.91	.00	-2.03	.13	.00	-2.93	.05
F4_D1	λ_2 (G2) = 0.4	.00	-1.59	.20	.00	-.78	.46	.93	-	-	.00	-.87	.42	.00	-1.27	.28
Reference F4#	λ_2 (G2) = 0.6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
F4_D3	λ_2 (G2) = 0.8	.00	2.16	8.67	.00	.66	1.94	.63	-	-	.00	.67	1.96	.00	1.50	4.46
Reference F5#	Int_2 (G2) = .00	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
F5_D2	Int_2 (G2) = .15	.00	-.96	.38	.00	-1.05	.35	.00	-.85	.43	.00	1.14	3.14	.00	1.27	3.57
F5_D3	Int_2 (G2) = .30	.00	-2.05	.13	.00	-2.15	.12	.00	-1.84	.16	.00	2.13	8.46	.00	2.26	9.59
F5_D4	Int_2 (G2) = .45	.00	-2.66	.07	.00	-1.83	.16	.00	-2.81	.06	.00	3.23	25.24	.00	3.88	48.46
F4_D1*F5_D2	Interaction effect	.78	-	-	.42	-	-	.82	-	-	.72	-	-	.10	-.45	.64
F4_D1*F5_D3	idem	.28	-	-	.00	.78	2.19	.78	-	-	.46	-	-	.75	-	-
F4_D1*F5_D4	idem	.30	-	-	.00	1.05	2.87	.04	.39	1.48	.10	-.36	.69	.08	-1.30	.27
F4_D3*F5_D2	idem	.10	-.69	.50	.69	-	-	.44	-	-	.58	-	-	.72	-	-
F4_D3*F5_D3	idem	.18	-	-	.13	-	-	.14	-	-	.49	-	-	.87	-	-
F4_D3*F5_D4	idem	.05	-.78	.46	.00	-1.03	.36	.54	-	-	.54	-	-	.49	-	-

Notes: (1) Negative coefficients are underlined (or e^B <= 0); non-significant coefficients are not printed (except for the constant term); N.R. means 'not relevant';
(2) #Cross-group invariance condition (with respect to this measurement parameter);
(3) e^B is the impact on the odds ratio;
(4) -2LL means 'minus 2 times loglikelihood'; CC means 'correct classification rate'; RCDS means 'rate of correct [statistical] decisions in sample'.

The conclusions based on the logistic regression analyses will be presented in the next paragraphs. These paragraphs are organised as follows: the influence of the different design factors on the correctness of the factor mean difference test will be discussed for every design factor separately. Results which relate to 3-indicator conditions will be discussed before any results related to 4-indicator conditions will be presented.

Design factor: type of distribution

The five logistic regression models which relate to 3-indicator conditions (see Table 4.4) shows that the type of distribution (i.e. F1) has (almost) no effect on the correctness of outcome of the factor mean difference test. The standard normal distribution (i.e. F1=1) was chosen as a reference distribution. Significant effects were obtained only in model 2 (F3=2) and model 4 (F3=4). These significant effects showed that when compared to the normal distribution, the uni-modal, left-skewed distribution may lead to an increase (F3=2) or a decrease (F3=4) in terms of the percentage of correct statistical conclusions based on the factor mean difference test. Inspection of the e^B coefficients (or $1/e^B$ if $e^B < 1$) showed that the effects of the type of distribution were rather small (e.g. $e^B = 1.18$ or $e^B = 0.86$). Values for e^B that did not differ much from one[49], indicated effects which were only marginally relevant. The corresponding logistic regression models presented in Table 4.5 (i.e. for 4-indicator conditions) shows that for both the uni-modal, left-skewed distribution and the bi-modal distribution, some effects were significant. Inspection of the e^B coefficients (or $1/e^B$ if $e^B < 1$) showed that these effects were also marginally relevant. The conclusion is that in all k-indicator conditions ($k=3,4$), the type of distribution of the indicators had only a minor effect on the correctness of the statistical conclusion based on the factor mean difference test. Obviously, the type of distribution could have been a more important factor when conditions with much smaller sample sizes (e.g. less than 100 observations per group), would have been included in the simulation experiment.

Design factor: sample size per group

Next, the effects of different sample sizes were examined. Table 4.4 shows that all effects related to sample size were negative and significant in all logistic regression models except for the third one (i.e. F3=3). The negative sign of the effect was the consequence of the fact that the largest sample size (i.e. 750 observations per group) was chosen as the reference condition.

When looking at positive (F3=4, F3=5) and negative discrepancy (i.e. F3=1, F3=2) cases in Table 4.4, it is clear that the effect gets more and more negative the smaller the sample size per group. This is logical as statistical theory tells us

[49] A multiplicative term equal to one (i.e. 1) has no effect in the (multiplicative) model predicting the odds ratio: $1 / (1 - p) = e^{\beta_0} e^{\beta_1 X_1} \ldots e^{\beta_i X_i} \ldots e^{\beta_k X_k}$

that 'power' (i.e. the chance of finding significant differences between factor means) is indirectly related to sample size. When inspecting the size of the e^B coefficients (or $1/e^B$ if $e^B < 1$), Table 4.4 reveals that the smaller sample sizes in particular (i.e. N=200 and N=300 per group) had a strong negative impact on the correctness of the factor mean difference test. The e^B coefficients which relate to conditions with small sample sizes (N=200 and N=300 per group) were much smaller (i.e. closer to zero) than the e^B coefficients which relate to conditions with larger sample sizes (e.g. N=400 or N=500 per group). As the e^B coefficients related to sample size effects were all smaller than 1, its reciprocal value (i.e. $1/e^B$) may be used as a basis for comparing effects. The $1/e^B$ coefficients which relate to conditions with smaller sample sizes (N=200 and N=300 per group) were much larger than the $1/e^B$ coefficients which relate to conditions with larger sample sizes (e.g. N=400 or N=500 per group). Larger values for $1/e^B$ imply a higher probability of an incorrect statistical conclusion based on the factor mean difference test. The reader should keep in mind that $1/e^B$ coefficients determine the reciprocal of the odds ratio (i.e. $[1-p]/p$), not the odds ratio itself (i.e. $p/[1-p]$).

When factor means do not differ at population level (i.e. F3=3), the reduced statistical power associated with smaller sample sizes favours the correct statistical conclusion, namely a failure to reject the hypothesis of equal factor means across populations. Evidence for this finding is found in Table 4.4. Notice that, in the third logistic regression model, the e^B coefficients became increasingly positive when sample size per group decreased. Conditions in which the sample size per group is very small (e.g. N=200 or N=300) have e^B coefficients which largely exceed one (e^B=4.16 and 2.79). Conditions in which the sample size per group is larger (e.g. N=400 and N=500) have e^B coefficients which are somewhat higher than one (e^B=1.99 and 1.35). Table 4.5 shows that exactly the same conclusions can be drawn for 4-indicator conditions. To conclude, the sample size had a strong effect on the correctness of the outcome of the factor mean difference test.

Design factors: non-invariant factor loadings and indicator intercepts

For the factors causing non-invariance conditions (i.e. factor loading [F4] and indicator intercept [F5]), the invariance setting was used as a 'reference'. The next paragraphs will discuss the results related to negative discrepancy cases (F3=1 or F3=2) and positive discrepancy cases (F3=4 or F3=5). In later paragraphs, the results related to 'no difference' cases (i.e. F3=3) will be presented.

Notice that in Table 4.4 the effect related to the first level of factor 4 was strongly negative in all regression models dealing with negative and positive discrepancy cases. This means that, in negative and positive discrepancy cases, a factor loading of 0.4 for the non-invariant indicator in the second group

(versus 0.6 in the first group) reduced the probability of obtaining a correct statistical conclusion based on the factor mean difference test. Table 4.4 also shows that, in negative discrepancy cases, a (positive) difference in the non-invariant indicator intercept (i.e. F5=2,3,4) decreased the probability of detecting a factor mean difference at population level (i.e. F3=1 or F3=2). In positive discrepancy cases (i.e. F3=4 or F3=5), the effect goes in the opposite direction (i.e. an increase rather than a decrease in the probability of detecting a factor mean difference at population level). Analogous results were obtained on the basis of C&RT analyses. Table 4.4 further reveals that a particular interaction effect, namely: F4=1 combined with F5=4, was found to be significant in most regression models. The sign of this interaction effect was opposite to the sign of the effects related to the effect of the different levels of factor 5 (i.e. the difference in the non-invariant indicator intercept). This finding implies that a smaller factor loading in group 2 can (partially) compensate for the upward effect on the estimated factor mean in group 2 which was caused by a (positive) difference in the non-invariant indicator intercept. Hence, this interaction effect was positive in negative discrepancy cases and negative in positive discrepancy cases. Recall that this interaction effect was also chosen as a splitting variable in some parts of the C&RTrees (see Appendix 4.5: T4, T5, and T10). Notice that, in regression model 2 of Table 4.4 (i.e. F3=2), the absolute size of the regression coefficient, which relates to the largest difference in the non-invariant indicator intercept (i.e. F5=4), was smaller than the absolute size of the regression coefficients reported for smaller differences in the indicator intercept (e.g. F5=3). The reason for this lies in the significance of additional interaction effects (F4=3 and F5=3; F4=3 and F5=4). These interaction effects were additional to the (main) effects of the non-invariant factor loading and the non-invariant indicator intercept. As shown in Table 4.4, these interaction effects became more negative when the interaction effect involved larger differences in the non-invariant indicator intercept (e.g. compare F5=4 with F5=3).

From Table 4.4 it is clear that, in the 'no difference' cases (i.e. F3=3), a difference in the non-invariant factor loading had no significant effect on the correctness of the outcome of the factor mean difference test. In contrast, a (positive) difference in the non-invariant indicator intercept did have a strong (negative) effect on the probability of drawing a correct statistical conclusion based on the factor mean difference test. In addition, there was also a positive effect associated with the interaction effect: F4=1 and F5=4.

A comparison between Tables 4.4 and 4.5 shows that all earlier conclusions with respect to the effects of non-invariant factor loadings and non-invariant indicator in 3-indicator conditions were also valid for 4-indicator conditions.

A closer inspection of the e^B coefficients (or, alternatively, $1/e^B$) may be helpful to determine the relative impact of the design factors on the correctness of the

outcome of the factor mean difference test. Without any doubt, a conclusion was made that factor 5 (i.e. a non-invariant indicator intercept) was -by far- the most influential design factor. As far as factor 5 is concerned, e^B coefficients were either close to zero (e.g. between 0.03 and 0.48 in negative discrepancy cases and 'no difference' cases) or largely in excess of one in positive discrepancy cases (i.e. typically between 3 and 63). Factor 4 (i.e. a non-invariant factor loading) and factor 2 (i.e. sample size per group) occupied a second position. As far as factor 4 is concerned, e^B coefficients ranged between 0.20 and 0.57, and between 2 and 9. As far as factor 2 is concerned, e^B coefficients ranged between 0.03 and 0.78, and between 1 and 4. Factor 1 (i.e. the type of distribution of the indicators) was, at best, only marginally relevant (i.e. almost no significant effects; e^B coefficients range between 0.82 and 1.14 when the effects were significant).

4.3.1.3. Conclusions

The C&RTree analyses and logistic regression analyses have shown that violations of the measurement invariance assumption across groups have a strong impact on the correctness of the outcome of the factor mean difference test. A difference in the non-invariant indicator intercept as large as (about) one tenth of the total length of the scale (a difference of 0.45 on a 5-point scale), or even smaller, has a strong impact on the percentage of correct statistical conclusions based on the factor mean difference test. The same is true for a difference in the non-invariant factor loading as large as 0.2 (factor loading in the reference group is 0.6). These findings are generally consistent across 3- and 4-indicator conditions. The effects of sample size (per group) were also found to be strong. Furthermore, the analyses have also shown that the type of distribution of the indicators has almost no impact on the probability of drawing correct statistical conclusions based on the factor mean difference test (at least when sample sizes are at least 200 observations per group). This is an indication that a treatment of the indicators, as if they were metric (even if they are in fact ordinal), is an analysis strategy that may work with 5-point Likert type of scales (even when the distribution is uni-modal, left-skew or symmetric, bi-modal). All conclusions are summarised in Table 4.6. The rank orders reported in Table 4.6 indicate the relative importance of factors 1, 2, 4, and 5 for all levels of factor 3. The rank orders were assigned on the basis of the size of the e^B coefficient (or $1/e^B$ if $e^B < 1$).

Table 4.6.
Overview of the determinants of the CORRECTNESS of the statistical difference test between factor means

Type of the effect determining the % of correct statistical test outcomes	Factor mean in group 2				
	= -0.30	= -0.15	= 0.00	= +0.15	= +0.30
F1: distribution of indicators	No effect	Uni-modal, left-skewed => + but marginal effect! [3]	No effect	No effect	No . effect
F2: sample size per group	+ [3]	+ [2]	- [2]	+ [2]	+ [2]
F4: non-invariant factor loading	+ [2]	+ [2]	No effect	+ [3]	+ [2]
F5: non-invariant indicator intercept	- [1]	- [1]	- [1]	+ [1]	+ [1]
Interaction effect(s) between F4 and F5.	Counter-balancing effect (F4- & F5+ => %C+) [4]	Same counter-balancing effect as in the first column [2]	Same counter-balancing effect as in the first column [3]	Counter-balancing effect (F4- & F5+ => %C-) [4]	Same counter-balancing effect as in the fourth column [4]

Explanation:

(1) Consider the cell [Interaction effect(s) between F4 and F5, F3=1] in this table: The notation *F4- & F5+ => %C+* denotes that *smaller* values for F4 (non-invariant factor loading is smaller in group 2 when compared to group 1) combined with *larger* values for F5 (i.e. non-invariant indicator intercept is larger in group 2 when compared to group 1) lead to an increase of the percentage of conditions in which a correct conclusion is made based on the factor means difference test;

(2) Rank orders are indicated between square brackets (1=strongest effect)

4.3.2. Robust and non-robust conditions

Robustness is only an interesting concept as far as non-invariance conditions (i.e. including partial invariance conditions) are concerned. The criteria for robustness were explained earlier on in the analysis section. A 99% confidence interval (based on the binomial distribution) is specified around the number of correct statistical conclusions of the (full) invariance condition which 'corresponds' with a specific non-invariance condition. If the number of correct statistical conclusions reported for the non-invariance condition falls within this confidence interval, the non-invariance condition is considered to be 'robust'. Otherwise, it is considered to be 'not robust'.

4.3.2.1. Descriptive results

Of all non-invariance conditions, (only) about 35% was found to be robust. This conclusion applies to both the 3- and 4-indicator conditions. Table 4.7 shows the percentage of robust cases for each level of factor 3. The percentage of robust cases was relatively high in large positive discrepancy cases (i.e. F3=5). In these cases, the percentage ranges between 65 and 75%. In all other cases, the percentage of robust cases was much smaller. When comparing the percentage of robust cases across 3- and 4-indicator conditions, no significant differences were found. Even though these differences were not significant, it is possible that actual differences may favour one of both measurement models (i.e. with either 3 or 4 indicators). The factor model with three indicators performed somewhat better in small negative discrepancy cases (i.e. F3=2), and in large positive discrepancy cases (i.e. F3=5). In all other cases, the factor model with 4 indicators performed slightly better. One may, therefore, conclude that (also with respect to robustness) none of both factor models outperformed the other model. Detailed tables such as Tables 4.3 and 4.4 are not provided as a supplement to Table 4.7. Such tables would not provide very useful information, taking into account the limited sample size on which the assessment of robustness of non-invariance conditions is based (the number of non-invariance conditions equaled 198 per level of factor 3).

Table 4.7.
Percentage of ROBUST non-invariance conditions

% of robust cases	Factor mean in group 2				
	= -0.30	= -0.15	= 0.00	= +0.15	= +0.30
K=3 indicators Overall: 35.5%	19.7	39.4	30.3	15.2	72.7
K=4 indicators Overall: 34.9%	28.3	26.3	34.9	18.7	66.2

Note: The percentage of robust non-invariance conditions is not significantly different across conditions with 3 and 4 indicators.

4.3.2.2. Influence of the design factors on the robustness of the factor mean difference test (against violations of the measurement invariance principle across groups)

C&RTree analyses

When all design factors were used in a C&RTree analysis, factor 3 popped up as the first factor to split on. This is shown in C&RTree T1 and T2 in Appendix 4.6. Large positive discrepancy conditions (i.e. F3=5) were separated from all other conditions (i.e. F3<>5) in the first sample split. Consistent with the results presented in Table 4.7, the percentage of robust non-invariance conditions was relatively large in conditions representing large positive discrepancy cases. Further down in both C&RTrees (T1 and T2), subsamples were formed based on the degree of non-invariance of the indicator intercept (i.e. factor 5). Larger differences in the non-invariant indicator intercept decreased the probability that the non-invariance condition was robust. The last two logistic regression models shown in Appendix 4.4 revealed that factor 3 is a very important determinant of the robustness of the non-invariance condition. In the next paragraphs C&RTrees will be presented for each level of factor 3.

F3=1

First of all, the focus is on large negative discrepancy cases (i.e. F3=1). The C&RTrees T3 and T4 in Appendix 4.6 show that the first important sample split was made using the third level of factor 4 (i.e. a non-invariant factor loading equal to 0.8 in group 2 [versus 0.6 in group 1]) as the variable to split the sample on. In C&RTree T3 the very first sample split was made using an interaction effect as the splitting variable (i.e. interaction effect: F4=3 and F5=2). This sample split seemed relatively unimportant because of the limited number of observations in the right branch of the tree (N=18). In C&RTree T4, the very first sample split was made using the third level of factor 4 (i.e. F4=3)

as the splitting variable. In the same tree, further sample splits were made using various degrees of non-invariance of the indicator intercept as splitting variables (i.e. splitting variable: F5=4 for the branch F4=3, and splitting variable: F5=2 for the branch F4<>3). C&RTree T4 clearly shows that a large non-invariant factor loading combined with a large non-invariant indicator intercept may lead to a very small percentage of robust cases (i.e. 16.7%). C&RTree T3 shows different sample splits, but they all seem to be relatively unimportant as indicated by the small score obtained for the measure of improvement.

F3=2

In small negative discrepancy cases (i.e. F3=2), the first sample split distinguished between conditions with very small sample sizes (N=200 per group) and all other conditions. This is shown in C&RTree T5 and T6. Small sample sizes seem to have a positive effect on the robustness of the non-invariance condition. Further down the tree, the sample was split using a couple of interaction effects between the non-invariant measurement parameters as splitting variables (i.e. the interaction effects: F4=1 & F5=4, and F4=3 & F5=2 in 3-indicator conditions, and the interaction effect: F4=3 & F5=2 in 4-indicator conditions). Apparently, a small non-invariant factor loading (in group 2) can partially compensate for the decrease in robustness due to a large non-invariant indicator intercept. Still further down the tree, most sample splits were made using various levels of the non-invariant indicator intercept or the non-invariant factor loading as splitting variables.

F3=3

In 'no difference' cases, successive sample splits were made using various degrees of non-invariance of the indicator intercept as splitting variables. This is shown in C&RTrees T7 and T8. The higher the difference in the non-invariant indicator intercepts, the smaller the probability that the non-invariance condition is robust against violations of the measurement invariance principle (across groups).

F3=4

In small positive discrepancy cases (i.e. F3=4), various levels of non-invariance of the indicator intercept were successively chosen as splitting variables. This is shown in C&RTrees T9 and T10. The higher the non-invariance of the indicator intercepts, the smaller the probability that the non-invariance condition is robust against violations of the measurement invariance principle (across groups).

F3=5

In large positive discrepancy conditions (i.e. F3=5), the first couple of sample splits were made using different sample sizes per group (i.e. F2) as splitting variables. This is shown in C&RTrees T11 and T12. In conditions with small sample sizes, a smaller percentage of robust non-invariance conditions was obtained. Further sample splits were made using the degree of non-invariance

of the factor loading (i.e. F4) as the splitting variable. A substantially smaller percentage of robust non-invariance conditions was reported in conditions with a non-invariant factor loading equal to 0.4 in the second group (whereas the corresponding factor loading was 0.6 in group one).

Logistic regression analyses

Tables 4.8 and 4.9 (see next pages) show the logistic regression models predicting the robustness of the factor mean difference test against violations of the measurement invariance principle for 3- and 4-indicator conditions, respectively.

Table 4.8.

Logistic regression models predicting the ROBUSTNESS of the statistical difference test (0/1) between factor means at population level against violations of the cross-group invariance assumption (one-factor model with 3 indicators)

3 indicators		Factor mean in group 2														
		= -0.30 -2LL=78.60 N=198 CCR: 91.4 % RRCS: 19.7 %			= -0.15 -2LL=138.54 N=198 CCR: 82.6 % RRCS: 39.4 %			= 0.00 -2LL=58.12 N=198 CCR: 94.4 % RRCS: 30.3 %			= 0.15 -2LL=82.64 N=198 CCR: 87.9 % RRCS: 15.2 %			= 0.30 -2LL=95.35 N=198 CCR: 90.9 % RRCS: 72.7 %		
BINARY VARIABLE	SETTING	P Value	B	e^B	P value	B	e^B	P value	B	e^B	P value	B	e^B	P value	B	e^B
Constant	Constant = 1	.99	1.35	N.R.	.47	-.97	N.R.	.99	1.55	N.R.	.99	-.66	N.R.	.92	.14	N.R.
Reference F1	Standard normal distribution	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
F1_D2	Uni-modal distr.	.73	-	-	.13	-	-	.06	1.60	4.93	.34	-	-	.53	-	-
F1_D3	Bi-modal distr.	.73	-	-	.44	-	-	.13	-	-	.33	-	-	.21	-	-
F2_D1	N=200/200	.36	-	-	.00	6.12	454.12	.01	3.20	24.60	.64	-	-	.00	-3.52	.03
F2_D2	N=300/300	1.00	-	-	.00	2.78	16.06	.01	3.20	24.60	.36	-	-	.05	2.77	15.94
F2_D3	N=400/400	.62	-	-	.00	3.61	37.12	.83	-	-	1.0	-	-	.05	2.77	15.94
F2_D4	N=500/500	.62	-	-	.11	-	-	.33	-	-	.64	-	-	.02	4.08	59.22
Reference F2	N=750/750	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
F2_D6	N=200/400	.14	-	-	.00	3.20	24.54	.19	-	-	.36	-	-	.11	-	-
F4_D1	λ_2 (G2) = 0.4	.97	-	-	.23	-	-	.99	-	-	.99	-	-	.01	-5.16	.01
Reference F4#	λ_2 (G2) = 0.6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
F4_D3	λ_2 (G2) = 0.8	1.00	-	-	.04	-1.81	.16	1.0	-	-	1.0	-	-	.57	-	-
Reference F5#	Int_2 (G2) = .00	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
F5_D2	Int_2 (G2) = .15	.97	-	-	.02	-2.99	0.05	.99	-	-	.99	-	-	.12	-	-
F5_D3	Int_2 (G2) = .30	.91	-	-	.55	-	-	.91	-	-	.92	-	-	.97	-	-
F5_D4	Int_2 (G2) = .45	.84	-	-	.39	-	-	.86	-	-	.86	-	-	.57	-	-
F4_D1* F5_D2	Interaction effect	.95	-	-	.84	-	-	.99	-	-	1.0	-	-	.52	-	-
F4_D1* F5_D3	idem	.98	-	-	.59	-	-	.99	-	-	.99	-	-	.00	7.95	2838.43
F4_D1* F5_D4	idem	.97	-	-	.01	3.88	48.53	.99	-	-	.99	-	-	.00	5.80	330.09
F4_D3* F5_D2	idem	.96	-	-	.00	4.44	84.64	1.0	-	-	.93	-	-	.11	-	-
F4_D3* F5_D3	idem	.93	-	-	.62	-	-	1.0	-	-	.99	-	-	.97	-	-
F4_D3* F5_D4	idem	.99	-	-	.47	-	-	.99	-	-	.99	-	-	.99	-	-

Notes: (1) Negative coefficients are underlined (or $e^B <= 0$); non-significant coefficients are not printed (except for the constant term); N.R. means 'not relevant';

(2) #Cross-group invariance condition (with respect to this measurement parameter);

(3) e^B is the impact on the odds ratio;

(4) -2LL means 'minus 2 times loglikelihood'; CCR means 'correct classification rate'; RRCS means 'rate of robust cases in sample' .

Table 4.9.
Logistic regression models predicting the ROBUSTNESS of the statistical difference test (0/1) between factor means at population level against violations of the cross-group invariance assumption (one-factor model with 4 indicators)

4 indicators		Factor mean in group 2														
		= -0.30 -2LL=89.83 N=198 CCR: 91.9 % RRCS: 28.3 %			= -0.15 -2LL=151.11 N=198 CCR: 83.8% RRCS: 26.3 %			= 0.00 -2LL=107.98 N=198 CCR: 86.9 % RRCS: 34.8 %			= 0.15 -2LL=72.97 N=198 CCR: 92.4 % RRCS: 18.7 %			= 0.30 -2LL=170.41 N=198 CCR: 80.8% RRCS: 66.2 %		
BINARY VARIABLE	SETTING	P value	B	e^B	P value	B	e^B	P value	B	e^B	P value	B	e^B	P value	B	e^B
Constant	Constant = 1	.65	.71	N.R.	.60	-.73	N.R.	.88	9.45	N.R.	.99	-.81	N.R.	.25	1.29	N.R.
Reference F1	Standard normal distribution	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
F1_D2	Uni-modal distr.	.21	-	-	.03	1.20	3.33	1.0	-	-	.30	-	-	.16	-	-
F1_D3	Bi-modal distr.	.21	-	-	.01	1.32	3.73	.38	-	-	.73	-	-	.82	-	-
F2_D1	N=200/200	.65	-	-	.00	3.39	29.64	.05	1.68	5.36	.02	2.42	11.20	.00	-2.95	.05
F2_D2	N=300/300	1.0	-	-	.14	-	-	.05	1.68	5.36	.07	1.90	6.66	.00	-2.00	.14
F2_D3	N=400/400	.35	-	-	.25	-	-	.23	-	-	.34	-	-	1.0	-	-
F2_D4	N=500/500	.37	-	-	1.0	-	-	.67	-	-	1.0	-	-	.11	-	-
Reference F2	N=750/750	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
F2_D6	N=200/400	.18	-	-	.07	1.41	4.09	.23	-	-	.34	-	-	.04	-1.38	.25
F4_D1	λ_2 (G2) = 0.4	.84	-	-	.19	-	-	.90	-	-	.99	-	-	.04	-2.50	.08
Reference F4#	λ_2 (G2) = 0.6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
F4_D3	λ_2 (G2) = 0.8	.29	-	-	.59	-	-	1.0	-	-	1.0	-	-	1.0	-	-
Reference F5#	Int_2 (G2) = .00	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
F5_D2	Int_2 (G2) = .15	.52	-	-	.04	-2.87	.06	.87	-	-	.98	-	-	.55	-	-
F5_D3	Int_2 (G2) = .30	.84	-	-	.01	-3.96	.02	.83	-	-	.91	-	-	1.0	-	-
F5_D4	Int_2 (G2) = .45	.00	-4.10	.02	.00	-3.35	.04	.69	-	-	.84	-	-	1.0	-	-
F4_D1* F5_D2	Interaction effect	.99	-	-	.28	-	-	.90	-	-	.96	-	-	.03	3.35	28.78
F4_D1* F5_D3	idem	.91	-	-	.32	-	-	.89	-	-	.99	-	-	.00	4.65	104.54
F4_D1* F5_D4	idem	.96	-	-	.41	-	-	.90	-	-	.99	-	-	.00	4.08	59.08
F4_D3* F5_D2	idem	.52	-	-	.02	3.18	23.97	.99	-	-	.92	-	-	.76	-	-
F4_D3* F5_D3	idem	.86	-	-	.44	-	-	.98	-	-	1.0	-	-	.77	-	-
F4_D3* F5_D4	idem	.92	-	-	.37	-	-	.99	-	-	.99	-	-	.68	-	-

Notes: (1) Negative coefficients are underlined (or $e^B <= 0$); non-significant coefficients are not printed (except for the constant term); N.R. means 'not relevant';
(2) #Cross-group invariance condition (with respect to this measurement parameter);
(3) e^B is the impact on the odds ratio;
(4) -2LL means 'minus 2 times loglikelihood'; CCR means 'correct classification rate'; RRCS means 'rate of robust cases in sample'.

Due to the small number of non-invariance conditions for each level of factor 3 (N=198 non-invariance conditions per level of factor 3), and the relatively large number of regression coefficients to be estimated (22 excluding the constant term) the logistic regression models in Tables 4.8 and 4.9 did not provide very reliable and useful information. Even though high percentages of correct classifications were obtained (all above 80%), many logistic regression models hardly reported any significant effects.

For some logistic regression models none of the regression coefficients were found to be significant (e.g. F3=1 & F3=4 for the 3-indicator conditions). Another logistic regression model reported only one significant effect (4-indicator conditions; F3=1), in particular: an effect of the largest degree of non-invariance of the indicator intercept (i.e. F5=4). The corresponding C&RTree analyses were much more informative in terms of the identification of impactful design factors when it comes to explaining the robustness of non-invariance conditions. Still other logistic regression models (i.e. 3- and 4-indicator conditions; F3=3 and F3=4) indicated primarily significant effects for small sample sizes (per group). This is remarkable as the corresponding C&RTrees (T7 to T10 in Appendix 4.6) did not produce any sample split using sample size per group as a splitting variable. The only exception to this was C&RTree T7 (3-indicator conditions; F3=3).

The C&RTrees, which were discussed before, showed that non-invariance of the indicator intercept (i.e. F5) is the most important factor in determining the percentage of robust non-invariance conditions. In some logistic regression models, the effects related to a non-invariant indicator intercept were not reported as significant. For these reasons, a decision was made to only discuss those logistic regression models which seemed to produce results which were (at least partially) in line with the results obtained by means of C&RT analyses. This implies that the discussion will be limited to one logistic regression model for small negative discrepancy cases (i.e. F3=2), and one logistic regression model for large positive discrepancy cases (i.e. F3=5).

F3=2

In small negative discrepancy cases (i.e. F3=2), small sample sizes (e.g. F2=1) had a strong positive effect on the robustness of the non-invariance condition. This is true for both 3- and 4- indicator conditions. This conclusion is supported by the second logistic regression model (i.e. F3=2) presented in Table 4.8 and Table 4.9, respectively. The corresponding C&RTrees (i.e. T5 and T6 in Appendix 4.6) show that the very first sample split was made using the smallest sample size (i.e. N=200 per group) as the splitting variable. Table 4.8 shows that, in 3-indicator conditions, there was a counterbalancing effect of a smaller non-invariant factor loading and a larger non-invariant indicator intercept in group 2 (when compared to group 1). A significant positive interaction effect (e.g. F4=1 & F5=4 in 3-indicator conditions) provided empirical evidence for this

finding. Note that this interaction effect was also present in the corresponding C&RTree (i.e. T5).

F3=5

As shown by the logistic regression models dealing with large positive discrepancy cases (i.e. F3=5), small sample sizes per group (e.g. N=200) substantially lower the probability of the non-invariance condition to be robust. Similarly, a smaller non-invariant factor loading in group 2 (i.e. a factor loading of 0.4 in group 2 versus a factor loading of 0.6 in group 1) led to a decrease in terms of the percentage of robust non-invariance cases. These conclusions are valid, both for 3- and 4-indicator conditions.

4.3.2.3. Conclusions

The C&RTree analyses and logistic regression analyses have shown that violations of the measurement invariance assumption across groups may have a very strong impact on the robustness of (simulated) non-invariance conditions. The extent to which non-invariance conditions are non-robust depends on which measurement parameters (i.e. factor loading and/or indicator intercept) fail to exhibit measurement non-invariance across groups. The influence of a non-invariant intercept is dominant when compared to a non-invariant factor loading in negative discrepancy cases and in small positive discrepancy cases. In large positive discrepancy cases, the effect of a non-invariant factor loading is relatively more outspoken than in all other non-invariance conditions. In negative discrepancy cases, a smaller non-invariant factor loading (in group 2) may partially compensate for the negative effect of a larger non-invariant indicator intercept on the robustness of the non-invariance condition (for instance, when sample size per group is small [i.e. F2=2]). The robustness of the non-invariance condition is also influenced by the size of the sample size in each group. This is true for small negative discrepancy cases and large positive discrepancy cases. The distribution of indicators does not affect the robustness of non-invariance conditions. All conclusions are summarised in Table 4.10. Table 4.10 also indicates the relative importance of factors 1, 2, 4, and 5 for all levels of factor 3. The assigned rank orders are based on the size of the e^B coefficient (or $1/e^B$ if $e^B < 1$).

Table 4.10.
Overview of the determinants of the ROBUSTNESS of the condition against violations of the measurement invariance principle.

Determinants of the % of robust cases	Factor mean in group 2				
	= -0.30	= -0.15	= 0.00	= +0.15	= +0.30
F1: distribution of indicators	No effect	No effect	No effect	No effect	No effect
F2: sample size per group	No effect	- [1]	No effect	No effect	+ [1]
F4: non-invariant factor loading	No effect-	No effect	No effect	No effect	+ [2]
F5: non-invariant indicator intercept	- [1]	- [3]	- [1]	- [1]	No effect
Interaction effect(s) between F4 and F5.	Synergetic effect (F4+ & F5+=> %R-) [1]	Counterbalancing effect (F4- & F5+=> %R+) [2]	No effect	No effect	No effect

Explanation:

(1) Consider the cell [Interaction effect(s) between F4 and F5, F3=1] in this table: The notation *F4+ & F5+ => %R-*) denotes that *larger* values for F4 (non-invariant factor loading is larger in group 2 when compared to group 1) combined with *larger* values for F5 (i.e. non-invariant indicator intercept is larger in group 2 when compared to group 1) lead to a decrease of the percentage of conditions which are robust against violations of the measurement invariance assumption across groups (i.e. more severe violations of the measurement invariance assumptions across groups for a more serious threat to the robustness of the difference test based on factor means);

(2) Rank orders are indicated between square brackets (1=strongest effect).

4.4. Final conclusions

The simulation study has shown that a non-invariant indicator may have a very strong impact on the percentage of correct statistical conclusions which are based on a statistical comparison between the estimated factor means in two populations. Of all simulated replications, about 65% resulted in a correct (statistical) outcome for the factor mean difference test.

A difference in the non-invariant indicator intercept as large as (about) one tenth of the total length of the scale (a difference of 0.45 on a 5-point scale) -or even smaller- strongly reduced the probability of drawing correct statistical conclusions based on a factor mean difference test. The same is true for a 0.2 difference in a non-invariant factor loading (the factor loading in the reference group being equal to 0.6). Sample size (per group) turned out to be another major determinant of the correctness of the factor mean difference test. The underlying distribution of the indicators did not exert a substantial influence on the correctness of the factor mean difference test (at least not with sample sizes of at least 200 observations per group). This finding is important as it shows that the treatment of ordinal data as if they were metric is not problematical (at least not for 5-point Likert types of scales with a left-skewed distribution or a symmetric bi-modal distribution). All of these conclusions apply equally well to 3- and 4-indicator conditions.

The main research question in this simulation study was to evaluate the extent to which non-invariance conditions are robust against violations of the measurement invariance principle (across groups). Non-invariance conditions were considered to be robust if the number of correct statistical conclusions fell within a 99% tolerance region around the number of correct statistical conclusions for the corresponding full invariance condition. Of all simulated non-invariance conditions, only about 35% turned out to be robust. The low overall percentage of robust non-invariance conditions shows that non-invariant measurement parameters (of one indicator across groups) have a very strong impact on the robustness of non-invariance conditions. In this simulation study, robust non-invariance conditions were rather exceptional.

Apart from a difference in factor means (at population level), the major determinant of the robustness of non-invariance conditions turned out to be the degree of non-invariance of the indicator intercept. This is true for all simulated non-invariance conditions, except for non-invariance conditions with a large positive discrepancy between factor means.

The effect of the non-invariant factor loading was somewhat more important in large positive discrepancy cases. In these cases, the percentage of robust non-invariance conditions was rather high (about 70%). The combination of: (1) a large difference in factor means at population level, and (2) the positive bias

due to a non-invariant indicator intercept was responsible for a small difference in the percentage of correct statistical conclusions between non-invariance conditions and their corresponding full invariance condition. As a consequence, a high percentage of robust non-invariance conditions were obtained.

A smaller factor loading (in group 2) could partially compensate for the bias due to a larger indicator intercept (in the same group). In addition to the effect of non-invariant measurement parameters, there was also an effect of sample size per group on the robustness of the non-invariance condition. This effect was found in small negative discrepancy cases and large positive discrepancy cases. The distribution of the indicators did not exert an influence on the robustness of non-invariance conditions. All conclusions regarding the factors determining the robustness of non-invariance conditions were consistent across 3- and 4-indicator cases.

In sum, this simulation study has shown that non-invariant measurement parameters form a serious threat to the correctness of a factor mean difference test between two populations. A non-invariant factor loading, and in particular: a non-invariant indicator intercept, have a strong impact on the percentage of correct statistical conclusions regarding the factor mean difference test. The degree of non-invariance (as simulated in this study) was severe enough to seriously affect the robustness of the factor mean difference test against violations of the measurement invariance principle (across groups). Furthermore, it does not seem to matter very much if one uses three or four indicators to measure the underlying (one-dimensional) factor. The results were highly consistent across 3- and 4-indicator conditions.

For these reasons, the general advice to applied researchers is to test for measurement invariance (across groups) prior to conducting any factor mean comparisons across groups (as described in Chapter 3). It is crucial that indicators which do not exhibit measurement invariance across groups are removed from the measurement model. Otherwise, factor mean comparisons across groups may not be meaningful at all.

Chapter 5. Measurement invariance assessment in a large-scale employee survey

"Approximation is the soul of science."

C. Glymour, R. Scheines, P. Spirtes, & K. Kelly

5.1. Introduction

The globalisation of the marketplace is arguably the most important challenge facing companies today (Yip, 1995). The rapid trend towards globalisation (Shenkar, 1995) affects all aspects of policy-making in (multinational) companies, including human resource (HR) management. Many of these companies have to implement HR practices globally to be successful in the global market (Erez, 1994).

To evaluate whether the global HR policy is effective, multinational companies may monitor its performance through global research tools, such as common appraisal performance systems, and global employee opinion surveys. A comparative analysis between countries allows HR professionals to distinguish between those aspects of the global HR policy that are effective in all countries, and those aspects that are not effective at all in some (or all) countries. Cultural differences between countries, for instance: in terms of the factors that determine one's motivation to work, may necessitate a local adaptation of the (global) HR policy.

As mentioned in the introductory chapter, cross-country comparisons are only meaningful from a substantive point of view if comparability of data is established across countries. In practice, it is hard to establish comparability of data if cross-country comparisons are to be based on (more abstract) factors such as: organisational commitment, immediate boss' support, thrust in managerial decisions, etc. Such (abstract) factors are typically measured by means of multi-items scales. If these multi-item scales do not exhibit 'measurement invariance' across countries, then any comparisons between countries based on the (country-mean) factor scores may be highly inaccurate, if not completely wrong.

The requirement of measurement invariance of multi-item scales across countries may be very unrealistic in empirical research. As explained in Chapter 3, a series of MACS models can be used to formally test the assumption of measurement invariance across countries. If the tests show that measurement invariance across countries is not established, it is worthwhile to ask the following question:

> *"How threatening is non-invariance of items (across countries) in terms of the adequacy of factor mean comparisons across countries?"*

In this chapter the following research questions are addressed:

(1) First of all, a formal test will be conducted on whether the multi-item measures (as used in a specific global employee opinion survey) exhibit measurement invariance across countries,

(2) Secondly, the impact of non-invariant items (if any) on factor mean comparisons across countries will be assessed.

A specific (statistical) procedure is proposed to address the latter research question.

In this chapter data from a particular global employee opinion survey (in a multinational company) is used to answer the two research questions mentioned above.

Readers should keep in mind that this research concerns a case study. The conclusions from this research cannot be generalised to other global employee opinion surveys. The results depend, amongst other factors, on the countries involved in the study, the factors studied, and the particular multi-item scales used to operationalise these factors.

5.2. Background

5.2.1. Measuring employees' job satisfaction

Global HR management faces the complex task of finding ways to improve performance of their employees to ensure staying competitive, while keeping them satisfied with their job and work environment.

Companies can benefit from having highly satisfied employees in several ways. As explained in Exhibit 5.1, employees who are satisfied with their job show a higher commitment towards the company, and are less likely to quit their job and the company. Provided that the majority of the employees perform well on their job, a low (voluntary) turnover rate leads to substantial cost savings (e.g. recruitment and training costs), and higher levels of productivity.

To increase employees' performance and keep them satisfied with their job, companies operating in global markets should collect data on employees' actual performance and the extent to which employees are satisfied with their job. The most popular way to collect data on employees' performance is through supervisory ratings (Bernardin and Vilanova, 1986; Borsman, 1991; Pulakos, 1997). Data on employees' job satisfaction is usually obtained by means of global employee opinion surveys.

In global employee opinion surveys, employees are typically confronted with several statements about themselves, their department, the (local) company or the multinational company (as a whole). They are asked to indicate to which extent they agree or disagree with the statements on a rating scale (typically a 5-point Likert-type of scale). The statements often provide information on their general attitudes (e.g. human values) and job-related attitudes, including job satisfaction. Oftentimes, an assumption is made that several statements measure one and the same underlying theoretical construct (i.e. a one-dimensional latent factor). These statements are said to constitute a (multi-item) measurement instrument for the factor under study. Depending on the research objectives, the study may focus on measuring factors which are known to be determinants of key constructs such as employee job satisfaction, organisational commitment, and turnover intent.

Exhibit 5.1.
Employees' job satisfaction, its antecedents and consequences

DEFINITION OF JOB SATISFACTION
Employee job satisfaction is defined as a positive emotional state resulting from the appraisal of one's job or job experiences (Locke, 1976).

ANTECEDENTS OF JOB SATISFACTION
There are two general categories of factors that are believed to influence employees' job satisfaction: demographic characteristics (e.g. age, gender, educational level, tenure) and work environment factors (Lambert et al., 2001). The research by Lambert and others (2001) has shown that work environment factors have a greater effect on job satisfaction than do demographic factors.

CONSEQUENCES OF JOB SATISFACTION
Employees' job satisfaction is known to be a key antecedent of voluntary turnover (Mobley et al., 1979; Price and Mueller, 1986; Williams and Hazer, 1986). Empirical studies have shown that: (1) the relationship between job satisfaction and voluntary turnover is negative and consistent, but also that (2) the percentage of explained variance is small (Locke, 1976). Mobley and his colleagues have argued that this relationship is mediated by intentions (Mobley at al., 1978, 1979). Most researchers are now convinced that intention to leave the company is the final cognitive step in the decision process of voluntary turnover (Steel and Ovalle, 1984). This explains why 'intention to leave' (the organisation) is a frequently encountered (dependent) variable in job satisfaction studies.

The causal relationship between employees job satisfaction and turnover intent is moderated* (not mediated!) by factors such as: alternative employment opportunity (Lambert et al., 2001), and organisational commitment (Cohen, 1993). Much empirical evidence is provided in the literature to support the hypothesis that high levels of job satisfaction positively influence one's commitment towards the organisation (Marsh and Manari, 1977; Mowday, Porter and Steers, 1982; Price and Mueller, 1986, Williams and Hazer, 1986; Martensen and Grønholdt, 2001).

MORE DETAILS?
The interested reader is encouraged to consult the meta-analysis study on the relationship between job satisfaction and organisational commitment / turnover intent by Tett and Meyer (1993), and on the relationship between organisational commitment and turnover intent in particular (e.g. Mathieu and Zajac, 1990; Cohen, 1993).

Note: *The variable Y is a moderator w.r.t. to the relationship between variable X and an outcome variable Z if the followings path-analytic relationships apply: (1) A -> Z, (2) Y -> Z, as well as (3) the interaction effect: X&Y -> Z. The variable Y is a mediator w.r.t. the relationship between variable X and outcome variable Z if the following path-analytic relationships apply: (1) X->Y, (2) Y->Z, and (3) X->Z being a non-significant path. (Baron and Kenny, 1986).

5.2.2. Cross-country comparisons

To evaluate if the (global) HR policy is effective in all countries in which it is implemented, it is key to compare the results of the (global) employee opinion survey across countries. A cross-country comparison of job satisfaction data helps to differentiate between countries with highly satisfied employees and countries with employees who are not satisfied at all. Additionally, cross-country comparisons may also be made with respect to factors such as organisational commitment and turnover intent. As explained in Exhibit 5.1, organisational commitment and turnover intent are (causal) consequences of employees' job satisfaction.

A cross-country comparison based on (mean levels of) job satisfaction does not offer insights as to how higher satisfaction levels can be established. Therefore, cross-country comparisons can also be made based on those factors which are known to be determinants of employees' job satisfaction[50] (e.g. specific work environment factors). Such a comparative analysis may provide cues as to how the (global) HR policy can be improved (or modified) to make employees within a given country more satisfied with their job. Provided that the global employee opinion survey is conducted at regular points in time, the effectiveness of changes in the HR policy may also be assessed by comparing data over time (i.e. before and after the change in HR policy).

5.2.3. The issue of comparability of data across countries

Cross-country comparisons are hampered by some methodological problems. The first problem concerns the fundamental choice between the etic and emic approach[51] (or an approach combining both approaches). Another problem concerns the meaningfulness of factor mean score comparisons across countries. These problems are explained in more detail in the next paragraphs.

[50] In case the determining factors of employees' job satisfaction are not known, a regression-type of analysis (or a path analysis) can be used to identify the factors influencing employees' job satisfaction.
[51] The terms 'etic' and 'emic' are explained in chapter 1.

Etic versus emic

As argued by Ryan and colleagues (1999), HR practitioners rely on the etic approach rather than on the emic approach. The etic approach allows them to quickly adapt HR practices in a global workforce (Ryan et al., 1999). The emic approach fails to do so as this approach is far too time-consuming and far too complex (see, for example, Ployhart et al., 2003). This may explain why, in this particular global employee opinion survey, an etic approach to multi-country research is adopted as well. More detailed information on the etic and emic approach was provided in Chapter 1 of this dissertation.

The requirement of measurement invariance across countries

Another critical issue concerns the meaningfulness of factor mean comparisons across countries. Such factor mean comparisons are only meaningful (from a substantive point of view) if measurement invariance across countries is established (see Ryan et al., 1999). The condition of measurement invariance implies that:

(1) translations of a measurement instrument are 'culturally appropriate' (Weech-Maldonado et al., 2001),

(2) the conceptual equivalence of different versions of the same measurement instrument is established (Hui and Triandis, 1985). Different versions of the same measurement instrument are conceptually equivalent if they have essentially the same meaning across cultures (countries).

Meeting both conditions (i.e. cultural appropriateness and conceptual equivalence across countries) is not sufficient to meaningfully compare factor means across countries. They are only a necessary condition (Drasgow, 1984, 1987). Comparisons of factor mean scores across countries are only meaningful, if, in addition, the measurement instruments used to operationalise these factors exhibit measurement invariance across countries.

Measurement invariance can generally be defined as the extent to which individuals with the same (latent) factor score have the same observed score (with the exception of differences due to differences in the reliability of instruments) (see Drasgow and Kanfer, 1985). Measurement invariance across countries is present if persons from different countries who have the same (latent) factor scores score identical on the observed variables which are supposed to measure these factor scores.

Without measurement invariance, interpreting cross-country differences in factor means, factor variances, and correlations with other variables may not be

meaningful. Lack of measurement invariance across countries implies that there is no common basis to compare data across countries. In such cases, factor mean differences may be the result of differences in measurement instruments across countries (cultures) rather than true differences across countries (cultures).

In a factor-analytic framework[52], measurement invariance implies that the mathematical properties (i.e. measurement parameters), which are needed to quantify the underlying construct (or factor), can be applied in a uniform way to all groups (countries) of interest. To date, there is no consensus in the methodological literature as to what the minimal set of (measurement) parameters should be that has uniform (i.e. identical) values across groups (countries). In this chapter, Meredith's (1993) strong definition of measurement invariance (across groups) is used as a reference. According to this definition, factor loadings and indicator intercepts of observed variables (i.e. indicators) should be identical across groups (countries). Unique variances (i.e. indicator unreliabilities) of indicators may, however, vary across countries. The same is true for factor means, factor variances, and factor covariances. Meredith's invariance condition is referred to as 'tau-invariance' (across groups) in this chapter. According to Meredith (1993), Little (1997) and Steenkamp and Baumgartner (1998) tau-invariance (across groups) is a prerequisite for the comparison of (latent) factor means.

[52] Confirmatory Factor Analysis (CFA) models are used in this chapter to test for measurement invariance of observed variables (items). An alternative approach would be to use Item Response Models (e.g. Reise et al., 1993; Maurer et al., 1998; Salzberger et al., 1999; Eid and Rauber, 2000; Facteau and Craig, 2001).

5.2.4. Positioning of this research

The most well-known cross-cultural work in the area of job attitudes is the work by Hofstede (1976, 1983, 1985). Hofstede analysed data collected from IBM employees in 69 countries. In later years, some empirical works on cross-cultural differences in job satisfaction were conducted (Candell and Hulin, 1986; Lincoln et al., 1981; Slocum and Topichak, 1972; Smith and Misumi, 1989; Spector and Wimalasiri, 1986). These studies are typically one-country (and some two-country) studies. In these studies, the issue of measurement invariance were not considered.

More recent work by Ryan and others (Ryan et al., 1999, 2000) is distinctive in that it tests for measurement invariance of measurement instruments in a multi-country setting. The papers by Ryan and co-authors (1999, 2000) serve as a source of inspiration for this chapter.

As mentioned in the introduction, this research centers around two research questions:

(1) Is the assumption of (cross-country) measurement invariance of multi-item scales (as used in this global employee opinion survey[53]) realistic?

(2) If not (realistic), to which extent do non-invariant items distort (i.e. bias) factor mean comparisons across countries?

In a supplementary analysis, it is assessed to which extent work environment factors under study determine global job satisfaction in the individual countries participating in the study.

[53] The reader who is interested in the cross-cultural applicability of performance appraisal systems rather than employee opinion surveys may consult two recent papers (e.g. Facteau and Craig, 2001; Ployhart et al., 2003).

The present research differs from the papers by Ryan and co-authors (1999, 2000) in that:

(1) The number of countries (16) is very large;

(2) This research introduces a special (statistical) procedure which can be used to investigate to which extent violations of the measurement invariance assumption (across countries) bias (estimated) the factor mean comparisons across countries.

    The large number of countries enables the researcher to use the individual countries as units of analysis when assessing the severity of the bias (caused by non-invariance of measurement parameters across countries). As shown further on in this chapter, the impact of the bias can be assessed by calculating a simple correlation coefficient (i.e. r) between factor mean score estimations as derived from two competing CFA models (the tau-invariance model [specifying measurement invariance across countries] and another, more realistic model [specifying non-measurement invariance across countries]). By statistically comparing the size of the correlation with one (i.e. 1), a conclusion can be made as to whether the impact of the bias is substantial (i.e. if r is significantly different from one) or not (i.e. if r is not significantly different from one),

(3) Finally, Meredith's (1993) (stronger) definition of the concept of measurement invariance (across groups) is adopted (i.e. requiring cross-group equality of factor loadings and indicator intercepts).

    Ryan et al. (1999), for instance, adopted a less stringent definition in which only the cross-group equality of factor loadings (or a subset of all factor loadings) is required.

5.3. Method

5.3.1. Sample

Data from respondents in 16 countries were collected at three points in time in 2002 (March, May, and September) within a multinational company. The respondents completed questionnaires which were sent to them either by normal mail or, alternatively, made accessible via the Internet. Because of respondents' privacy issues, no demographic characteristics were made available to the author. The list of countries included: Belgium (N=932), France (N=1152), Germany (N=1668), Hungary (N=1061), Italy (N=1711), The Netherlands (N=1360), Russian Federation (N=1180), Sweden (N=960), United Kingdom (N=3826), Canada (N=1084), United States (N=6700), Brazil (N=9397), Mexico (N=2549), Australia (N=1313), Israel (N=1802), and South Africa (N=822). The overall response rate across countries was 76.4%[54]. The sample sizes indicated between brackets indicate the total number of observations in each country. Due to the large number of missing values and the use of list wise deletion of missing data in the analytical procedures, sample sizes reported in the analysis section may be smaller than the sample sizes listed above.

5.3.2. Measures

A global team consisting of members from the multinational company and the research agency developed survey items in English for the global employee opinion survey study. The survey items focused on factors of key importance to the global HR policy within the multinational company (e.g. employees' remuneration, effectiveness of one's direct boss, clarity of the business strategy, etc.). The global team derived benefit from the expert knowledge of experts within the multinational company and the research agency on each of these factors of key importance. Regional survey leaders determined which translations were needed for their region. The research agency supervised and monitored 29 different translations which were checked by local survey co-ordinators using the English questionnaire as the basis for comparison.
Except for a small number of country-specific questions, all survey questions were common across all countries. The survey included 102 questions (items) on employees' opinions. Of these 102 items only 19 items have been used in this paper. All these items were assumed to adequately represent one specific work environment factor. The following work environment factors were considered:

Factor 1: Fair remuneration (3 items*)
Factor 2: Supporting role of people within the department (3 items*)

[54] Response rates per country have not been made available to the author.

Factor 3: Immediate boss' support (3 items*)
Factor 4: Clarity of strategy (3 items*)
Factor 5: Confidence in managerial decisions (2 items*)
Factor 6: Organisational and managerial efficiency (2 items*)
Factor 7: Environmental and societal responsibility (2 items*)

Note: * see Appendix 5.1

These 7 work environment factors were considered to be of key importance to the multinational company. All items are listed in Appendix 5.1. The items were scored on five-point Likert-type scales. The extent to which the employee agreed (or disagreed) with a statement was scored using the following response scale: 1=Disagree, 2=Tend to disagree, 3=Neither agree, nor disagree, 4=Tend to agree and 5=Agree. Other items were scored on a 5-point evaluation scale using the categories: 1=Very poor, 2=Poor, 3=Fair, 4=Good and 5=Very good.

Employee job satisfaction was determined by just one question: *'Considering everything, how satisfied are you with your job?'*. The response scales for this global measure of employee satisfaction were: 1=Very dissatisfied, 2=Dissatisfied, 3=Neither satisfied, nor dissatisfied, 4=Satisfied and 5=Very satisfied. Because the factor 'job satisfaction' is measured by just one question, one could set the factor score equal to the score obtained on the question measuring (general) employee satisfaction.[55] The global measure of employee satisfaction may be used as the dependent variable when assessing the extent to which these seven work environment factors determine global employee job satisfaction levels in every individual country.

As there was no generally agreed (weighted) measure of job satisfaction (Bedeian et al., 1992; Ironson et al., 1989) the global measure may be the best choice to use. By using the global measure biased results due to making universal assumptions about the weights of various facets of the job satisfaction construct may be avoided. The assumption of equal weights across countries may seriously bias the test results (Lambert et al., 2001). Scarpello and Campbell (1983) demonstrated that a 5-point global measure of job satisfaction is a reliable and adequate measure for the job satisfaction construct.

[55] When using structural equation modelling (SEM) one may also specify that the single indicator has limited reliability (see for instance: Jöreskog and Sörbom [1993, p. 37] who assume that the single indicator has a reliability of 0.85 [instead of 1.0]). In appendix 5.6, the influence of the work environment factors on employee job satisfaction is determined using an SEM approach. In this appendix, the stability of statistical results is assessed by specifying alternative levels of indicator reliability (i.e. 1.0 and 0.85) for the single indicator.

5.3.3. Analyses

The software Mplus version 2 (Muthén and Muthén, 1999, 2003) was used for all analyses discussed in this chapter.

5.3.3.1. Sequence of CFA models to be evaluated

The first model which was tested was a Confirmatory Factor Analysis (CFA) model[56] which imposed the assumed 7-factor structure onto the data. This model was evaluated using different samples. The first sample was the total sample including all observations from 16 countries (N=25018). Next, the same model was evaluated using data from each individual country. A measurement scale was created for all seven factors by fixing the factor loading of one of its indicator variables to one in each country. This indicator variable was referred to as the 'reference indicator'. The reference indicators are listed in Appendix 5.1. Using reference indicators is a common procedure to scale factors in a CFA framework (see Williams and Thomson, 1986; Bollen, 1989).

Provided that the data fit the 7-factor model well (i.e. evidence for *factorial* invariance across countries), a hierarchical sequence of nested statistical models (e.g. Vandenberg and Lance, 2000) can be used to assess *measurement* invariance across groups (countries).

It is common practice to start with a baseline model in which no parameters (i.e. factor loadings, indicator intercepts, unique variances, factor means, and factor variances and covariances), except for the factor loading of the reference indicator, are constrained to be equal across groups/countries. This model is referred to as the 'congeneric factor invariance model'. For identification purposes the factor means in the first group/country (i.e. Belgium) are always fixed to zero. The congeneric factor invariance model is graphically depicted in Figure 5.1.

[56] This model does not include factor (and item) mean structures.

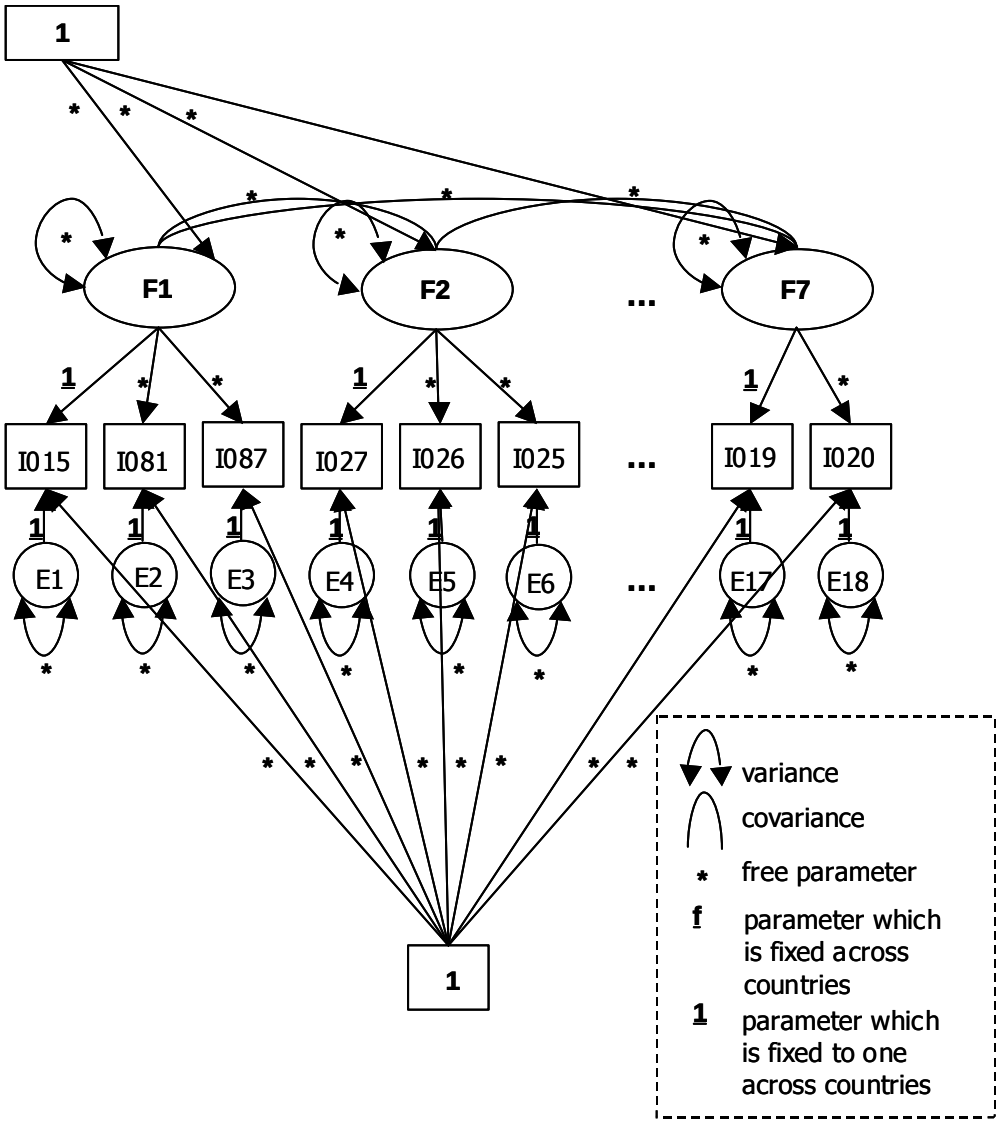


Figure 5.1.
The congeneric factor invariance model

Note: The observed variables are listed in Appendix 5.1.

The second model in the sequence constrains all factor loadings to be identical across groups/countries while all other parameters (i.e. indicator intercepts, unique variances, factor means, and factor variances and covariances) are freely estimated. This model is called the 'metric invariance model' (across groups). The metric invariance model is shown in Figure 5.2.

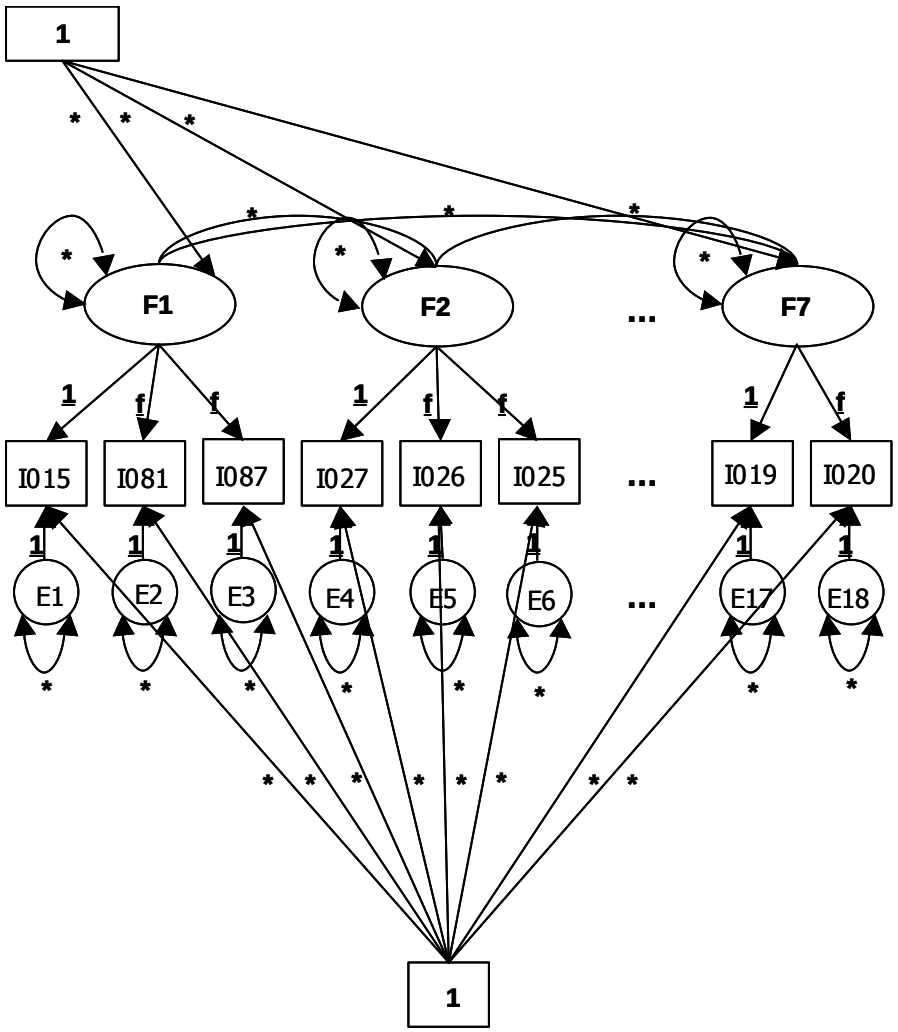


Figure 5.2.
The metric invariance model

Note: The observed variables are listed in Appendix 5.1.

Model fit statistics of the metric invariance model (for example the Chi-squared statistic) may be statistically compared with the congeneric factor invariance model (i.e. the baseline model). A non-significant difference in the fit statistic (e.g. the Chi-square statistic), given the difference in degrees of freedom between both models, favours the metric invariance (i.e. the more restricted) model over the congeneric factor invariance (i.e. the less restricted) model.

The third model in the sequence, the tau-invariance model, constrains both the factor loadings and the indicator intercepts to be identical across groups/countries. All other parameters (i.e. unique variances, factor means, and factor variances and covariances) are freed across groups/countries. The tau-invariance model is shown in Figure 5.3.

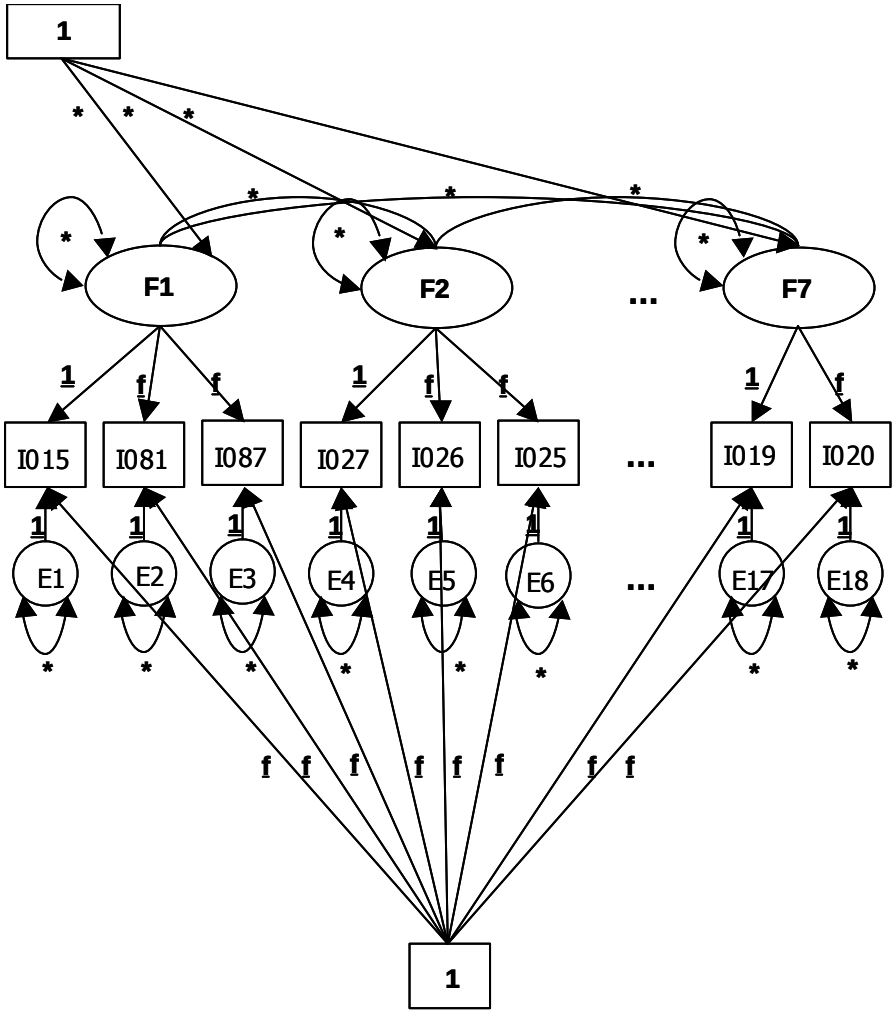


Figure 5.3.
The tau-invariance model

Note: The observed variables are listed in Appendix 5.1.

Analogously, a non-significant difference in the fit statistic (e.g. Chi-square), given the difference in degrees of freedom between the metric invariance model and the tau-invariance model, favours the tau-invariance model (i.e. the more restrictive of both models).

Theoretically, the equality of unique variances may also be tested in addition to factor loadings and indicator intercepts (i.e. assuming that reliabilities of indicator variables are identical across groups). This model, known as the 'parallel invariance model' (across groups), is overly restrictive in terms of the need for establishing measurement instruments which exhibit measurement invariance across groups (see, for example, Hittner, 1995)[57].

5.3.3.2. Assessment of model fit

To assess the fit of the Confirmatory Factor Analysis models, the 'classical' Chi-square statistic (plus its corresponding degrees of freedom) can be considered. The classical Chi-square statistic assesses the magnitude between the discrepancy between the actual covariance matrix of the observed variables and the covariance matrix implied by the hypothesised model (Hu and Bentler, 1995). Because of its known sensitivity to sample size, it will often produce highly significant values with large sample sizes as used in this study. These significant values may be due to differences of trivial size between both the covariance matrices. For this reason, a reliance on the classical Chi-square statistic was suspended in this study and four alternative measures of model fit which were less sensitive to sample size were considered instead.

The first two alternative measures of model fit are the Comparative Fit Index (CFI) by Bentler (1990) and the Tucker-Lewis Index (TLI), also called the Bentler-Bonnet Non-Normed Fit Index (NNFI). Both measures of model fit indicate the proportion by which the hypothesised model improves fit compared to a null model in which all observed variables are uncorrelated (i.e. the 'independence model'). One important difference between both indices is the following: the CFI index takes on values between 0 and 1, whereas the TLI index may –occassionally- fall outside the [0,1] range. By convention, both measures of model fit should be equal to or greater than 0.90 to accept the model.

The third alternative measure of model fit is the Root Mean Square Error of Approximation (RMSEA) as proposed by Steiger (1990). It is a per-degree-of-freedom measure of the estimated discrepancy between the covariance matrix between the observed variables as implied by the hypothesised model and the

[57] Hittner (1995) argues that the probability of establishing such an extreme form of measurement invariance would be very unlikely given that unique indicator variances consist largely of random error.

covariance matrix at population level. As shown by Hu and Bentler (1999), an RMSEA of 0.06 or less indicates good model fit. An RMSEA between 0.06 and 0.08 indicates a reasonable error of approximation.

The fourth alternative measure of model fit is the Standardised Root Mean Square Residual (SRMR). SRMR is the average of standardised differences between the actual covariances between the observed variables and the covariances as implied by the hypothesised model. SRMR is zero when model fit is perfect. SRMR values lower than (or equal to) 0.05 are preferable.

These four alternative measures of global model fit are commonly used in Confirmatory Factor Analyses. In addition to measures of global model fit, alternative models can also be compared on a statistical basis.

To statistically compare alternative measurement invariance models (such as the congeneric factor invariance model, the metric, and the tau-invariance model), the *Chi-square difference statistic* can be used (i.e. calculating the difference in Chi-square between both models, and the difference in degrees of freedom for both models) (see, for instance, Byrne et al., 1989; Reise et al., 1993; Steenkamp and Baumgartner, 1998). Recently, a simulation study was presented by Cheung and Rensvold (2002). In their study they assessed the usefulness of many other measures of model fit (in addition to Chi-square) when statistically comparing alternative CFA models specifying different levels of measurement invariance across groups. Their study shows that the difference in Comparative Fit Index (CFI) between nested invariance models is a more reliable (and robust) measure of model fit than the classical Chi-square difference test.

5.4. Results

5.4.1. The hypothesised 7-factor structure

The proposed 7-factor structure was tested using all observations from 16 countries, as well as the observations from each country separately. Measures of model fit are shown in Table 5.1.

Table 5.1.
CFA models[1] specifying a seven-factor structure

Country	Sample size	Chi-square statistic (+ d.f.)	Proba-bility	CFI	TLI	RMSEA[2]	SRMR
All countries	N=25018	5415 (114)	0.00	0.968	0.957	0.043 / 0.039	0.029
Belgium	N=658	278 (114)	0.00	0.958	0.944	0.047 / 0.040	0.039
France	N=788	242 (114)	0.00	0.974	0.965	0.038 / 0.031	0.031
Germany	N=1107	420 (114)	0.00	0.955	0.940	0.049 / 0.044	0.038
Hungary	N=707	243 (114)	0.00	0.970	0.960	0.040 / 0.036	0.034
Italy	N=1247	447 (114)	0.00	0.959	0.945	0.048 / 0.041	0.035
Netherlands	N=895	326 (114)	0.00	0.956	0.941	0.046 / 0.039	0.037
Russian federation	N=632	245 (114)	0.00	0.962	0.949	0.043 / 0.033	0.040
Sweden	N=488	235 (114)	0.00	0.968	0.957	0.047 / 0.041	0.038
U.K.	N=2620	846 (114)	0.00	0.963	0.950	0.050 / 0.045	0.037
Canada	N=739	273 (114)	0.00	0.976	0.967	0.043 / 0.035	0.032
United States	N=4496	1455 (114)	0.00	0.963	0.951	0.051 / 0.045	0.034
Brazil	N=6206	1193 (114)	0.00	0.970	0.960	0.039 / 0.035	0.027
Mexico	N=1821	482 (114)	0.00	0.968	0.957	0.042 / 0.036	0.033
Australia	N=919	336 (114)	0.00	0.968	0.958	0.046 / 0.040	0.038
Israel	N=1063	275 (114)	0.00	0.973	0.963	0.036 / 0.029	0.027
South Africa	N=632	365 (114)	0.00	0.943	0.924	0.059 / 0.049	0.047

Notes:
(1) [1]These models do not include a factor mean structure;
(2) [2]The second figure for RMSEA measure is based on Satorra-Bentler's scaled Chi-square statistic which corrects for model misfit due to non-normal distributed data.

Apart from highly significant Chi-square statistics, Table 5.1 shows that the hypothesised 7-factor structure was consistent with the covariance and mean

structures obtained from most countries. In all countries, the CFI and TLI statistics exceeded 0.90. Both the RMSEA measure which does not correct for multivariate non-normality of the data and the RMSEA measure which does correct for this violation did not exceed the critical value of 0.06. The SRMR values were consistently lower than 0.05. It is implied that the hypothesised 7-factor structure can be conceived as an adequate representation of the data on work environment variables as obtained in the survey study. The correlations between the seven work environment factors are shown in Appendix 5.2.

In the next paragraphs, measurement invariance across countries was assessed.

5.4.2. Measurement invariance of work environment factors

Next, different CFA models were tested to assess measurement invariance of items across countries. The baseline model was the congeneric factor invariance model (i.e. M1) in which no parameters (except for the factor loading of the reference indicator) were constrained to be equal across countries. Subsequent models were: the metric invariance model (i.e. factor loadings were constrained across countries), referred to as M2, and the tau-invariance model (i.e. both factor loadings and indicator intercepts were constrained across countries), which is referred to as M3.

Table 5.2.
Multi-group CFA models[1] to test for measurement invariance across 16 countries

ESTIMATED CFA MODELS

	Sample size	Chi-square statistic (+ d.f.)	Proba-bility	CFI	TLI	RMSEA[2]	SRMR
M1: Congeneric factor invariance model	N=25018	7662 (1824)	0.00	0.965	0.954	0.045 / 0.040	0.034
M2: Metric invariance model	N=25018	9103 (1989)	0.00	0.958	0.948	0.048 / 0.042	0.044
M3: Tau-invariance model	N=25018	16829 (2154)	0.00	0.913	0.901	0.066 / 0.059	0.056
M3': Partial tau-invariance model	N=25018	12948 (2123)	0.00	0.936	0.926	0.057 / 0.051	0.050

Notes:
(1) [1]These models do include a factor (and item) mean structure;
(2) [2]The second figure for RMSEA measure is based on Satorra-Bentler's scaled Chi-square statistic which corrects for model misfit due to non-normal distributed data;
(3) M1: no parameters are fixed across countries;
(4) M2: all factor loadings are fixed across countries;
(5) M3: all factor loadings and indicator intercepts are fixed across countries;
(6) M3alt: all factor loadings and most indicator intercepts are fixed across countries {exceptions are: i081 [i.e. an indicator of F1] and i025 [i.e. an indicator of F2] (in all countries), and i020 [i.e. an indicator of F7] (in Brazil)}.

Table 5.2. (continued)
Multi-group CFA models[1] to test for measurement invariance across 16 countries

DIFFERENCE TESTS BETWEEN NESTED CFA MODELS

	Sample size	Difference in Chi-square statistic (+ d.f.)	Proba-bility	Diffe-rence in CFI			
Difference between M2 and M1	N=25018	1441# (165)	0.00	-0.007	-	-	-
Difference between M3 and M2	N=25018	7726# (165)	0.00	-0.045	-	-	-
Difference between M3' and M2	N=25018	3845# (134)	0.00	-0.022	-	-	-

Notes:
#An alternative to computing the classical Chi-square difference statistic is to calculate a difference statistic which is based on the Satorra-Bentler scaled Chi-square statistic (see Satorra, 2000, and Satorra and Bentler, 1999). The Satorra-Bentler difference Chi-square statistic (as reported in these papers) follows a Chi-square distribution, asymptotically. For these model comparisons this difference statistic is: 1038 (model M2 versus model M1), 6382 (model M3 versus model M2), and 3220 (model M3' versus model M2). Changing to this difference statistic does not lead any major changes in the conclusions as all related probabilities are also equal to 0.00. The formula for calculating the Satorra-Bentler chi-square difference test is: $(SB\_X^2_n - SB\_X^2_c) / \{(DF\_SB_n * SF_n - DF\_SB_c * SF_c) / (DF\_SB_n - DF\_SB_c)\}$ where: SB_X^2 indicates the Satorra-Bentler corrected X^2 statistic, DF_SB the degrees of freedom, and SF the Satorra-Bentler scaling factor (as reported by Mplus). The subscripts '$_n$' and '$_c$' are used to refer to the nested (i.e. most restricted) model and the comparison model (i.e. the less restricted model).

The measures of global model fit reported in Table 5.2 (other than Chi-square) revealed that the congeneric factor invariance model and the metric invariance model represented models with an adequate fit (CFI>=0.90; TLI>=0.90; RMSEA <=0.06; SRMR <=0.05). The tau-invariance model did not show adequate fit as SRMR was somewhat too high for the tau-invariance model (SRMR = 0.056).

Table 5.2 presents a statistical comparison between alternative CFA models specifying a different level of measurement invariance across groups/countries.

The difference in CFI between the congeneric factor invariance model (M1) and the metric invariance model (M2) is equal to minus 0.007. As this value is smaller than the critical difference of (minus) 0.01 as suggested by Cheung and Rensvold (2002), the (more restrictive) metric invariance model should be selected instead of the (less restrictive) congeneric factor invariance model. In other words, the first model comparison provides empirical support for the assumption of equal factor loadings of indicators across countries.

In the next step, a comparison of the model fit of the tau-invariance model (M3) and the metric invariance model (M2) was made. The difference in CFI is minus 0.045 which – obviously – exceeds the critical difference of minus 0.01. So, based on the difference test in CFI, a conclusion can be made that the assumption of equal indicator intercepts across countries is not realistic. The earlier finding that the SRMR exceeded 0.05 also suggests that the assumption of equality of all indicator intercepts is not tenable. The EPC values (i.e. Expected Parameter Values) as reported by Mplus showed that indicator intercepts were expected to change by (maximally) 0.50 when freeing specific indicator intercepts across countries.

An additional CFA model, labelled M3', is shown in Table 5.2. It is a partial tau-invariance model in which the intercept of indicator i081 (i.e. an indicator of F1) and indicator i025 (i.e. an indicator of F2) were allowed to vary across all countries. The intercept of indicator i020 (i.e. an indicator of F7) in Brazil is also different from the corresponding indicator intercept in other countries. The indicator intercepts of these three indicators (i.e. i081, i025, i020) were found to be non-invariant (i.e. biased) across (at least some) countries. The inspection of 'modification indices' (see Exhibit 5.2) calculated when considering the tau-invariance model provided evidence for their non-invariance across countries.

Exhibit 5.2.
Modification indices

Modification indices indicate the expected improvement in the model Chi-square statistic when a constraint on one model parameter (i.e. 1 degree-of-freedom) is released. To release a cross-group restriction on a measurement parameter (i.e. an indicator intercept or a factor loading) signifies that:

(1) The origin of the scale in a particular group is not further required to be identical to its scale used in other groups (if the freed measurement parameter is an indicator intercept) OR
(2) The indicator in this group is not further required to be equally sensitive to changes in the underlying factor across groups as it may be the case in other groups (if the freed measurement parameter is a factor loading).

If a measurement parameter of an indicator is freed across all groups, then it is as if that indicator is no longer used for quantification of the construct. As long as there are still two or more (invariant) indicators left, it is still possible to derive reliable (estimated) factor (mean) scores for a unidimensional factor using a measurement model (see, for example, Bollen and Lennox, 1991). With only one indicator left, the only option is to set the factor score equal to the indicator score (apart from an arbitrarily chosen constant) for all observations. In this case, the estimated measurement error does not provide any substantial improvement in terms of factor score estimation. As a consequence, all factor mean comparisons depend only on that indicator's value.

The partial-tau invariance model (M3' in Table 5.2) shows good global model fit (CFI>=0.90; TLI>=0.90; RMSEA <=0.06; SRMR =0.05). A statistical comparison between this model and the metric invariance model would still favour the metric invariance model (the difference in CFI is minus 0.022) over the partial tau-invariance model. Because of good global model fit (as indicated by all relevant measures of global model fit), the partial tau-invariance model is considered to be an adequate model to describe the data from 16 countries.

In sum, the partial tau-invariance model was selected as a basis for making factor mean comparisons across countries. However, the tau-invariance model will also be considered in this study. This (inadequate) model makes it possible to assess the bias from making unrealistic assumptions about the invariance of all factor loadings and all indicator intercepts across groups (i.e. assumed tau-invariance). The difference in (estimated) factor means as obtained from the (inadequate) tau-invariance model and the (adequate) partial tau-invariance model indicate the extent to which biased results may be obtained when falsely assuming tau-invariance.

The estimated factor means per country as derived from both the tau-invariance model and the partial tau-invariance model are listed in Appendix 5.3. Appendix

5.4 shows the estimated intercepts of indicators that were found to be non-invariant across countries according to the partial tau-invariance model. Appendix 5.5 shows the indicator reliabilities for both the tau-invariance model and the partial tau-invariance model.

5.4.3. Bias due to non-invariance of indicators: Assessing its impact

The partial tau-invariance model (M3') specifies identical factor loadings and indicator intercepts across countries for all indicators measuring the following work environment factors:

(1) Factor 3 (i.e. 'my immediate boss' support'),
(2) Factor 4 (i.e. 'clarity of strategy'),
(3) Factor 5 (i.e. 'confidence in managerial decisions'),
(4) Factor 6 (i.e. 'organisational and managerial efficiency')

As far as these factors are concerned, factor means across countries can be meaningfully compared.

Due to the non-invariance of specific indicator intercepts (i.e. i081, i025, and i020), there may be problems as far as the other work environment factors are concerned:

(1) Factor 1 [i.e. 'fair remuneration'],
(2) Factor 2 [i.e. 'supporting role of people within the department'],
(3) Factor 7 [i.e. 'environmental and societal responsibility']

In the next paragraphs, an assessment is made as to the extent that a cross-country comparison of means for factors 1, 2, and 7 provides (strongly) biased results due to falsely assuming tau invariance of all indicators across countries.

Figure 5.4 provides a pictorial representation of the difference in factor means for F1 as estimated by the tau-invariance model (indicated as as 'TAU') and the partial tau-invariance model (indicated as 'partial TAU').

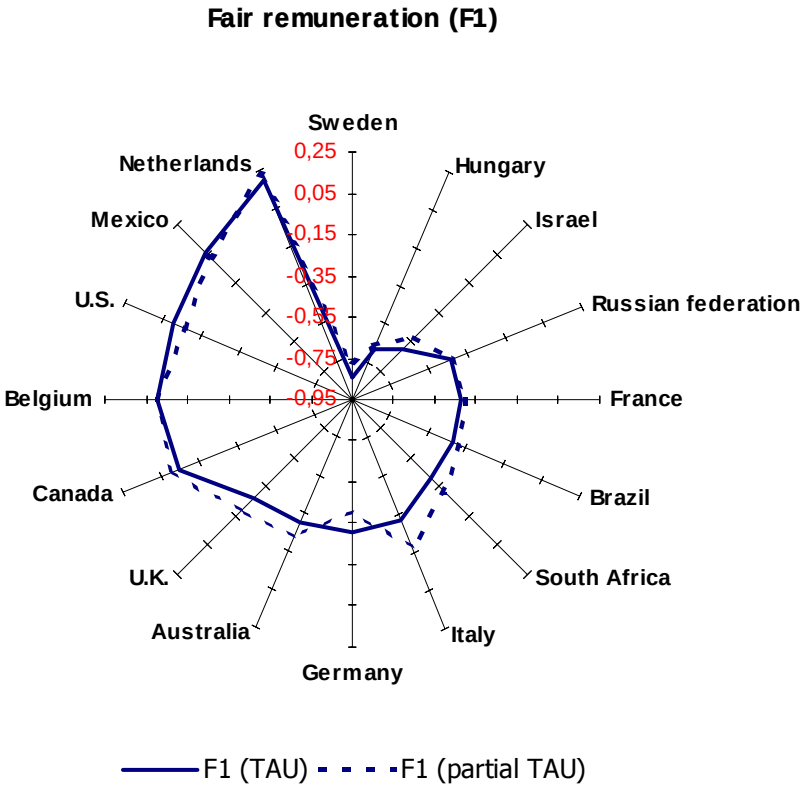


Figure 5.4.
Estimated factor means derived for F1 using two different measurement invariance models

Notes:
(1) The countries are ordered (clockwise) in increasing order of their mean score on F1 as derived from the tau-invariance model;
(2) Belgium = country of reference [i.e. factor mean = 0];
(3) The partial tau-invariance defines a country-specific intercept for indicator i081.

In Figure 5.4, the countries are ordered according to the size of the estimated factor mean under the tau-invariance model. The Pearson correlation coefficients between the estimated factor mean for F1 as derived from both models were very high, specifically 0.98 (i.e. not significantly different from one[58] at $\alpha = 0.10$). This implies that, despite the significant difference in the intercept of indicator i087, both models lead to (almost) identical conclusions regarding a cross-country factor mean comparison on the first factor (i.e. F1). The non-invariant measurement parameter does not lead to a significant bias in a factor mean comparison of F1 across countries.

[58] A z-test is used to test for the significance of the difference between a correlation coefficient and a specified value (e.g. 1) (see Kanji, 1993, p. 34).

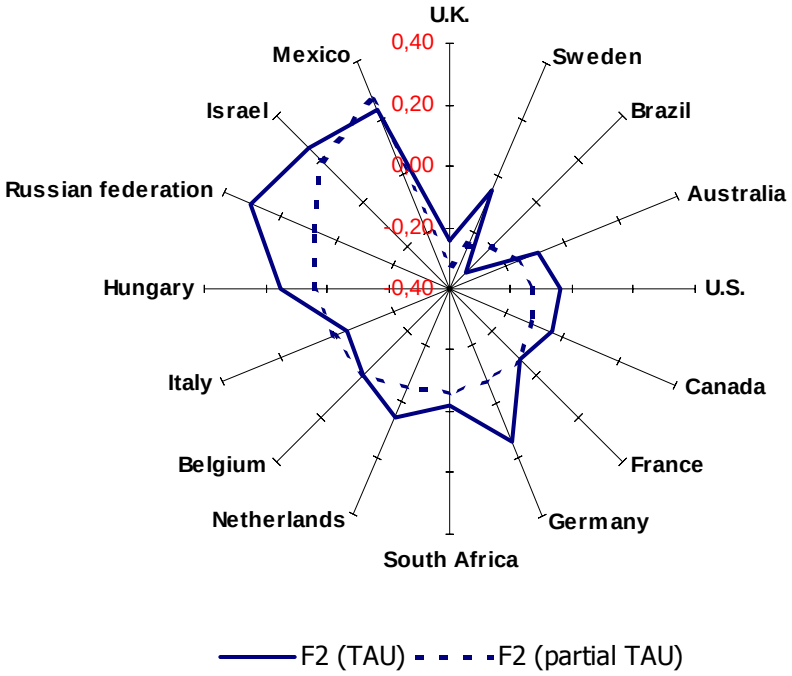


Figure 5.5.
Estimated factor means for F2 using two different measurement invariance models

Notes:
(1) The countries are ordered (clockwise) in increasing order of their mean score on F2 as derived from the tau-invariance model;
(2) Belgium = country of reference [i.e. factor mean = 0];
(3) The partial tau-invariance defines a country-specific intercept for indicator i025.

Figure 5.5 shows the difference in estimated factor means for factor 2 (i.e. F2) using the tau-invariance model and the partial tau-invariance model. The Pearson correlation coefficient between the estimated factor means as derived from both models was 0.84 (i.e. significantly different from one at $\alpha = 0.01$). Cross-country comparisons based on the estimated mean score for factor 2 may differ substantially depending on the model that is used for estimation purposes.

The substantial amount of bias due to the non-invariance of the indicator intercept i025 is remarkable as the relative variation in the non-invariant indicator intercept of F2 (i025) is not much larger than the relative variation in the non-variant indicator intercept of F1 (i.e. i087). The coefficient of variation as reported in Appendix 5.4 (Table A.5.4/2) is 0.058 and 0.052, for F2 and F1 respectively. Figure 5.5 shows that, in an absolute sense, the estimated factor means for F2 are strongly biased for countries like Sweden, Germany, and Russian Federation, while, for other countries, the bias is not that strong (e.g. France, Italy, Israel, Mexico, etc.). Figure 5.4 shows that, as far as F1 is concerned, the bias is relatively small in all countries. This may explain why, compared to the non-invariant indicator i087, the bias in factor score estimation caused by the non-invariant indicator i025 is so strong. Further on in this chapter, some further clarification will be given as to the extent to which a classification of 16 countries (in quartiles) will be misleading for various levels of the correlation between the true model and the approximative model.

Taking into account the better model fit of the partial tau-invariance model, the estimated factor means for F2 as derived from this model are considered to be more trustworthy than the estimated factor means for F2 as derived from the tau-invariance model.

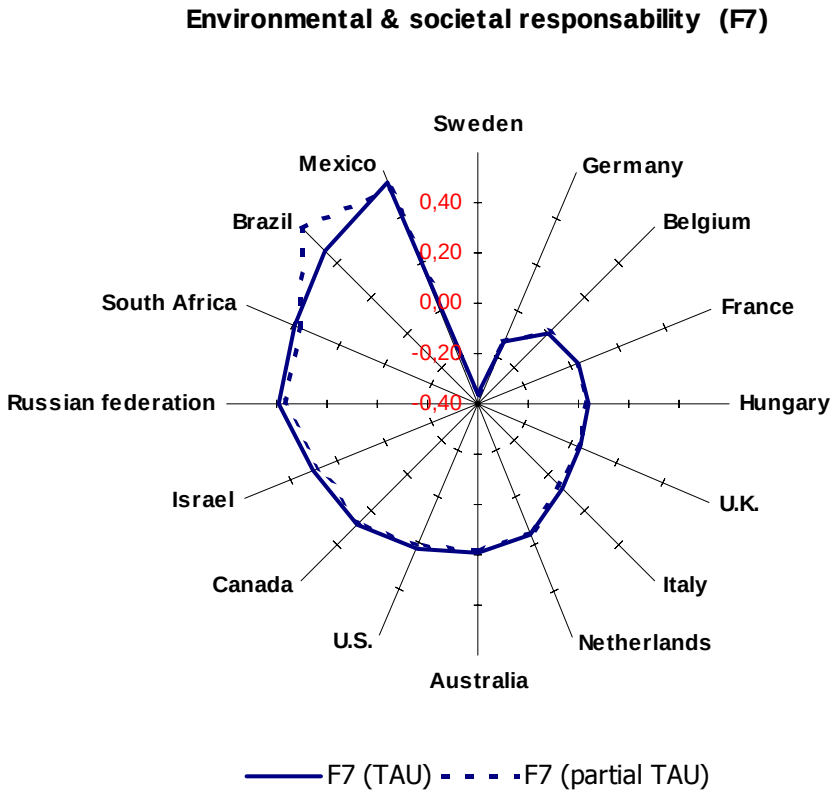


Figure 5.6.
Estimated factor means for F7 using two different measurement invariance models

Notes:
(1) The countries are ordered (clockwise) in increasing order of their mean score on F7 as derived from the tau-invariance model;
(2) Belgium = country of reference [i.e. factor mean = 0];
(3) The partial tau-invariance defines a country-specific intercept for indicator i020.

Factor 7 is measured by only two indicators i019 and i020. All factor loadings and indicator intercepts were found to be equal across countries, except for the intercept of indicator i020 in Brazil. As shown in Figure 5.6, the estimated factor mean for F7 as derived from both competing models were significantly different in Brazil (only). The Pearson correlation coefficient between the estimated factor means as derived from both models was 0.99 (i.e. not significantly different from one at $\alpha = 0.10$). Therefore, it is concluded that the biasing effect of the non-invariant parameter is negligible.

In summary, apart from F2, the estimated factor means are very similar if the tau-invariance model or the partial tau-invariance model is used.

To get a better understanding of the extent to which a relative positioning of 16 countries based on estimated factor mean scores may be misleading, a small simulation study was set up using the software Mplus. In the simulation study, factor means were estimated under the 'true model', as well as under an 'approximative model'. The actual correlation coefficient between the factor means as estimated in both models (at population level) was a design factor in the simulation study. Data for 16 countries were generated. Given the estimated factor means in both models, the countries were classified in quartiles (i.e. top 4 countries, bottom 4 countries, 4 countries just above the median, and 4 countries just below the median). The average 'difference' (across 16 countries) in terms of the assigned quartile of the distribution under the true model and the approximative model served as an indicator of the unreliability of classifications of the approximative model. One hundred replications were simulated for each value of the correlation coefficient between the 'true model' and the 'approximative model'. The results of this simulation study are presented in Table 5.3.

Table 5.3. Results of a small simulation study

Correlation between the true model and the approximative model (at population level)	r=0.80	r=0.85	r=0.90	r=0.95	r=0.99
Average difference computed on the basis of the variables indicating quartile membership under both models (across 100 replications of 16 countries)	0.74	0.67	0.56	0.39	0.20
Standard error	0.24	0.22	0.18	0.15	0.13

Table 5.3 shows that, if the correlation between both models is 0.85, one may expect (on average) to report a misclassification in a higher or lower quartile with every classification of 1.5 countries (i.e. 0.67 * 1.5 = 1.0). As a consequence, 11 out of 16 countries are expected to be missclassified. Obviously, this expectation is only valid as far as the less severe misclassifications are concerned. It is also possible to have less (but more serious) misclassifications. When estimating the factor means for F2, the correlation between the tau-invariance model and the partial tau-invariance model was 0.84 (i.e. close to 0.85). As far as the quantification of F2 is concerned, only four countries would be 'misclassified'. The degree of classification is thus less than may be expected on the basis of a correlation coefficient between both models of about 0.85.

In Appendix 5.6, a supplementary analysis is shown in which the predictive power of the seven work environment factors on job satisfaction were assessed in all countries, individually. The conclusions from this analysis are also included in Appendix 5.6.

5.5. Conclusions

In this research two important research questions were addressed:

(1) Is the assumption of (cross-country) measurement invariance of multi-item scales as used in this global employee opinion survey realistic?

(2) If not (realistic), to which extent does measurement non-invariance across (certain) countries distort (i.e. bias) factor mean comparisons across countries?

The analyses have shown that the assumption of measurement invariance across countries is realistic only as far as (all) factor loadings of indicators are concerned. Based on Meredith's notion of 'measurement invariance' (Meredith, 1993), the invariance of factor loadings across countries is not sufficient as indicator intercepts should be invariant across countries as well. Even though the majority of the indicator intercepts turned out to be invariant across countries, some were clearly non-invariant. Therefore, the conclusion may be that (with respect to the multi-item scales used in this research) the assumption of measurement invariance across countries is not realistic.

When assessing the impact of the non-invariant indicators on factor mean estimations, there is no real problem. For six out of seven factors (i.e. excluding factor 2 [i.e. 'supporting role of people within the department']) the bias due to the non-invariant indicator intercepts is not significant.

As far as this study is concerned, the conclusion is that falsely assuming tau-invariance across countries would not lead to biased results in terms of factor mean comparisons across countries (except for factor means of the second factor). Obviously, a thorough assessment of measurement invariance (across groups), as well as an assessment of (potentially) biased factor mean comparisons due to non-invariance of (measurement) parameters, remains necessary in every new study. The present study has shown how this can be done using a Confirmatory Factor Analysis approach.

Chapter 6. Two additional case studies dealing with international consumer research

"In the ball-room of interpretation the quick-step is off track."

A.Boomsma

6.1. Introduction

Chapter 5 dealt with an international employee survey study. In Chapter 5, the central research question was to investigate whether or not the assumption of measurement invariance of scales across countries is tenable. The results showed that this assumption was not realistic. Next, it was investigated how biased the (estimated) factor mean scores were, when sources of measurement non-invariance were ignored by the researcher. The analyses showed that this type of bias was not substantial for all most factors modelled in the study (i.e. all but one factor).

In this chapter, similar analyses will be run, using different data. The data, which will be used in this chapter, relate to consumers residing in different countries from all over the world. In particular, existing data from two applied consumer research projects will be used. A specialised (global) unit of 'Research International Ltd.', a major market research company, was in charge of both applied consumer research projects. The clients were Unilever (i.e. study no. 1) and another multinational whose identity is not revealed because of confidentiality reasons (i.e. study no. 2). The aim of these studies was twofold:

(1) to investigate the needs of consumers in 'the global market';
(2) to derive 'brand personality profiles' in all participating countries.

In both studies, consumers were segmented into 'global consumer segments' based on their specific consumer needs (e.g. expectations regarding products' performance, and expectations regarding the benefits products will deliver to the consumer).

The clients of these studies were mainly interested in the results from: (1) a (global) consumer segmentation, and (2) a (global) brand mapping exercise. In this Ph.D. research the aim is to examine to which extent *multi-item* scales (as used in these commercial studies) exhibit measurement invariance across countries.

Unfortunately, the questionnaires of both studies did not include many (pre-defined) multi-item scales. This may surprise the reader as many multi-items scales are available for use in marketing and market research. Bruner and Hensel (1997) and Bearden and Netemeyer (1999), for instance, introduced a large number of multi-item scales to marketing and consumer researchers. The reality is, however, that in applied market research (especially in *commercial* market research) multi-item scales are not often used.

Working with multi-item scales implies that: (1) the factors being studied are well-defined, and (2) that multiple items are formulated for each factor under study. These conditions are often not fulfilled in applied consumer research practice. Many items may be included in a questionnaire, but the researcher seldomly makes explicit the mutual relationships between the items (and between the items and the factors they intend to measure). At the very best, factors may be extracted on the basis of a set of items during the analysis phase. An exploratory factor analysis is often used for this purpose. Afterwards, consumers may be segmented into (global) consumer segments based on their factor scores.

Exploratory factor analysis, followed by a segmentation analysis (using the derived factor scores as segmentation variables), is a popular approach in applied consumer research (see, for instance, Wedel and Kamakura, 1998, p. 243 & p. 248; Vriens et al., 1999, p. 20; Lilien and Rangaswamy, 2003, p. 84). Such an approach is problematical as the implicit assumption of cross-country comparability of data (e.g. due to measurement non-invariance of scales/items across countries) is not formally tested. At least, it is recommended to investigate the stability of factor structures across countries before consumers are segmented into (global) consumer segments. As explained in Chapter 1, some exploratory techniques (e.g. Procrustean analysis) may be used for this purpose.

In this research, only those scales were considered for which there were multiple statements included in the questionnaire. In the next section, the two consumer research projects are briefly introduced.

6.2. Two global consumer research projects

Study no.1: 'Pesto' (Unilever, 1999)

Pesto is the name for a global consumer research project conducted by Unilever's SCC category in 1999. SCC stands for 'Spreads and Cooking products Category'. The aim of the study was to propose a market segmentation model for all 'Yellow Fats' (YF) products. The name 'Yellow Fats' is used within Unilever to refer to all spreads and cooking products. More specifically, the objectives of the study were outlined as follows:

(1) to provide a quantified, consumer needs- based market model linking brands, product attributes, consumer attitudes and values.
(2) to create a new positioning map for Unilever and competitor brands that facilitates portfolio and positioning management.
(3) to compare consumers across countries and business regions.
(4) to identify gaps/opportunities in the Unilever portfolio.

(Unilever internal report, 2000)

Sixteen countries from all over the world participated in this global study. The list of countries included: France, Germany, Netherlands, Portugal, Sweden, Spain, United Kingdom, Hungary, Poland, Russia, Turkey, Brazil, Chile, Peru, United States of America, and South Africa. These sixteen countries represented about 75% of the global YF volume. Based on this global research project, the SCC category proposed four 'global brand positionings'. Unilever senior management reduced the size of its YF brand portfolio substantially on the basis of the results from this study.

About 1000 respondents per country were interviewed using face-to-face (i.e. interviewer led) interviews. All respondents were between 18-69 years of age, were mainly responsible for preparing meals, and were users of YF products. The samples were nationally representative in 13 out of 16 countries. In Brazil, Chile, and Peru, only respondents from a couple of big cities participated in the research (i.e. an urban population). In most countries, more than 90 per cent of all respondents were female. Exceptions were: Germany, Sweden, and the USA. In these countries, 20, 20, and 17 per cent of all respondents were male, respectively.

Study no. 2: International consumer study on drinks (i.e. the 'drinks study')

The aim of the second study, an international 'drinks study', was to derive global consumer segments based on specific consumer needs. In addition, the relevance of Liquid Refreshment Beverages (LRB) was assessed with respect to:

(1) specific consumer groups (e.g. consumer groups characterised by demographic or psychographic variables)
(2) specific consumption occasions
(3) specific consumer needs

In addition, brand personality profiles were derived on the basis of the data from this study.

Seven countries participated in this global study. The list of countries included: Brazil, Italy, Saudi Arabia, Thailand, Poland, India, and Hungary. Representative samples from all countries were drawn. The sample size varied between 600 and 1200 respondents. The data for this study were gathered by means of face-to-face (i.e. interviewer led) interviews. Within each country, half of the respondents were male, whereas the other half were female. Forty per cent of all male and female respondents were aged between 16 and 24 years. Sixty per cent were older than 25 years. Young consumers between 16 and 24 years of age are considered to be an important consumer segment as far as Liquid Refreshment Beverages are concerned.

6.3. Method

Multi-item scales

As mentioned in the introduction of this chapter (i.e. Section no. 6.1), only a very limited number of multi-item scales will be used to assess measurement invariance across countries. These multi-item scales are shown in Exhibit 6.1 and Exhibit 6.2.

Exhibit 6.1.
Multi-item measures used in the first study (i.e. the 'Pesto' study)

Factor no. 1: Naturalness of the product

Q16A10* My preferred product# is pure and natural.
Q16A11: My preferred product# is rich in vitamins and minerals.
Q16A28: My preferred product# is made from natural ingredients.

Factor no. 2: Emotional benefit 'feeling like a better cook/host'

Q16A01* My preferred product# makes me feel like a better cook.
Q16A16: My preferred product# makes me feel like a better host.

Factor no. 3: Functional benefits of the product (for one's health)

Q16A03* My preferred product# helps to control cholesterol.
Q16A09: My preferred product# is a low fat product.

The items indicated with an '*' will serve as reference items in the Covariance Structure models presented further on in this chapter. In the context of this study, 'my preferred product' (indicated by '#') refers to a (yellow) 'fat'.

All items are scored on 5-point Likert type of disagree/agree scales.

Exhibit 6.2.
Multi-item measures used in the second study (i.e. the 'drinks study')

Factor no. 1: Added vitamins and minerals

B15*: I would prefer drinks with minerals added.
B16: I would prefer drinks with vitamins added.

Factor no. 2: Carefulness with ingredients

B3*: I read the ingredient labels on food products carefully.
B4: I am grateful to eat and serve drinks that have a lot of good nutrients.

Factor no. 3: Willingness to try new beverages

B28*: I like to try a lot of new and different types of beverages.
B29: I am often the first one I know to try new beverages.

Factor no. 4: Physical activity

P02*: I exercise to maintain my weight.
P06: I do not feel right unless I exercise everyday.
P09: My sports or exercise activity provides me with an energy boost to help me get through the day.

In this study, all items are scored on end-labelled 6-point scales (1=strongly disagree, 6=strongly agree).

Most of the factors, which are shown in Exhibit 6.1 and Exhibit 6.2, are measured by means of two items only. Ideally, more than two items would be included to operationalise each factor (e.g. 5 to 6 items) in the questionnaire. For model identification purposes, it is recommended to include at least three (or four[59]) indicators per factor in the final model (i.e. after items showing non-measurement invariance across countries have been removed from the model).

Statistical analyses

Just as in Chapter 5, (Mean- and) Covariance Structure models will be used to assess measurement invariance of multi-item scales across countries. First of all, a Covariance Structure model (without a mean-structure) is run which imposes the proposed 3- (or 4-) factor structure onto the data. If these covariance structure models are not rejected, a test for measurement invariance across countries can be computed. Measurement invariance testing is done using the hierarchical sequence of model tests which has been proposed earlier on (see Chapter 3 and Chapter 5). The process of model evaluation and model testing is the same as in Chapter 5.

[59] Netemeyer, Bearden, and Sharma (2003, p. 146) argued that at least four indicators per factor are needed as one-factor models with only three indicators are just identified, and little insight is provided for measurement fit (see also Clark and Watson, p. 317). If, however, two or more factors are combined in one structural equation model (i.e. a k-factor model), it is possible to use less than four indicators per factor (e.g. three). Two indicators per factor should be considered an absolute minimum (see Steenkamp & Baumgartner, 1998) in any cross-group analysis.

6.4. Results

6.4.1. The hypothesised 3- (or 4-) factor structure

Study no.1: 'Pesto'

Table 6.1 shows the results of model fit in each individual country. The CFI and TLI scores were sufficiently high (i.e. > 0.90) in all countries, except for Sweden. The Standardised Root Mean Residual (SRMR) was sufficiently small in all countries. Inspection of the (two) RMSEA statistics showed that there were several countries for which the hypothesised 3-factor structure was not plausible (i.e. RMSEA exceeds 0.06). For this reason, a conclusion can be made that the 3-factor structure does not hold for all 16 countries. Therefore, a decision was made to proceed only with a limited number of countries, specifically those countries for which at least one of the RMSEA statistics was smaller than 0.06.

The implication of the decision was that measurement invariance of multi-item scales will only be tested across the following 9 countries: USA, South Africa, Poland, Brazil, Germany, France, Spain, Peru, and Chile.

Table 6.1.
CFA models[1] specifying a three-factor structure (the 'Pesto' study)

Country	Sample size	Chi-square statistic (+ d.f.)	Proba-bility	CFI	TLI	RMSEA[2]	SRMR
All countries	N=16132	655.5 (11)	0.00	0.982	0.966	0.060 / 0.052	0.020
USA	N=994	52.1 (11)	0.00	0.986	0.973	0.061 / 0.047	0.019
South Africa	N=1000	55.8 (11)	0.00	0.983	0.967	0.064 / 0.052	0.024
Turkey	N=1041	74.0 (11)	0.00	0.956	0.915	0.074 / 0.067	0.030
Poland	N=1050	42.8 (11)	0.00	0.984	0.970	0.052 / 0.042	0.022
Brazil	N=1000	49.6 (11)	0.00	0.984	0.969	0.059 / 0.049	0.021
Russian Federation	N=1011	79.7 (11)	0.00	0.959	0.922	0.079 / 0.069	0.033
Germany	N=1003	63.8 (11)	0.00	0.977	0.957	0.069 / 0.059	0.035
Netherlands	N=1020	64.8 (11)	0.00	0.976	0.954	0.069 / 0.060	0.028
France	N=991	41.8 (11)	0.00	0.984	0.969	0.053 / 0.045	0.024
United Kingdom	N=1007	106.5 (11)	0.00	0.960	0.923	0.093 / 0.079	0.035
Sweden	N=1000	126.6 (11)	0.00	0.943	0.890	0.102 / 0.086	0.039
Spain	N=1005	28.2 (11)	0.00	0.989	0.980	0.039 / 0.031	0.020
Peru	N=1000	62.5 (11)	0.00	0.961	0.926	0.068 / 0.057	0.029
Chile	N=1007	40.9 (11)	0.00	0.985	0.972	0.052 / 0.042	0.023
Portugal	N=1000	66.9 (11)	0.00	0.970	0.942	0.071 / 0.060	0.031
Hungary	N=1003	78.7 (11)	0.00	0.967	0.937	0.078 / 0.069	0.033

Notes:
(1) [1]These models do not include a factor mean structure;
(2) [2]The second figure for RMSEA measure is based on Satorra-Bentler's scaled Chi-square statistic which corrects for model misfit due to non-normal distributed data.

Study no. 2: the 'drinks' study

Similarly, the plausibility of the 4-factor structure was tested using the data from the 'drinks study'. The results are displayed in Table 6.2.

Table 6.2.
CFA models[1] specifying a four-factor structure (the 'drinks study')

Country	Sample size	Chi-square statistic (+ d.f.)	Proba-bility	CFI	TLI	RMSEA[2]	SRMR
All countries	N=6850	86.3 (21)	0.00	0.996	0.993	0.021 / 0.018	0.011
Brazil	N=1063	63.1 (21)	0.00	0.978	0.963	0.043 / 0.039	0.024
Italy	N=1191	67.5 (21)	0.00	0.984	0.973	0.043 / 0.039	0.023
Saudi Arabia	N=1006	97.3 (21)	0.00	0.974	0.956	0.059 / 0.053	0.035
Thailand	N=1094	99.4 (21)	0.00	0.958	0.929	0.058 / 0.046	0.030
Poland	N=1035	104.0 (21)	0.00	0.964	0.938	0.062 / 0.054	0.027
India	N=593	69.4 (21)	0.00	0.954	0.921	0.062 / 0.052	0.037
Hungary	N=868	55.3 (21)	0.00	0.982	0.969	0.043 / 0.038	0.029

Notes:
(1) [1]These models do not include a factor mean structure;
(2) [2]The second figure for RMSEA measure is based on Satorra-Bentler's scaled Chi-square statistic which corrects for model misfit due to non-normal distributed data.

The results presented in Table 6.2 do not give rise to believe that the 4-factor structure would not hold in any of the seven participating countries (i.e. CFI>0.90; TLI>0.90; at least one RMSEA<0.06; SRMR<0.05). As a consequence, measurement invariance of these scales will be tested across all seven countries.

6.4.2. Measurement invariance of multi-item scales

Study no.1: 'Pesto'

Table 6.3 shows the MACS model comparisons using the data from 9 countries from the 'Pesto' study. All model fit statistics of the congeneric factor invariance model showed a good fitting model (CFI>0.90; TLI>0.90; RMSEA<0.06; SRMR<0.05). Based on the chi-square difference statistic, the metric invariance model would be rejected in favour of the congeneric factor invariance model. However, when the difference in CFI was computed and compared with the critical value of -0.01 (Cheung and Rensvold, 2002), the metric invariance model was accepted (see Table 6.3). As shown in Table 6.3, the model fit statistics of the metric invariance model indicated a good fitting model. In the second model comparison, the tau-invariance model had to be rejected in favour of the metric invariance model. The difference in CFI equaled –0.024, which is more negative than the critical value for this difference (i.e. –0.01). Inspection of the model fit statistics of the tau-invariance model also led to the conclusion that this model is not plausible. Attempts to release some parameter constraints across groups (after inspecting the modification indices) did not lead to a good fitting model which is 'close to' the tau-invariance model. It is concluded that the multi-item scales do not exhibit the psychometric conditions which are required for meaningfully comparing (estimated) factor mean scores across countries.

Table 6.3.
Multi-group CFA models[1] to test for measurement invariance across 9 countries (the 'Pesto' study)

ESTIMATED CFA MODELS							
	Sample size	Chi-square statistic (+ d.f.)	Proba-bility	CFI	TLI	RMSEA[2]	SRMR
M1: Congeneric factor invariance model	N=9050	437.6 (99)	0.00	0.982	0.966	0.058 / 0.048	0.025
M2: Metric invariance model	N=9050	521.8 (131)	0.00	0.980	0.971	0.054 / 0.046	0.032
M3: Tau-invariance Model	N=9050	1249.1 (163)	0.00	0.956	0.949	0.071 / 0.063	0.042

Notes:
(1) M1: no parameters are fixed across countries;
(2) M2: all factor loadings are fixed across countries;
(3) M3: all factor loadings and indicator intercepts are fixed across countries.

Table 6.3. (continued)
Multi-group CFA models[1] to test for measurement invariance across 9 countries (the 'Pesto' study)

DIFFERENCE TESTS BETWEEN NESTED CFA MODELS

	Sample size	Difference in Chi-square statistic (+ d.f.)	Proba-bility	Diffe-rence in CFI			
Difference between M2 and M1	N=9050	84.2# (32)	0.00	-0.002	-	-	-
Difference between M3 and M2	N=9050	727.3# (32)	0.00	-0.024	-	-	-

Notes:
(1) [1]These models do not include a factor (and item) mean structure;
(2) [2]The second figure for RMSEA measure is based on Satorra-Bentler's scaled Chi-square statistic which corrects for model misfit due to non-normal distributed data;
(3) #An alternative to computing the classical Chi-square difference statistic is to calculate a difference statistic which is based on the Satorra-Bentler scaled Chi-square statistic (see Satorra, 2000, and Satorra and Bentler, 1999). The Satorra-Bentler difference Chi-square statistic (as reported in these papers) follows a Chi-square distribution, asymptotically. For these model comparisons this difference statistic is: 70.6 (model M2 versus model M1), 402.9 (model M3 versus model M2). Changing to this difference statistic does not lead any major changes in the conclusions as all related probabilities are also equal to 0.00. The formula for calculating the Satorra-Bentler chi-square difference test is: $(SB\_X^2_n - SB\_X^2_c) / \{(DF\_SB_n * SF_n - DF\_SB_c * SF_c) / (DF\_SB_n - DF\_SB_c)\}$ where: SB_X^2 indicates the Satorra-Bentler corrected X^2 statistic, DF_SB the degrees of freedom, and SF the Satorra-Bentler scaling factor (as reported by Mplus). The subscripts '$_n$' and '$_c$' are used to refer to the nested (i.e. most restricted) model and the comparison model (i.e. the less restricted model).

Study no. 2: The 'drinks study')

The model fit statistics of the congeneric invariance model (as displayed in Table 6.4) showed that the congeneric factor invariance model is a good fitting model (CFI>=0.973; TLI=0.953; RMSEA=0.053 (or 0.046); SRMR=0.029). The first model comparison favoured the metric invariance model over the congeneric factor invariance model (i.e. the difference in CFI=-0.007). The model fit statistics of the metric invariance model also showed a good model fit (CFI=0.966; TLI=0.951; RMSEA=0.055 (or 0.047); SRMR=0.035). The second model comparison showed that the tau-invariance model was rejected in favour of the metric invariance model. Inspection of the model fit statistics of the tau-invariance model also led to the conclusion that this model was not plausible. Just as in the 'Pesto study', attempts to find a better fitting model which is 'close to' the tau-invariance model were not successful. The conclusion is, once again, that the multi-item scales do not exhibit the psychometric conditions which are required for meaningfully comparing (estimated) factor mean scores across countries.

Table 6.4.
Multi-group CFA models[1] to test for measurement invariance across 7 countries (the 'drinks study')

ESTIMATED CFA MODELS							
	Sample size	Chi-square statistic (+ d.f.)	Proba-bility	CFI	TLI	RMSEA[2]	SRMR
M1: Congeneric factor invariance model	N=6850	556.0 (147)	0.00	0.973	0.953	0.053 / 0.046	0.029
M2: Metric invariance model	N=6850	694.4 (177)	0.00	0.966	0.951	0.055 / 0.047	0.035
M3: Tau-invariance Model	N=6850	1249.1 (207)	0.00	0.931	0.915	0.072 / 0.064	0.045

Notes:
(1) M1: no parameters are fixed across countries;
(2) M2: all factor loadings are fixed across countries;
(3) M3: all factor loadings and indicator intercepts are fixed across countries;

Table 6.4. (continued)
Multi-group CFA models[1] to test for measurement invariance across 7 countries (the 'drinks study')

DIFFERENCE TESTS BETWEEN NESTED CFA MODELS

	Sample size	Difference in Chi-square statistic (+ d.f.)	Proba-bility	Diffe-rence in CFI			
Difference between M2 and M1	N=6850	138.4# (30)	0.00	-0.007	-	-	-
Difference between M3 and M2	N=6850	554.7# (30)	0.00	-0.035	-	-	-

Notes:
(1) [1]These models do include a factor (and item) mean structure;
(2) [2]The second figure for RMSEA measure is based on Satorra-Bentler's scaled Chi-square statistic which corrects for model misfit due to non-normal distributed data;
(3) #An alternative to computing the classical Chi-square difference statistic is to calculate a difference statistic which is based on the Satorra-Bentler scaled Chi-square statistic (see Satorra, 2000, and Satorra and Bentler, 1999). The Satorra-Bentler difference Chi-square statistic (as reported in these papers) follows a Chi-square distribution, asymptotically. For these model comparisons this difference statistic is: 94.9 (model M2 versus model M1), 471.8 (model M3 versus model M2). Changing to this difference statistic does not lead any major changes in the conclusions as all related probabilities are also equal to 0.00. The formula for calculating the Satorra-Bentler chi-square difference test is: $(SB\_X^2_n - SB\_X^2_c) / \{(DF\_SB_n * SF_n - DF\_SB_c * SF_c) / (DF\_SB_n - DF\_SB_c)\}$ where: SB_X^2 indicates the Satorra-Bentler corrected X^2 statistic, DF_SB the degrees of freedom, and SF the Satorra-Bentler scaling factor (as reported by Mplus). The subscripts '$_n$' and '$_c$' are used to refer to the nested (i.e. most restricted) model and the comparison model (i.e. the less restricted model).

6.5. Conclusions

In this chapter, measurement invariance of multi-item scales as used in two global consumer studies was assessed. The results have shown that the multi-item scales do not meet the necessary criteria of tau-invariance across countries. Instead, evidence was found for a weaker form of measurement invariance: metric invariance. As concluded from the simulation research presented in Chapter 4, this weaker form of measurement invariance is not sufficient to compare (estimated) factor mean scores across countries.

Chapter 7. General conclusions and discussion

"Criticism is the most powerful weapon in any methodology of science."

P.B. Medawar

"To avoid criticism do nothing, say nothing, be nothing"

E. Hubbard (1856-1915),

American editor/publisher & writer

In this dissertation, the question was addressed as to how meaningful (and reliable) cross-country comparisons are when such comparisons are based on (estimated) factor mean scores at country-level. A comparison of factor mean scores across countries is very common in international management research. International management research comprises a very wide spectrum of areas of research (e.g. globalisation of the marketing mix, organisational behaviour, cross-cultural communication, etc.). A discussion as to how international comparisons are made within all of these areas would not be feasible. For this reason, the focus in this dissertation was on two particular areas of international management research: global consumer research and global research in international Human Resource (HR) management. These areas were selected as they are of particular importance to multinational companies.

The main methodological objective of this dissertation was to investigate the extent to which violations of the principle of measurement invariance across groups (e.g. nations or cultures) lead to wrong conclusions regarding factor mean comparisons across groups. To date, there is no consensus in the methodological literature as to what level of measurement invariance across groups is required before such cross-group comparisons at factor-level are meaningful. Some authors, for instance William Meredith (1993), firmly stated that factor loadings and indicator intercepts of all indicators need to be invariant across groups (i.e. 'tau-invariance' across groups). Others have advocated less stringent conditions, such as the invariance of factor loadings across groups (i.e. 'metric invariance' across groups) (e.g. Duane Alwin and David Jackson, 1981) or the invariance of only a subset of all factor loadings across groups (i.e. 'partial metric invariance') (e.g. Barbara Byrne et al., 1989). References to many papers were provided in Section 3.1 of Chapter 3. Metric (or partial metric) invariance across groups does not require the invariance of indicator intercepts across groups.

As indicated in Chapter 1, the effect of measurement non-invariance on factor mean comparisons across groups was assessed using a Confirmatory Factor

Analysis framework. Two important assumptions were made. First of all, it was assumed that all indicators (i.e. observed variables) are metric, at least from an analysis point of view (see Chapter 3). This assumption is in line with the theory of asymptotic robustness of normal-theory based estimators, as advocated by, for instance, Michael Browne and Albert Satorra (for references: see Chapter 3). Secondly, it was assumed that all indicators were assumed to be 'reflective indicators' of the construct/factor they intended to measure. As explained in Chapter 2, this signifies that the construct/factor is seen as the 'common cause' of all indicators which are assumed to measure that particular factor.

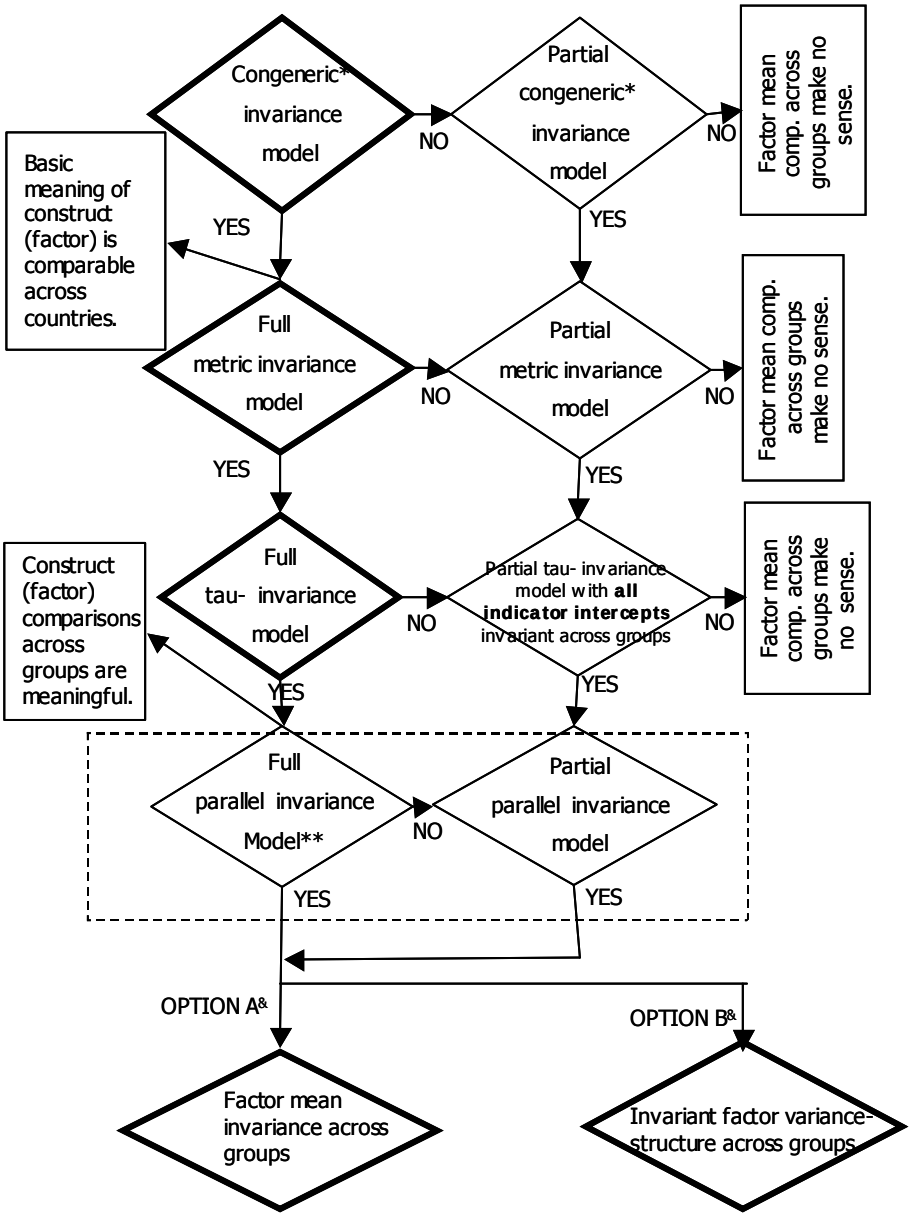
The evaluation as to how threatening violations of the measurement invariance principle are, was based on a simulation study (i.e. Chapter 4). In addition, case studies were presented in Chapter 5 and Chapter 6. The case study in Chapter 5 concerned a global employee survey study (i.e. HR management), whereas the two case studies in Chapter 6 concerned international consumer studies.

The simulation study in Chapter 4 showed that invariance of indicator intercepts across groups is an important requirement whenever the researcher intends to make cross-country comparisons based on (estimated) factor mean scores. In the simulation study small percentages of correct statistical conclusions based on the factor mean difference test were reported for cases in which one indicator intercept is non-invariant across countries. Other factors such as non-invariance of factor loadings and sample size (per group) were also found to have an impact on the correctness of the factor mean difference test. The impact of both of these factors was, however, much smaller than the impact of a non-invariant indicator intercept.

In the simulation study, determinants of the (degree of) robustness of simulated non-invariance conditions were also assessed. Robustness of a non-invariance condition was operationalised as follows: if the percentage (number) of correct statistical conclusions of the non-invariance condition does not differ significantly from the percentage (number) of correct statistical conclusions of the corresponding full invariance condition, the non-invariance condition is considered to be robust (see Chapter 4). Otherwise, the non-invariance condition is not considered to be robust. The results of the simulation study have shown that a non-invariant indicator intercept across groups exerted a strong negative influence on the robustness of the non-invariance condition (i.e. conditions with a non-invariant indicator intercept and/or a non-invariant factor loading). Based on the results of the simulation study, it is concluded that non-invariance of an indicator intercept is the most important factor in determining the robustness of the non-invariance condition.

In sum, the simulation study has shown that non-invariance of the indicator intercept across groups has a large influence on both the percentage of correct statistical conclusions of the factor mean difference test, and the robustness of non-invariance conditions. This implies that the condition of partial tau-

invariance of indicators across groups forms a serious threat when comparing (estimated) factor means across groups. As a consequence, Figure 3.1 should be modified. The modified figure is shown in Figure 7.1.



(Continuation of this figure: see Figure 3.2)

Figure 7.1.
Modified hierarchical sequence of MACS model tests (recommended sequence)

Notes (with Figure 7.1.):

A dashed box indicates an optional hypothesis test (conduct this test if it is relevant from theory)

*Also referred to as 'configural invariance model';

**Two different versions of the parallel invariance test have been presented in the text (one being more restrictive than the other).

As shown in Figure 7.1, factor mean comparisons are not considered reliable if one (or more) indicator intercepts substantially differ(s) across countries. In addition, case studies were analysed in Chapter 5 and Chapter 6.

The case study in Chapter 5 concerned an international employee survey study. The analyses in Chapter 5 showed that, strictly speaking, three out of seven factors did not exhibit full tau-invariance across countries. A partial tau-invariance model (with some non-invariant indicator intercepts across countries) seemed to be a more realistic model. Additional analyses showed, however, that falsely assuming tau-invariance across countries would not lead to biased results in terms of factor mean comparisons across countries (except for factor means of the second factor), at least not for the dataset analysed in Chapter 5.

Additional case studies were presented in Chapter 6. These case studies concerned two international consumer segmentation studies. The analyses showed that, in both consumer studies, indicators of the constructs (under study) did not exhibit measurement invariance across countries.

As explained in Chapter 6, only a very limited number of multi-item scales were available for analysis. In consumer segmentation studies, and in particular: in commercial consumer segmentation studies, multi-item scales are not that common. Practical limitations such as budget constraints and limitations in terms of the interviewing time often lead to a decision to select only one item to measure a particular construct. One typically uses a 'battery of items'. Often each item in the battery measures a completely different aspect of the phenomena being studied (e.g. a specific consumer need, a specific consumer attitude, a specific consumer value, etc.). In the case studies presented in Chapter 6, there were only a very limited number of constructs which were measured by more than one item. In these cases, the number of items did not exceed two (most of the times).

The fact that indicators used in the two international consumer studies did not exhibit measurement invariance across countries did not surprise the author. International consumer segmentation studies (as executed by market research agencies) are typically 'ad-hoc studies' for which there are generally no 'standard scales' available. This is an important point of difference with global employee opinion surveys. In employee opinion surveys, research agencies typically set up questionnaires using a large pool of items which have been used

repeatedly in previous studies. In some cases, these items form (true) multi-item scales. In addition, one may expect that the 'factorial structure' of these items is well-understood by the researchers as: (1) the items have repetitively been used as input for exploratory factor analyses, and (2) the correlational structure of these items have often been investigated as well. This is certainly not the case for the international consumer studies as they are typically designed to be conducted only once. Practitioners could try to improve the quality of the scales used by following 'best-practice guidelines' for scale development. Examples can be found in: Hinkin (1995, 1998), DeVellis (2003), and Netemeyer et al. (2003).

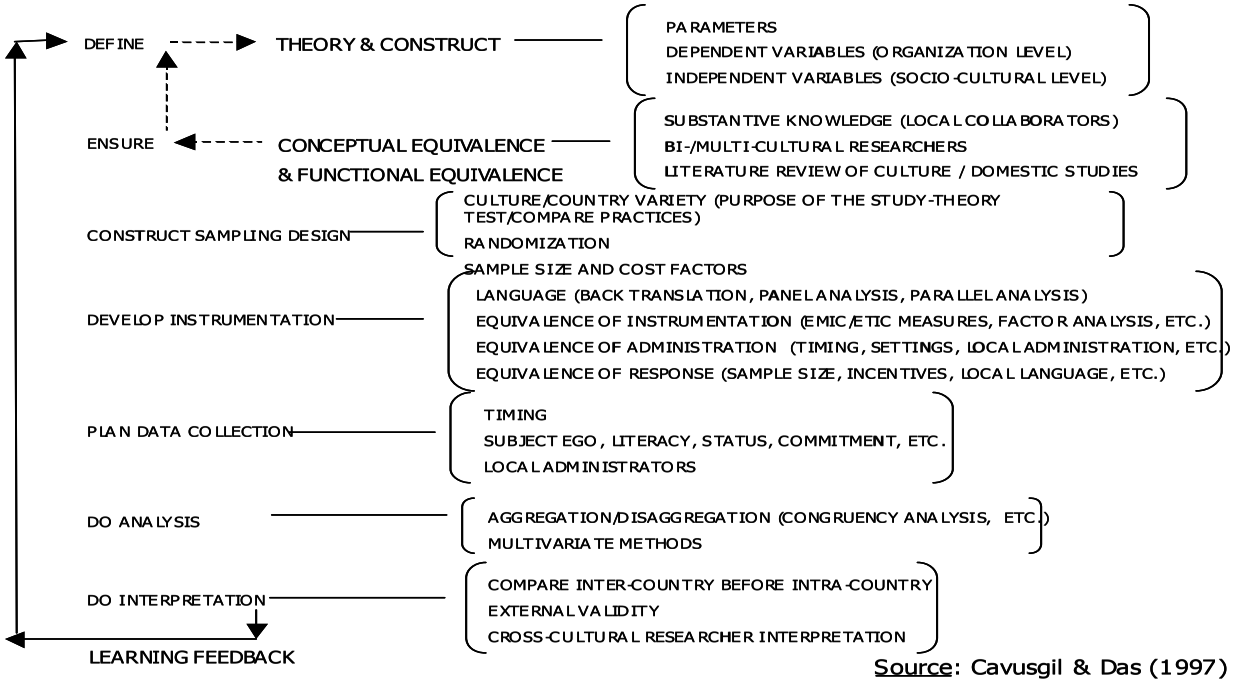
Obviously, drawing general conclusions based on only two case studies should be done with a skeptical eye. The author's reflections concerning international consumer studies and employee opinion studies are based on: (1) the author's working experience in the (market) research industry, and (2) empirical evidence from the case studies presented in this dissertation. There is no absolute guarantee that other researchers would concur with these reflections concerning common practices in applied research.

This dissertation has shown that using multi-item scales which meet high psychometric measurement properties (i.e. ideally showing tau-invariance across groups) is a prerequisite for making meaningful cross-country comparisons based on (estimated) factor mean scores. Whenever multi-item measurement scales do not exhibit these high measurement properties (in particular: tau-invariance across countries), it is critical to evaluate what risk is involved in terms of making wrong statistical conclusions based on factor mean comparisons across countries. An inspection of the simulation results (as presented in Chapter 4), and a bias assesment (as proposed in Chapter 5) provide useful tools for making such an evaluation.

Appendices

Appendix 1.1. Cavusgil and Das' framework

(for creating sound cross-cultural research methodology designs)



Source: Cavusgil & Das (1997)

Appendix 1.2. Basic IRT models

(Source: Van Zessen & De Beuckelaer, 2000)

Table A.1.2/1.
Basic Item Response Models

	Normal Ogive model by Lord (1952)	Two-Parameter Logistic model by Birnbaum (1957)	One-Parameter Logistic model by Rasch (1960)	Three-Parameter Logistic model by Birnbaum (1968)
Underlying Distribution	Normal	Logistic	Logistic	Logistic
Adjustment with regard to the 3-parameter logistic model	Not applicable	$\gamma_i=0$	$\alpha_i=a$ $\gamma_i=0$	None
Number of parameters	2	2	1	3
Parameters	α_i, β_i	α_i, β_i	β_i	$\alpha_i, \beta_i, \gamma_i$

Statistical models:

Normal ogive model by Lord (1952):

$$P_i(\theta) = \int_{-\infty}^{\alpha_i(\theta-\beta_i)} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz$$

Two-Parameter Logistic model by Birnbaum (1957):

$$P_i(\theta) = \frac{e^{\alpha_i(\theta-\beta_i)}}{1 + e^{\alpha_i(\theta-\beta_i)}} \text{ with } \alpha_i > 0$$

One-Parameter Logistic model by Rasch (1960):

$$P_i(\theta) = \frac{e^{(\theta-\beta_i)}}{1 + e^{(\theta-\beta_i)}} \text{ with } \sum \beta_i = 0$$

Three-Parameter Logistic model by Birnbaum (1968):

$$P_i(\theta) = \gamma_i + (1 - \gamma_i)\frac{e^{\alpha_i(\theta-\beta_i)}}{1 + e^{\alpha_i(\theta-\beta_i)}} \text{ with } \alpha_i > 0 \text{ and } 0 \leq \gamma_i < 1$$

Notation used:

- α: discrimination parameter; this parameter indicates how well the ICC discriminates between people with adjacent positions on the latent variable (i.e. latent trait);
- β: location parameter; this parameter refers to the level of difficulty of an item or the 'positiveness' of an item;
- γ: the pseudo-chance level (sometimes named: guessing parameter), a lower boundary. Even with a very low value on the latent variable (trait) the probability for a positive answer is at least γ;
- i: subscript i is used to refer to a particular item (in the questionnaire);
- z: a normal deviate from a distribution with mean β_i and standard deviation $1/\alpha_i$;
- Θ: the value on the latent variable (trait).

$P_i(X=k|\Theta)$, or shortly $P_i(\Theta)$: the probability that a random selected subject with a value on the latent variable (trait) equal to Θ gives a 'favourable' answer to item i (i.e. agreeing with item i).
Recall that all of the basic IRT models assume dichotomous (i.e. 0/1) response categories.

Appendix 2.1. Examples of latent constructs in the management literature

'Consumer Ethnocentrism (CETSCALE)' (Shimp and Sharma, 1987; see also Yoo, 2002)
(with indicators: 17 statements)

'General attitude toward advertising' (Durvasula et al., 1993)
(indicators of general attitude toward advertising: good/bad; positive/negative; favourable/unfavourable [all sematic differential pairs])

'Physical distribution service quality' (consisting of 6 unidimensional subfactors) (Bienstock et al., 1997; see also Bearden and Netemeyer, 1999)
(with indicators: 30 statements)

'Vanity: 4 trait aspects of vanity [i.e. 4 unidimensional factors]' (Netemeyer et al., 1995; see also Bearden and Netemeyer, 1999)
(with indicators: 21 statements in total)

'Buying impulsiveness scale' (Rook and Fisher, 1995; see also Bearden and Netemeyer, 1999)
(with indicators: 9 items)

'Opinion leadership and information seeking' (Reynolds and Darden, 1971)
(with indicators: 8 items in total [i.e. 5+3])

'Consumer innovativeness: 2 subfactors: consumer independent judgment-making and consumer novelty seeking' (Manning, Bearden, and Madden, 1995)
(with indicators: 14 items [i.e. 6+8])

Appendix 2.2. Examples of emergent constructs in the management literature

(this overview is partly based on Diamantopoulos and Winklhofer, 2001)

'Group heterogeneity' (Jarley et al., 1997)
(with indicators: difference on race; difference on gender; and difference on occupation)

'Job embeddedness' (Mitchell et al., 2001)
(with indicators: attain fit; form linkages; and make sacrifices regarding the organisation and the community)

'Career success' (Judge & Bretz, 1994)
(with indicators: salary, job level, and number of promotions)

'Market orientation' (Sandvik and Sandvik, 2003; Jaworski and Kohli, 1993)
(with indicators: 32 items)

'Perceived coercive power' (in a marketing channel) (Gaski and Nevin, 1985)
(with indicators: delay delivery; delay warranty claims; take legal action; refuse to sell; charge high prices; deliver unwanted products)

'Advertising expenditures' (bank) (McKee et al., 1989)
(with indicators: television; radio; newspaper; all media in total)

'Convenience' (shopping) (Lumpkin and Hunt, 1989)
(with indicators: delivery to home; telephone in order; transportation to store; convenient parking; location close to home; variety of stores close together)

'Company resource sharing' (Burke, 1984)
(with indicators: plant and equipment; production personnel; sales force; distribution channels; management services; research and development facilities; research and development personnel)

'Health information sources' (nonpersonal) (Moorman and Matulich, 1993)
(with indicators: advertisements; books, magazines, or pamphlets about health; newspapers; television and radio programming; product labels)

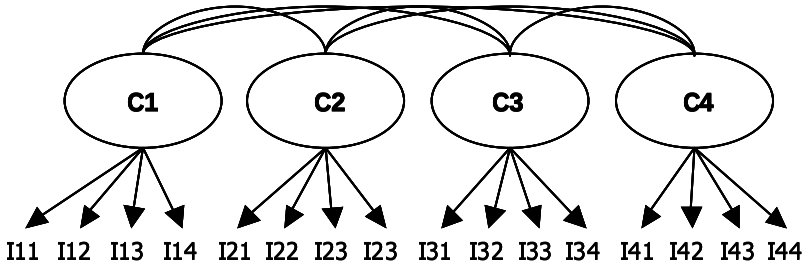
'Ecological awareness' (Richins and Dawson, 1992)
(with indicators: recycle newspapers used at home; recycle glass jars and bottles used at home; intentionally eat meatless meals; contribute to ecological or conservation organisations)

'Company reputation' (Goldberg and Hartwick, 1990)
(with indicators: with its employees; with financial investors; with the U.S. public; with the Canadian public)

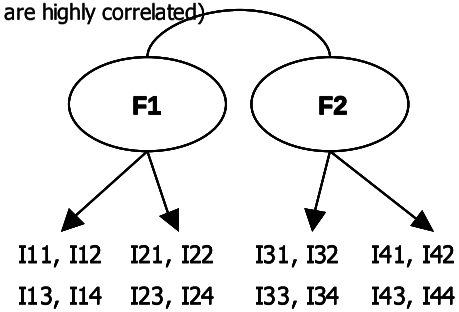
'Coviewing television' (parent/child) (Carlson and Grosshart, 1988)
(with indicators: watch television with my children on weekdays; Saturdays; Sundays)

Appendix 2.3. Bagozzi and Edwards' measurement models

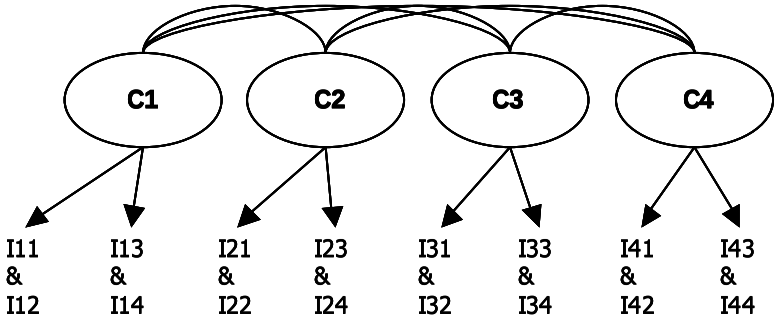
Model A: Total disaggregation model



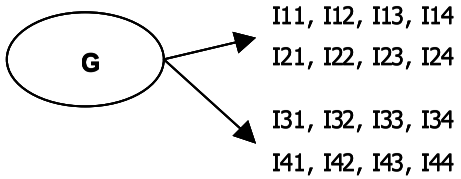
Model C: Partial aggregation model
(assumes that pairs C1 and C2 and C3 and C4 are highly correlated)



Model B: Partial disaggregation model



Model D: Total aggregation model
(assumes F1 and F2 are highly correlated)



Note: **C**i = component of scale; **I**ij = item j of component i; **F**k = facet of scale; **G**= overall scale

Source: Bagozzi & Edwards (1998)

Appendix 2.4. Constructs wrongly perceived as latent constructs

Cohen et al. (1990) argued that the following constructs were wrongly* perceived as latent constructs (*They are more adequately described as emergent constructs).

'Stressful change events'
(with indicators: been sexually attacked; family and parent stress; accident and illness events; family relocation events)

'Life change'
(with indicators: number of undesirable events; number of events producing at least moderate life change)

'Illness'
(with indicators: number of illnesses; respiratory problems or illnesses)

Appendix 2.5. MIMIC modelling: An alternative to multigroup MACS modelling?

Bengt Muthén introduced the MIMIC model in 1989 (Muthén, 1989). Some authors have used the MIMIC model (Multiple Indicators / Multiple Causes) as an alternative to the MACS model when comparing latent factor means across populations (see, for instance, Rubio et al. [2003] and Rivera & Satorra [2002]).

In the MIMIC approach, a global analysis is conducted (i.e. using data from multiple populations) rather than separate analyses (i.e. using data from each population). Part of the MIMIC model is an ordinary (latent) factor model including multiple indicator variables. When dealing with two populations, an extra indicator variable is included to specify to which population each observation belongs. In cases in which the number of populations exceeds two, say K (K>2), K-1 indicator variables are used to specify population membership. A direct path is specified from these (extra) indicator variable(s) to the latent variable. These direct paths model differences in latent factor means across populations. Significant path coefficients provide strong empirical support to assume that the (estimated) factor mean differences across populations are substantial. From a 'causal perspective' these indicator variables may be considered to be 'causes' of the latent variable (i.e. the 'multiple causes' part of the MIMIC model).

A major shortcoming of the MIMIC model is that the (latent variable) indicators are implicitely assumed to exhibit 'measurement invariance' across populations. The reason is that all model parameters which are related to the measurement model (i.e. factor loadings, indicator intercepts, unique variances) are estimated using all data from multiple populations. It is, however, possible to test for the stability of some model parameters across multiple populations. For this purpose, a direct path would be specified from the indicator variables (used to model population membership) to the individual indicators of the latent variable. A significant path would indicate that that particular indicator (of the latent variable) does not exhibit measurement invariance across populations. The problem with this approach is that only non-invariance due to differences in indicator intercepts (across populations) can be detected. Differences in factor loadings of indicators across populations cannot be detected when adopting a MIMIC modelling approach. In the author's Ph.D. dissertation, measurement non-invariance across groups has been operationalised as non-invariance of indicator intercepts **and** / or non-invariance of factor loadings. The inadequacy of the MIMIC model to detect differences in factor loadings across groups explains why the author did not really consider the MIMIC modelling approach to be a suitable alternative to the multigroup MACS approach (for instance in the simulation study presented in Chapter 4).

Appendix 4.1. Experimental plan

F0: Two separate simulation studies are run (i.e. factor 0). In study no. 1 three indicators are used to score the factor; in study no. 2 four indicators are used.				
Factor name	Levels	Number of possible values	Implemented in ...	Additional remarks
F1: Type of threshold model (*) (used to generate ordinal category scores)	5-point scale **1**: Standard normal distribution (i.e. no threshold model) **2**: Uni-modal (4), left-skewed distribution. **3**: Bi-modal (2&4), symmetrical distribution; For details: see additional remarks.	3	Mplus model specification file	(at level 2): Thresholds are chosen so that: 8% score 1; 12% score 2; 15% score 3; 40% score 4; 25% score 5; The actual threshold values are: -1.40; -.84; -.38; +.68 (i.e. z-scores). (at level 3): Thresholds are chosen so that: 15% score 1; 30% score 2; 10% score 3; 30% score 4; 15% score 5; The actual threshold values are: -1.04; -.13; +.13; +1.04 (i.e. z-scores).
F2: No. of observations / group	Group1 Group2 **1**: 200 200; **2**: 300 300; **3**: 400 400; **4**: 500 500; **5**: 750 750; **6**: 200 400	6	Mplus model specification file	Except for the sixth condition all conditions represent balanced conditions (i.e. the same number of observations in both groups).

Factor name	Levels	Number of possible values	Implemented in ...	Additional remarks
F3: Factor mean difference	**1**: Factor mean (G2)= -0.30; **2**: Factor mean (G2)= -0.15; **3**: Factor mean (G2)= 0.00; **4**: Factor mean (G2)=+0.15; **5**: Factor mean (G2)=+0.30. Note: Factor mean (G1)= 0.00 (for all levels)	5	Mplus data generation file	
F4: Inequality of factor loadings across groups	In G1 the factor loadings are fixed as follows: λ_1: 0.7; λ_2: 0.6; λ_3: 0.6; λ_4: 0.5 (=λ_3 in study no. 2 [with 3 indicators]). In G2 the factor loading of the second indicator (i.e. λ_2) equals: 0.4 (exp. cond. no. **1**); 0.6 (exp. cond. no. **2**); 0.8 (exp. cond. no. **3**). All other indicators have identical factor loadings across groups.	3	Mplus data generation file	The second condition indicates the invariant condition. All factor loadings are equal across groups (i.e. including indicator no. 2). Reliability of the indicators equals: 0.24 (λ_i=0.4), 0.33 (λ_i=0.5), 0.41 (λ_i=0.6), 0.49 (λ_i=0.7) or 0.56 (λ_i=0.8) Reliability coefficient for indicator i = 1 – (Error variance / λ_i^2 + Error variance)

Factor name	Levels	Number of possible values	Implemented in ...	Additional remarks
F5: Inequality of indicator intercepts across groups	1: Indicator 2 has the same intercept in both G2 and G1; 2: Indicator 2 has an intercept that is 0.15 higher in G2 than in G1; 3: Indicator 2 has an intercept that is 0.30 higher in G2 than in G1; 4: Indicator 2 has an intercept that is 0.45 higher in G2 than in G1.	4	Mplus data generation file	The first level indicates the invariant condition (i.e. no inequalities in indicator intercepts across groups).

Factor name	Levels	Number of possible values	Implemented in ...	Additional remarks
Summary statistics		1080 experimental conditions(**) (full factorial)	2 different files to manipulate data generation process in Mplus software	

Fixed settings:

(1) One-dimensional factor model with 4 indicators (study no. 1) or 3 indicators (study no. 2);
(2) Factor scores are compared across 2 groups: G1 and G2;
(3) Indicators are categorical (*) and measured on a 5-point scale Likert-type of scale (type of scale that is typically used in market research when measuring one's [dis]agreement with certain 'statements');
(4) All indicators (except, possibly, the second one) have identical factor loadings and indicator intercepts across groups;
(5) 50 replications (generated data files) are generated for all k experimental conditions [k=1080] (**);
(6) The variance of the factor is fixed to 1 in both groups;
(7) Error variances of indicators are fixed to 0.51;
(8) The factor mean score in G1 is fixed to zero (i.e. $\kappa_1=0$). The factor mean score in G2 (i.e. κ_2) are expressed as deviations from the factor mean score in G1;
(9) Maximum Likelihood Estimates with Robust Standard Errors and Mean- and Variance Adjusted X^2 statistic (MLMV) are calculated in this simulation study (i.e. assuming continuous variables).

Appendix 4.2. Programs used in the simulation process

Step no. + Preparation (for) / Execution (in) Mplus + No. of (additional) files involved	Program name	Functionality	Input files required	Output files generated
1. Preparation (60 files)	MC3/4_DGF.exe	Fixing mean- and covariance structure* according to the model parameters in each experimental condition (* assuming that indicators follow a standard normal distribution).	None	MC_*.dat (factor3, factor4, factor5)
2. Preparation (21600 files)	MC3/4_GRD.exe	Creates Mplus input files (one for each experimental condition) to generate raw data files. In this second step the threshold model (i.e. factor 1) is imposed on all 3/4 indicators.		F123450.inp (factors 1 to 5 + repetition)
3. Preparation (1 file)	MC3/4_BT.exe	Creates a DOS BATCH file to process step 2 in Mplus.	None	MC3/4_BAT1.bat
4. Execution (43200 files)	MC3/4_BAT1.bat	Execution of step 2 (i.e. creation of raw data files).	MC_*.dat F123450.inp	F123450.out (useless) DF_123450.rdf (raw data files)
5. Preparation (21600 files)	MC3/4_ANA.exe	Creates Mplus input files to estimate measurement parameters of a one-factor model, ignoring possible violations of the measurement invariance principle (across groups).		A123450.inp

Step no. + Preparation (for) / Execution (in) Mplus + No. of (additional) files involved	Program name	Functionality	Input files required	Output files generated
6. Preparation (1 file)	MC3/4_BT2.exe	Creates a DOS BATCH file to process step 5 in Mplus.	None	MC3/4_BAT2.bat
7. Execution (21600 files)	MC3/4_BAT2.bat	Execution of step 5 in Mplus (estimation of measurement model not taking into account violations of the measurement invariance principle).	DF_123450.rdf A123450.inp	A123450.out
8. Preparation (5 files)	MC3/4_ROF.exe	Extracts all relevant output from Mplus output files.	A123450.out	MC3/4M.out (factor means) MC3/4V.out (factor variances) MC3/4L.out (factor loadings) MC3/4I.out (indicator intercepts) MC3/4F.out (model fit)
ALL STEPS: 108067 files (*2, i.e. for the three- and four-indicator model)				

Note: all programs are written in the programming language TURBO PASCAL

Appendix 4.3. Examples of Mplus files (used for the simulation)

MC_111.dat (one-factor model with three indicators)

```
0.00 0.00 0.00
1.00
0.42 0.87
0.35 0.30 0.76
-0.12 -0.12 -0.12
1.00
0.28 0.67
0.35 0.20 0.76
```

F111110.inp (one-factor model with three indicators)

```
TITLE: MonteCarlo Simulation
       MACS/ F1:1;F2:1;F3:1;F4:1;F5:1
       ONE non-invariant indicator out of THREE
MONTECARLO:
   FILE IS MC_111.dat;
   NAMES are y1-y3;
   NCUTS = 3*4;
   CUTPOINTS = 3 (-1.40 -.84 -.38 +.68);
   ESTIMATE = 3*4;
   NOBSERVATIONS = 200 200;
   NGROUPS = 2;
   NREPS = 1;
   SEED = 38641 ;
   SAVE = DF_111110.rdf;
ANALYSIS:
   … etc.
```

A111110.inp (one-factor model with three indicators)

```
TITLE: MonteCarlo Simulation: Analysis
       MACS/ F1:1;F2:1;F3:1;F4:1;F5:1
       ONE non-invariant indicator out of THREE
DATA:
    FILE IS DF_111110.rdf;
    TYPE IS INDIVIDUAL;
VARIABLE: NAMES are y1 Y2 y3 G;
         GROUPING = G (1=g1 2=g2);
ANALYSIS:
   TYPE IS MEANSTRUCTURE;
   ESTIMATOR IS MLMV;
   ITERATIONS = 2500;
   CONVERGENCE = .000001;
MODEL: f BY y1-y3;
       f@1.0;
       f BY y1*;
```

Appendix 4.4. Additional logistic regression models

		CORRECTNESS of the statistical difference test						ROBUSTNESS of the statistical difference test					
		Case: 3 indicators -2LL=54774.89 N=54000 CCR: 75.0 % RCDS: 65.2 %			Case: 4 indicators -2LL=52560.50 N=54000 CCR: 76.4 % RCDS: 67.3 %			Case: 3 indicators -2LL=958.66 N=1080 CCR: 77.9 % RRCS: 35.5 %			Case: 4 indicators -2LL=982.34 N=1080 CCR: 76.9 % RRCS: 34.9 %		
BINARY VARIABLE	SETTING	P value	B	e^B	P value	B	e^B	P value	B	e^B	P value	B	e^B
Constant	Constant = 1	.00	1.05	N.R.	.00	1.69	N.R.	36	8.80	N.R.	.18	8.11	N.R.
Reference F1	Standard normal distribution	-	-	-	-	-	-	-	-	-	-	-	-
F1_D2	Uni-modal distr.	.50	-	-	.63	-	-	.63	-	-	.03	.41	1.51
F1_D3	Bi-modal distr.	.86	-	-	.69	-	-	.92	-	-	.10	.32	1.38
F2_D1	N=200/200	.00	-.62	.54	.00	-.79	.45	.03	.64	1.89	.28	-	-
F2_D2	N=300/300	.00	-.37	.69	.00	-.49	.61	.00	.89	2.44	.78	-	-
F2_D3	N=400/400	.00	-.20	.82	.00	-.30	.74	.00	1.00	2.72	.18	-	-
F2_D4	N=500/500	.00	-.15	.86	.00	-.18	.83	.25	-	-	.89	-	-
Reference F2	N=750/750	-	-	-	-	-	-	-	-	-	-	-	-
F2_D6	N=200/400	.00	-.48	.62	.00	-.59	.55	.25	-	-	.99	-	-
F3_D1	F. Mean G2 = -0.30	.00	.31	1.37	.00	.11	1.12	.01	-.68	.51	.12	-	-
F3_D2	F. Mean G2 = -0.15	.00	-1.01	.36	.00	-1.79	.17	.04	.49	1.63	.04	-.51	.60
Reference F3	F. Mean G2 = 0.00	-	-	-	-	-	-	-	-	-	-	-	-
F3_D4	F. Mean G2 = 0.15	.00	1.20	3.32	.00	.47	1.60	.00	-1.04	0.35	.00	-1.04	.35
F3_D6	F. Mean G2 = 0.30	.00	2.93	18.78	.00	2.32	10.13	.00	2.18	8.84	.00	1.60	4.95

Notes:

(1) Negative coefficients are underlined (or $e^B <= 0$); non-significant coefficients are not printed (except for the constant term); N.R. means 'not relevant';
(2) # Cross-group invariance condition (with respect to this measurement parameter);
(3) e^B is the impact on the odds ratio;
(4) 2LL means 'minus 2 times loglikelihood'; CCR means 'correct classification rate'; RCDS means 'rate of correct [statistical] decisions in sample'; RRCS means 'rate of robust cases in sample' .

		CORRECTNESS of the statistical difference test						ROBUSTNESS of the statistical difference test					
		Case: 3 indicators -2LL=54774.89 N=54000 CCR: 75.0 % RCDS: 65.2 %			Case: 4 indicators -2LL=52560.50 N=54000 CCR: 76.4 % RCDS: 67.3 %			Case: 3 indicators -2LL=958.66 N=1080 CCR: 77.9 % RRCS: 35.5 %			Case: 4 indicators -2LL=982.34 N=1080 CCR: 76.9 % RRCS: 34.9 %		
BINARY VARIABLE	SETTING	P value	B	e^B	P value	B	e^B	P value	B	e^B	P value	B	e^B
F4_D1	λ_2 (G2) = 0.4	.00	-.76	.47	.00	-.87	.42	.31	-	-	.14	-	-
Reference F4#	λ_2 (G2) = 0.6	-	-	-	-	-	-	-	-	-	-	-	-
F4_D3	λ_2 (G2) = 0.8	.00	.68	1.98	.00	.69	2.00	.37	-	-	.24	-	-
Reference F5#	Int_2 (G2) = .00	-	-	-	-	-	-	-	-	-	-	-	-
F5_D2	Int_2 (G2) = .15	.00	-.24	.79	.00	-.20	.82	.31	-	-	.15	-	-
F5_D3	Int_2 (G2) = .30	.00	-.83	.44	.00	-.70	.50	.22	-	-	.09	-10.40	.00
F5_D4	Int_2 (G2) = .45	.00	-1.01	.36	.00	-.98	.38	.26	-	-	.09	-10.31	.00
F4_D1* F5_D2	Interaction effect	.06	.13	1.14	.38	-	-	.32	-	-	.13	-	-
F4_D1* F5_D3	idem	.00	.46	1.58	.00	.41	1.50	.27	-	-	.12	-	-
F4_D1* F5_D4	idem	.00	.75	2.12	.00	.58	1.79	.29	-	-	.13	-	-
F4_D3* F5_D2	idem	.08	-.14	.87	.41	-	-	.33	-	-	.20	-	-
F4_D3* F5_D3	idem	.00	-.26	.77	.00	-.22	.81	.33	-	-	.17	-	-
F4_D3* F5_D4	idem	.00	-.44	.64	.00	-.40	.67	.39	-	-	.23	-	-

Notes:

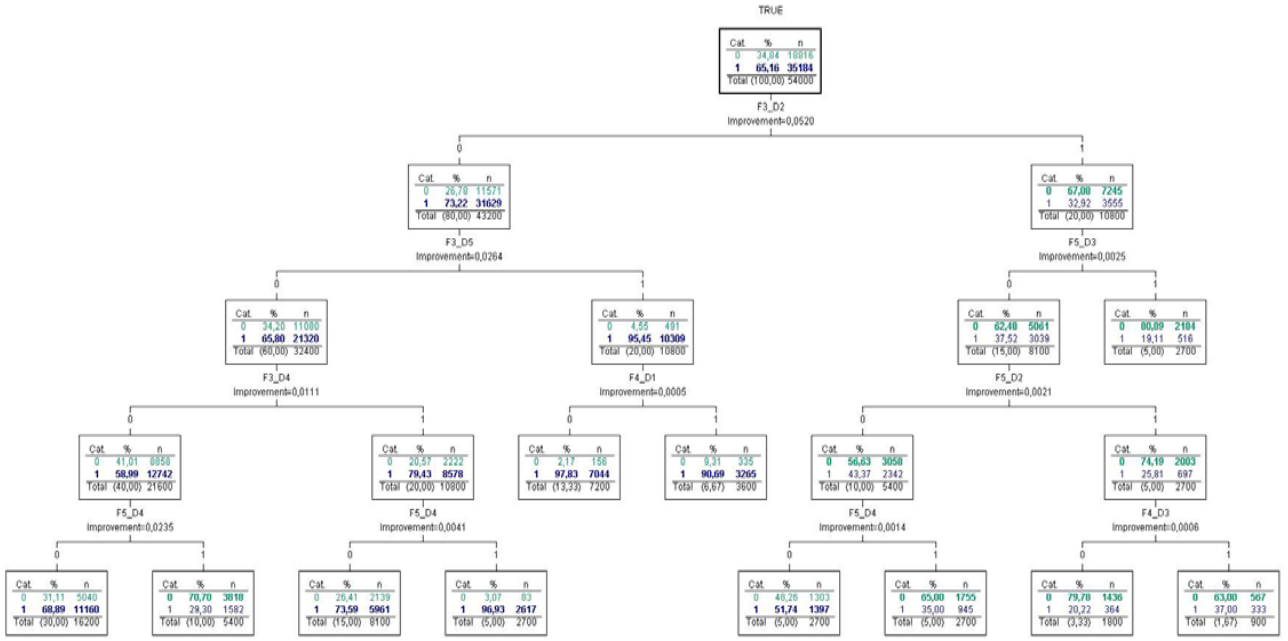
(1) Negative coefficients are underlined (or $e^B <= 0$); non-significant coefficients are not printed (except for the constant term); N.R. means 'not relevant';
(2) # Cross-group invariance condition (with respect to this measurement parameter);
(3) e^B is the impact on the odds ratio;
(4) 2LL means 'minus 2 times loglikelihood'; CCR means 'correct classification rate'; RCDS means 'rate of correct [statistical] decisions in sample'; RRCS means 'rate of robust cases in sample' .

Appendix 4.5. C&RTrees for correct statistical conclusions

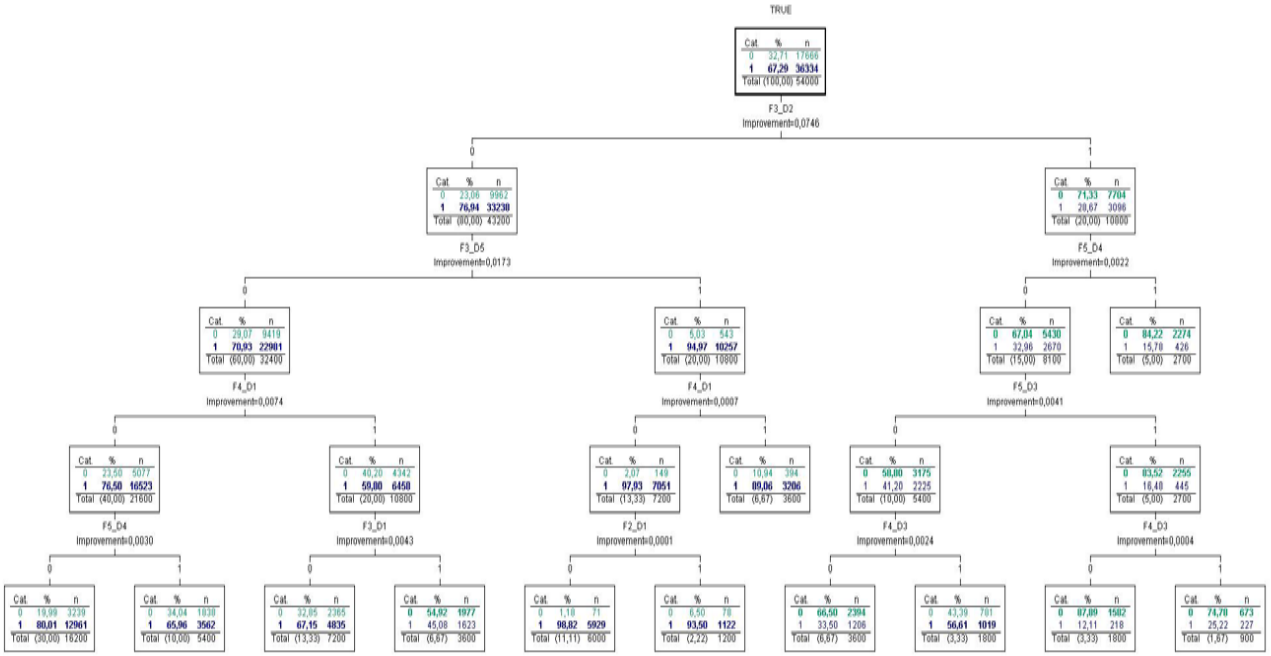
Technical details:

Algorithm: CART
Impurity measure: Gini

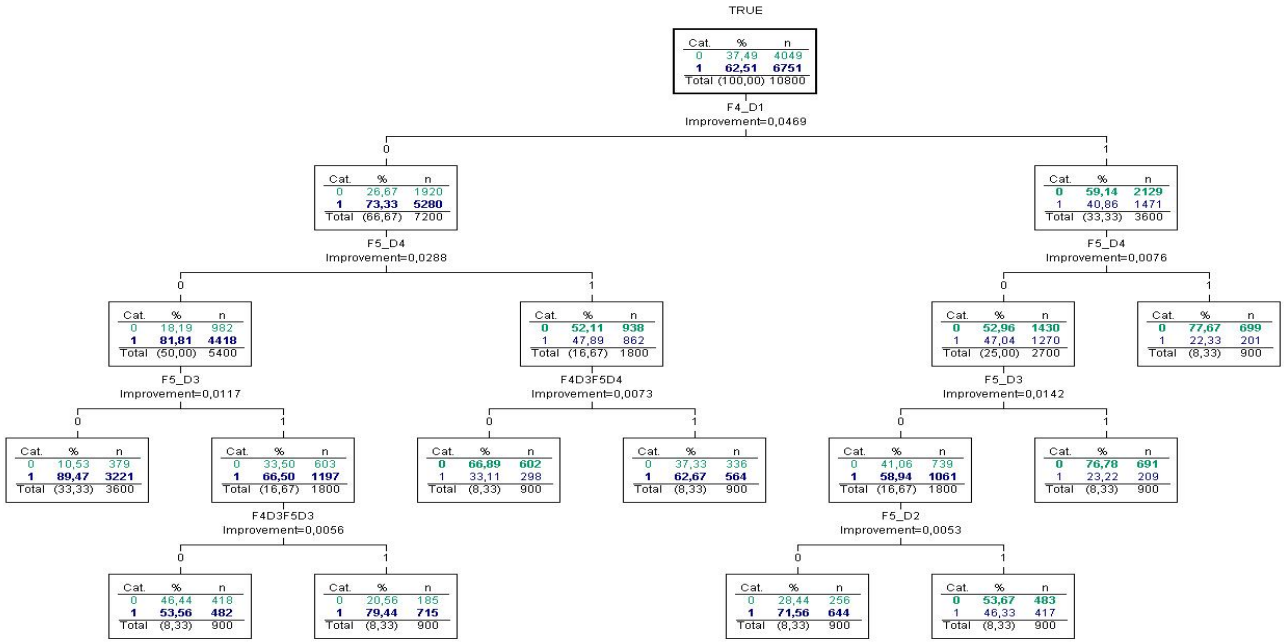
Minimum number of observations for parent node: 1500
Minimum number of observations for child node: 900
Maximum tree depth: 5 (or 4 for tree 1 & 2)



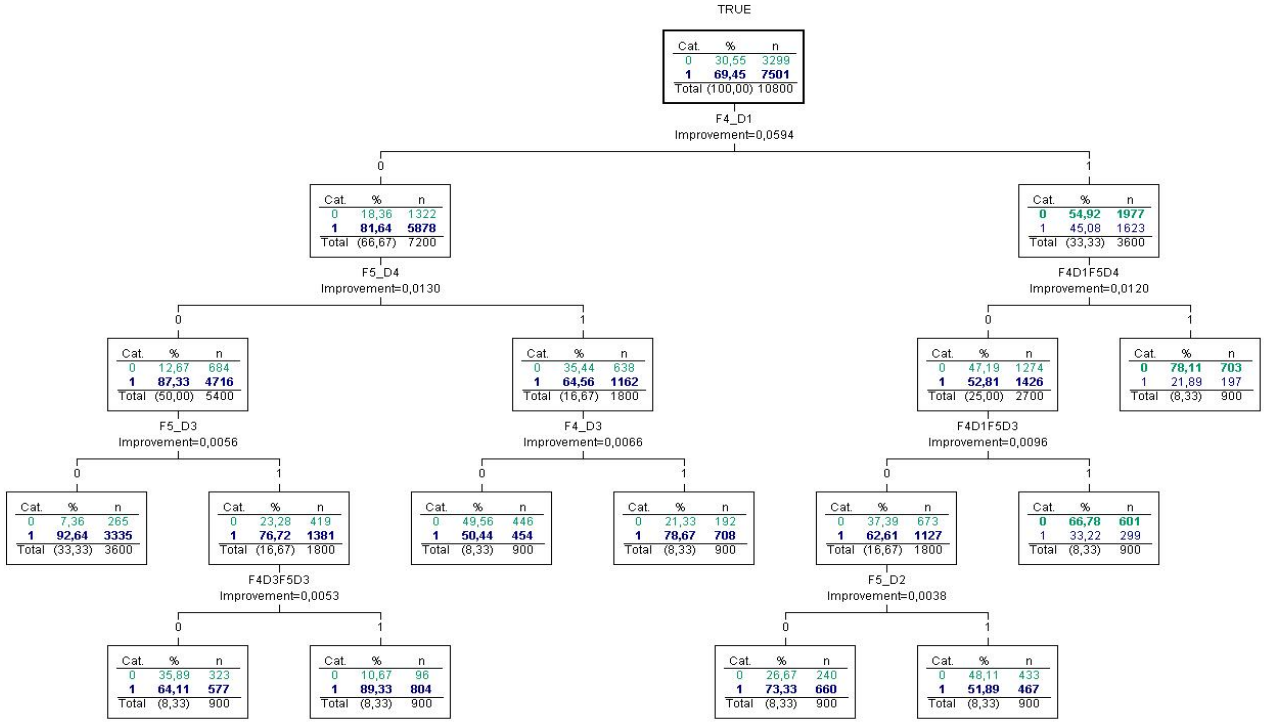
Correct(true):T1 / 3 indicators – F3: all levels



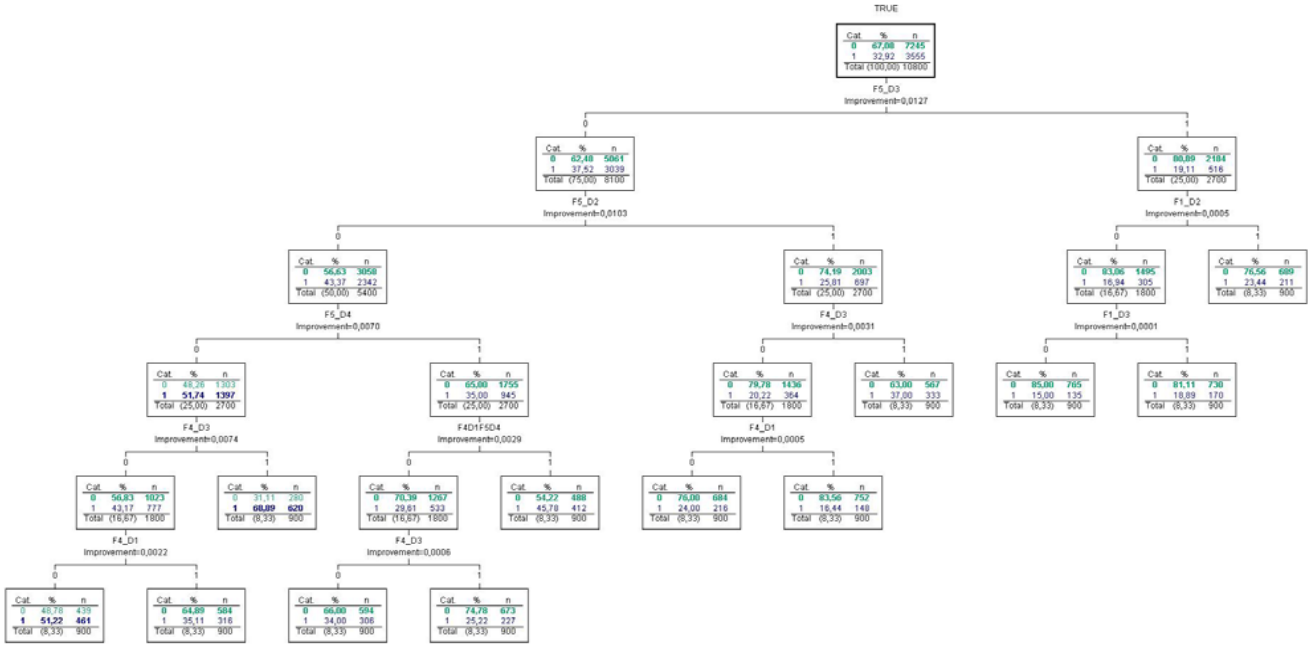
Correct(true):T2 / 4 indicators – F3: all levels



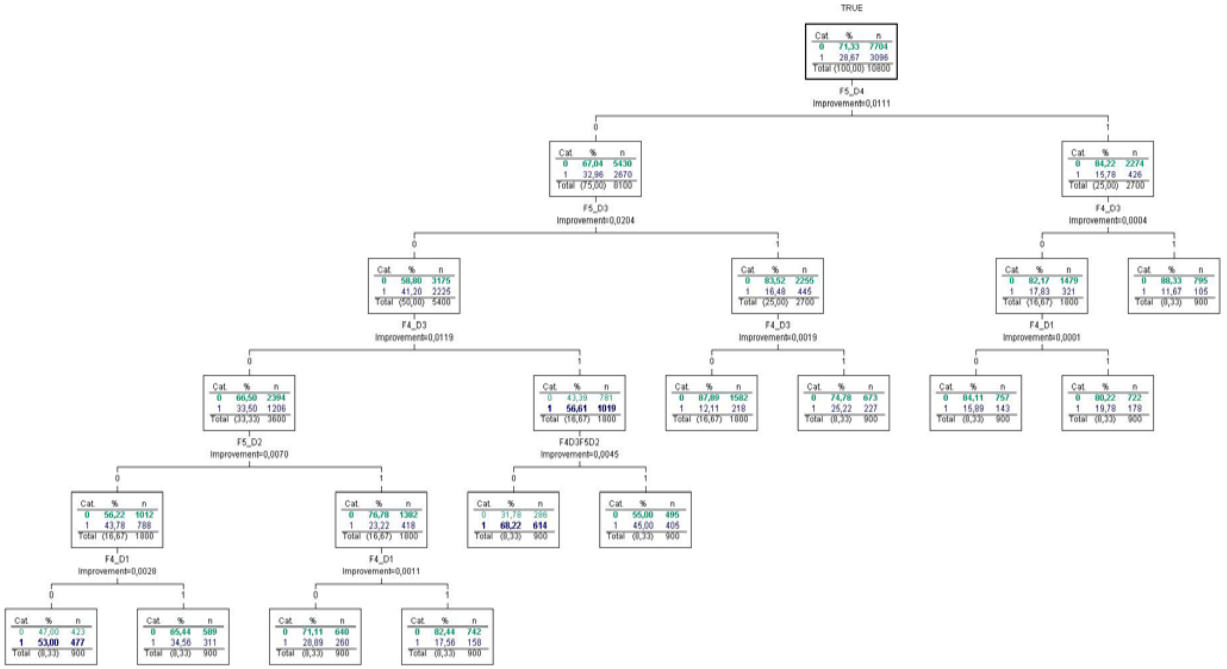
Correct(true):T3 / 3 indicators – F3=1



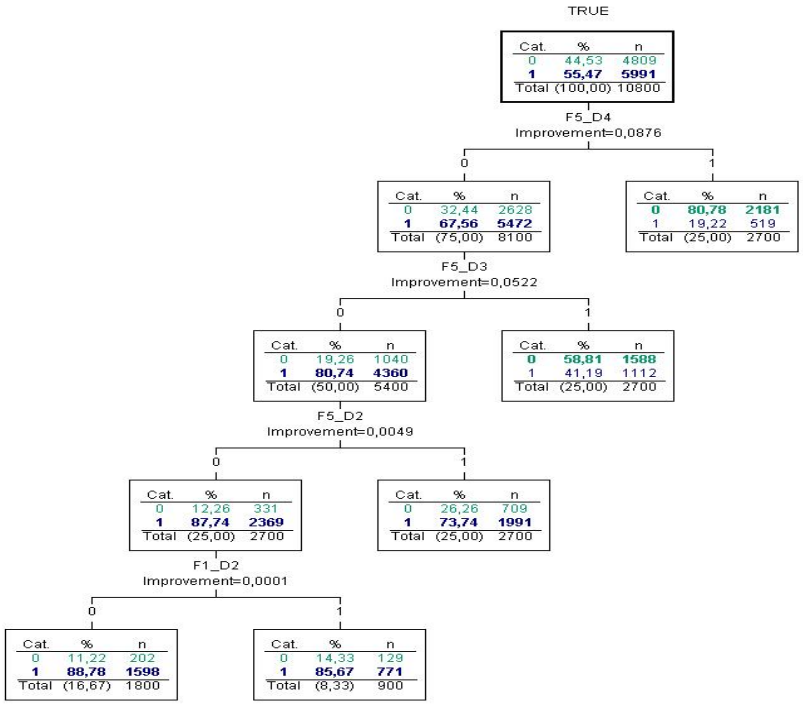
Correct(true):T4 / 4 indicators – F3=1



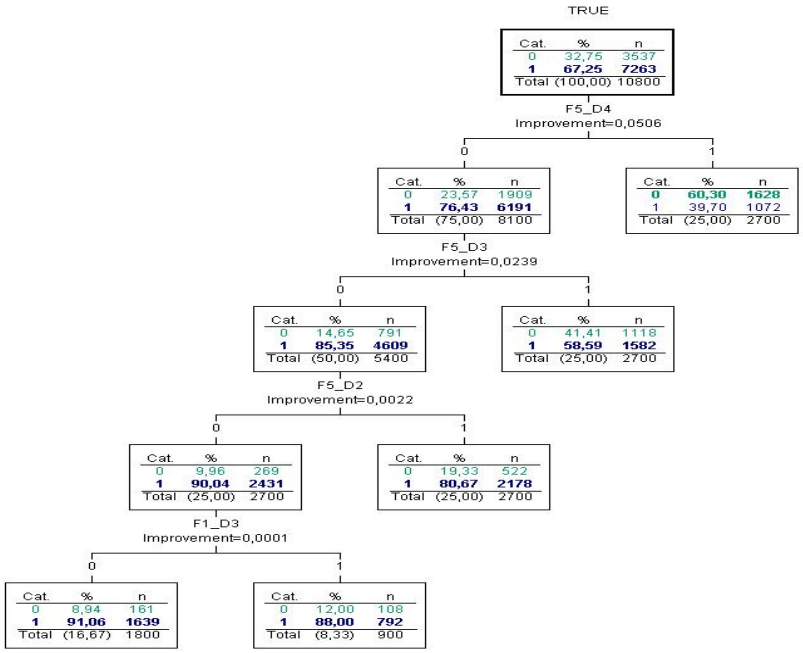
Correct(true):T5 / 3 indicators – F3=2



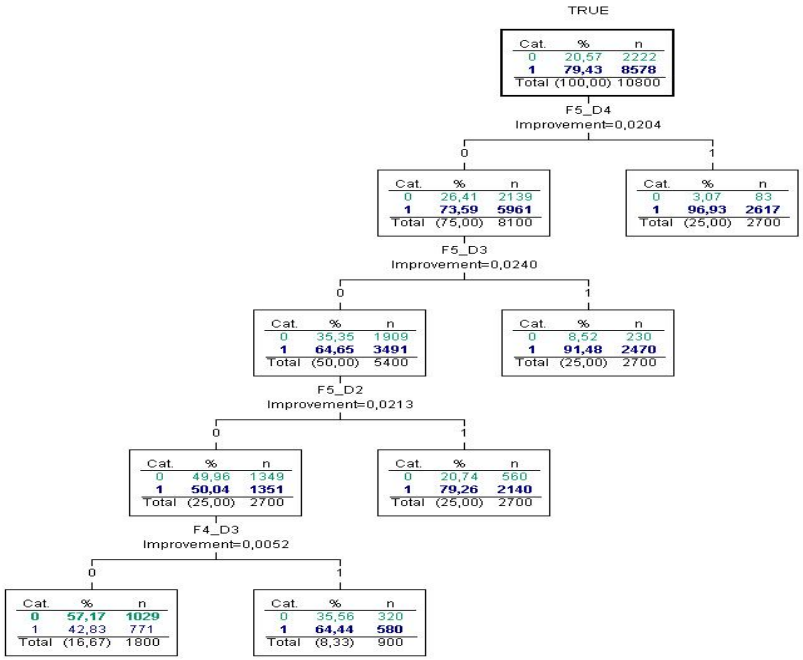
Correct(true):T6 / 4 indicators – F3=2



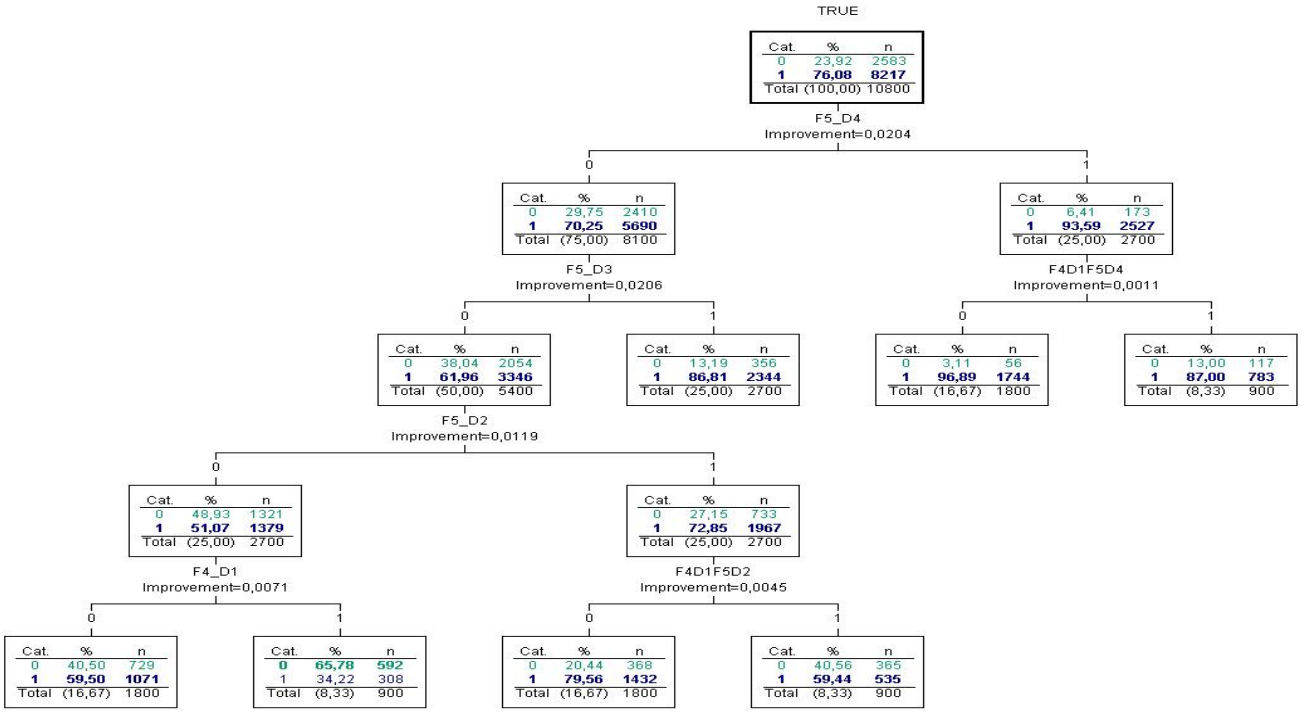
Correct(true):T7 / 3 indicators – F3=3



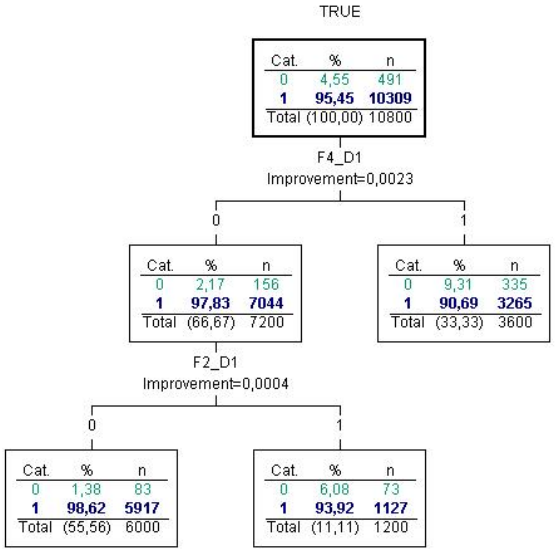
Correct(true):T8 / 4 indicators – F3=3



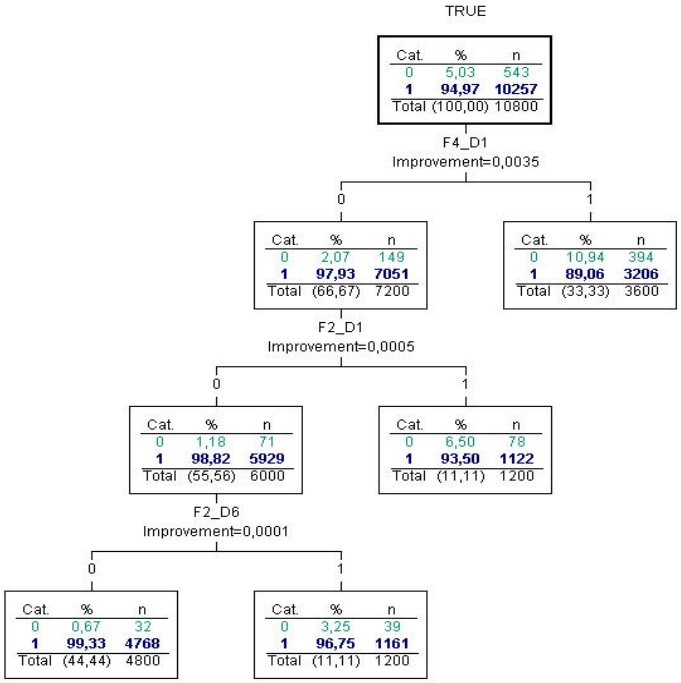
Correct(true):T9 / 3 indicators – F3=4



Correct(true):T10 / 4 indicators – F3=4



Correct(true):T11 / 3 indicators – F3=5



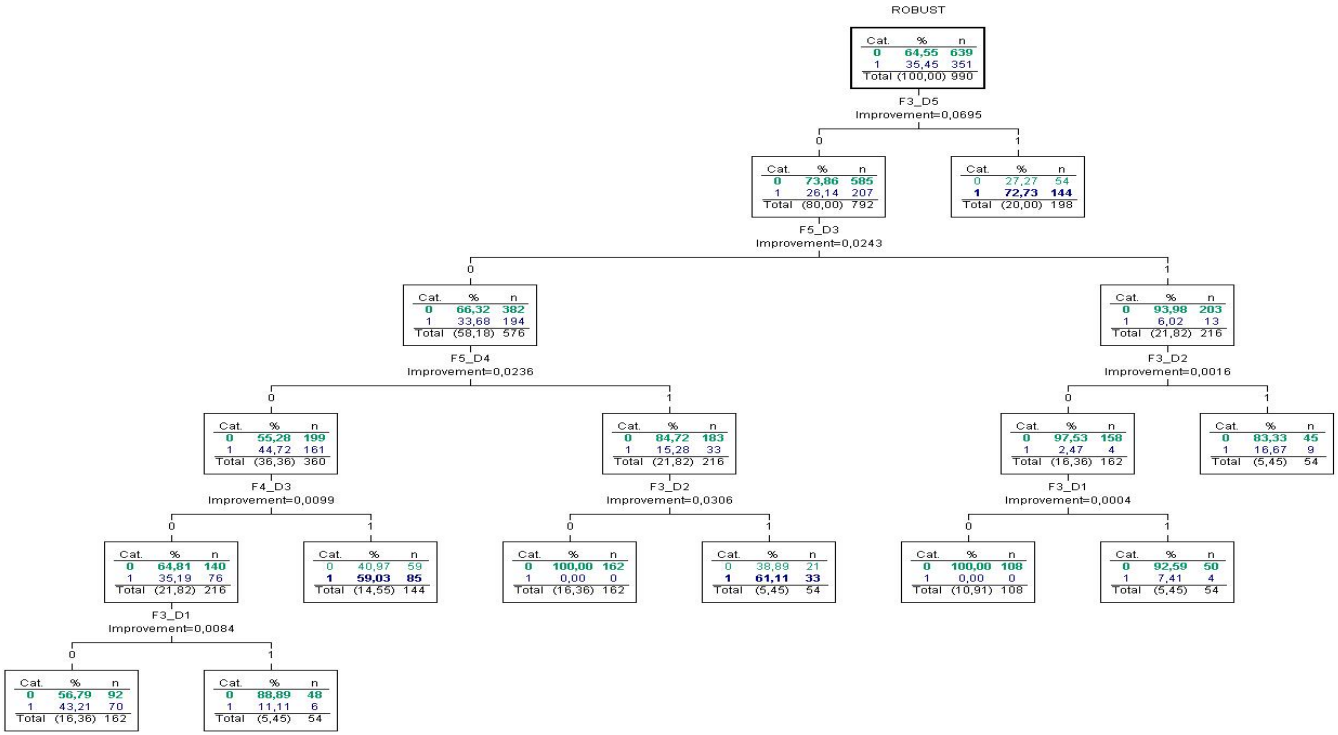
Correct(true):T12 / 4 indicators – F3=5

Appendix 4.6. C&RTrees for robust cases

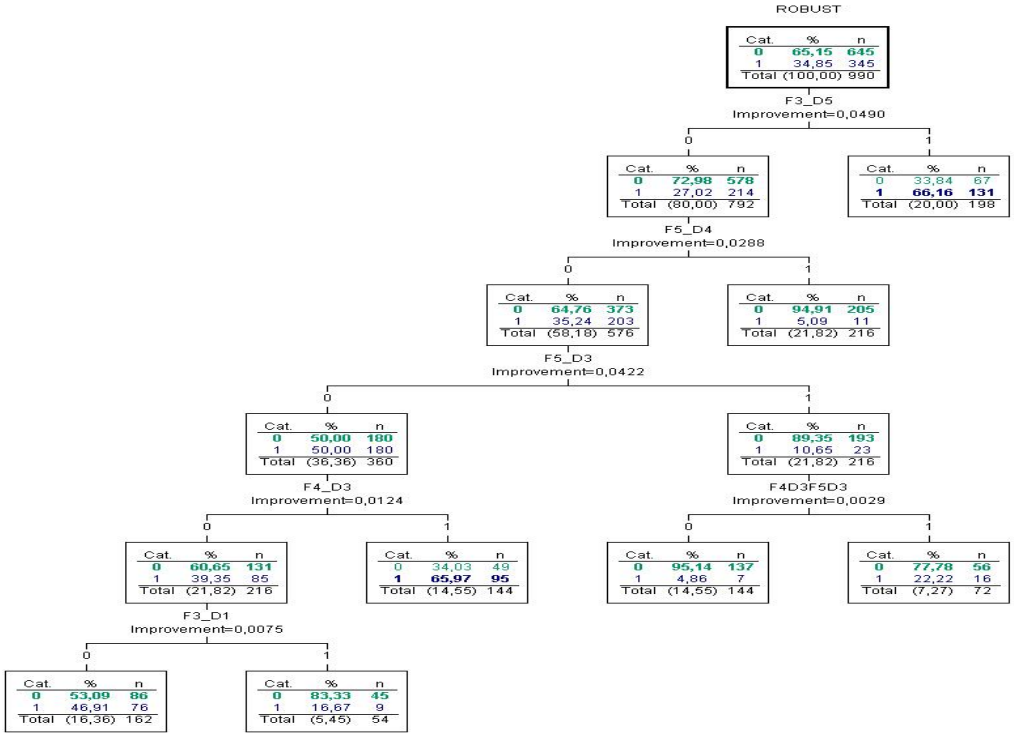
Technical details:

Algorithm: CART
Impurity measure: Gini

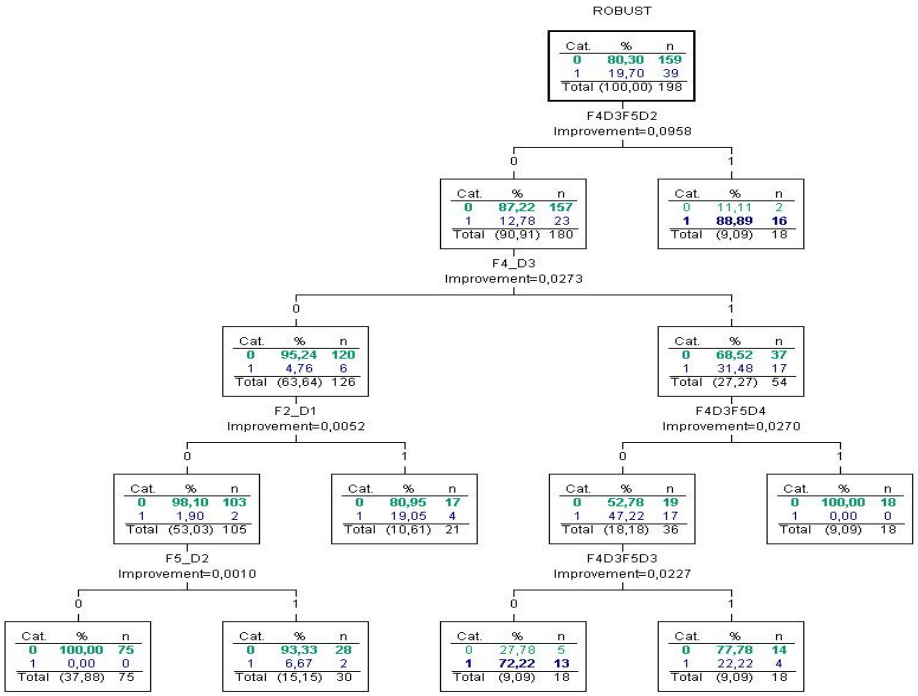
Minimum number of observations for parent node: 30
Minimum number of observations for child node: 15
Maximum tree depth: 5



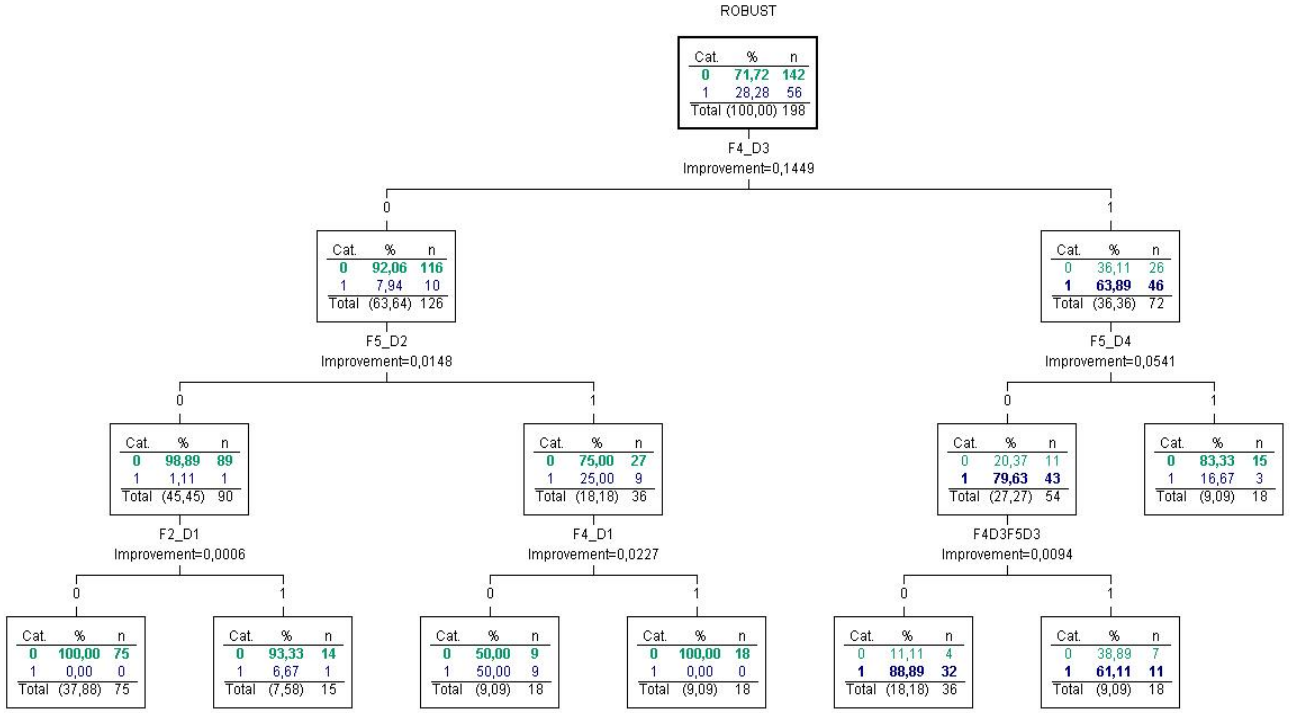
Robust:T1 / 3 indicators – F3: all levels



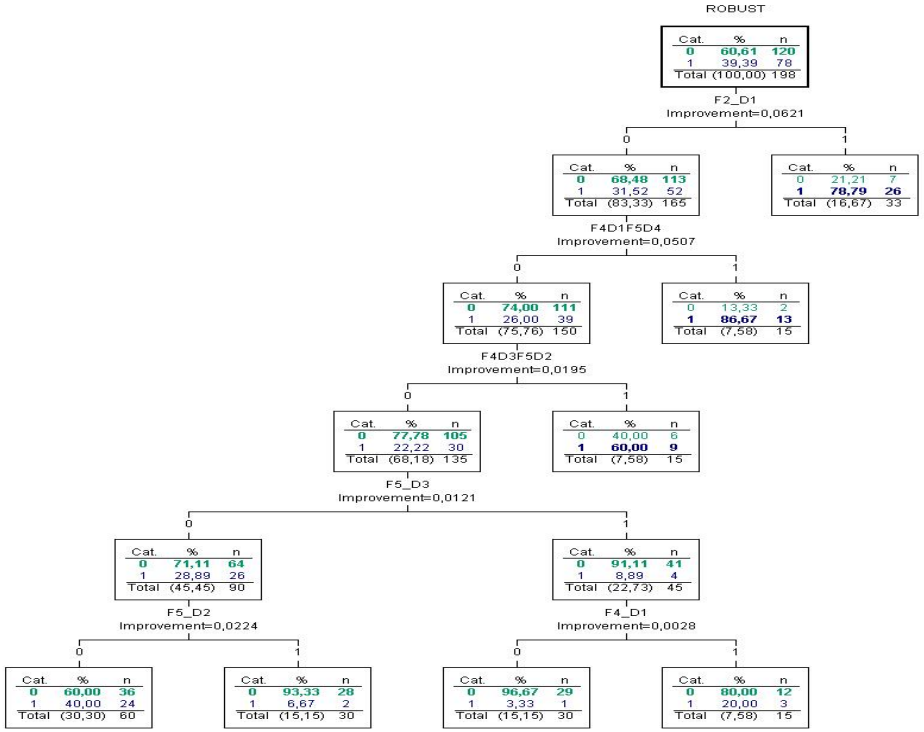
Robust:T2 / 4 indicators – F3: all levels



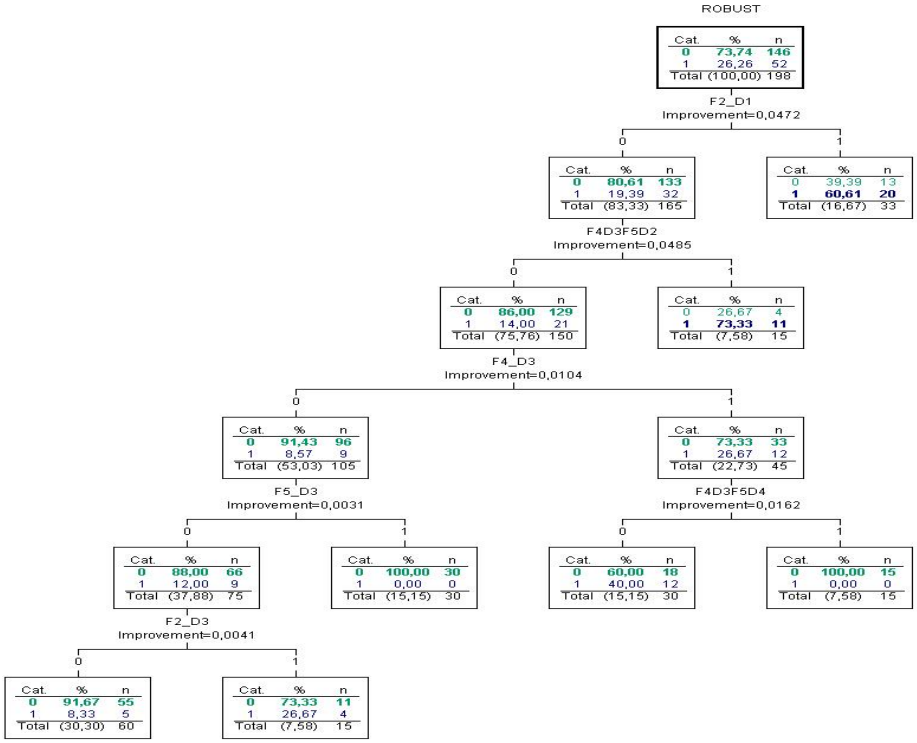
Robust:T3 / 3 indicators – F3=1



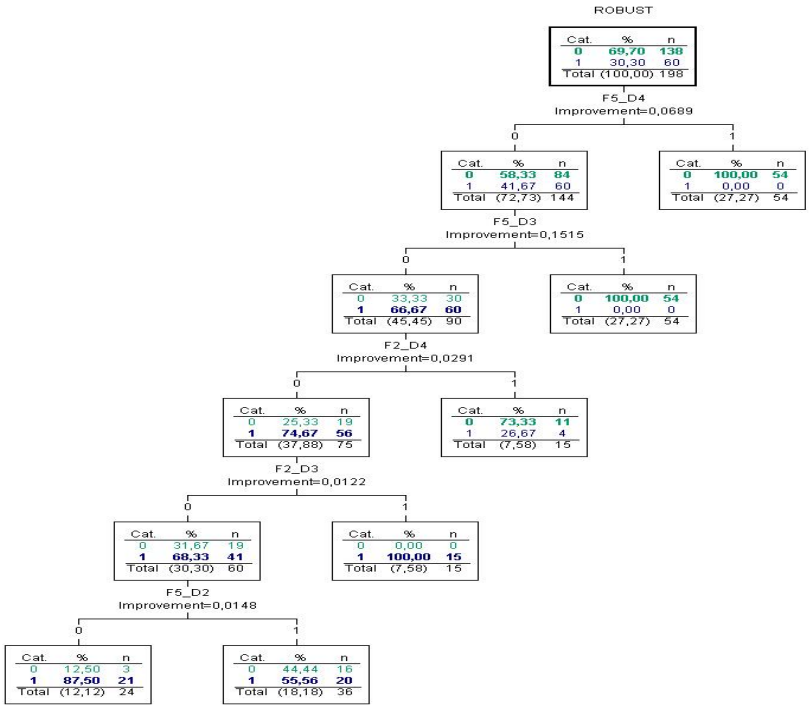
Robust:T4 / 4 indicators – F3=1



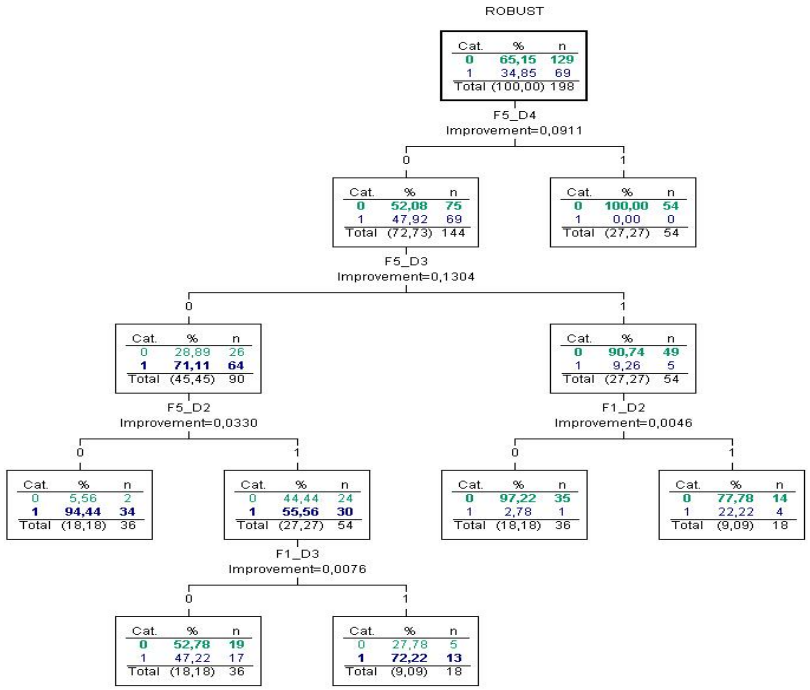
Robust:T5 / 3 indicators – F3=2



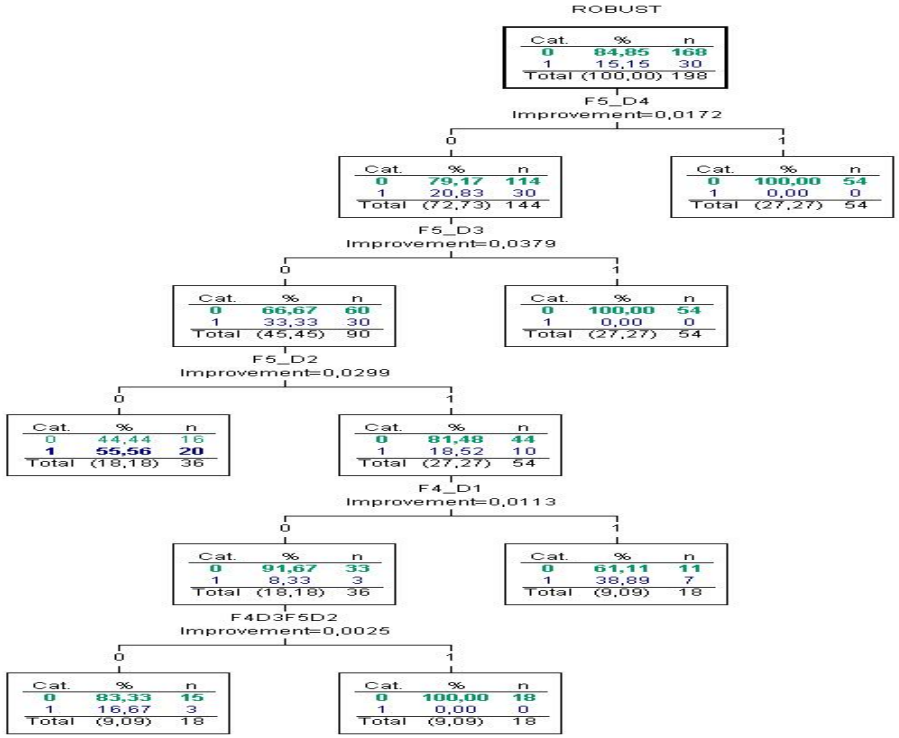
Robust:T6 / 4 indicators – F3=2



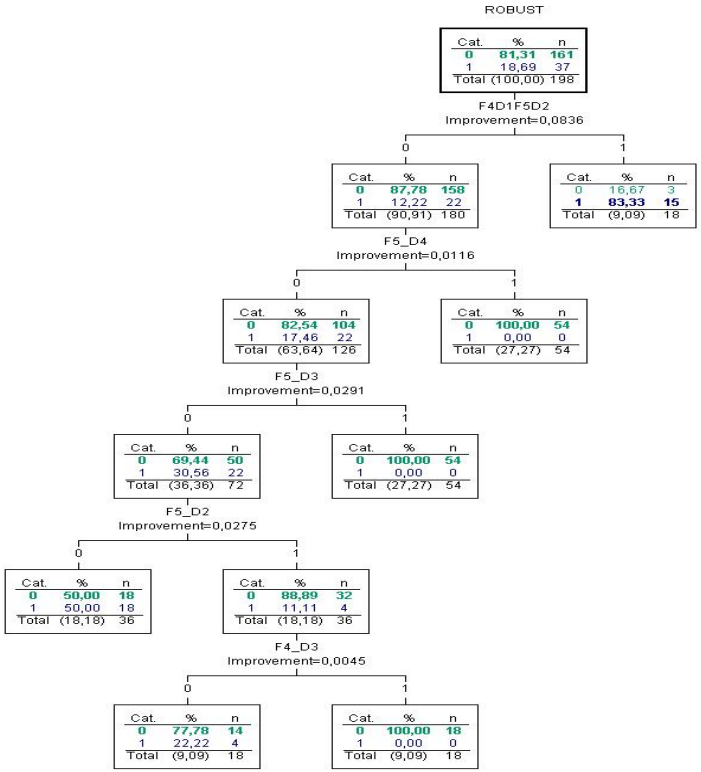
Robust:T7 / 3 indicators – F3=3



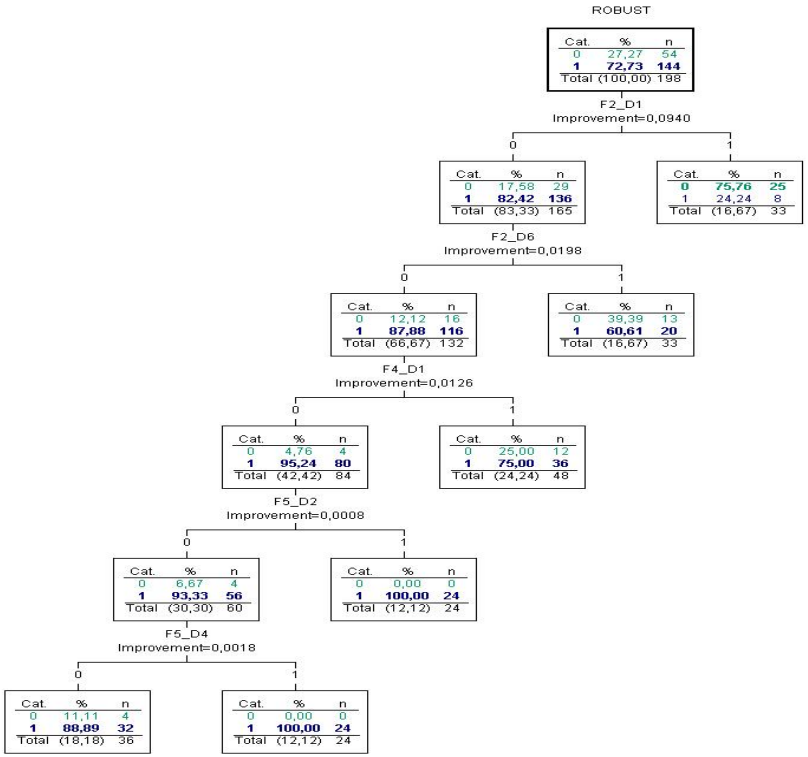
Robust:T8 / 4 indicators – F3=3



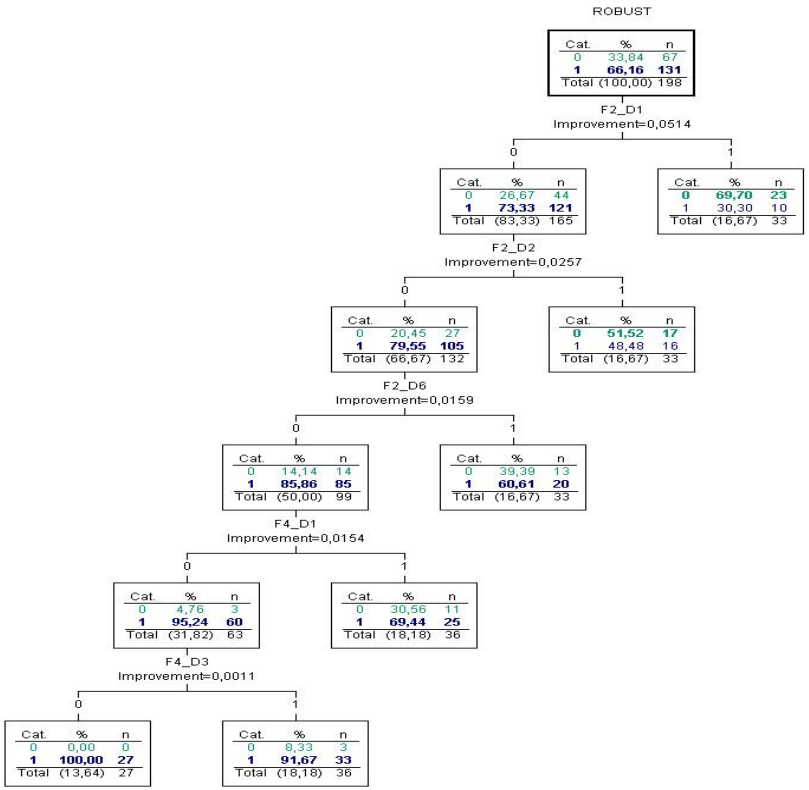
Robust:T9 / 3 indicators – F3=4



Robust:T10 / 4 indicators – F3=4



Robust:T11 / 3 indicators – F3=5



Robust:T12 / 4 indicators – F3=5

Appendix 5.1. Items used in the analyses

Work environment factors (7 factors)

Factor 1: Fair remuneration

I015*: From what I hear, our pay is as good as or better than the pay in other similar companies.
(5 pt. Agreement/Disagreement [A/D] scale)
I081: How do you rate your total benefits programme?
(5 pt. evaluation scale)
I087: How good a job is the organisation doing in matching pay to performance?
(5 pt. evaluation scale)
(* = reference item)

Factor 2: Supporting role of people within the department

I027*: In my department people provide each other with useful feedback.
(5 pt. A/D scale)
I026: In my department people do not accept mediocrity in their work.
(5 pt. A/D scale)
I025: In my department people usually do what they say they will.
(5 pt. A/D scale)

Factor 3: My immediate boss' support

I033*: My immediate boss gives me regular feedback on my performance.
(5 pt. A/D scale)
I048: My immediate boss communicates clearly.
(5 pt. A/D scale)
I065: I feel my immediate boss coaches me when I need it.
(5 pt. A/D scale)

Factor 4: Clarity of strategy

I006*: I have a clear understanding of the goals and objectives of my department.
(5 pt. A/D scale)
I007: I have a clear understanding of the goals and objectives of my organisation.
(5 pt. A/D scale)
I008: I have a clear understanding of the goals and objectives of the multinational as a whole.
(5 pt. A/D scale)

Factor 5: Confidence in managerial decisions

I045*: I have confidence in the decisions made by managers of my organisation.
(5 pt. A/D scale)
I046: I have confidence in the decisions made by managers of my business group / region.
(5 pt. A/D scale)

Factor 6: Organisational and managerial efficiency

I076*: In your judgement, how does this organisation compare with its competitors on responding rapidly to changes in the market?
(5 pt. A/D scale)
I086: How good are managers in your organisation doing in developing simple and fast processes from supplier through to consumer?
(5 pt. Evaluation scale)

Factor 7: Environmental and societal responsibility

I019*: I believe that my organisation is environmentally responsible.
(5 pt. A/D scale)
I020: I believe that my organisation is a socially responsible member of the community.
(5 pt. A/D scale)

Job satisfaction

Global measure of employee job satisfaction (identical to the item value)

I099: Considering everything, how satisfied are you with your job.
(5 pt. Evaluation scale)

Appendix 5.2. Correlations between work environment factors

Correlations between work environment factors (based on observations from all countries)

Table A.5.2/1
Correlations between 7 factors

Correlation	F1	F2	F3	F4	F5	F6	F7
F1	1.00	0.32	0.39	0.25	0.44	0.25	0.20
F2	0.32	1.00	0.51	0.25	0.44	0.18	0.19
F3	0.39	0.51	1.00	0.32	0.63	0.24	0.20
F4	0.25	0.25	0.32	1.00	0.40	0.15	0.18
F5	0.44	0.44	0.63	0.40	1.00	0.34	0.30
F6	0.25	0.18	0.24	0.15	0.34	1.00	0.15
F7	0.20	0.19	0.20	0.18	0.30	0.15	1.00

Notes: (1) Model used: 7-factor CFA model without a mean structure;
(X^2=5415, d.f.=114, p=0.00; CFI=0.968; TLI=0.957; RMSEA=0.043/0.039; SRMR=0.029)
(2) Data used: all observations from 16 countries (N=25018).

The highest correlations are between F3 (Immediate boss' support) and F5 (Confidence in Managerial Decisions) [r=0.63], and between F2 (Supporting role of people within the department) and F3 (Immediate boss' support) [r=0.51]. The high correlation between F3 and F5 may indicate that (a strong) personal support from one's immediate boss increases one's confidence in managerial decisions. The high correlation between F2 and F3 may indicate that (a strong) personal support from an immediate boss and support from colleagues tend to co-occur.

Appendix 5.3. Country-specific (estimated) factor means

TAU-INVARIANCE MODEL

Table A.5.3/1
Estimated factor means in 16 countries (tau-invariance model)

Factor means	Belgium	France	Germany	Hungary	Italy	Nether-lands	Russian federation	Sweden
F1	0.00	-0.42	-0.31	-0.69	-0.32	+0.20	-0.43	-0.84
F2	0.00	-0.07	+0.14	+0.15	-0.04	+0.05	+0.31	-0.05
F3	0.00	-0.21	-0.20	+0.35	+0.11	0.00	+0.30	-0.10
F4	0.00	-0.16	-0.16	0.00	-0.31	+0.05	+0.06	-0.44
F5	0.00	-0.09	-0.88	+0.26	-0.05	-0.08	+0.19	-0.32
F6	0.00	-0.22	-0.45	-0.15	+0.32	-0.17	+0.03	-0.38
F7	0.00	+0.03	-0.13	+0.04	+0.07	+0.16	+0.39	-0.36

Notes: (1) Model used: Tau-invariance model (CFA model with a mean structure)
(X^2=16829, d.f.=2154, p=0.00; CFI=0.913; TLI=0.901; RMSEA=0.066/0.059; SRMR=0.056);
(2) Data used: all observations from 16 countries (N=25018).

Table A.5.3/1 (continued)
Estimated factor means in 16 countries (tau-invariance model)

Factor Means	U.K.	Canada	U.S.	Brazil	Mexico	Australia	Israel	South Africa
F1	-0.28	-0.04	0.00	-0.42	+0.05	-0.31	-0.60	-0.41
F2	-0.24	-0.04	-0.04	-0.32	+0.23	-0.09	+0.25	-0.02
F3	-0.33	0.00	+0.02	-0.04	+0.05	0.00	+0.35	-0.04
F4	-0.27	-0.23	-0.13	-0.03	+0.12	-0.13	-0.04	+0.10
F5	-0.43	0.00	-0.08	+0.14	+0.17	-0.25	+0.14	+0.14
F6	-0.32	-0.25	-0.17	+0.18	+0.24	-0.13	+0.20	+0.04
F7	+0.04	+0.28	+0.22	+0.46	+0.55	+0.19	+0.30	+0.39

Notes: (1) Model used: Tau-invariance model (CFA model with a mean structure)
(X^2=16829, d.f.=2154, p=0.00; CFI=0.913; TLI=0.901; RMSEA=0.066/0.059; SRMR=0.056);
(2) Data used: all observations from 16 countries (N=25018).

PARTIAL TAU-INVARIANCE MODEL

Table A.5.3/2
Estimated factor means in 16 countries (partial tau-invariance model)

Factor Means	Belgium	France	Germany	Hungary	Italy	Nether-lands	Russian federation	Sweden
F1	0.00	-0.40	-0.40	-0.67	-0.18	+0.22	-0.42	-0.78
F2	0.00	-0.07	-0.07	+0.04	+0.01	-0.05	+0.08	-0.25
F7	0.00	+0.02	-0.13	+0.03	+0.06	+0.15	+0.37	-0.37

Notes: (1) Model used: Partial tau-invariance model (CFA model with a mean structure) (X^2=12948, d.f.=2123, p=0.00; CFI=0.936; TLI=0.926; RMSEA=0.057/0.051; SRMR=0.050);
(2) Data used: all observations from 16 countries (N=25018);
(3) Freed parameters: indicator intercept of i081 (F1) in all countries; indicator intercept of i025 (F2) in all countries; indicator intercept of i020 (F7) in Brazil.

Table A.5.3/2 (continued)
Estimated factor means in 16 countries (partial tau-invariance model)

Factor Means	U.K.	Canada	U.S.	Brazil	Mexico	Australia	Israel	South Africa
F1	-0.19	0.00	-0.07	-0.38	+0.03	-0.23	-0.52	-0.32
F2	-0.32	-0.11	-0.13	-0.21	+0.26	-0.16	+0.19	-0.06
F7	+0.04	+0.27	+0.21	+0.58	+0.52	+0.18	+0.28	+0.37

Notes: (1) Model used: Partial tau-invariance model (CFA model with a mean structure) (X^2=12948, d.f.=2123, p=0.00; CFI=0.936; TLI=0.926; RMSEA=0.057/0.051; SRMR=0.050);
(2) Data used: all observations from 16 countries (N=25018);
(3) Freed parameters: indicator intercept of i081 (F1) in all countries; indicator intercept of i025 (F2) in all countries; indicator intercept of i020 (F7) in Brazil.

Appendix 5.4. Variable indicator intercepts (partial tau-invariance model)

Table A.5.4/1
Variable indicator intercepts across 16 countries

Indicator intercepts	Belgium	France	Germany	Hungary	Italy	Nether-lands	Russian federation	Sweden
I081 (F1)	3.90	3.85	4.34	3.85	3.45	3.84	3.88	3.72
I025 (F2)	3.74	3.75	4.15	4.00	3.58	3.98	4.23	4.19
I020 (F7)	4.06	4.06	4.06	4.06	4.06	4.06	4.06	4.06

Notes: (1) Model used: Partial tau-invariance model (CFA model with a mean structure) (X^2=12948, d.f.=2123, p=0.00; CFI=0.936; TLI=0.926; RMSEA=0.057/0.051; SRMR=0.050);
(2) Figures which are underlined are constrained to be equal across groups.

Table A.5.4/1 (continued)
Variable indicator intercepts across 16 countries

Indicator Intercepts	U.K.	Canada	U.S.	Brazil	Mexico	Australia	Israel	South Africa
I081 (F1)	3.65	3.75	4.08	3.83	3.93	3.69	3.68	3.68
I025 (F2)	3.92	3.91	3.95	3.36	3.65	3.89	3.90	3.82
I020 (F7)	4.06	4.06	4.06	3.54	4.06	4.06	4.06	4.06

Notes: (1) Model used: Partial tau-invariance model (CFA model with a mean structure) (X^2=12948, d.f.=2123, p=0.00; CFI=0.936; TLI=0.926; RMSEA=0.057/0.051; SRMR=0.050);
(2) Figures which are underlined are constrained to be equal across groups.

Table A5.4/2
Coefficient of variation of variable indicator intercepts

	Coefficient of variation (across countries)
I081 (F1)	0.052
I025 (F2)	0.058

Appendix 5.5. Indicator reliabilities

Table A.5.5/1
Indicator reliabilities

Indicator	Reliability coefficient (from Classical Test Theory)
I015 (F1)	0.44 / same value
I081 (F1)	0.36 / same value
I087 (F1)	0.60 / 0.62
I025 (F2)	0.51 / same value
I026 (F2)	0.39 / 0.41
I027 (F2)	0.47 / 0.49
I033 (F3)	0.56 / same value
I048 (F3)	0.65 / same value
I065 (F3)	0.67 / same value
I006 (F4)	0.51 / same value
I007 (F4)	0.81 / same value
I008 (F4)	0.45 / same value
I045 (F5)	0.77 / same value
I046 (F5)	0.66 / same value
I076 (F6)	0.31 / same value
I086 (F6)	0.50 / same value
I019 (F7)	0.58 / 0.54
I020 (F7)	0.60 / same value

Notes: The first coefficient represents the reliability coefficient of the tau-invariance model; the second coefficient represents the reliability coefficient of the partial tau-invariance model.

Appendix 5.6. Determinants of employees' job satisfaction in all countries (based on a separate model for each country in which the individual path coefficients are estimated)

Table A.5.6/1

	U.S. (N= 4496)	U.K. (N=2620)	Italy (N=1247)	France (N=788)	Brazil (N=6206)	Mexico (N=1821)	Nether-lands (N=895)	Sweden (N=488)
R^2	0.47*	0.45*	0.39*	0.29*	0.44*	0.44*	0.34*	0.33*
F1	**0.30** [0.28] (12.7)	**0.37** [0.31] (10.5)	**0.17** [0.17] (3.5)	**0.18** [0.19] (3.7)	**0.27** [0.31] (16.8)	**0.33** [0.32] (9.4)	**0.18** [0.17] (3.8)	**0.13** [0.14] (2.0)
F2	**0.06** [0.06] (2.8)	*n.s.*	**0.18** [0.15] (2.9)	*n.s.*	*n.s.*	*n.s.*	*n.s.*	**0.21** [0.19] (3.0)
F3	**0.14** [0.16] (8.9)	**0.13** [0.14] (6.1)	**0.23** [0.23] (5.1)	**0.13** [0.18] (3.7)	**0.17** [0.21] (9.4)	**0.11** [0.12] (4.4)	**0.20** [0.26] (6.2)	**0.12** [0.16] (2.6)
F4	**0.10** [0.08] (4.4)	**0.18** [0.14] (5.7)	**0.15** [0.12] (3.6)	**0.19** [0.18] (4.2)	**0.08** [0.06] (4.2)	*n.s.*	**0.16** [0.12] (2.7)	*n.s.*
F5	**0.12** [0.13] (5.2)	**0.11** [0.12] (3.7)	**0.12** [0.11] (2.7)	*n.s.*	*n.s.*	*n.s.*	**0.10** [0.12] (2.3)	*n.s.*
F6	**0.28** [0.19] (7.7)	**0.23** [0.12] (3.2)	*n.s.*	**0.39** [0.22] (3.1)	**0.34** [0.22] (10.4)	**0.41** [0.23] (4.9)	*n.s.*	*n.s.*
F7	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*	**0.11** [0.06] (2.1)	*n.s.*	*n.s.*

Notes:
All results presented in this table are estimations under the assumption that the indicator of employee job satisfaction (i.e. the dependent variable) is perfectly reliable. An additional analysis in which the reliability of the indicator was set equal to 0.85 showed that all results obtained were 'stable' (i.e. none of the significant effects became non-significant, and none of the non-significant effects became significant).

*R^2 when specifying a reliability of 0.85 for the dependent variable: U.S. (0.55); U.K. (0.52) Italy (0.48); France (0.35); Brazil (0.54); Mexico(0.54); Netherlands (0.42); Sweden (0.39);

(1) *n.s.* = not significant;
(2) Unstandardised regression coefficients are printed in **bold**. They may be used to make comparisons of effects BETWEEN countries. Standardised regression coefficients are printed between square brackets. They may be used to rank the effects WITHIN a given country according to their relative importance;
(3) Numbers between rounded brackets represent t-values;
(4) F1=Fair remuneration; F2=Supporting role of people within the department; F3=Immediate boss' support; F4=Clarity of strategy; F5=Confidence in managerial decisions; F6=Organisational and managerial efficiency; F7=Environmental and societal responsibility.

Table A.5.6/1 (continued)

	Australia (N=919)	Canada (N=739)	South Africa (N=632)	Germany (N=1107)	Hungary (N=707)	Belgium (N=658)	Russian Federation (N=632)	Israel (N=1063)
R^2	0.40*	0.46*	0.35*	0.34*	0.42*	0.39*	0.38*	0.37*
F1	**0.20** [0.21] (4.2)	**0.36** [0.31] (5.8)	**0.33** [0.33] (4.3)	**0.24** [0.21] (4.7)	**0.27** [0.36] (6.9)	**0.29** [0.33] (6.5)	**0.60** [0.53] (5.6)	**0.42** [0.40] (7.3)
F2	*n.s.*	*n.s.*	**0.12** [0.12] (2.0)	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*
F3	**0.15** [0.17] (3.6)	**0.13** [0.14] (2.6)	**0.15** [0.17] (3.5)	**0.20** [0.24] (6.0)	**0.17** [0.22] (5.2)	**0.29** [0.35] (6.0)	**0.17** [0.18] (3.0)	**0.18** [0.16] (3.1)
F4	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*
F5	*n.s.*	**0.33** [0.33] (2.7)	*n.s.*	**0.09** [0.11] (2.4)	**0.09** [0.12] (2.0)	*n.s.*	*n.s.*	*n.s.*
F6	**0.53** [0.27] (2.7)	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*
F7	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*

Notes:
All results presented in this table are estimations under the assumption that the indicator of employee job satisfaction (i.e. the dependent variable) is perfectly reliable. An additional analysis in which the reliability of the indicator was set equal to 0.85 showed that all results obtained were 'stable' (i.e. none of the significant effects became non-significant, and none of the non-significant effects became significant).

*R^2 when specifying a reliability of 0.85 for the dependent variable: Australia (0.47); Canada (0.54); South Africa (0.43); Belgium (0.48); Germany (0.40); Hungary (0.51); Russian Federation (0.44); Israel (0.44);

(1) *n.s.* = not significant;

(2) Unstandardised regression coefficients are printed in **bold**. They may be used to make comparisons of effects BETWEEN countries. Standardised regression coefficients are printed between square brackets. They may be used to rank the effects WITHIN a given country according to their relative importance;

(3) Numbers between rounded brackets represent t-values;

(4) F1=Fair remuneration; F2=Supporting role of people within the department; F3=Immediate boss' support; F4=Clarity of strategy; F5=Confidence in managerial decisions; F6=Organisational and managerial efficiency; F7=Environmental and societal responsibility.

Conclusions (from Table A.5.6/1)

Due to the biasing effect of the non-invariant intercept of indicator i025 on factor mean estimations, the partial tau-invariance model was used to assess the predictive power of the seven work environment factors on employees' job satisfaction.

Table 1 in Appendix 5.6 shows the percentage of explained variance in all countries. The average percentage of explained variance across countries is 39 percent (under the assumption that the indicator of employee job satisfaction is perfectly reliable).

Factor 1 (i.e. fair remuneration) and factor 3 (i.e. immediate boss' support) determine employee job satisfaction in all countries. The standardised regression coefficients in Table 1 show that in some countries fair remuneration (F1) is the most important determinant of job satisfaction (e.g. U.S., U.K., Brazil, Mexico, South Africa, Hungary, and in particular, the Russian Federation and Israel). In other countries, the immediate boss' support (F3) is the most important determinant (e.g. Italy, and The Netherlands).

Factor 4 (i.e. clarity of strategy) and factor 6 (i.e. organisational and managerial efficiency) are determinants of job satisfaction in 6 out of 16 countries. In the U.S., U.K., France, and Brazil, both factors determine job satisfaction. In Italy and the Netherlands, only clarity of strategy (F4) is a determinant of job satisfaction, whereas in Mexico and Australia, only organisational and managerial efficiency (F6) are determinants.

Confidence in managerial decisions (F5) is a determinant in the U.S., Canada, U.K., Italy, The Netherlands, Germany, and Hungary. The supporting role of people within the department (F2) is a determinant of job satisfaction in the U.S., Italy, Sweden, and South Africa.

Environmental and societal responsibility (F7) cannot be considered to be a determinant in any country (except for Mexico, where it is marginally significant, probably just because of the large sample size in Mexico).

When making such comparisons between countries, the reader should be aware that the large number of determinants in the U.S. and U.K. may be the result of the larger sample sizes in these countries. As a consequence, the results should be interpreted with caution.

References

' "There is more to life than what you read in books", said Weary.

"You'll find that out".'

K. Vonnegut

A

Alden, D.L., Steenkamp, J.-B.E.M., and Batra, R. (1999). Brand positioning through advertising in Asia, North America, and Europe: The role of the global consumer culture. Journal of Marketing, 63, 75-87.

Alwin, D.F., and Jackson, D.J. (1981). Applications of simultaneous factor analysis to issues of factorial invariance. In D.J. Jackson and E.F. Borgatta (Eds.), Factor analysis and measurement in sociological research (pp. 249-279). Beverly Hills, CA: Sage.

Anderson, J.C., and Gerbing, D.W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. Psychological Bulletin, 103, 411-423.

Andrich, D. (1978a). Application of a psychometric rating model to ordered categories which are scored with successive integers. Applied Psychological Measurement, 2, 4, 581-594.

Andrich, D. (1978b). A rating formulation for ordered response categories. Psychometrika, 43, 561-573.

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

B

Bagozzi, R.P. (1991). Structural equation modeling in marketing research. In W.D. Neal (Ed.), First annual Advanced Research Techniques Forum (pp. 335-379). Chicago, IL: American Marketing Association (AMA).

Bagozzi, R.P., and Edwards, J.R. (1998). A general approach for representing constructs in organizational research. Organizational Research Methods, 1, 1, 45-87.

Bagozzi, R.P., and Fornell, C. (1982). Theoretical concepts, measurements, and meaning. In C. Fornell (Ed.), A second generation of multivariate analysis (vol. 1). New York, NY: Praeger.

Bagozzi, R.P., and Phillips, L.W. (1982). Representing and testing organisational theories: A holistic construal. Administrative Science Quarterly, 27, 459-489.

Baligh, H.H. (1994). Components of culture: Nature, interconnections, and relevance to the decisions on the organizational structure. Management Science, 40, 1, 14-27.

Baron, R.M., and Kenny, D.A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology, 51, 6, 1173-1182.

Barrett, P.T. (1986). Factor comparison: An examination of three methods. Personality and Individual Differences, 7, 327-340.

Bartunek, J.M., and Franzak, F.J. (1988). The effects of organizational restructuring on frames of reference and cooperation. Journal of Management, 14, 4, 579-592.

Bauer, E. (1989). Ubersetzungsprobleme und Ubersetzungsmethoden bei einer multinationalen Marktforschung. In GfK Jahrbuch der Absatz- und Verbrauchsforschung, 2, 174-205.

Baumgartner, H., and Steenkamp, J.B.E.M. (2001). Response styles in marketing research: A cross-national investigation. Journal of Marketing Research, 38, 143-156.

Bearden, W.O., and Netemeyer, R.G. (1999). Handbook of marketing scales: Multi-item measures for marketing and consumer behavior research. Thousand Oaks, CA: Sage.

Bedeian, A., Ferris, G., and Kacmar, K. (1992). Age, tenure, and job satisfaction: A tale of two perspectives. Journal of Vocational Behavior, 40, 33-48.

Bentler, P.M. (1968). Alpha-maximized factor analysis (alphamax): Its relation to alpha and canonical factor analysis. Psychometrika, 33, 3, 335-345.

Bentler, P.M. (1990). Comparative fit indexes in structural models. Psychological Bulletin, 107, 238-246.

Bernardin, H.J., and Vilanova, P. (1986). Performance appraisal. In E. Locke (Ed.), Generalizing from laboratory to field settings (pp. 43-62). Lexington, MA: Lexington Books.

Berry, J.W. (1969). On cross-cultural comparability. International Journal of Psychology, 4, 2, 119-128.

Berry, J.W. (1989). Imposed etics – emics - derived etics: The operationalisation of a compelling idea. International Journal of Psychology, 24, 721-735.

Berscheid, E., Snyder, M., and Omoto, A.M. (1989). The Relationship Closeness Inventory: Assessing the closeness of interpersonal relationships. Journal of Personality and Social Psychology, 57, 792-807.

Bienstock, C.C., Mentzer, J.T., and Bird, M.M. (1997). Measuring physical distribution service quality. Journal of the Academy of Marketing Science, 25, 1, 31-44.

Billiet, J.B., and McClendon, McKee J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. Structural Equation Modeling: An Interdisciplinary Journal, 7, 4, 608-629.

Birnbaum, A. (1957). Efficient design and use of tests of a mental ability for various decision-making problems. Series report no. 58-16. Project no. 7755-23. Texas: USAF School of Aviation Medicine, Randolph Air Force Base.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M.R. Novick (Eds.), Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Blalock, H.M. (1964). Causal inferences in nonexperimental research. Chapell-Hil: University of North Carolina Press.

Blau, G., Merriman, K., Tatum, D.S., and Rudmann, S.V. (2001). Antecedents and consequences of basic versus career enrichment benefit satisfaction, Journal of Organizational Behavior, 22, 669-688.

Boddewyn, J.J., and Iyer, G. (1999). International business research: Beyond déjà vu. Management International Review, 39, 2, 161-184.

Bollen, K.A. (1982). A confirmatory factor analysis of subjective air quality. Evaluation Review, 6, 521-535.

Bollen, K.A. (1984). Multiple indicators: Internal consistency or no necessary relationship. Quality and Quantity, 18, 377-385.

Bollen, K.A. (1989). Structural equations with latent variables. New York, NY: J. Wiley.

Bollen, K.A. (2002). Latent variables in psychology, Annual Review of Psychology, 53, 605-634.

Bollen, K.A., and Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. Psychological Bulletin, 110, 2, 305-314.

Bollen, K.A., and Ting, K. F. (1993). Confirmatory tetrad analysis. In P. Marsden (Ed.), Sociological methodology 1993 (pp. 147-175). Washington, DC: American Sociological Association (ASA).

Bollen, K.A., and Ting, K. F. (2000). A tetrad test for causal indicators. Psychological Methods, 5, 1, 3-22.

Boomsma, A. (2003). Covariantiestructuuranalyse I & II. University of Groningen, The Netherlands. (textbook for the course on Structural Equation Modelling):

Borg, I., and Groenen, P. (1997). Modern multidimensional scaling: Theory and applications. New York, NY: Springer.

Borg, I., and Leutner, D. (1985). Measuring the similarity of MDS configurations. Multivariate Behavioral Research, 20, 325-334.

Borsboom, D., Mellenbergh, G.J., and van Heerden, J. (2003). The theoretical status of latent variables. Psychological Review, 110, 2, 203-219.

Borsman, W.C. (1991). Job behavior, performance, and effectiveness. In M.D. Dunnette, and L.M. Hough (Eds.), Handbook of industrial and organizational psychology (vol. 2, pp. 271-326). Palo Alto, CA: Consulting Psychologists Press.

Braun, M. (2000). Evaluation der Äquivalenz eines gemeinsamen Satzes an Indikatoren in der Interkulturell vergleichende Sozialforschung. ZUMA How-to-Reihe no. 3. Mannheim, Germany: ZUMA

Braun, M., and Scott, J. (1998). Multidimensional scaling and equivalence: Is having a job the same as working? Zuma Nachrichten Spezial, 3 (special issue on 'Cross-Cultural Survey Equivalence' edited by J.A. Harkness), 129-144.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, P.J. (1984). Classification and regression trees. Belmont, CA: Wadsworth.

Brennan, R.L. (2001). Generalizability theory. New York, NY: Springer-Verlag.

Broderick, A.J. (1999). Testing for metric equivalence using confirmatory factor analysis: A consumer involvement study. Working paper no. RP 9903. Birmingham, United Kingdom: Aston Business School.

Brokken, F.B. (1983). Orthogonal Procrustes rotation maximizing congruence. Psychometrika, 48, 343-352.

Browne, M.W. (1982). Covariance structures. In D.M. Hawkins (Ed.), Topics in applied multivariate analysis. London, United Kingdom: Cambridge University Press.

Browne, M.W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. British Journal of Mathematical and Statistical Psychology, 37, 62-83.

Browne, M.W., and Arminger, G. (1995). Specification and estimation of Mean- and Covariance- Structure Models. In Arminger, G., Clogg, C., and Sobel, M.E. (Eds.), Handbook of statistical modeling for the social and behavioral sciences (pp. 185-249) New York, NY: Plenum Press.

Bruner, G., and Hensel, P. (1997). Marketing scales Handbook: A compilation of multi-item measures. (2nd ed.) Chicago: American Marketing Association.

Burke, M.C. (1984). Strategic choice and marketing managers: An examination of business-level marketing objectives. Journal of Marketing Research, 21, 345-359.

Burt, C.L. (1948). The factorial study of temperamental traits. British Journal of Psychology, Statistical Section, 1, 178-203.

Byrne, B.M. (1989). A primer of LISREL: Basic applications and programming for confirmatory factor analytic models. New York, NY: Springer-Verlag.

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

Byrne, B.M., and Watkins, D. (2003). The issue of measurement invariance revisited. Journal of Cross-Cultural Psychology, 34, 2, 155-175.

C

Camilli, G., and Shepard, L.A. (1994). MMSS Methods for identifying biased test items. Thousand Oaks, CA: Sage.

Candell, G.L., and Hulin, C.L. (1986). Cross-language and cross-cultural comparisons in scale transitions: Independent sources of information item nonequivalence. Journal of Cross-Cultural Psychology, 17, 4, 417-440.

Carlson, L., and Grosshart, S. (1988). Parental style and consumer socialisation of children. Journal of Consumer Research, 15, 77-94.

Cattell, R.B. (1949). A note on factor invariance and the identification of factors. British Journal of Psychology, 2, 134-138.

Cattell, R.B., and Baggaley, A.R. (1960). The salient variable similarity index for factor matching. British Journal of Statistical Psychology, 13, 33-46.

Cattell, R.B., Balcar, K.R., Horn, J.L., and Nesselroade, J.R. (1969). Factor matching procedures: An improvement of the s index; with tables. Educational and Psychological Measurement, 29, 781-792.

Cavusgil, S.T., and Das, A. (1997). Methodological issues in empirical cross-cultural research: A survey of the management literature and a framework. Management International Review, 37, 1, 71-96.

Cermak, G.W. (1983). An experimental test of two models of attribute integration. General Motors Research Publication no. GMR-4386. Warren, Michigan.

Chan, D. (1998). The conceptualization and analysis of change over time: An integrative approach incorporating longitudinal mean and covariance structures analysis (LMACS) and multiple indicator latent growth modeling (MLGM). Organizational Research Methods, 1, 421-483.

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

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

Cheung, G.W., and Rensvold, R.B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. Journal of Cross-Cultural Psychology, 31, 2, 187-212.

Cheung, G.W., and Rensvold, R.B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. Structural Equation Modeling: An Interdisciplinary Journal, 9, 2, 233-255.

Chin, W.W. (1995). Partial least squares is to LISREL as principal component analysis is to common factor analysis. Technology Studies, 2, 315-319.

Chin, W.W. (1998). The partial least squares approach to structural equation modelling. In G.A. Marcoulides (ed.). Modern methods for business research (pp. 295-336). Mahwah, NJ: Lawrence Erlbaum.

Chou, C.-P., and Bentler, P.M. (1995). Estimates and tests in structural equation modeling. In Hoyle, R.H. (Ed.), Structural equation modeling: Concepts, issues and applications (pp. 37-55). Newbury Park, CA: Sage

CIM (Chartered Institute of Marketing) (1999). Syllabus text of the Diploma Paper no. 1 on International Marketing Strategy. London (U.K.): BPP Publishing.

Clark, L.A., and Watson, D. (1995). Constructing validity: Basic issues in scale development. Psychological Assessment, 7, 3, 309-319.

Cliff, N. (1966). Orthogonal rotation to congruence. Psychometrika, 31, 33-42.

Clogg, C.C., and Goodman, L.A. (1985). Simultaneous latent structure analysis in several groups. Sociological Methodology, 15, 81-110.

Cohen, A. (1993). Organizational commitment and turnover: A meta-analysis. Academy of Management Journal, 36, 5, 1140-1157.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46.

Cohen, P.J., Cohen, J., Teresi, J., Marchi, M., and Velez, C.N. (1990). Problems in the measurement of latent variables in structural equations causal models. Applied Psychological Measurement, 14, 2, 183-196.

Cohen, A.S., Kim, S.H., and Baker, F.B. (1993). Detection of differential item functioning in the graded response model. Applied Psychological Measurement, 17, 335-350.

Cole, D.A., and Maxwell, S.E. (1985). Multitrait-multimethod comparisons across populations: A confirmatory factor analytic approach. Multivariate Behavioral Research, 20, 389-417.

Cole, D.A., Maxwell, S.E., Arvey, R., and Salas, E. (1993). Multivariate group comparisons of variable systems: MANOVA and structural equation modeling. Psychological Bulletin, 114, 1, 174-184.

Cooke, D.J., Kosson, D.S., and Michie, C. (2001). Psychopathy and ethnicity : Structural, item, and test generalizability of the psychopathy checklist – revised (PCL-R) in Caucasian and African American participants. Psychological Assessment, 13, 4, 531-542.

Costner, H.L. (1969). Theory, deduction, and the rules of correspondence. American Journal of Sociology, 75, 245-263.

Craig, C.S, and Douglas, S.P. (2000). International Marketing Research. New York, NY: J. Wiley.

Crocker, L., and Algina, J. (1986). Introduction to classical and modern test theory. Fort Worth, TX: Harcourt Brace.

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

Curran, P.J., West, S.G., and Finch, J.F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. Psychological Methods, 1, 16-29.

D

Davidson, A.R., Jaccard, J.J., Triandis, H.C., Morales, M.L., and Diaz-Guerrero, R. (1976). Cross-cultural model testing: Toward a solution of the etic-emic dilemma. International Journal of Psychology, 11, 1, 1-13.

De Beuckelaer, A. (2002). Comparison of Construct Mean Scores Across Populations: A Conceptual Framework (pp. 175-182). In S. Nishisato; Y. Baba, H. Bozdogan, and K. Kanefuji (Eds.), Measurement and Multivariate Analysis, Tokyo, Japan: Springer-Verlag.

DeVellis, R.F. (1991). Scale development: Theories and applications. Newbury Park, CA: Sage.

De Vera, M.V. (1985). Establishing cultural relevance and measurement equivalence using emic and etic items. Unpublished dissertation. Urbana, IL: University of Illinois.

DeVellis, R.F. (2003). Scale development : Theory and applications. (2nd edition). Thousand Oaks, CA: Sage.

Devins, G.M., Beiser, M., Dion, R., Pelletier, L.G., and Edwards, R.G. (1997). Cross-cultural measurements of psychological well-being: The psychometric equivalence of Cantonese, Vietnamese, and Laotian translations of the affect balance scale. American Journal of Public Health, 87, 794-799.

Diamantopoulos, A., and Winklhofer, H.M. (2001). Index construction with formative indicators: An alternative to scale development. Journal of Marketing Research, 18, 269-277.

Douglas, S.P., and Craig, C.S. (1992). Advances in international marketing. International Journal of Research in Marketing, 9, 3, 291-318.

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

Drasgow, F. (1987). Study of measurement bias of two standardized psychological tests. Journal of Applied Psychology, 72, 19-29.

Drasgow, F., and Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. Journal of Applied Psychology, 70, 662-680.

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

Durvasula, S., Lysonsky, S., and Watson, J. (2001). Does vanity describe other cultures? A cross-cultural examination of the vanity scale. The Journal of Consumer Affairs, 35, 1, 180-199.

Du Toit, M. (Ed.) (2003). IRT from SSI. Lincolnwood, IL: Scientific Software International (SSI).

E

Edwards, J.R., and Bagozzi, R.P. (2000). On the nature and direction of relationships between constructs and measures. Psychological methods, 5, 2, 155-174.

Efron, B., and Tibshirani, R. (1993). An introduction to the bootstrap. New York, NY: Chapman and Hall.

Eid, M., Langeheine, R., and Diener, E. (2003). Comparing typological structures across cultures by multigroup latent class analysis. Journal of Cross-Cultural Psychology, 34, 2, 195-210.

Eid, M., and Rauber, M. (2000). Detecting measurement invariance in organisational surveys. European Journal of Psychological Assessment, 16, 1, 20-30.

Ellis, B.B., Becker, P., and Kimmel, H.D. (1993). An item response theory evaluation of an English version of the Trier Personality Inventory (TPI). Journal of Cross-Cultural Psychology, 24, 133-148.

Ellis, B.B., Minsel, B., and Becker, P. (1989). Evaluation of attitude survey translations: An investigation using item response theory. International Journal of Psychology, 24, 665-684.

Embretson, S.E. (1996). The new rules of measurement. Psychological Assessment, 8, 4, 341-349.

Embretson, S.E., and Reise, S.P. (2000). Item response theory for psychologists. Mahwah, NJ: Erlbaum.

Engelen, R.J.H., and Eggen, T.J.H.M. (1993). Equivaleren. In T.J.H.M. Eggen, and P.F. Sanders (Eds.), Psychometrie in de praktijk (pp. 179-238). Arnhem, The Netherlands: CITO.

Erez, M. (1994). Toward a model of cross-cultural industrial and organizational psychology. In H.C. Triandis, M.D. Dunnette, and L.M. Hough (Eds.), Handbook of

industrial and organizational psychology (vol. 4, pp. 559-608). Palo Alto, CA: Consulting Psychologists Press.

F

Facteau, J.D., and Craig, S.B. (2001). Are performance appraisal ratings from different sources comparable? Journal of Applied Psychology, 86, 2, 215-227.

Fornell, C. (1982). A second generation of multivariate analysis: An overview. In C. Fornell (Ed.), A second generation of multivariate analysis (vol. 1, pp. 1-21). New York, NY: Praeger.

Fornell, C., and Bookstein, F.L. (1982). Two structural equation models: LISREL and PLS applied to consumer exit-voice theory. Journal of Marketing Research, 19, 440-452.

Fornell, C., and Cha, J. (1994). Partial Least Squares. In R.P. Bagozzi (Ed.), Advanced methods of marketing research (pp. 52-78). Cambridge, MA: Blackwell.

Fornell, C., Rhee, B.-D., and Yi, Y. (1991). Direct regression, reverse regression, and covariance structure analysis. Marketing Letters, 2, 3, 309-320.

G

Gaski, J.F., and Nevin J.R. (1985). The differential effects of exercised and unexercised power sources in a marketing channel. Journal of Marketing Research, 22, 130-142.

Gielens, K., and Dekimpe, M.G. (2001). Do international entry decisions of retail chains matter in the long run? International Journal of Research in Marketing, 18, 3, 235-259.

Glas, C.A.W., and Verhelst, N.D. (1993). Een overzicht van itemresponsemodellen. In T.J.H.M. Eggen, and P.F. Sanders (Eds.), Psychometrie in de praktijk (pp. 179-238). Arnhem, The Netherlands: CITO.

Glymour, C., Scheines, R., Spirtes, P, and Kelly, K. (1987). Discovering causal structure. Orlando, FL: Academic Press.

Goffman, E. (1974). Relations in public. New York, NY: Harper and Row.

Goldberg, L.R. (1990). An alternative 'Description of Personality': The Big-Five factor structure. Journal of Personality and Social Psychology, 59, 6, 1216-1229.

Goldberg, M.E., and Hartwick, J. (1990). The effects of advertiser reputation and extremity of advertising claim on advertising effectiveness. Journal of Consumer Research, 17, 172-179.

Goodenough, W.H. (1971). Culture, language, and society. Reading, MA: Addison-Wesley.

Goodman, L. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika, 61, 215-231.

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.

Green, R.T., and Alden, D.L. (1988). Functional equivalence in cross cultural consumer behavior: The case of gift giving in Japan. Psychology and Marketing, 5, 2, 159-172.

Groenen, P. (2002). Modern multidimensional scaling. Lecture notes of the 2002 Spring Seminar. Cologne, Germany: Zentralarchiv für Empirische Sozialforschung.

Grunert, K.G., Brunso K., and Bisp, S. (1993). Food-related life style: Development of a cross-culturally valid instrument for market surveillance. MAPP working paper no. 12. Aarhus, Denmark: Aarhus School of Business.

Grunert, S.C., Grunert, K.G., and Kristensen, K. (1994). Une méthode d' etimation de la validité interculturelle des instruments de mesure: Le cas de la mesure des valeurs de consomateurs par la liste des valeurs LOV. Recherches et Applications en Marketing, 8, 4, 5-28.

Guadagnoli, E., and Velicer, W. (1991). A comparison of pattern matching indices. Multivariate Behavioral Research, 26, 2, 323-343.

H

Hambleton, R.K., and Kanjee, A. (1995). Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptations. European Journal of Psychological Assessment, 11, 147-157.

Hambleton, R.K., and Swaminathan, H. (1985). Item response theory, principles and applications. Dordrecht, The Netherlands: Kluwer-Nijhof.

Hancock, G.R, and Mueller, R.O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. du Toit, and D. Sörbom (Eds.), Structural Equation Modeling: Present and Future – Festschrift in honor of Karl Jöreskog (pp. 195-216). Licolnwood, IL: Scientific Software International (SSI).

Harris, P.R., and Moran, R.T. (1987). Managing cultural differences (2nd edition). Houston: Gulf.

Hassan, S.S., and Katsanis, L.P. (1994). Global market segmentation strategies and trends. In E. Kaynak, and S.S. Hassan (Eds.), Globalization of consumer markets: Structures and strategies (pp. 47-63). New York, NY: International Business Press.

Hassan, S.S., and Kaynak, E. (1994). The globalizing consumer market: Issues and concepts. In E. Kaynak, and S.S. Hassan (Eds.), Globalization of consumer markets: Structures and strategies (pp. 19-25). New York, NY: International Business Press.

Hauser, R.M., and Goldberger, A.S. (1971). The treatment of unobservable variables in path analysis. In H.L. Costner (Ed.), Sociological Methodology 1971 (pp. 81-117). San Francisco: Jossey-Bass.

Heise, D.R. (1972). Employing nominal variables, induced variables, and block variables in path analysis. Sociological Methods and Research, 1, 147-173.

Helfrich, H. (1999). Beyond the dillema of cross-cultural psychology: Resolving the tension between etic and emic approach. Culture and Psychology, 5, 2, 131-153.

Hinkin, T.R. (1995). A review of scale development practices in the study of organizations. Journal of Management, 21, 967-988.

Hinkin, T.R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. Organizational Research Methods, 1, 104-121.

Hittner, J.B. (1995). Factorial validity and equivalency of the alcohol expectancy questionnaire tension-reduction subscale across gender and drinking frequency. Journal of Clinical Psychology, 51, 563-576.

Hofstede, G. (1976). Nationality and espoused values of managers. Journal of Applied Psychology, 61, 148-155.

Hofstede, G. (1983). The cultural relativity of organizational practices and theories. Journal of International Business Studies, 14, 75-89.

Hofstede, G. (1985). The interaction between national and organizational value systems. Journal of Management Studies, 22, 347-357.

Hollis, M., and Muthén, B. (1987). Structural covariance models with categorical data: An illustration involving the measurement of political attitudes and belief systems. Paper presented at the meeting of the American Political Science Association, Chicago, IL (September 1987).

Horn, J.L., and McArdle, J.J. (1992). A practical and theoretical guide to measurement invariance in aging research. Experimental Aging Research, 18, 117-144.

Horn, J.L., McArdle, J.J., and Mason, R. (1983). When is invariance not invariant: A practical scientist's look at the ethereal concept of factor invariance. Southern Psychologist, 1, 179-188.

Hosmer, D.W., and Lemeshow, S. (1989). Applied logistic regression. New York, NY: J. Wiley.

Hu, L., and Bentler, P.M. (1995). Evaluating Model Fit. In R.H. Hoyle (Ed.). Structural Equation Modeling. Concepts, issues, and applications (pp. 76-99). Thousand Oaks, CA: Sage.

Hu, L., and Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling: An Interdisciplinary Journal, 6, 1, 1-55.

Hu, L., Bentler, P.M., and Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? Psychological Bulletin, 112, 351-362.

Huberty, C.J., and Morris, J.D. (1989). Multivariate analysis versus multiple univariate analyses. Psychological Bulletin, 105, 2, 302-308.

Hui, C.H. (1990). Work attitudes, leadership styles, and managerial behaviors in different cultures. In R.W. Brislin (Ed.), Applied Cross-cultural psychology, Newbury Park, CA: Sage.

Hui, C.H., and Triandis, H.C. (1983). Multistrategy approach to cross-cultural research: The case of locus of control. Journal of Cross-Cultural Psychology, 14, 65-83.

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

Hulin, C.L. (1987). A psychometric theory of evaluations of item and scale translations: Fidelity across languages. Journal of Cross-Cultural Psychology, 18, 115-142.

Hulin, C.L., Drasgow, R., and Komocar, J. (1982). Applications of item response theory to analysis of attitude scale translations. Journal of Applied Psychology, 67, 6, 818-825.

Hulin, C.L., and Mayer, L.J. (1985). Psychometric equivalence of a translation of the Job Descriptive Index into Hebrew. Journal of Applied Psychology, 71, 83-94.

Hulland, J. (1999). Use of partial least squares (PLS) in strategic management research. Strategic Management Journal, 20, 195-204.

Hurley, J., and Cattell, R.B. (1962). The Procrustes program: Producing direct rotation to test a hypothesized factor structure. Behavioral Science, 7, 258-262.

I

Igra, A. (1979). On forming variable set composites to summarize a block recursive model. Social Science Research, 8, 253-264.

Inglehart, R. (1997). Modernization and postmodernization: Cultural, economic, and political change in 43 societies. Princeton University Press.

Inglehart, R., and Baker, W. (2000). Modernization, cultural change and the persistence of traditional values. American Sociological Review, 65, 1, 19-51.

Ironson, G., Smith, P., Brannick, M., Gobson, W., and Paul, K. (1989). Construction of a job in general scale: A comparison of global composite and specific measures. Journal of Applied Psychology, 74, 193-200.

J

Jagpal, H.S. (1981). Measuring joint advertising effects in multiproduct firms. Journal of Advertising Research, 21, 1, 65-69.

Jarley, P., Fiorito, J., and Delaney, J.T. (1997). A structural, contingency approach to bureaucracy and democracy in US national unions. Academy of Management Journal, 40, 831-861.

Jaworski, B.J., and Kohli, A.K. (1993). Market orientation: Antecedents and consequences. Journal of Marketing, 57, 53-70.

Johnson, T.P. (1998). Approaches to equivalence in cross-cultural and cross-national survey research. Zuma Nachrichten Spezial, 3, 1-40 (special issue on 'cross-cultural survey equivalence' edited by J.A. Harkness).

Johnson, T.P., O'Rourke, D., Chavez, N., Sudman, S., Warnecke, R., Lacey, L., and Horn, J. (1997) Social Cognition and responses to survey questions among culturally diverse populations. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin (Eds.). Survey measurement and process quality (pp. 87-113). New York: J. Wiley.

Jöreskog, K.G. (1969). A general approach to confirmatory maximum likelihood factor analysis. Psychometrika, 34, 183-202.

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

Jöreskog, K.G. (1973). Analysis of covariance structures. In P.R. Krishnaiah (Ed.), Multivariate Analysis – III (pp. 263-285). New York, NY: Academic Press.

Jöreskog, K.G., and Goldberger, A.S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. Journal of the American Statistical Association, 70, 631-639.

Jöreskog, K.G., and Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. Multivariate Behavioral Research, 36, 3, 347-387.

Jöreskog, K.G., and Sörbom, D. (1993). Lisrel 8: Structural Equation Modeling with the SIMPLIS Command Language. Lincolnwood, IL: Scientific Software International.

Judge, T.A., and Bretz, R.D. (1994). Political influence behavior and career success. Journal of Management, 20, 43-65.

Judge, T.A., Locke, E.A., Durham, C.C., and Kluger, A.N. (1998). Dispositional effects on job and life satisfaction: The role of core evaluations. Journal of Applied Psychology, 83, 1, 17-34.

K

Kanji, G.K. (1993). 100 Statistical tests. Thousand Oaks, CA: Sage.

Kaplan, D., and George, R. (1995). A study of the power associated with testing factor mean differences under violations of factorial invariance. Structural Equation Modeling: An Interdisciplinary Journal, 2, 2, 101-118.

Katerberg, R., Smith, F.J., and Hoy, S. (1977). Language, time, and person effects on attitude scale translations. Journal of Applied Psychology, 62, 4, 385-391.

Katigbak, M.S., Church, A.T., and Akamine, T.X. (1996). Cross-cultural generalizability of personality dimensions: Relating indigenous and imported dimensions in two cultures. Journal of Personality and Social Psychology, 70, 1, 99-114.

Kerlinger, F.N. (1986). Foundations of behavioral research. New York, NY: Hold, Rinehart, and Winston.

Kiers, H.A.L. (1990). SCA: A program for simultaneous components analysis. Groningen, The Netherlands: IEC ProGamma.

Kiers, H.A.L., and ten Berge, J.M.F. (1989). Alternating Least Squares Algorithms for simultaneous component analysis with equal component weight matrices for all populations. Psychometrika, 54, 3, 467-473.

Kim, J.O., and Mueller, C.W. (1978). Factor analysis. Beverly Hills, CA: Sage.

Kluckhohn, F.R., and Strodtbeck, F.L. (1961). Variations in value orientations. Westport, CT: Greenwood Press.

Korth, B., and Tucker, L.R. (1976). Procrustes matching by congruence coefficients. Psychometrika, 41, 531-535.

Kroontz, H. (1980). The management theory jungle. Academy of Management Review, 5, 175-187.

Kühnel, S. (1988). Testing MANOVA designs with LISREL. Sociological Methods and Research, 16, 4, 504-523.

Kumar, V. (2000). International Marketing Research. Upper Saddle River, NJ: Prentice-Hall.

L

Labouvie, E.W. (1980). Identity versus equivalence of psychological measures and constructs. In L.W. Poon (Ed.). Aging in the 1980's (pp. 493-502). Washington, DC: American Psychological Association (APA).

Lambert, E., Hogan, N.L., and Barton, S.M. (2001). The impact of job satisfaction on turnover intent: A test of a structural measurement model using a national sample of workers. The Social Science Journal, 38, 233-250.

Langeheine, R. (1980). Appropriate norms and significance tests for the Lingoes-Borg Procrustean Individual Differences Scaling PINDIS, Technical Report no. 39. Kiel, Germany: Universität Kiel, Institut für Pädagogik der Naturwissenschaften.

Langeheine, R. (1982). Statistical evaluations of measures of fit in the Lingoes-Borg Individual Differences Scaling. Psychometrika, 47, 427-442.

Lastovicka, J.L. (1982). On the validation of lifestyle traits: A Review and illustration. Journal of Marketing Research, 19, 126-138.

Lazarsfeld, P.F., and Henry, N.W. (1968). Latent structure analysis. Boston, MA: Houghton and Mifflin.

Leutner, D., and Borg, I. (1983). Zufallskritische Beurteilung der Übereinstimmung von Faktor- Und MDS- Konfigurationen. Diagnostica, 24, 320-335.

Levitt, T. (1983). The globalization of markets. Harvard Business Review, 61, 92-102.

Liang, J. (1986). Self-reported physical health among aged adults. Journal of Gerontology , 41, 248-260.

Lilien, G.L., and Rangaswamy, A. (2003). Marketing engineering: Computer-assisted marleting analysis and planning. (2nd edition). Prentice Hall.

Lincoln, J.R., Hanada, M., and Olson, J. (1981). Cultural orientations and individual reactions to organizations: A study of employees of Japanese-owened firms. Administrative Science Quarterly, 26, 93-115.

Little, T.D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. Multivariate Behavioral Research, 32, 53-76.

Locke, E. (1976). The nature and causes of job satisfaction. In M. Dunnell (Ed.), Handbook of industrial and organisational psychology (pp. 1297-1349). Chicago, IL: Rand-McNally.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. Psychological Reports, 3, 1-18.

Long, J. Scott (1997). Regression models for categorical and limited dependent variables. Thousand Oaks, CA: Sage.

Lord, F.M. (1952). A theory of test scores. Psychometric Monograph 7.

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

Lord F.M., and Novick, M.R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Lumpkin, J.R., and Hunt, J.B. (1989). Mobility as an influence on retail patronage behavior of the elderly: Testing conventional wisdom. Journal of the Academy of Marketing Science, 17, 1-12.

M

MacCallum, R.C., and Austin, J.T. (2000). Applications of structural equation modelling in psychological research. Annual Review of Psychology, 51, 201-226.

MacCallum, R.C., and Browne, M.W. (1993). The use of causal indicators in covariance structure models: some practical issues. Psychological Bulletin, 114, 3, 533-541.

Mahajan, V., and Muller, E. (1994). Innovation diffusion in a borderless global market: Will the 1992 unification of the European Community accelerate diffusion of new ideas, products, and technologies? Technological Forecasting and Social Change, 45, 221-235.

Maheswaran, D., and Shavitt, S. (2000). Issues and new directions in global consumer psychology. Journal of consumer Psychology, 9, 2, 59-66.

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

Manning, K.C., Bearden, W.O., and Madden, T.J. (1995). Consumer innovativeness and the adoption process. Journal of Consumer Psychology, 4, 4, 329-345.

Marsden, P.V. (1982). A note on block variables in multi-equation models. Social Science Research, 11, 127-140.

Marsh, H.W. (1990). Self Description Questionnaire II: A theoretical and empirical basis for the measurement of multiple dimensions of adolescent self-concept. A test manual and a research monograph. San Antonio, TX: Psychological Corporation.

Marsh, H.W. (1994). Confirmatory factor analysis models of factorial invariance: A multifaceted approach. Structural Equation Modeling: An Interdisciplinary Journal, 1, 5-34.

Marsh, H.W., and Grayson, D. (1990). Public/catholic differences in the high school and beyond data: A multi-group structural equation modelling approach to testing mean differences. Journal of Educational Statistics, 5, 199-235.

Marsh, H.W., and Grayson, D. (1994). Longitudinal stability of latent means and individual differences: A unified approach. Structural Equation Modeling: An Interdisciplinary Journal, 1, 2, 317-359.

Marsh, H.W., and Hocevar, D. (1985). Application of confirmatory factor analysis to the study of self-concept: First- and higher order factor models and their invariance across groups. Psychological Bulletin, 97, 562-582.

Marsh, R.M., and Manari, H. (1977). Organizational commitment and turnover: A prediction study. Administration Science Quarterly, 22, 57-75.

Martensen, A., and Grønholdt, L. (2001). Using employee satisfaction measurement to improve people management: An adaptation of Kano's quality types. Total Quality Management, 12, 7 & 8, 949-957.

Martinez, Z.L., and Toyne, B. (2000). What is international management, and what is its domain? Journal of International Management, 6, 1, 11-28.

Masters, G.N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 2, 149-172.

Mathieu, J.E., and Zajac, D.M. (1990). A review and meta analysis of the antecedents, correlates, and consequences of organizational commitment. Psychological Bulletin, 108, 2, 171-194.

Maurer, T.J., Raju, N.S., and Collins, W.C. (1998). Peer and subordinate performance appraisal measurement equivalence. Journal of Applied Psychology, 83, 5, 693-702.

Mavondo, F., Gabbott, M., and Tsarenko, M. (2003). Measurement invariance of marketing instruments: An implication across countries. Journal of Marketing Management, 19, 523-540.

McCutcheon, A.L. (1987). Latent Class Analysis. Beverly Hills, CA: Sage.

McKee, D.O., Varadarajan, P.R., and Pride, W.M. (1989). Strategic adaptability and firm performance: A market contingent perspective. Journal of Marketing, 53, 21-35.

Mellenbergh, G.J. (1994). A unidimensional latent trait model for continuous item responses. Multivariate Behavioral Research, 29, 223-236.

Mellenbergh, G.J. (1995). Conceptual notes on models for discrete polytomous item responses. Applied Psychological Measurement, 19, 1, 91-100.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. Psychometrika, 58, 525-543.

Messick, S. (1995). Validity of psychological assessment. American Psychologist, 50, 741-749.

Miller, J., Slomczynski, K.M., and Kohn, M.L. (1985). Continuity of learning generalization: The effect of job on men's intellective process in the United States and Poland. American Journal of Sociology, 91, 593-615.

Millsap, R.E., and Everson, H.T. (1993). Methodology review: Statistical approaches for assessing measurement bias. Applied Psychological Measurement, 17, 3, 297-334.

Millsap, R.E., and Hartog, S.B. (1988). Alpha, beta, and gamma change in evaluation research: A structural equation approach. Journal of Applied Psychology, 73, 574-584.

Mitchell, T.R., Holtom, B.C., Lee T.W., Sablynski, C.J., and Erez, M. (2001). Are there methodological and substantive roles for affectivity in job diagnostic survey relationships? Journal of Applied Psychology, 81, 795-805.

Mitra, D., and Golder, P.N. (2002). Whose culture matters? Near-market knowledge and its impact on foreign market entry timing. Journal of Marketing Research, 39, 3, 350-365.

Mobley, W., Griffeth, R., Hand, H., and Meglino, B. (1979). Review and conceptual analysis of the employee turnover process. Psychological Bulletin, 86, 493-522.

Mobley, W., Horner, W., and Hollingsworth, A. (1978). An evaluation of the precursors of hospital employee turnover. Journal of Applied Psychology, 63, 408-414.

Mooney, C.Z., and Duval, R.D. (1993). Bootstrapping. A nonparametric approach to statistical inference Series: Quantitative Applications in the Social Sciences no. 95. London, United Kingdom: Sage.

Moorman, C., and Matulich, E. (1993). A model of consumers' preventive health motivation and health ability. Journal of Consumer Research, 20, 208-228.

Mowday, R.T., Porter, L.W, and Steers, R.M. (1982). Employee-organization linkages. New York, NY: Academic Press.

Mulaik, S.A. (1975). Confirmatory factor analysis. In D.J. Amick, and H.J. Walberg (Eds.), Introductory multivariate analysis (pp. 170-207). Berkeley, CA: McCutchan.

Mullen, M.R. (1995). Diagnosing measurement equivalence in cross-national research. Journal of International Business Studies, 26, 3, 573-596.

Muraki, E. (1992). A generalised partial credit model: Application of an EM algorithm. Applied Psychological Measurement, 16, 2, 159-176.

Muraki, E., and Bock, R.D. (1997). PARSCALE: Item based test scoring and item analysis for graded open-ended exercises and performance tasks. Chicago, IL: Scientific Software International Inc.

Muthén, B.O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika, 49, 115-132.

Muthén, B.O. (1989). Latent variable modelling in heterogeneous populations. Psychometrika, 54, 4, 557-585.

Muthén, B.O., and Christoffersson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. Psychometrika, 46, 407-419.

Muthén, L.K., and Muthén, B.O. (1999). Mplus, the comprehensive modeling program for applied researchers (User's guide – 2nd printing). Los Angeles, CA: Muthén and Muthén.

Muthén, L.K., and Muthén, B.O. (2003). Mplus version 2.13 Addendum to the Mplus user's guide. Los Angeles, CA: Muthén and Muthén.

Myers, M.B., Calantone, R., Page, T.J., and Taylor, C.R. (2000). An application of multiple-group causal models in assessing cross-cultural measurement equivalence. Journal of International Marketing, 8, 4, 108-121.

N

Nesselroade, J.R., and Thompson, W.W. (1995). Selection and related threats to group comparisons: An example comparing factorial structures of higher and lower ability groups of adult twins. Psychological Bulletin, 117, 2, 271-284.

Netemeyer, R.G., Bearden, W.O., and Sharma, S. (2003). Scaling Procedures: Issues and applications. Thousand Oaks, CA: Sage.

Netemeyer, R.G., Burton, S., and Lichtenstein, D.R. (1995). Trait aspects of vanity: Measurement and relevance to consumer behavior. Journal of Consumer Research, 21, 612-626.

Nevis, E.C. (1983). Cultural assumptions and productivity. Sloan Management Review, Spring, 11-29.

Nunnally, J.C. (1978). Psychometric theory (2nd edition). New York, NY: McGraw-Hill.

Nunnally, J.C., and Bernstein, I.H. (1994). Psychometric theory (3rd edition). New York, NY: McGraw-Hill.

O

Olsson, U.H., Finch, J.F., and Curran, P.J. (1995). The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and nonnormality. Structural Equation Modeling: An Interdisciplinary Journal, 7, 557-595.

Oort, F.J. (1994). Potentiele schenders van de ééndimensionaliteit van psychologische meetinstrumenten. Nederlands Tijdschrift voor de Psychologie, 49, 35-46.

Orley, Dr. J. (1993). Study protocol for the World Health Organisation project to develop a quality of life assessment instrument (WHOQOL). Quality of Life Research, vol. 2.

P

Parasuraman, A., Berry, L.L., and Zeithaml, V.A. (1988). SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. Journal of Retailing, 64, 1, 14-40.

Parasuraman, A., Berry, L.L., and Zeithaml, V.A. (1994). Alternative scales for measuring service quality: A comparative assessment based on psychometric and diagnostic criteria. Journal of Retailing, 70, 3, 201-230.

Parker, P.M., and Tavassoli, N.T. (2000). Homeostasis and consumer behavior across cultures. International Journal of Research in Marketing, 17, 1, 33-53.

Paunonen, S.V. (1997). On chance and factor congruence following orthogonal Procrustes rotation. Educational and Psychological Measurement, 57, 1, 33-59.

Pike, K.L. (1954). Emic and etic standpoints for the description of behavior. In K.L. Pike (Ed.), Language in relation to a unified theory of the structure of human behavior (pp. 8-28). Glendale, IL: Summer Institute of Linguistics.

Pike, K.L. (1971). Language in relation to a unified theory of the structure of human behavior. The Hague, The Netherlands: Mouton.

Ployhart, R.E., Wiechmann, D., Schmitt, N., Sacco, J.M., and Rogg, K. (2003). The cross-cultural equivalence of job performance ratings. Human Performance, 16, 1, 49-79.

Popper, K.G. (1959). The logic of scientific discovery. New York, NY: Basic Books.

Porter, M.E. (1980). Competitive strategy: Techniques for analyzing industries and competitors. New York, NY: Free Press.

Ployhart, R.E., Wiechmann, D., Schmitt, N., Sacco, J.M., and Rogg, K. (2003). The cross-cultural equivalence of job performance ratings. Human Performance, 16, 1, 49-79.

Price, J., and Mueller, C. (1986). Absenteeism and turnover among hospital employees. Greenwich, CT: JAI Press.

Przeworski, A., and Teune, H. (1970). The logic of comparative social inquiry. New York, NY: J. Wiley.

Pulakos, E.D. (1997). Ratings of job performance. In D.L. Whetzel, and G.R. Wheaton (Eds.), Applied measurement methods in industrial psychology (pp. 291-317). Palo Alto, CA: Davies-Black Publishing.

R

Radloff, L.S. (1977). The CES-D scale: A self-report depression scale for research in the general population. Applied Psychological Measurement, 1, 385-401.

Raju, N.S., Laffitte, L.J., and Byrne, B.M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. Journal of Applied Psychology, 87, 517-529.

Raju, N.S., Van der Linden, W., and Fleer, P. (1995). An IRT-based internal measure of test bias with applications for differential item functioning. Applied Psychological Measurement, 19, 353-368.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: Danish Institute for Educational Research.

Reise, S., Widaman, K.F., and Pugh, R.H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. Psychological Bulletin, 114, 3, 552-566.

Retzer, J., and Fusso, T. (1999). Prediction in structural equation models involving latent constructs: LISREL versus partial least squares approaches. Paper presented at the Advanced Research Techniques Forum of the American Marketing Association (AMA), Santa Fé, New Mexico (June 1999).

Reynolds, F.D., and Darden, W.R. (1971). Mutually adaptive effects of interpersonal communication. Journal of Marketing Research, 8, 449-454.

Richins, M.L., and Dawson, S. (1992). A consumers values orientation for materialism and its measurement: Scale development and validation. Journal of Consumer Research, 19, 303-316.

Riordan, C.R., and Vandenberg, R.J. (1994). A central question in cross-cultural research: Do employees of different cultures interpret work-related measures in an equivalent manner? Journal of Management, 20, 643-671.

Rivera, P., and Satorra, A. (2002). Analysing group differences: A comparison of SEM approaches. In G.A. Marcoulides, and I. Moustaki (Eds.). Latent variable and latent structure models. (pp. 85-104). New Jersey: Lawrence Erlbaum Associates.

Robert, C., Probst,T.M., Martocchio,J.J, Drasgow, F., and Lawler, J.J. (2000). Empowerment and continuous improvement in the United States, Mexico, Poland, and India: Predicting fit on the basis of the dimesions of power distance and individualism. Journal of Applied Psychology, 85, 5, 643-658.

Rock, D.A., Werts, C.E., and Flaugher, R.L. (1978). The use of analysis of covariance structures for comparing the psychometric properties of multiple variables across populations. Multivariate Behavioral Research, 13, 403-418.

Rodgers, W. (1999). The influence of conflicting information on novice and loan officers' actions. Journal of Economic Psychology, 20, 2, 123-145.

Rook, D., and Fisher, R.J. (1995). Normative influences on impulsive buying behavior. Journal of Consumer Research, 22, 305-313.

Rubio, D.M., Berg-Weger, M., Tebb, S.S., Rauch, S.M. (2003). Validating a measure across groups: The use of MIMIC models in scale development. Journal of Social Service Research, 29, 3, 53-67.

Ryan, A.M., Chan, D., Ployhart, R.E., and Slade, A.L. (1999). Employee attitude surveys in a multinational organization: Considering language and culture in assessing measurement equivalence. Personnel Psychology, 52, 37-58.

Ryan, A.M., Horvath, M., Ployhart, R.E., Schmitt, N., and Slade, A.L. (2000). Hypothesizing differential item functioning in global employee opinion surveys. Personnel Psychology, 53, 531-562.

S

Salzberger, T., Sinkovics, R.R., and Schlegelmilch, B.B. (1999). Data equivalence in cross-cultural research: A comparison of Classical Test Theory and Latent Trait Theory based approaches. Australasian Marketing Journal, 7, 2, 23-38.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph no. 17.

Sandvik, I.L., and Sandvik, K. (2003). The impact of market orientation on product innovativeness and business performance. International Journal of Research in Marketing, 20, 4, 355-376.

Satorra, A. (1992). Asymptotic robust inferences in the analysis of mean and covariance structures. In P.V. Marsden (Ed.). Sociological Methodology 1992 (vol. 22) (pp. 249-278). Oxford, England: Blackwell Publishers.

Satorra, A., and Bentler, P. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye, and C.C. Clogg (Eds.), Latent variable analysis: applications to developmental research (pp. 399-419). Newbury Park, CA: Sage.

Satorra, A., and Bentler, P. (1999). A scaled difference chi-square test statistic for moment structure analysis. University of California, Los Angeles (UCLA) (UCLA Statistics Series no. 260).

Satorra, A. (2000). Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In R.D.H. Heijmans, D.S.G. Pollock, and A. Satorra (Eds.). Innovations in multivariate statistical analysis. A festschrift for Heinz Neudecker (pp. 233-247). London: Kluwer Academic Publishers.

Scarpello, V., and Campbell, J.P. (1983). Job satisfaction: Are all the parts there? Personnel Psychology, 36, 577-600.

Schaffer, B.S., and Riordan, C.M. (2003). A review of cross-cultural methodologies for organizational research: A best-practices approach. Organizational Research Methods, 6, 2, 169-215.

Schaie, K.W., and Hertzog, C. (1985). Measurement in the psychology of adulthood and aging. In J.E. Birren and K.W. Schaie (Eds.), Handbook of the psychology of aging (2nd ed., pp. 61-92). New York, NY: Van Nostrand Reinhold.

Schaubroeck, J., and Green, S. G. (1989). Confirmatory factor analytic procedures for assessing change during organizational entry. Journal of Applied Psychology, 74, 892-900.

Schmit, M.J., Kihm, J.A., and Robie, C. (2000). Development of a global measure of personality. Personnel Psychology, 53, 1, 153-193.

Schmitt, N. (1982). The use of analysis of covariance structures to assess beta and gamma change. Multivariate Behavioral Research, 17, 343-358.

Scholderer, J. (2004). Cross-cultural validity of the food-related lifestyles instrument (FRL) within Western Europe. Appetite, 42, 197-213.

Schuler, R.S., and Jackson, S.E. (1996). Human resource management: Positioning for the 21st century (6th edition) USA: West Publishing Company.

Schuler, R.S., and Rogovsky, N. (1998). Understanding compensation practice variations across firms. The impact of national culture. Journal of International Business Studies, 29, 159-159-177.

Schwab, D.P. (1980). Construct validity in organizational behavior. In L.L. Cummings, and B.M. Staw (Eds.). Research in organizational behavior (vol. 2, pp. 3-43). Greenwich , CT: JAI Press.

Schwarz, 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., and Sagiv, L. (1995). Identifying culture-specifics in the content and structure of values. Journal of Cross-Cultural Psychology, 26, 92-116.

Sharma, S., and Wheathers, D. (2003). Assessing generalizability of scales used in cross-national research. International Journal of Research in Marketing, 20, 287-295.

Shavelson, R.J., Hubner, J.J., and Stanton, G.C. (1976). Validation of construct interpretations. Review of Educational Research, 46, 407-441.

Shenkar, O. (1995). Global perspectives on human resource management. Upper Saddle River, NJ: Prentice-Hall.

Shimp, T.A., and Sharma, S. (1987). Consumer ethnocentrism: Construction and validation of the CETSCALE. Journal of Marketing Research, 24, 280-289.

Singh, J. (1995). Measurement issues in cross-national research. Journal of International Business Studies, 26, 597-619.

Singh, J. (2004). Tackling measurement problems with item response theory: Principles, characteristics, and assessment, with an illustrative example. Journal of Business Research, 57, 184-208.

Sirohi, N., McLaughlin, E.W., and Wittink, D.R. (1998). A model of consumer perceptions and store loyalty intentions for a supermarket retailer. Journal of Retailing, 74, 2, 223-245.

Slocum, J.W. Jr, and Topichak, P.M. (1972). Do cultural differences affect job satisfaction? Journal of Applied Psychology, 56, 177-178.

Smith, A.M., and Reynolds, N. (2001). Measuring cross-cultural service quality. International Marketing Review, 19, 5, 450-481.

Smith, P.B. (2004). Acquiescent response bias as an aspect of cultural communication style. Journal of Cross-Cultural Psychology, 35, 1, 50-61.

Smith, P.B., and Mitsumi, J. (1989). Japanese management – A sun rising in the West? In Cooper, C.L., and I.T. Robertson (Eds.), International review of industrial and organizational psychology (pp. 330-369). Chicester, UK: J. Wiley.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. British Journal of Mathematical and Statistical Psychology, 27, 229-239.

Sörbom, D. (1978). An alternative to the methodology for analysis of covariance. Psychometrika, 43, 3, 1978.

Sörbom, D. (1981). Structural equation models with structured means. In K.G. Jöreskog, and H. Wold (Eds.), Systems under indirect observation: Causality structure, and prediction (pp. 183-195). Amsterdam, The Netherlands: North Holland.

Sparrow, P.R., Schuler, R.S., and Jackson S.C. (1994). Convergence or divergence: human resource practices and policies for competitive advantages worldwide. The International Journal of Human Resource Management, 5, 2, 267-299.

Spector, P.E., and Wimalasiri, J. (1986). A cross-cultural comparison of job satisfaction dimensions in the United States and Singapore. International Review of Applied Psychology, 35, 147-158.

Spirtes, P., Scheines, R., Meek, C., and Glymour, C. (1994). Tetrad II : Tools for causal modeling. [User's manual]. New Jersey: Erlbaum.

Steel, R., and Ovalle, N. (1984). A review and meta-analysis of research on the relationship between behavioral intentions and employee turnover. Journal of Applied Psychology, 69, 673-686.

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

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

Steenkamp, J-B.E.M., and ter Hofstede, F. (2002). International market segmentation: Issues and perspectives. International Journal of Ressearch in Marketing, 19, 185-213.

Steenkamp, J.-B.E.M., and van Trijp, H.C.M. (1996). Quality guidance: A consumer-based approach to food quality improvement using partial least squares. European Review of Agricultural Economics, 23, 195-215.

Steiger, J.H. (1990). Structural model evaluation and modification: An interval estimation approach. Multivariate Behavioral Research, 25, 173-180.

Steinberg, L., and Thissen, D. (1995). Item response theory in personality research. In P.E. Stout and S.T. Fiske (Eds.), Personality research, methods, and theory: A Festschrift honoring Donald W. Fiske (pp. 161-181). Hillsdale, NJ: Erlbaum.

Stelzl, I., and Schnabel, K. (1992). The two-group MANOVA problem with unequal covariance matrices: A simulation study comparing Hotelling's T^2 to the LISREL approach. Methodika, 6, 54-75.

Suppes, P. (1970). A probabilistic theory of causation. Amsterdam, The Netherlands: North-Holland.

T

Takane, Y., and de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. Psychometrika, 52, 393-408.

Taris, T.W., Bok, I.A., and Meijer, Z.Y. (1998). Assessing stability and change of psychometric properties of multi-item concepts across different situations: A general approach. Journal of Psychology, 132, 301-316.

Taylor, S., Beechler, S., and Napier, N. (1996). Toward an integrative model of strategic human resource management. Academy of Management Review, 21, 4, 959-985.

Temple, A. (1997). Watch your tongue: Issues in translation and cross-cultural research. Sociology, 31, 3, 607-618.

Ter Hofstede, F, Steenkamp, J.-B.E.M., and Wedel, M. (1999). International market segmentation based on consumer-product relations. Journal of Marketing Research, 36, 1-17.

Tett, R.P., and Meyer, J.P. (1993). Job satisfaction, organizational commitment, turnover intention, and turnover: Path analyses based on meta-analytic findings. Personnel Psychology, 46, 259-293.

Thissen, D. (1991). MULTILOG: Multiple categorical item analysis and test scoring using item response category (Version 6). Chicago, IL: Scientific Software International Inc.

Thissen, D., and Steinberg, L. (1988). Data analysis using iteme response theory. Psychological Bulletin, 104, 385-395.

Thissen, D., Steinberg, L., and Gerard, M. (1986). Beyond group-mean differences: The concept of item bias. Psychological Bulletin, 99, 118-128.

Thissen, D., Steinberg, L., and Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer, and H.I. Braun (Eds.), Test validity (pp. 147-169). Hillsdale, NJ: Erlbaum.

Thompson, B. (1984). Canonical correlation analysis: Uses and interpretation. Newbury Park, CA: Sage.

Tomás, I.M., González-Romá, V., and Benito, J.G. (2000). Teoría de respuesta al ítem análisis factorial confirmatorio: Dos métodos para analizar la equivalencia psicométrica en la traducción de cuestionarios. Psicothema, 12, 2, 540-544.

Torrington, D. (1994) (Ed.). International human resource management. Think globally, act locally. Hempel Hempstead: Prentice-Hall.

Traub, R.E. (1994). Reliability for the social sciences: Theory and applications (vol. 3). Thousand Oaks, CA: Sage.

Triandis, H.C. (1972). The Analysis of Subjective Culture. New York, NY: J. Wiley.

Triandis, H.C., Berry, J.W., Bristin, R.W., Draguns, J.G., Heron, A., Lambert, W.W., and Lonner, W. (Eds.) (1980, 1981, 1985). Handbook of Cross-Cultural Psychology (Six volumes). Boston, MA: Allyn and Bacon.

Triandis, H.C., and Marin, G. (1983). Etic plus emic versus pseudoetic: A test of the basic assumption of contemporary cross-cultural psychology. Journal of Cross-Cultural Psychology, 14, 489-500.

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

Tung, R., and Punnett, B.J. (1993). Research in international human resource management. In D. Wong-Rieger, and F. Rieger (Eds.), International management research: Looking to the future (pp. 35-53). Berlin, Germany: Walter de Gruyter.

U

Ueltschy, L.C., Laroche, M., Tamilia, R.D., Yannopoulos, P. (2004). Cross-cultural invariance of measures of satisfaction and service quality. Journal of Business Research, 57, 901-912.

Unilever (2000). Brand alignment through consumer segmentation. (Executive summary report published by the Unilever's Spreads and Cooking Products' Category).

Usunier, J.-C. (1996). Marketing across cultures. Hemel Hempstead: Prentice-Hall.

Usunier, J.-C. (1998). International and Cross-Cultural Management Research. London, United Kingdom: Sage.

V

Vandenberg, R.J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. Organizational Research Methods, 5, 2, 139-158.

Vandenberg, R.J., and Lance, C.E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organisational research. Organizational Research Methods, 3, 1, 4-70.

Vandenberg, R.J., and Self, R.M. (1993). Assessing newcomers' changing commitment to the organization during the first 6 months of work. Journal of Applied Psychology, 78, 557-568.

Vandenberghe, C., Stinglhammer, F., Bentein, K., and Delhaise, T. (2001). An examination of the cross-cultural validity of the multidimensional model of commitment in Europe. Journal of Cross-Cultural Psychology, 32, 3, 322-347.

Van de Pol, F., Langeheine, R., and De Jong, W. (1996). PANMARK 3. User's manual. Panel analysis using Markov chains – A latent class analysis program. Voorburg, The Netherlands.

Van de Vijver, F.J.R., and Leung, K. (1997). Methods and data analysis of comparative research. In J.W. Berry, Y.H. Poortinga, and J. Pandey (Eds.), Handbook of cross-cultural psychology (pp. 257-300). Chicago, IL: Allyn and Bacon.

Van de Vijver, F.J.R., and Poortinga, Y.H. (1982). Cross-cultural generalization and universality. Journal of Cross-Cultural Psychology, 13, 387-408.

Van Herk, H. (2000). Zijn in marktonderzoek gevonden culturele verschillen echt? Het bepalen van equivalentie in international onderzoek middels detectie van bias. In A.E. Bronner et al. (Eds.), 2002 Ontwikkelingen in het marktonderzoek. Jaarboek 2002 MarktOnderzoekAssociatie. Haarlem, The Netherlands: de Vrieseborch publications.

Van Herk, H., Poortinga, Y.H., and Verhallen, T.M.M. (2004). Response styles in rating scales: Evidence of method bias in data from 6 EU countries. Journal of Cross-Cultural Psychology, 35, 3, 346-360.

Van Herk, H., Poortinga, Y.H., and Verhallen, T.M.M. (2005?). Equivalence of survey data: Relevance for international marketing. To appear in European Journal of Marketing.

Van Zessen, K., and De Beuckelaer, A. (2000). The use of item response models to investigate the cross-cultural applicability of attitudinal statements. Unilever Research and Development Research Report no. VD 00 0258. Vlaardingen, The Netherlands: Unilever Research and Development (R&D) Vlaardingen.

Voss, K.E., Stem Jr., D.E., Johnson, L.W., and Arce, C. (1996). An exploration of the comparability of semantic adjectives in three languages. A magnitude estimation approach. International Marketing Review, 13, 5, 44-58.

Vriens, M., van der Scheer, H., and Cary, M.S. (1999). Market segmentation: Selective review and application guidelines. Series in Marketing Research Techniques no. 2. Publication of the "Management Sciences Group" of Research International Ltd.

W

Wallentin, F. (2004). Confirmatory Factor Analysis with ordinal variables: A simulation study. Paper presented at the conference of the Society for Multivariate Analysis in the Behavioural Sciences (SMABS), Jena, Germany, July 2004.

Wang, C.L. (1996). The evolution of international consumer research: A historical assessment from the 1960s to mid- 1990s. Journal of Euromarketing, 5, 1, 57-81.

Wasti, S.A., Bergman, M.E., Glomb, T.M., and Drasgow, F. (2000). Test of the cross-cultural generalizability of a model of sexual harassment. Journal of Applied Psychology, 85, 5, 766-778.

Watkins, D. (1989). The role of confirmatory factor analysis in cross-cultural research. International Journal of Psychology, 24, 685-701.

Wedel, M., and Kamakura, W.A. (1998). Market segmentation: Conceptual and methodological foundations. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Weech-Maldonado, R., Weidmer, B.O., Morales, L.S., and Hays, R.D. (2001). Cross-Cultural Adaptation of Survey Instruments: The CAHPS Experience. In M.L. Cynamon, and R.A. Kulka (Eds.), Proceedings of the Seventh Conference on Health Survey Research Methods, Hyattsville, Maryland.

Welkenhuysen-Gybels, J., Billiet, J., and Cambré, B. (2003). Adjustment for acquiescence in the assessment of the construct equivalence of Likert-type scores. Journal of Cross-Cultural Psychology, 34, 6, 702-722.

West, S.G., Finch, J.F., and Curran, P.J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In Hoyle, R.H. (Ed.), Structural equation modeling: Concepts, issues and applications (pp. 56-75). Newbury Park, CA: Sage.

Westphal, J.D., and Zajac, E.J. (2001). Decoupling policy from practice: The case of stock repurchase programs. Administrative Science Quarterly, 46, 202-228.

Williams, L.J., Edwards, J.R., and Vandenberg, R.J. (2003). Recent advances in causal modelling methods for organizational and management research. Journal of Management, 29, 6, 903-936.

Williams, L., and Hazer, J. (1986). Antecedents and consequences of satisfaction and commitment in turnover models: A re-analysis using latent variable structural equation models. Journal of Applied Psychology, 71, 219-231.

Williams, R., and Thomson, E. (1986). Normalisation issues in latent variable modelling. Sociological Methods and Research, 15, 1 & 2, 24-43.

Wold, H. (1982). Systems under direct observations using PLS. In C. Fornell (Ed.), A second generation of multivariate analysis (vol. 2). New York, NY: Praeger.

Wold, H. (1985). Systems analysis by partial least squares. In P. Nijkamp, H. Leitner, and N. Wrightley (Eds.), Measuring the unmeasurable. Dordrecht, The Netherlands: Martinus Nijhoff Publishers.

Wrigley, C.S., and Neuhaus, J.O. (1955). The matching of two sets of factors. American Psychologist, 10, 418-419.

Y

Yaprak, A. (2003). Measurement problems in cross-national consumer research: The state-of-the-art and future research directions. In S.C. Jain (Ed.), Handbook of research in international marketing. Edward Elgar Publishing.

Yip, G.S. (1995). Total global strategy. Englewood Cliffs, NJ: Prentice-Hall.

Yoo, B. (2002). Cross-group comparisons: A cautionairy note. Psychology and Marketing, 19,4, 357-368.

Yoo, B., and Donthu, N. (2001). Developing and validating a multidimensional consumer-based brand equity scale. Journal of Business Research, 52, 1-14.

Yoo, B., and Donthu, N. (2002). Testing cross-cultural invariance of the brand equity creation process. Journal of Product and Brand Management, 11, 6, 380-398.

Samenvatting (summary in Dutch)

Dit proefschrift behandelt een specifiek methodologisch probleem (meerbepaald een 'meetprobleem') waarmee kwantitatieve onderzoekers geconfonteerd worden wanneer ze internationaal vergelijkend onderzoek uitvoeren. Dit methodologische probleem wordt in de volgende paragraaf nader omschreven. Het internationaal onderzoek binnen de managementwetenschappen neemt (als toepassingsdomein) in dit proefschrift een centrale plaats in.

Het methodologische probleem dat in dit proefschrift onderzocht wordt, treedt op telkens wanneer een onderzoeker landen wil vergelijken op basis van hun (gemiddelde) score op zogenaamde 'constructen' (ook wel 'factoren' genoemd). Constructen zijn abstracte concepten die vaak 'geoperationaliseerd' (d.w.z. gemeten) worden middels een aantal vragen in een vragenlijst. Vaak nemen deze vragen de vorm aan van zogenaamde 'beweringen'. Deze beweringen worden vertaald en opgenomen in alle vragenlijsten die in de verschillende landen gebruikt worden. De respondenten worden geacht aan te geven in welke mate zij het eens dan wel oneens zijn met de individuele beweringen. Vaak wordt een 5-puntenschaal gebruikt om hun antwoorden te registreren. Wanneer een schaal gebruikt wordt die samengesteld is uit meerdere beweringen per (te meten) construct, moet voldaan zijn aan het principe van 'meetinvariantie' (of 'schaalinvariantie') over alle landen heen. Er is voldaan aan dit principe indien de (te schatten) 'parameters' in het meetmodel dezelfde waarde hebben in alle landen die in het onderzoek participeren. Indien niet aan deze voorwaarde voldaan is, treedt het 'methodologische probleem' op waar in de inleidende paragraaf naar gerefereerd werd.

Alhoewel er verschillende statistische methoden bestaan die toelaten om gemiddelde constructscores per land te schatten (zie hoofdstuk 1), wordt in dit proefschrift enkel gebruik gemaakt van zeer specifieke 'covariantie-structuurmodellen' (met name covariantie-structuurmodellen waaraan informatie betreffende gemiddelde-structuren is toegevoegd). Op dit moment worden, binnen de managementwetenschappen, deze specifieke covariantiestructuur-modellen veelvuldig gebruikt om -in een internationale context- gemiddelde constructscores te schatten. In hoofdstuk 3 van het proefschrift worden deze modellen in detail beschreven. In hoofdstuk 2 wordt ingegaan op een belangrijke voorwaarde voor de toepassing van covariantie-structuurmodellen. De scores op de gestelde 'beweringen' (die geacht worden een construct te meten) dienen immers, vanuit 'oorzakelijk perspectief', het 'gevolg' te zijn van de mate waarin de respondent 'scoort' op het te meten construct. Een

meetstructuur met een dergelijk oorzakelijk verband tussen de geobserveerde variabelen (de beweringen) en het construct wordt een 'latente (meet)structuur' genoemd (zie hoofdstuk 2). Soms komt het voor dat constructen geoperationaliseerd worden als een welbepaalde (gewogen) optelsom van geobserveerde variabelen. In dat geval gaat de causale relatie tussen de geobserveerde variabelen en het construct in de tegenovergestelde richting (d.w.z. van geobserveerde variabelen naar het construct). Een dergelijke meetstructuur wordt ook wel een 'emergente (meet)structuur' genoemd. Zoals eerder aangegeven, wordt in dit proefschrift uitsluitend met latente meetstructuren gewerkt. In hoofdstuk 2 wordt uitgebreid ingegaan op verschillende typen meetmodellen.

De belangrijkste onderzoeksvraag in dit proefschrift is hoe betrouwbaar vergelijkingen tussen groepen/landen zijn indien niet (of niet volledig) voldaan is aan het principe van meetinvariantie over de groepen/landen heen. In feite wordt dus onderzocht welke de minimale set van meetparameters (b.v. factorladingen / regressiecoefficienten, intercepten, en 'errorvarianties') zijn die gelijk moeten zijn over de groepen/landen heen, wil men, op basis van constructgemiddelden, zinvolle vergelijkingen kunnen maken tussen de groepen/landen. Tot op heden biedt de methodologische literatuur geen sluitend antwoord op deze vraag.

De hoger genoemde onderzoeksvraag wordt in hoofdstuk 4 beantwoord op basis van simulatieonderzoek. In het simulatieonderzoek worden verschillende datasets aangemaakt. De datasets verschillen van elkaar in de mate waarin de verschillende meetparameters voldoen aan de eis van meetinvariantie over de groepen/landen heen. Voor elke gesimuleerde dataset zijn de gemiddelde constructscores (op populatieniveau) per groep/land op voorhand vastgelegd. Bijgevolg is eveneens gekend of de hypothese die stelt dat de constructgemiddelden identiek zijn over de groepen/landen heen (d.w.z. de nulhypothese) al dan niet zou moeten verworpen worden. De resultaten van het onderzoek tonen aan dat vergelijkingen over groepen/landen heen erg onbetrouwbaar worden indien de intercepten een verschillende waarde hebben in de verschillende groepen/landen. Deze resultaten impliceren dat de onderzoekers/methodologen die (in de literatuur) gepleit hebben voor 'zwakkere vormen' van meetinvariantie het zeker niet bij het rechte eind hadden. In hoofdstuk 3 en hoofdstuk 4 wordt dieper ingegaan op deze methodologische discussie. In het algemeen kan (op basis van deze simulatiestudie) gesteld worden dat de factorladingen en in het bijzonder de intercepten gelijk moeten zijn over de groepen/landen heen, wil men, op basis van geschatte constructgemiddelden, zinvolle vergelijkingen kunnen maken tussen groepen / landen.

Naast het bestuderen van deze methodologische onderzoeksvraag wordt in dit proefschrift eveneens aandacht besteed aan praktijk-toepassingen binnen de managementwetenschappen. Omdat internationaal onderzoek binnen de

managementwetenschappen een zeer uitgebreid gebied is (zie hoofdstuk 1), wordt slechts dieper ingegaan op twee deelgebieden die, met name voor multinationale ondernemingen, erg belangrijk zijn. Het betreft enerzijds internationaal onderzoek betreffende Human Resource management en anderzijds internationaal consumentenonderzoek. In hoofdstuk 5 wordt nagegaan of schalen die gebruikt werden in maar liefst 16 (!) landen voldoen aan het principe van meetinvariantie over de landen heen. Deze schalen werden gebruikt om constructen te meten die uiterst relevant zijn voor het evalueren van het HR beleid van een multinationale onderneming. Alhoewel de schalen niet volledig voldoen aan het hoger gestelde criterium van meetinvariantie (d.w.z. gelijke factor ladingen en gelijke intercepten over de landen heen), blijkt (voor deze dataset) dat de landen-vergelijkingen op basis van constructgemiddelden toch betrouwbaar zijn (zelfs indien men ten onrechte zou veronderstellen dat volledig voldaan is aan het meetinvariantie principe [over de landen heen]).

In hoofdstuk 6 worden nog twee bijkomende gevalstudies besproken. Beide gevalstudies zijn gebaseerd op data afkomstig van internationaal consumentenonderzoek. Een gerenomeerd internationaal marktonderzoeks-bureau voerde deze onderzoeken uit. Telkens betreft het onderzoek in meerdere landen (respectievelijk in 16 en 7 landen) waarin globale consumenten-segmenten werden geïdentificeerd, alsook gekeken werd naar de positionering van verschillende merken en producten. Na onderzoek van de gebruikte schalen (die samengesteld waren uit twee of meer beweringen) bleek overduidelijk dat geen van alle schalen ook maar enigzins voldoet aan het criterium van meetinvariantie over de landen heen. Het gevolg hiervan is dat zinvolle vergelijkingen tussen landen niet kunnen gemaakt worden, tenminste niet op basis van de gemiddelde constructscore per land.

Ondanks een aantal praktische problemen op het gebied van de (commercële) onderzoekspraktijk (o.a. budget- en tijdsrestricties) blijkt uit de eerste gevalstudie dat het niet onmogelijk is om schalen te construeren die: (1) samengesteld zijn uit meerdere beweringen en (2) die ook (bijna) voldoen aan de 'strenge meetinvariantie-eis'. Wellicht is het dan wel nodig dat de schalen herhaaldelijk gebruikt worden in meerdere landen zodat inzicht verkregen wordt in de onderlinge relaties tussen de verkregen antwoorden op de verschillende beweringen. Dit inzicht is cruciaal voor de kwaliteitsverbetering van de gebruikte meetinstrumenten. Indien noodzakelijk zouden de schalen aangepast kunnen worden. Beweringen waarvan de relatie tot de andere beweringen afhankelijk is van het land, kunnen zodoende geschrapt worden omdat ze vanuit internationaal perspectief niet 'neutraal' zijn. Van sommige vragen die gebruikt werden in de eerste gevalstudie (internationaal onderzoek in HR management) is geweten dat ze reeds eerder gebruikt werden. Dit zou de hoge kwaliteit van deze schalen kunnen verklaren.

Op basis van dit proefschrift kan gesteld worden dat de eis wat betreft de gelijkheid van intercepten en factorladingen een belangrijke voorwaarde is voor geldig internationaal vergelijkend onderzoek (waarin landen vergeleken worden op basis van hun gemiddelde constructscores). Het testen van deze eis is in dit type onderzoek dan ook absoluut noodzakelijk. De testen die hiervoor geschikt zijn, staan uitvoerig beschreven in hoofdstuk 3. Verder toont dit onderzoek aan dat, mits de nodige inspanningen, het mogelijk is om schalen te construeren die, vanuit internationaal vergelijkend perspectief, een hoge 'psychometrische kwaliteit' waarborgen.