Data Equivalence in Cross-Cultural Research: A Comparison of Classical Test Theory and Latent Trait Theory Based Approaches

Thomas Salzberger
Department of Marketing
Wirtschaftsuniversität Wien,

Rudolf R. Sinkovics
Department of International Marketing and Management
Wirtschaftsuniversität Wien,

Bodo B. Schlegelmilch
Chair of International Marketing and Management
Wirtschaftsuniversität Wien,
Vienna,
Austria

ABSTRACT
Given the increasing importance of international business, cross-cultural research becomes more and more relevant to marketing academics and practitioners. This paper illustrates the difficulties in achieving equivalence when conducting marketing research across borders. It opens with a general typology of equivalence issues in cross-cultural research and, subsequently, focuses specifically on data equivalence. Recent studies either disregard data equivalence at all or they predominantly suggest the use of simultaneous confirmatory factor analysis (CFA) for establishing data equivalence. The Latent Trait Theory (LTT), based on a different measurement paradigm, offers an alternative which promises to overcome many of the problems inherent in CFA. This paper contrasts the advantages and disadvantages of both approaches and illustrates their application by means of a simulated data set. In conclusion, a call is made for the incorporation of equivalence issues in scale development by quantitative as well as qualitative analyses.
INTRODUCTION

While marketing researchers appear to agree on the importance of replications as genuine source of scientific knowledge (Collins 1985; Rosenthal and Rosnow, 1984), there is a dearth of replication studies in the marketing literature (Hubbard and Armstrong, 1994). However, international marketing seems to be more open to replications. Indeed, the international marketing literature contains a large variety of comparisons between different countries and cultures. Typically, researchers adopt a measurement instrument developed within the frame of reference of one specific country. Mostly the instrument is based on US research and subject to its cultural idiosyncrasies. Scales developed in the US are designed to fit the US culture and therefore are commonly ethnocentric (Chikudate, 1997). Subsequently, scales are often translated and applied in another country with a different cultural system. Unfortunately, many replications largely disregard at least some of the traps of cross-cultural research and fail to apply appropriate analyses ensuring the comparability of results.

When crossing cultural borders, the very meaning of a scale may change and classical quality indicators such as reliability (Davis, Douglas and Silk, 1981) and validity, may be strongly influenced by cultural factors. Consequently, the state-of-the-art in international marketing research calls for appropriate equivalence and comparability checks. Dealing with equivalence issues insufficiently bears the risk that measurement inequivalence may severely diminish the usefulness of findings and, in the worst case scenario, render results totally meaningless (Douglas and Craig, 1992; Manrai and Mainrai, 1996; Sinkovics, Salzberger and Holzmüller, 1998).

Given the crucial importance of measurement equivalence, this paper initially attempts to shed light on the potential pitfalls involved in conducting comparative research across countries. Subsequently, it outlines statistical methods and techniques suitable to test for measurement equivalence.

DIMENSIONS OF EQUIVALENCE IN CROSS-CULTURAL RESEARCH

Comparative research has a long tradition in sociological, anthropological and social sciences. These contributions have been incorporated in many marketing textbooks with widely varying degrees of comprehensiveness (e.g., Burns and Bush, 1995; Douglas and Craig 1983; Holzmüller, 1995; Jain, 1993; Toyne and Walters, 1989; Usunier, 1996). In the context of international marketing research, Bauer (1989) offers a particularly useful classification of equivalence. Figure 1 depicts this scheme extended by the issue of measurement equivalence.

The diverse facets of equivalence are closely linked and, in most cases, there will be no single source unequivocally responsible for a possible lack of data comparability. At the problem definition stage, the equivalence of research topics represents the minimum requirement for eventually meaningful quantitative comparisons across cultural borders. If the construct under investigation has a different function in the cultures under scrutiny and/or is conceptually inequivalent, research comparing intensities on this construct ceases to be meaningful. However, qualitative investigations might still provide interesting insights into cultural divergence. Once the problem has been successfully defined cross-culturally, further equivalence aspects have to be considered during data collection. For instance, equivalence of research methods requires the selection of culturally appropriate data collection techniques (Casley and Lury, 1981; Kracmar, 1971) as well as adequate steps to minimise culture-specific response distortions (culture-specific response behaviour).

The equivalence of research units affects the definition of the total populations in the cultures investigated and the way the samples are drawn. The fundamental problem involved is the trade-off between intra-culturally representative samples on the one hand and cross-culturally comparable samples on the other. Finally, the equivalence of research administration has to ensure that there are no specific variables (e.g., time-related factors like climate, political events like elections, and so forth) in either culture that might impair the quality of the data collected. Following data collection, during the data preparation stage, care has to be taken that the data are equally handled (e.g., equivalent response translations). However, if the questionnaire does not contain any open-ended questions, the design of equivalent response categories is mostly an issue of equivalence of research methods.
The ultimate goal of quantitative cross-cultural research lies in the equivalence of data. It is always an empirical issue whether data equivalence is achieved or not. To this end, statistical procedures suitable to test for measurement equivalence have to be carried out. If data equivalence does not hold, then the research should be confined to qualitative interpretation rather than conducting meaningless quantitative comparisons across cultures. Data equivalence may be seen as a function of all equivalence aspects. In fact, from problem definition to data preparation, all the researcher can do and should do, is taking care of all facets of equivalence which can only be assessed qualitatively. However, these efforts do not dispense the researcher from investigating data equivalence by quantitative means since data still may suffer from, e.g., culture specific response bias – despite all efforts undertaken. Thus, given the central importance of measurement equivalence, the remainder of the paper will focus on this issue.

**MEASUREMENT EQUIVALENCE IN CROSS-CULTURAL RESEARCH**

The cross-cultural measurement equivalence of a scale can only be examined by comparing data sets collected from different cultures. The possible outcome of the analysis, namely that the data does not support comparability, might cause a reluctance among researchers to engage in rigorous testing of data equivalence. This is unfortunate, as non-equivalence should in itself be a highly valued research result with far-reaching consequences for subsequent studies.

Cross-cultural marketing research may utilise two fundamentally different approaches for the investigation of measurement equivalence. The first approach is based on the multiple group confirmatory factor analysis (CFA) (Horn and McArdle, 1992; Jöreskog, 1971; Sörbom, 1974). It has gained much popularity since it has been introduced to marketing research (Mullen, 1995; Singh, 1995; Steenkamp and Baumgartner, 1998). However,
application studies still disregard these basic principles more often than not. The second approach advocated in this paper rests on the paradigm of Latent Trait Theory (LTT, also referred to as Item-Response Theory, IRT) in general (Hambleton, Swaminathan and Rogers, 1991; Lord, 1974) and on Rasch modeling in particular (Rasch, 1960). Latent Trait Theory comprises a comprehensive family of models. Some of them have been introduced to marketing research (Balasubramian and Kamakura, 1989; Singh, 1995; Singh, Howell and Rhoads, 1990), however, to our best knowledge, no application to international marketing research as regards measurement equivalence has been published so far. Subsequently, both approaches are discussed and illustrated by means of a simulated data set.

**Multiple Group Structural Equation Modeling Approach (CFA Approach)**

When introducing the CFA approach to marketing research, Singh (1995), Mullen (1995) and Steenkamp and Baumgartner (1998) are using, at least in part, different terminology. We decided to adopt exclusively the terminology used by Steenkamp and Baumgartner (1998) when summarising the approach.

The basic idea of this type of analysis is a structural identity of the construct intended to be measured across cultures, which is necessary and sufficient for conducting mean comparisons. Consequently, the measurement model, including the relations between latent variables and observed variables, has to be defined first as a confirmatory factor analysis model - driven by the substantive theory underlying the research project. This step should be undertaken in any quantitative empirical study anyway and represents no additional effort. The measurement model has to fit the data within any of the cultures under scrutiny. Unfortunately, many cross-cultural research studies still stop at this stage. However, to examine cross-cultural equivalence, these separate models have to be analysed simultaneously in order to test for cross-cultural invariance of the measurement model.

In this context, three increasingly stringent levels of cross-cultural invariance may be distinguished, each allowing different types of comparisons: configural invariance, metric invariance, and scalar invariance (Horn and McArdle, 1992; Steenkamp and Baumgartner, 1998). The basic form of invariance, configural invariance, pertains to separate confirmatory factor analyses of the same measurement model based on covariance matrices. Commonly, one assumes that each manifest variable relates to just one latent variable (factor), in line with Thurstone's concept of simple structure (Thurstone, 1940). Technically, this means that the loadings are estimated, when items are to constitute indicators of a factor (salient loadings). Otherwise, the loadings are fixed to zero (non-salient loadings) although any other arbitrary value might be chosen instead. In CFA, factors may be correlated (being an issue of discriminant validity) and thereby account for inter-item correlations across factors.

Configural invariance requires this structure to be the same across cultures, i.e. the non-salient loadings correspond (being fixed to zero in all cultures) while the salient loadings may vary in terms of their magnitude. Thus, configural invariance holds, if the model fits (evaluated by fit statistics as the chi-square, which is simply the sum of all intra-cultural chi-squares, and practical fit indices like the RMSEA, the TLI and many others) (see Hair, Anderson, Tatham and Black, 1995).

If configural invariance is not supported (indicated by a poor model fit), the scale has not been strictly replicated rather led the replication to culture specific constructs. It should be noted once again that from a scientific point of view this might nevertheless be a valuable outcome. An acceptance of non-replicability should certainly be preferred over dubious and unjustified mean comparisons.

Data showing configural invariance can further be tested for metric invariance. On this level, the salient loadings are constrained to be equal across cultures. Since the model of metric invariance is nested into the model of configural invariance (the two models differ only by the number of equality constraints), the fit of the metric invariance model may be determined relative to the fit of the configural invariance model using the chi-square difference test. Thus, configural invariance is the baseline model against which other more stringent models are tested.

Given metric equivalence, the same scale units of the latent variable(s) have been established and differences within separate cultures may be compared, e.g., differences between various age groups within each
culture. However, even full metric invariance is not the ultimate prerequisite for mean comparisons across cultures for additive biases may still be present. The reason is that the magnitude of the factor loadings does not show if some items in one sample are affected by a culture-specific bias, such as a higher tendency of affirmative answers (acquiescence-bias). The comparison of means makes sense only when scalar invariance holds. This form of invariance requires metric invariance and, in addition, equal origins of measurement. The origins of measurement are equal if, and only if, a score of, e.g., 2 in culture A means the same of the property being measured as a 2 in culture B. What looks to be self-evident, may be untrue if a bias affects observed scores in that, e.g., a 2 in culture A parallels a 3 in culture B. Figure 2 illustrates a situation in which the same observed score in two cultures \((X_A, X_B)\) indicates different levels of the latent score in culture A \((T_A)\) and culture B \((T_B)\), respectively. This invariance is due to noninvariant item intercepts \((I_A, I_B)\).

In order to test for scalar invariance, the factor analysis model has to be extended by the inclusion of item intercepts. The item intercept is simply the constant of the regression of the observed item score on the score of the latent variable. Since the origin of the scale of the latent variable is totally arbitrary, it has to be fixed. This is achieved by setting the intercept of one item to zero (or any other convenient value). The intercepts of all other items are then estimated relative to this value. In substance, the item intercept might be regarded as a “difficulty” parameter of the item. The higher the intercept value, the higher is the item score of a person given a fixed score on the latent variable of that person. This interpretation explains why item intercepts are at least as important as loadings and why they must not be ignored in cross-cultural research. It clearly makes a difference when an item’s difficulty depends on cultural aspects. If that is the case, the fit of the scalar invariance model decreases due to the additional equality constraints concerning item intercepts.

In practice, the large variety of cultural facets and habits affecting response patterns makes it rather unlikely that full metric or scalar invariance, i.e. equivalence of all items in a scale, can be found in many empirical settings. The concept of partial invariance accounts for some items lacking invariance (Steenkamp and Baumgartner, 1998) as long as the scale of the latent variable can be defined equally across cultures. Strictly speaking, only two items suffice to identify a latent dimension across cultures although from a construct validity perspective more invariant indicators are certainly preferred. Specifically, the loading and intercept of one item is fixed across cultures (in most cases to one and zero, respectively), for defining the units and the origin of the latent variable. Consequently, the invariance of this item may not be tested but is rather assumed. Therefore, at least one other item has to be identified as a cross-culturally invariant indicator of that dimension. Items that are not invariant are linked to this scale by item specific estimates of the loading and the intercept. For practical purposes, the modification indices reported by structural equation modeling software (e.g., LISREL, EQS, AMOS) help to identify non-equivalent items (for a detailed guide see Steenkamp and Baumgartner 1998). It should be noted that the raw scores of partially invariant constructs must not be compared across cultures (because they do not account for non-invariance) but latent means should be referred to. This may be done by setting an additional equality constraint on the means of the factor scores.

For empirical marketing research, a key question is which level of invariance has to be met for what type of comparison to be meaningful. The major concern in replicating a scale should initially lie on conceptual aspects, i.e. the equivalence of research topics. It follows that the main emphasis should be placed on configural invariance. Where configural invariance can be achieved, but metric and/or scalar invariance cannot be achieved, a researcher has an indicator of the potential transferability of the concept. Still, assuming that higher levels of invariance could not be met, the operationalisation of the scale items would need to be improved substantially.
Latent Trait Theory Based Approach  
(LTT Approach)

The Latent Trait Theory (LTT) offers a promising alternative to Classical Test Theory (CTT), the underlying paradigm of almost all measurement models employed in marketing research, including the CFA approach. While CTT models the item score as a linear regression on the latent score, LTT refers to observable responses to items as the outcome of a comparison of a parameter describing the person's location on a latent dimension and, at least, one item parameter characterising the "difficulty" of the item. In the context of attitudinal measurement, difficulty may not be taken literally. The item parameter rather refers to the amount of the property the item stands for, i.e., the location of the item. This means, that a more difficult item requires a more extreme attitude to agree with the item while for a less difficult item a less extreme attitude already leads to the expression of agreement. The peculiarity of LTT models lies in the fact that both the person parameter and the item location parameter are scaled onto the same dimension. Consequently, persons and items are directly comparable.

LTT includes a comprehensive family of models which differ in various ways. First, the question arises, how many person parameters are needed. Certainly, a multidimensional construct, grounded in substantive theory, asks for more than one person parameter while most of the LTT models, like the Rasch model (Rasch 1960) or the Birnbaum model (Birnbaum, 1968), proceed on unidimensionality. However, we rarely aim at covering several dimensions within one item. Even for the measurement of complex phenomena, the research goal would be to measure all dimensions separately, to aim at high discriminant validity and to provide several distinct unidimensional measurement instruments or constructs. Thus, we may set up separate unidimensional scales, analyse them using LTT models and correlate person parameters.

Second, we have to choose how many item parameters we want to model. In any case, an item location parameter is needed. Superficially, it seems reasonable to define additional parameters, such as discrimination parameters or guessing parameters (as is the case in the Birnbaum model). While these models aim at accounting for data properties (e.g., varying item discrimination), the Rasch modeling approach goes the other way: from principles of how measurement might be accomplished, a measurement model, the Rasch model, has been deduced which specifies how data have to look like if they were to constitute measurement. This different view explains why the Rasch model and the Birnbaum model differ in more than just one or two additional parameters. They differ in the fundamental definition of measurement and disciplines of either approach almost never come to the same conclusions.

The basic principle that guided Rasch (1960) is what he called specific objectivity. Specific objectivity states that person parameters have to be independent of specific items and, likewise, item parameters have to be independent of specific persons. Moreover, the measure for one person must not depend on what other persons are assessed. Rasch showed that the so called Rasch model is the only model that complies with this principle. The Rasch model is limited to only one item parameter, the item location parameter. Item discrimination parameters would destroy specific objectivity and render parameters sample dependent. Moreover varying item discrimination threatens construct validity and may account for hidden multidimensionality (see Wright and Stone 1979). For these reasons, the Rasch model (Rasch, 1960), also referred to as the one parameter logistic model, will be discussed in more detail.

The dichotomous Rasch model

The basic Rasch model (Rasch, 1960) refers to dichotomously scored items measuring the same dimension, i.e. it hypothesises unidimensionality (Molenaar, 1995). The relationship of the latent dimension (covering person as well as item location parameters) and the observed item response is represented by a logistic function based on the comparison of the person and the item parameter.

Equation 1. The Rasch Model (Rasch 1960)

\[ P_i(+ | \beta_v, \delta_i) = \frac{e^{(\beta_v - \delta_i)}}{1 + e^{(\beta_v - \delta_i)}} \quad i = 1, 2, \ldots, n \]

where

\[ P_i(+ | \beta_v, \delta_i) \] is the probability that a randomly chosen examinee \( v \) with attitude \( \beta_v \) agrees to the item \( i \) given \( \delta_i \),

\[ \delta_i \] is the item \( i \) location parameter

\[ \beta_v \] is the person \( v \) parameter
n is the number of items in the test
e is a transcendental number with the value 2.718 (correct to three decimals)

The equation of the Rasch model describes a curve, the so-called item characteristic curve (ICC) or item response function (IRF). This curve defines the probability of agreement in accordance with the location of the person underlying latent dimension (β), e.g., Ethnocentrism, and in accordance with the item location (δ), i.e. the degree of Ethnocentrism that is expressed by the item's content. Thus, each dichotomous item is represented by one ICC.

The definition of this model, i.e. equal item discrimination, implies that all these ICCs are parallel. Figure 3 illustrates the ideal shape of the function.

**Figure 3. Ideal Item characteristic curves (ICCs) for two items**

![Diagram showing the probability of agreement for two items](image)

Figure 3 demonstrates that for item 1, the probability of agreement rises strictly monotonously and asymptotically approaches one with increasing attitude. The ICC for an item with higher location has the same shape but is shifted to the right. For any given person location on the latent dimension, the probability of agreement to item 2 is lower than that for item 1. It should be noted that for intersecting ICCs, like in the Birnbaum model, this principle does not hold.

As mentioned above, the estimation of item parameters does not depend on the sample composition. Thus, for any sample split, item parameters have to be the equal (except for random estimation error). The same applies to groups defined by their cultural affiliation. Whether or not the person parameters differ depending on the culture or follow highly different distributions, the item parameters have to be equal, i.e. not significantly different, provided measurement invariance holds.

In practice, some technical aspects have to be observed. The scale of the parameters is a distance scale, i.e. all probabilities remain unchanged when a constant is added to all parameters while any multiplication would change the probabilities. Thus, only the origin has to be defined (Hambleton et al, 1991), analogous to the scale definition of the latent variable in CFA. It is crucial that this scale definition is conducted in the same fashion for all samples. This can be achieved by fixing the mean of the item parameters (or one item parameter) to zero. The alternative way of fixing the person parameters to zero would be tantamount to projecting differences, if any, into the item parameters. That's why the latter approach is totally inappropriate. It should be noted that fixing item parameters neither affects raw data in any way, nor does it result in any loss of information inherent in standardisation procedures.

**Generalisation of the dichotomous model for polytomous data**

The outlined model may be generalised by introducing further parameters for various purposes (Anderson, 1995; Fischer, 1995; Glas and Verhelst, 1995; Hambleton et al, 1991; von Davier and Rost, 1995). For empirical marketing research applications, generalisations regarding the number of item answer categories are most interesting. Dichotomous scales are not very widespread, instead multicategorical, i.e. polytomous scales are used. In principle, the answers to polytomous scales may be dichotomised (e.g., Ellis and Kimmel, 1992; Singh, 1996). However, apart from the problem of dichotomisation of odd-numbered categories and the fact that this would always represent a loss of information, collapsing categories is, from a theoretical point of view, very problematic in Latent Trait Theory (see Andrich, 1988). To avoid these problems, LTT-models for polytomously scored items should be used. Fortunately, the dichotomous Rasch model may be generalised straightforward without any additional assumptions and without loss of the special properties of the model, i.e., primarily, specific objectivity. In the case
of dichotomous items, the difficulty parameter can be viewed as a threshold between agreement and disagreement, i.e. the (only) two categories. By the same token, the thresholds between each pair of (adjacent) categories of a multycategorical item, e.g., fully agree and somewhat agree, may be modelled by the dichotomous Rasch model resulting in m-1 dichotomous models for an item with m response categories. Subsequently, these models are combined by a simple multiplication since all thresholds are modelled independently from each other. However, the mean of all threshold parameters of an item may still be regarded as an overall location parameter. Each category may then be described by an ICC-like curve: the category characteristic curve (CCC). Consequently, each item is characterised by a set of CCCs. There are several models concerning polytomously scored items differing in the number of constraints on the threshold differences. The interested reader is referred to Andrich (1978b) for the rating scale model, which assumes equal distances between adjacent categories across all items, and to Masters (1982) and Andrich (1988) for the partial credit model, which does not impose any restrictions on the threshold parameters at all.

**Model testing and cross-cultural invariance**

The rationale of applying the Rasch model to testing for measurement equivalence goes back to the basic principle of specific objectivity. Given the cross-cultural validity of the model, the item parameters estimated from samples of different culture may only differ by chance. Where the item-parameters do differ, the model does not hold for that item and invariance has to be considered falsified (Hambleton et al., 1991). Non-invariant items do not mean the same to respondents from different cultures and are said to show Differential Item Functioning (DIF) (Lord 1980, p. 212). Thus, they are inappropriate for cross-cultural comparisons and should be omitted from the measurement instrument. However, the invariant set of items allows any comparative interpretation including mean comparisons, because the same scale refers to different cultures. Strictly speaking, in the Rasch approach there is no parallel to partial invariance in CFA. The reason is that linking DIF items to the same scale across cultures would mean that a person measure would depend on the group the person belongs to. This, however, would violate specific objectivity.

Practically, the approach starts with the parameter estimation in each culture, preferably by conditional maximum likelihood estimation which retains specific objectivity. Then the significance of the differences between parameter estimates is assessed by Lord's chi-square (Hambleton et al., p.111) test for each item. In the case of polytomous items, the overall location parameter can be used for that purpose. After the exclusion of a non-invariant item, the remaining items should be re-calibrated since any non-invariant item alters the scale definition and, consequently, all other item parameters.

**Comparison of the CFA Approach and the LTT Approach**

The CFA approach and the LTT approach differ substantially in terms of the underlying measurement paradigm. Whenever paradigmatic foundations are concerned, in the present case the very definition of measurement, a comparison of competing approaches applied to a specific empirical setting may never provide a basis on which a decision in favour of either approach may be made. The reason is that the paradigm has to be chosen first and the approach for testing for equivalence follows. Specifically, the Rasch model claims to define how reasonable data sets should look like if they were to constitute measurement. Consequently, data may never falsify the Rasch model as such. The same is true for CTT models for they may also never be falsified as such.

Thus, the comparison of the approaches outlined has to consider the strengths and weaknesses of the underlying measurement theories. The foundation of Classical Test Theory is the well-known basic definition of the observed score as a combination of true score and measurement error. This definition implies at least two severe problems. First, it does not explain how measurement is accomplished, instead it treats an observed score as if it constitutes a measurement. Second, the basic equation of CTT represents a tautology and may never be falsified empirically, i.e. the item score may always be represented by the sum of a true score and an error score. Moreover, the observed score as a sum of several item scores is referred to as a continuous variable (Gullikson, 1950; Steyer and Eid, 1993) without explaining why this would be justified, given the qualitative nature of responses to dichotomous manifest items. Dealing with polytomous items, this issue is even more problematic in that summing up item scores.
requires interval scale properties which are, especially in cross-cultural research, highly questionable.

In contrast, the Rasch model states how responses and response pattern have to look like, i.e. the prediction of the model regarding item response behaviour may be falsified empirically. Thus, LTT states how measurement is accomplished in principle, namely by an interaction of item and person parameters, and leaves the question whether measurement is accomplished to empiricism. Since the model predicts individual response behaviour, any type of response bias may be easily detected by person fit statistics which compare the actual and the expected response patterns.

In terms of scale properties, in LTT, for the items answer categories, ordinal scaling is sufficient and, in contrast to CTT, represents a testable hypothesis of the data (Andrich, 1978a; Andrich, 1978b). This is important, since interval scale properties are particularly questionable in cross-cultural research (Mullen, 1995). Thus, LTT is able to handle settings in which it is doubtful whether the distances between two adjacent categories are equal, or even in the supposed order, and culturally universal.

Another important issue is the question of interdependencies between item and person parameters. Due to specific objectivity, the item characteristics are independent of person properties and vice versa while in CTT (and also in, e.g., the Birnbaum model) person and item parameters may not be separated. Actually, item location, or difficulty, parameters are hard to define in CTT. With the item intercept parameter, the CFA approach implies a definition of item difficulty. Unfortunately, this parameter is not independent of the slope of the regression line, i.e. the factor loading, whenever observed item scores are bounded within the relatively narrow limits of frequently used rating scales. If the content of an item is very hard to endorse compared to the content of the average item in the test, the observed score that would result from the regression equation would lie out of the range of actually possible category codes for many respondents. In such a case, many examiners with very different attitudes would choose one and the same extreme category (e.g., fully disagree). Due to this floor or ceiling effect, the correlation of this item with other items in the instrument would be underestimated and, thus, the factor loadings would also be underestimated. The meaning of item difficulty, operationalised as the item intercept would then be severely reduced. In practice, the item would (falsely!) be regarded as improper or bad. In fact, the only thing that is “wrong” with the item is that its difficulty is right above or below average.

For the CFA approach, this means that the test for metric and the test for scalar invariance are not fully independent whenever item difficulties vary substantially within each culture. As Singh (1996) pointed out, the factorial model is only valid if all items are equally “difficult”. This fact may be taken into consideration in test construction (explicitly or implicitly by discarding items with extreme difficulties). However, equal or very similar difficulty of all items would mean that the items in the instrument represent equal or very similar levels of attitude, tantamount to a rather narrow effective range of measurement. In contrast, the Rasch model suggests a wide range of item locations spreading a real dimension rather than establishing a point of measurement.

Although no empirical setting may help us decide which paradigm to follow, a simulation study shall illustrate either approach. Moreover the study will show, how the competing approaches cope with the problem of scale definition. As stated above, invariance testing by CFA requires choosing one item that is regarded as culturally invariant for the purpose of the latent variable’s scale definition. When at least one additional item turns out to be invariant, a common cross-culturally valid scale has been established. If no other item is cross-culturally invariant, Steenkamp and Baumgartner (1998) suggest choosing a different item for scale definition, since the one chosen originally might not be cross-culturally invariant and, therefore, unsuitable to establish a common scale. The question remains whether the choice of an item lacking invariance will always and necessarily be detected. Assume we are to measure a unidimensional construct by a 10 item scale within two different cultures. If five items are equally biased in one culture, i.e. affected by the same additive bias in reference to the other culture, and the remaining five items are truly invariant, then choosing one of the biased items would result in a seemingly invariant set of five items while the truly invariant items are falsely set out as biased. Thus, in this situation, we have a fifty percent chance to end up with the right conclusion and with the same probability we would make a wrong inference concerning data equivalence. Therefore, the quantitative assessment of data equivalence has to be accompanied with a qualitative examination. Before quantitative analyses are
carried out, hypotheses about which items may and which may not be invariant should be formulated. These hypotheses should be based on empirical experience and/or plausibility considerations. After the quantitative analysis, the results have to be investigated in terms of meaningfulness and plausibility.

How would the Rasch model based analysis cope with the situation of five unbiassed and five uniformly biased items? If the mean of the item locations is set to zero, then all items seem to be biased at first. A simple visual inspection of difficulties from both cultures plotted against each other would show that two sets of five items each may be unbiassed. The Rasch model does not help decide which set is truly invariant. The same, of course, is true for the CFA approach. The reason is that all we can ever detect are biases that affect differences between items. Consequently, there is no way to detect a uniform additive bias that renders all items invariant. The relative advantage of the Rasch model approach over the CTT based approach lies in the fact that the initial analysis clearly shows that a set of items is affected by a uniform bias and that qualitative considerations are indispensable.

SIMULATION STUDY

The purpose of the simulation study is to demonstrate the different approaches for the detection of additive measurement bias. To this end, simulated data have been generated using the random value function of SPSS. First, a normally distributed true score variable, using 400 respondents, has been computed with a variance of 1.2. Subsequently, ten manifest variables have been generated on the basis of this true score variable. Thus, any manifest variable is a linear combination of the true score, an item intercept value and a normally distributed error component. Hence, all items are perfectly unidimensional. For the error term a variance of 0.8 has been chosen for any item, i.e. each item has the same reliability.

In order to introduce an additive measurement bias, the data set has been randomly split into two halves. Therefore, there is no difference in terms of true scores between the two groups, i.e. the two cultures. Whereas for items one to five intercept values have been chosen invariantly over groups, intercept values for items six to ten differ by a constant between the two groups. Therefore items one to five are not affected by any bias whereas items six to ten are biased. Finally, all manifest variables have been rounded to the nearest integer and bounded within a hypothetical response scale ranging from one to five. Only a limited number of item values had to be bounded since the intercept values have been selected in order to minimise the need for bounding. Insofar as the data generated complies with the CTT model more than real data would ever do.

In terms of invariance testing by the confirmatory factor analysis approach, configural and full metric invariance is fulfilled. Scalar invariance is given for items one to five but an additive bias prevails for items six to ten. Therefore scalar invariance is limited to partial invariance for items one to five. Correspondingly, on the basis of LTT, items one to five have cross-culturally identical ICCs, whereas the ICCs for the remaining items differ in that the curves are shifted due to the additive bias.

CFA Approach

The CFA approach\(^1\) starts with the examination of configural invariance. As expected, the model fits the data quite well ($\chi^2=83.18$, df=70, $p=.13$). This fit statistic provides a baseline against which further invariance levels are tested (see Table 1 for complete results). Imposing constraints on the magnitude of salient loadings, we test for metric invariance and get a fit of $\chi^2=94.41$ (df=79, $p=.11$). The decrease in fit ($\Delta\chi^2=11.23$, df=9, $p=.26$) is insignificant, therefore full metric invariance holds in accordance with the true structure of the data. Subsequently, a test of full scalar invariance is carried out. What we would expect is a significant deterioration in fit. In fact, the $\chi^2$ increases dramatically and reaches 313.8 (df=88, $p<.01$). This drop in model fit is significant at any reasonable level we might select ($\Delta\chi^2=230.62$, df=18, $p<.01$). Thus, the analysis calls for testing for partial scalar invariance. If we screen the modification indices, we find that item one shows the largest modification index in terms of item intercepts. So, we lift the constraint of scalar invariance for that item and choose item six for scale definition purposes. Proceeding this way, we have to lift intercept equality constraints for another four items (items two to five) until ending up with a model of partial scalar invariance which

---

\(^1\) All analyses were carried out using LISREL 8 (Jöreskog and Sörbom 1993).
has a fit that is no longer significantly worse compared to the baseline model of configural invariance ($\Delta \chi^2 = 13.09$, df=13, p=.44).

As a result of the CFA analysis, five items are cross-culturally fully equivalent in terms of scalar invariance while the equivalence of another five items is limited to metric invariance. However, the outcome does not correspond to the true level of item equivalence. Whereas the analysis suggests items six to ten being invariant, items one to five are truly invariant. Thus, the model equates the scales using the wrong level. Therefore, testing for the equivalence of latent means, which is carried out by imposing a further equality constraint on the mean of the latent variable across groups, comes to the conclusion that there is a significant difference between groups on the construct investigated ($\Delta \chi^2 = 74.94$, df=14, p<.01). This demonstrates that the choice of the item for scale definition purposes, which is supposed to be invariant, is not a trivial issue. The simulation study shows that even four other items showing full scalar invariance do not guarantee that a meaningful solution has been found. As a consequence, the choice has to be made on the basis of additional qualitative judgements rather than following the mere values of modification indices. The result of such an analysis might be that items one to five are disguised items which are less likely affected by additive biases than undisguised items. Based on this assumption, we might maintain equality constraints for intercepts of items one to five and lift constraints referring to items six to ten. Finally, we get a model of partial scalar invariance with items one to five fully equivalent across groups ($\Delta \chi^2 = 14.72$, df=13, p=.33) – corresponding with the algorithm of data generation. Consequently, the test for equal latent means shows no significance ($\Delta \chi^2 = 14.76$, df=14, p=.39).

The results demonstrate that the CFA approach should not be carried out blindly. Particularly, the question of whether an item is suitable for scale definition purposes should be based on both quantitative indicators (as the modification index), as well as qualitative considerations. Ignoring qualitative judgements might result in incorrect and totally misleading conclusions as is demonstrated by the simulation study. Unfortunately, the CFA approach does not provide any “warning”, i.e. we might not recognise that there is another equally appropriate solution. It is rather a question of chance whether the scale of the latent variable is defined correctly or not.

**Table 1: CFA results of the simulation study**

<table>
<thead>
<tr>
<th>type of model</th>
<th>via v1</th>
<th>via v6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\chi^2$</td>
<td>df</td>
</tr>
<tr>
<td><strong>full invariance</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Configural</td>
<td>83.18</td>
<td>70</td>
</tr>
<tr>
<td>Metric</td>
<td>94.41</td>
<td>79</td>
</tr>
<tr>
<td>$\Delta$ configural - metric</td>
<td>11.23</td>
<td>9</td>
</tr>
<tr>
<td>Scalar</td>
<td>313.8</td>
<td>88</td>
</tr>
<tr>
<td>$\Delta$ configural - scalar</td>
<td>230.62</td>
<td>18</td>
</tr>
<tr>
<td><strong>partial invariance</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scalar</td>
<td>97.9</td>
<td>83</td>
</tr>
<tr>
<td>$\Delta$ configural - partial scalar</td>
<td>14.72</td>
<td>13</td>
</tr>
<tr>
<td>latent means</td>
<td>97.94</td>
<td>84</td>
</tr>
<tr>
<td>$\Delta$ configural - partial scalar/latent means</td>
<td>14.76</td>
<td>14</td>
</tr>
</tbody>
</table>

**Latent Trait Theory Based Approach**

Using the Rasch Model for polytomous data, the LTT based analysis of the simulated data set starts with item calibrations carried out separately in each group.\(^2\) Since we generated a five point rating scale, we get four threshold parameters for each item. The item parameters were independently scaled within each group by fixing their mean to zero. If all items were cross-culturally invariant, then all item parameter estimates in group A

\(^2\) All LTT analyses were performed using WINMIRA (van Davier 1995).
and in group B should correspond irrespective of the sample composition or any distributional characteristics of either group. In order to compare the estimated parameters, we plot the overall item locations, i.e. the means of the four threshold parameters of each item, in group A against the estimates in group B. Given cross-cultural invariance, all pairs of estimates should lie on a straight line with a 45 degree angle passing through the origin. Since five out of ten items of the simulated data are not cross-culturally invariant, we do not expect to get this result. Indeed, none of the ten items lie on the 45 degree straight line through the origin (see Figure 4). However, the subset of items one to five does lie on a straight line that is shifted upwards. The subset of the remaining items, i.e. items six to ten, also lies on a straight line that is parallel to the former and shifted downwards.

Thus, the LTT based approach unambiguously leads us to the conclusion that there prevails an additive bias. Moreover, the plot suggests that discarding either items one to five or items six to ten may lead to an invariant set of items.

Therefore, the analysis proceeds by eliminating items six to ten and re-calibrating the remaining items within each group. Figure 5 illustrates the results. Now, all item parameter estimates lie on a straight line through the origin indicating that all items are functioning identically across groups, i.e. the item parameter estimates would certainly not differ significantly. Alternatively, we could remove items one to five and recalibrate items six to ten. The outcome is depicted by Figure 6. Since all item parameter estimates lie on a straight line through the origin, we would come to the same conclusion: all items are cross-culturally invariant and therefore valid indicators across cultures.

Summarising the results of the CFA and the LTT based approach of testing for data equivalence, we demonstrated that, in any case, a qualitative judgement has to be made whether the result is plausible or not. This is especially true, if there are different models fitting equally well but suggesting different conclusions in terms of which items are cross-culturally invariant and which are not. The relative advantage of the LTT approach, besides its theoretical merits, lies in the fact that even the analysis of invariance of all items clearly shows that there are two separate sets of items and that there are two competing solutions. In contrast, the CFA analysis puts us on the wrong track initially and it is highly questionable whether we would have tested another model not knowing the true level of data equivalence. The reason, of course, is that the LTT based approach defines the latent scale by the mean of all item parameters whereas the CFA approach relies on one item.
Simulated Data Versus Real Data

We decided to illustrate the approaches suggested by simulated data rather than real data sets. The reason lies in the fact that one has to know the data properties in order to evaluate different methods. We need to know which level of invariance actually prevails, if we want to make statements on the quality of the approach applied. Specifically, the comparison of methods and the decision which method is the better one requires knowledge of how the results should look like and which conclusion is the right one. Therefore, real data sets would be totally inappropriate, since we would not know a priori what level of equivalence is actually achieved. The same applies to mean differences. We must know whether there are true differences or not, before evaluating how effective an approach is in detecting differences and similarities.

Admittedly, the simulation study illustrates only a narrow cut of reality. In most cases, non-invariant items will be affected by non-uniform biases, i.e. there will be no alternative subset featuring partial scalar invariance. However, it is important to know that this might be the case nevertheless. Moreover, the data set was generated in a way that complies with CTT. In most empirical settings in domestic as well as international marketing research, this will simply not be the case and, consequently, the LTT approach will be the more appropriate one due to its superiority in terms of the underlying measurement theory.

CONCLUSIONS

This paper attempts to illustrate the importance of equivalence in cross-country research and provides a comparison of two methods suitable for establishing and testing measurement equivalence, namely the CFA approach and the LTT approach. The comparison of these two methods reveals a number of advantages and disadvantages. In general, the LTT approach offers a superior avenue to measuring latent constructs in general and in cross-cultural settings in particular. Among others, this is due to the basic properties of the Rasch Model. The Rasch Model allows testing of whether items follow the model or not. Furthermore there is an independence of person and item parameters. Finally, in terms of response scales, the Rasch Model comes with less stringent assumptions, thus permitting the use of ordinal scales.

Notwithstanding the advantages of LTT compared to Classical Test Theory, there seem to be a number of practical aspects which limit the application of an LTT approach in most empirical settings. First of all, to date the theoretical foundations of LTT have not really been adopted in cross-cultural marketing research. While some of the shortcomings of CTT might be well known by marketing researchers, knowledge about possible research avenues on how to overcome those problems is not yet widespread. The computing procedures involved, which are unfamiliar to many researchers, represent a further obstacle to application. Last but not least, LTT-models are very stringent models and empirical data often falsifies the proposed model even within the same culture. As mentioned above, the Rasch Model will be rejected for items substantially indicating more than one dimension and therefore being under- or overdiscriminating. Other LTT models account for unequal discrimination but lead to other problems like sample dependencies of parameters and possible lack of construct validity.

Thus, from a practical point of view, researchers are likely to turn to Classical Test Theory, i.e. the CFA approach. Considering the fact that most cross-cultural comparisons neglect equivalence issues at least as far as measurement equivalence is concerned, the application of this approach is a clear improvement and enhances the
quality of conclusions. However, in the long run, the application of LTT models seems to be a far more promising research avenue.

Moreover, there is an interesting way of combining both approaches. The CFA approach offers opportunities to extend the model beyond mere measurement models by introducing structural relationships between latent constructs. Thus, the principle of constraining parameters cross-culturally may also be applied to antecedent and dependent constructs and their relationships with the construct in mind. Consequently, it is the structural equation modeling approach that offers the opportunity to test for cross-culturally equivalent nomological validity. However, since LTT measurement models are superior to CTT based approaches, the measurement models should be replaced by LTT models.

For an international marketing research practitioner, the paper illustrates two issues: First, the importance of testing scales for their cross-cultural invariance before engaging in comparative research across countries. Second, the complexities of testing for equivalence in cross-cultural research with either multiple group structural equation modelling or with Latent Trait Theory. This causes a potential dilemma, namely should the researcher ignore the equivalence issues or should (s)he engage in time consuming and complex statistical analyses. Since none of these options is satisfactory, this paper calls for the incorporation of the equivalence issue in scale development. Thus, instead of conducting a scale development exercise exclusively in one culture, scale developers should, ideally, test the applicability of their scale across different cultures and, ultimately, provide appropriate data on the cross-cultural equivalence of their scales.

This, of course, also opens up plenty of opportunities for future research contributions in that a number of existing scales need to be analysed for their cross-cultural suitability. A second research avenue emerges from questioning the desirability of confining studies to culturally universal, i.e. etic items. Arguably, such approach is akin to identifying the lowest common denominator and neglects culture-specific facets. In essence, data equivalence refers to internal validity, i.e. it ensures that data are accomplished in a comparable way. It can not, however, replace external validation of the scale in any culture regarded. Researchers may wish to develop comparative techniques that focus on a culturally invariant core of a marketing concept as well as on culture specific components. Of course, this does not negate the need to check for measurement equivalence, as culturally invariant items would have to be distinguished from culture specific items.

Finally, a key result of the simulation study is that data equivalence is a quantitative as well as a qualitative issue. When conducting quantitative cross-cultural marketing research, we suggest carefully regarding all aspects of equivalence prior to and during data collection. Subsequently, hypotheses should be developed in terms of how likely an item might be affected by biases. As to data equivalence, quantitative analyses are absolutely indispensable. Since many researchers are familiar with CFA techniques and scale development is primarily based on CTT, the CFA based approach is likely to meet a favourable response. In the long run, LTT based approaches will enrich the methodological instruments of marketing researchers. However, this will depend on the pace of penetrating the scientific community in our discipline.

REFERENCES


Wright, Benjamin D. and Mark Stone (1979), *Best Test Design*. Chicago: Mesa.