PSIHOLOGIJA, 2010, Vol. 43 (2),
© 2010 by the Serbian Psychological Association
UDC 159.92:159.072.594 ; 316.722:159.072.594
DOI:10.2298/PSI1002121K

Researching Measurement Equivalence in Cross-Cultural Studies

Miloš Kankaraš and Guy Moors
Tilburg University, The Netherlands

In cross-cultural comparative studies it is essential to establish equivalent measurement of relevant constructs across cultures. If this equivalence is not confirmed it is difficult if not impossible to make meaningful comparison of results across countries. This work presents concept of measurement equivalence, its relationship with other related concepts, different equivalence levels and causes of inequivalence in cross-cultural research. It also reviews three main approaches to the analysis of measurement equivalence – multigroup confirmatory factor analysis, differential item functioning, and multigroup latent class analysis – with special emphasis on their similarities and differences, as well as comparative advantages.

Key words: *measurement equivalence, cross-cultural studies, factor analysis, differential item functioning, latent class analysis;*

An ever-increasing line of cross-cultural research in psychology is produced in recent decades. Cross-national, cross-cultural and multi-lingual studies are used to compare a wide scope of opinions, attitudes, values, and abilities among different cultural groups. This increased interest in comparative studies may be related to the rapid changes our societies are going through (Van de Vijver, 1998). Economical, social, political and technological globalization and increasing migration are eroding centuries-old national boundaries, transforming local and regional phenomena into global ones. In Europe, integration processes and eastward expansion of European Union additionally triggered substantial amount of comparative research.

In spite of this considerable interest in cross-cultural studies, there is still no generally accepted way of dealing with issues specific to this kind of research. Since the same instruments (i.e. questionnaires, inventories, tests, etc.) are used for all involved groups, it is frequently simply assumed that obtained results are comparable among groups. This assumption of results' comparability, although critical for valid comparison, is often not tested, with researchers

Corresponding author: m.kankarash@uvt.nl

usually only focusing on the difference in average scores of the two or more cultural groups. However, as each cultural context reflects a constellation of many factors, processes, and attributes, the same set of questions or assignments may have different meaning for people from different cultures, i.e. it may measure somewhat different constructs in each culture. If this happens validity of the conclusions from such comparative research is in question. Therefore, a fundamental concern in any cross-cultural research is ensuring equivalence (i.e. comparability) when testing for cross-cultural differences (Hui & Triandis, 1985). In methodology this comparability, called *measurement equivalence*[2], is defined as "whether or not, under different conditions of observing and studying phenomena, measurement operations yield measures of the same attribute" (Horn & McArdle, 1992, p. 117).

The aim of this work is to introduce the theoretical background and quantitative methodological procedures in measurement equivalence research. The concept of measurement equivalence, its relation with other forms of equivalence, different equivalence levels and causes of inequivalence in cross-cultural research are presented in a first section. In a second section we show the main approaches to the analysis of measurement equivalence, their differences and similarities, as well as comparative advantages.

1. THE CONCEPT OF MEASUREMENT EQUIVALENCE

1.1 Measurement equivalence

The term "equivalence" has been used in a wide variety of disciplines representing different concepts and acquiring different meanings. Johnson (1998) found more than 50 specific terms used to indicate various forms of equivalence which he divided into two broad groups:

– *Interpretative equivalence* comprises all types of equivalences that are primarily concerned with the similarities and differences in interpretation or meaning of measured constructs across cultures. Conceptual equivalence defined as the possibility to meaningfully discuss constructs within each culture of interest (Hui & Triandis, 1985) belongs to this group, as well as functional equivalence that denotes the degree to which the concepts serves similar functions within each involved society (Singh, 1995).

Similarity in interpretation is a necessary but not sufficient condition for comparability of results: in order to validly compare results one also needs to establish equivalent measurement procedures. *Procedural equivalence* refers to those types of equivalence that are dealing with measures and procedures used in cross-cultural studies. Measurement equivalence (Horn & McArdle, 1992) and item equivalence (Hui & Triandis, 1985) belong to this group as both are

2 Measurement equivalence is also known as *measurement invariance* or *structural equivalence* (van de Vijver, 1998).

focused on the degree of similarity of measurement procedures across cultures. Interpretative equivalence is a precondition to procedural equivalence since equivalence in measurement procedures is not possible without equivalence in interpretations of measured constructs.

Thus, measurement equivalence implies that a same measurement instrument used in different cultures measures the same construct. In other words, measurement equivalence of cross-cultural results is established when the resulting differences across cultures in answers on test or questionnaire items are due only to the cross-cultural differences in measured constructs. If, on the other hand, respondents' answers reflect not only their position on the construct that is being measured, but are also influenced with additional factors and considerations that are different across cultures, the results will be measurement *inequivalent*.

In ability testing, for example, the problem of measurement equivalence has been present from the very beginning. IQ tests, designed to measure people's intelligence, have a long history of misuse, especially in the early days of mental testing (Gould, 1981). These tests, developed and standardized in USA were used to assess "intelligence" of individuals and groups of different races, nationalities and languages. Resulting lower scores of these groups are then interpreted as confirmation of the group differences in intellectual abilities which in some cases had dire consequences for these people. In one such example, an American psychologist Henry Goddard tested the IQ of immigrants from South and Eastern Europe in Ellis Island in 1910s, finding that large portions of their population are 'feeble-minded' (Goddard, 1917). Goddard's work resulted in dramatically higher deportation rates (Hothersall, 1995). Although the most notable of these methodological abuses occurred in early days of mental testing (the first half of 20th century), it is still common to find IQ studies that have largely ignored the question of measurement equivalence (Lynn, 2006).

In personality and attitudinal measurement the problem of measurement equivalence is equally important. Although there has been much research on the similarity of personality factor structures across cultures (Church & Lonner, 1998), few studies performed a thorough assessment of measurement equivalence. The notion of measurement equivalence is closely related to the concepts of *item bias* and *differential item functioning* (DIF) that streams from the framework of item response theory (IRT) and are intensively studied in the field of educational testing (Van de Vijver & Leung, 1997). DIF or item bias is defined as differences in answer probabilities for respondents with equal latent disposition. For example, if we (hypothetically) measure "the intensity of supportive behavior towards sport" in Serbia and England and ask a question about how often one attends football matches, the probability of answering, for example, "once a week" could be lower for a football fan in Serbia with the same attitude toward football as a counterpart from England, since attendance in football matches is much lower in Serbia than in England. In other words, this indicator would be biased if used in these two countries for comparative purposes.

1.2 Various levels of equivalence

From a measurement perspective, there are a number of different hierarchically linked types of equivalence that assume increasingly stronger level of measurement comparability across cultures. The three most important levels of measurement equivalence are configural, metric and scalar equivalence, although additional levels may be operationalized and investigated (Vandenberg & Lance, 2000). These levels are ordered hierarchically, in the sense that higher equivalence levels presuppose lower ones. Higher equivalence levels are harder to obtain as they provide a stronger test of cross-cultural equivalence, but also allow a more extended form of cross-cultural or cross-time comparison.

Configural equivalence. The basic level of methodological equivalence is "configural equivalence" (Horn & McArdle, 1992; Steenkamp & Baumgartner, 1998) also called "factor equivalence" and "structural equivalence" (Van de Vijver & Leung, 1997). Configural equivalence implies similarity of data configurations or structures across cultures, i.e. it assesses if the set of observed indicators (e.g. questionnaire items) has the same pattern (structure, configuration) of existing and non-existing relationships (e.g. factor loadings) with the construct to be measured across cultures. Thus, at this level of equivalence, it is not necessary that these relationships have exactly the same strength but that the same set of questions is related to same concepts in each culture.

Metric (measurement unit) equivalence: Configural equivalence does not indicate that respondents from different cultures assign the same meaning to questions, i.e. it does not allow for straightforward comparison of results. Metric, "measurement unit", or factorial equivalence is a more stringent form of equivalence as it subsumes configural equivalence and additionally assumes that the relationship between observed indicators and latent concepts is equal across groups (Singh, 1995; Cheung & Rensvold, 2000). In other words, metric equivalence implies the equality of the measurement units or intervals of the scale on which the latent concept is measured across cultural groups (Steenkamp & Baumgartner, 1998; Van de Vijver & Leung, 1997). In statistical terms it is operationalized as inter-group equality of slope parameters that measure the relationship between latent and observed variables. For example, in the context of factor analysis this would imply that the factor loadings of each item are equal across groups, i.e. that a questionnaire or test items are understood in a similar way in different cultures.

This level of equivalence implies that the instrument measures the same latent construct in all of the cultural groups under investigation. Thus, metric equivalence represents a necessary and sufficient condition for comparison of difference scores (e.g. mean-corrected scores) across countries. It also enables valid comparison of relationships of the latent variable with other variables of interest (Steenkamp & Baumgartner, 1998).

However, even with equal measurement units latent variable scores can still be uniformly biased upward or downward as they do not necessary share the same origin of the scale (Van de Vijver & Leung, 1997). In case of such

additive bias respondents with the same latent disposition but from different cultural group might have systematically higher or lower observed values. This possibility of additive bias prevents metric equivalence from enabling for full score comparability, i.e. it is not a sufficient condition for comparison of country/ cultural group means (Meredith, 1993).

Scalar equivalence: In order to establish complete measurement equivalence and to enable full comparison of country scores, including country means, it is necessary that the scales of the latent construct have the same origin. When measures also have a common origin across groups, they are considered to have scalar equivalence (Meredith, 1993; Steenkamp & Baumgartner, 1998) or calibration equivalence (Mullen, 1995). The distinction between measurement unit and scalar equivalence is important in cross-cultural research as in most cases the main research question of these studies pertains the use and comparison of mean scores across cultural groups.

It is important to note that equivalence of the parameters for *all items* is not necessary for substantive analyses to be meaningful (Byrne, Shavelson, & Muthén, 1989; Steenkamp & Baumgartner, 1998). Ideally, most of the items will be equivalent across countries because in that case latent means are estimated more reliably, i.e. they are based on many cross-culturally comparable items. However, as metric and/or scalar equivalence are unlikely in many situations, researchers can resort to *partial equivalence* as a compromise between full measurement equivalence and complete lack of measurement equivalence. Cross-national comparisons can be made in a valid way if at least two items per construct are equivalent. One item per scale (the so called "marker" item) has to be fixed to define the scale of each latent construct. However, in order to test equivalence of the marker item one more items needs to be equivalent. Thus, partial equivalence enables a researcher to control for a limited number of violations of the equivalence requirements and to proceed with substantive analysis of cross-cultural data (Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000).

1.3 Causes of inequivalence

The main sources of inequivalence in cross-cultural research are different forms of biases (Van de Vijver & Leung, 1997; Van de Vijver, 1998). Biases occur when obtained results systematically misrepresent true scores on the measured construct. Thus, biases indicate a treat to validity of instruments and can decrease the level of equivalence thus impairing the comparability of cross-cultural data. There are two main forms of biases: construct and method bias; former refers to differences in compared constructs while later denotes differences in the process of measurement.

Construct bias indicates dissimilarity of constructs in investigated cultures. It is present when an instrument measures constructs that differ or only partially overlap across cultures. Thus, it is not possible to establish interpretative

equivalence with these kind of phenomena as they do not share a same meaning across cultures, i.e. their comparison parallels that of comparing "apples and oranges" (Johnson, 1998). Construct bias will usually increase as the cultural distance is wider and when a given instrument is more saturated into a specific culture. Construct bias affects all levels of measurement equivalence and in most part prevents quantitative cross-cultural comparisons.

Method bias represents all kinds of biases that originate from methodological and procedural aspects of a cross-cultural study. Method bias is further divided into three subtypes of bias (Van de Vijver & Leung, 1997):

– First, there may be *sample bias* that stands for all differences in characteristics of samples from different cultures that can influence results. This type of bias is especially important when comparing highly divergent groups when random sampling can lead to dissimilar groups in terms of background variables. Matching samples procedure provides better control for the influence of background variables, but it can impair representativeness of the sample results to given populations.
– The second type of method bias is *instrument bias*. It is caused by characteristics of an instrument to which individuals from different cultural groups react in consistently dissimilar way. This type of bias includes differences in stimulus familiarity (which is especially important in mental testing), social desirability and response styles (that are more important in personality measurement). Different stimulus familiarity is one of the main sources of inequivalence in cross-cultural administration of intelligence tests, as their content is in most cases heavily saturated with academic material whose familiarity can vary greatly across cultures. Likewise, in opinion and attitude research, social desirability can impair validity of cross-cultural findings as different cultures can have different notions of social desirability that can influence their answers.

  A common form of instrument bias in attitudinal research is *response bias* that refers to "a systematic tendency to respond to a range of questionnaire items on some basis other than the specific item content" (Paulhus, 1991, p. 17). Response bias can seriously distort not only measurement of attitudes but the effects of covariates on these attitudes too. There are two main forms of response bias: extreme response bias, when a respondent systematically chooses extreme answer categories irrespective of specific item content, and agreement (acquiescence) bias that represents the tendency to choose agreeing answer categories without regard to item content (Cheung & Rensvold, 2000). These response biases are especially common when Likert-type rating scales are used and they present direct treat to measurement equivalence to the degree in which cultures differ in incidence of these response tendencies. For example, it is found that people of Mediterranean region are more prone to extreme response bias than people living in north Europe, same as when

African-American and Latino-American are compared with Americans of European origin (Greenleaf, 1992; Hui & Triandis, 1989).

– The third type of method bias is *administration bias* induced by various procedural aspects of the data collection, such as interviewer characteristics, testing facilities, communication problems, etc. For example, it is sometimes the case that the administration procedure can use locally inappropriate modes of communication or violate other social norms in one culture. Likewise, the conditions in which the data are collected can vary considerably across countries. For instance, in a study with the Ravens colored matrices in Nigeria, "children were tested on porches, in entrance rooms, or under trees" by untrained personnel (Fahrmeier, 1975, p. 282). This is very different from testing conditions of their counterparts in the West to whom they are routinely compared (Wicherts, 2007).

Method biases do not affect interpretative equivalence since this type of equivalence implies only that the same construct is measured across cultures. As long as no direct score comparison are conducted across cultures, the presence of method bias does not halt the interpretability of data within each cultural group independently. However, method biases can seriously impair all forms of measurement equivalence and bring validity of cross-cultural comparisons into question. For example, if the prevalence of agreement bias differs systematically across cultures, it will conceal or distort the real underlying cross-cultural differences on the measured construct. This type of bias, hence, will reduce the level of equivalence from scalar to metric equivalence and, consequently, it will not be possible to validly compare mean scores across countries. On the other hand, extreme response bias affects both metric and scalar equivalence and can further reduce the level of measurement equivalence to configural equivalence (Cheung & Rensvold, 2000).

2. ANALYSIS OF MEASUREMENT EQUIVALENCE

2.1 Main approaches to analysis of measurement equivalence

Among several approaches for testing measurement equivalence of cross-cultural data that have been suggested, the most prominent are multigroup confirmatory factor analysis (MCFA) (Steenkamp & Baumgartner 1998; Vandenberg & Lance, 2000) and methods for detecting differential item functioning (DIF) developed in the context of item response theory (IRT) (Raju, Laffitte, & Byrne, 2002). A third, less well-known but promising approach that combines multiple-group latent class analysis (Clogg & Goodman, 1984; McCutcheon, 2002) with latent class factor analysis (LCFA; Magidson & Vermunt, 2001) has been recently introduced (Moors, 2004; Kankaraš & Moors, 2009). The three approaches share a common core, i.e. defining a measurement model by comparing the latent structure for several groups in a single model.

2.1.1 MCFA

MCFA (Jöreskog, 1971) is the prevailing methodological approach for measurement equivalence assessment (Byrne et al., 1989; Meredith, 1993; Steenkamp & Baumgartner, 1998, Vandenberg & Lance, 2000). MCFA is a parametric, linear approach which assumes that both the latent construct and observed variable (e.g. item scale) are of continuous nature. It basically investigates whether the factor loadings, intercepts and error variances of a given model are equal across groups. Assuming that the factor structure is the same for all groups (i.e. configural equivalence), a multi-group CFA model implies the following linear regression model for item *k* for someone belonging to group *g*:

$$E(y_k \mid \Theta, g) = \tau_k^g + \sum_{l=1}^{L} \lambda_k^g \Theta_l \,. \quad (1)$$

Here, $E(y_k \mid \Theta, g)$ represents the expected score on item *k* given the latent variable Θ and group *g*; τ_k^g denotes the intercept for item *k* in group *g* while λ_k^g stands for factor loading of item *k* in group *g* on the latent variable Θ_l. When these factor loadings are equal across groups ($\lambda_k^1 = \lambda_k^2 = ... = \lambda_k^G$) metric equivalence is achieved. In order to establish scalar equivalence, however, it is necessary that both factor loadings and item intercepts are equal across groups ($\lambda_k^1 = \lambda_k^2 = ... = \lambda_k^G$ and $\tau_k^1 = \tau_k^2 = ... = \tau_k^G$). The more parameters we can restrict to be equal across groups, the more equivalent results are across groups (e.g. cultures). Hence, the method involves selecting the most parsimonious model, with as many equality restrictions as possible, without harming the fit of the data.

The fact that (a) models assuming continuous latent and observed variables are prevalent; (b) that the procedure for MCFA is relatively straightforward and well elaborated; and (c) that software is readily available – i.e. LISREL (Jöreskog & Sörbom, 1996), AMOS (Arbuckle, 2003), EQS (Bentler, 1995), MPlus (Muthén & Muthén, 1998) – has contributed to making MCFA the method of choice for most researches in wide variety of disciplines (Vandenberg & Lance, 2000; Steenkamp & Baumgartner, 1998).

However, in spite of its popularity this approach still entails some issues that need to be resolved. Most important is the fact that most rating scales define ordinal level data, something the method does not fully account for (Lubke & Muthén, 2004; Meade & Lautenschlager, 2004). Researchers who do not like to accommodate on this issue can choose among the following approaches.

2.1.2 DIF (IRT)

A second approach to the issue of measurement equivalence has been developed within the framework of item response theory. At the heart of this approach is the analysis of item bias, i.e. differential item functioning (DIF)

across groups (Thissen, Steinberg, & Wainer, 1988; Raju et al., 2002). DIF occurs when respondents from different countries with an equal position on the latent construct (e.g. equal knowledge or equal attitude) have different scores on the instrument items. This approach deals with ordinal response variables (both dichotomous and polytomous) while assuming continuous, normal distribution of latent constructs. It posits a nonlinear, logistic relationship between the latent construct and the observed score at the item level. Hence, different from MCFA this method retains the measurement level of the observed indicators.

IRT models for dichotomous responses are most popular. However, models for polytomous items are increasingly used (Raju et al., 2002). Let's consider the following multiple group IRT model for polytomous, ordinal items (Masters, 1982) – as presented in equation 2. The log of odds of selecting category *s* of item *k* instead of category *s-1* given a person's latent trait Θ and membership of group *g* is assumed to have the following form (Bock & Zimovski, 1997):

$$\log\left[\frac{P(y_k = s \mid \Theta, g)}{P(y_k = s-1 \mid \Theta, g)}\right] = a_k^g(\Theta - b_k^g) \quad (2)$$

for $2 \leq s \leq Sk$, where *s* denotes one of the Sk categories of variable y_k. Here, a_k^g is the slope for group *g* and item *k* – which is called the "discrimination" parameter in IRT-language – and b_k^g is the "intercept" or location for group *g*, item *k*, and category *s*. In IRT the latter is defined as the "difficulty" parameter. Hence, both difficulty and/or discrimination parameters may vary across groups and cause inequivalence or "differential item functioning". When DIF is present only in the location parameters b_k^g, it is called uniform DIF. Nonuniform DIF occurs when slope parameters a_k^g differ across groups.

DIF analysis is well suited for discrete, nominal or ordinal observed variables. It is able to determine different levels of measurement equivalence and to detect various forms of response biases. It provides psychometric information at the item response level and is particular good for the analysis of individual items. The most important limitation of DIF analysis in the context of cross-cultural research is that it involves pairwise comparison and hence limits its applicability to situations in which only two groups are compared.

2.1.3 MLCA

The aforementioned approaches, MCFA and IRT, have in common that they assume a continuous distribution of the latent variable. The third approach, i.e. the multi-group latent class analysis (MLCA) is probably a more general framework since it defines latent variables as categorical. Although discretized continuous variables can be defined, most MLCA models assume either

nominal of ordinal latent variables. Conceptually this approach is analogous to the MCFA as it investigates the equivalence of relationships between a set of observed variables (either nominal or ordinal) with one or more latent constructs (Hagenaars, 1990; Moors, 2004).

Depending on whether the latent variable is assumed to be nominal or ordinal, there are two main variants of the model:

– The first is the multiple group extension of the standard latent class model (Clogg & Goodman, 1985; Hagenaars, 1990; McCutcheon, 2002), which defines the latent construct as a nominal variable divided in two or more latent classes.
– The second type of MLCA relies on the multiple group extension of the latent class factor analysis, which represents a restricted latent class model with one or more ordinal latent variables. It defines the latent constructs as discrete ordinal variables with fixed and equidistant category scores (Magidson & Vermunt, 2001; Moors, 2004; Kankaraš & Moors, 2009).

A MLCA model for ordinal indicators is presented in equation 3:

$$\log\left[\frac{P(y_k = s \mid \Theta, g)}{P(y_k = s-1 \mid \Theta, g)}\right] = \alpha_{ks}^{g} + \sum_{l=1}^{L} \beta_{kl}^{g}\,\Theta_l \quad (3)$$

Here, α_{ks}^{g} are item– and category-specific intercepts and β_{kl}^{g} item– and factor-specific slopes. As can be seen, each of these can be assumed to differ across groups. The situation in which a set of α_{ks}^{g} parameters differs across groups is sometimes referred to as a "direct effect" because such a model can also be defined by including the grouping variable (i.e. where groups represent different cultures) as a nominal predictor in the model for item *k*. Such direct effects are present when group differences in item responses cannot fully be explained by group differences in the latent factors. Group differences can also be found in the β_{kl}^{g} parameters. This is referred to as "interaction effects" as such group differences occur when the relationship between item responses and latent factors is modified by the group membership, i.e. by the interaction effect of the grouping variable and the latent factor concerned.

An important advantage of the nonparametric approach used in the MLCA is that it avoids possible biases invoked by inappropriate and unverifiable assumptions about the distribution of the latent variable(s) (Vermunt, 2005; Vermunt & Van Dijk, 2001). A second advantage is that MLCA, contrary to the other approaches, does not require that at least two items in a scale need to be equivalent for identification purposes. A (minor) disadvantage, however, is that

in the MLCA approach the measurement of the construct may become more elusive since factor scores are represented on a logarithmic scale.

2.2 Similarities and differences between approaches to the analysis of measurement equivalence

In explaining the three approaches we already pointed to some similarities as well as differences between them. We should bear in mind that they come from different methodological fields and have a somewhat different focus, that they make different assumptions regarding the nature of values, and label parameters differently. This has, consequently, led to rather isolated practices of measurement equivalence research which was usually constrained to the specific terminology and methods characteristic for a given methodological framework.

However, apart from their apparent differences, the three approaches share many common elements that are often overlooked, from the theoretical assumptions about measurement models to the model parameters and measurement procedures employed (Kankaraš et al., in press). These similarities stream from the fact that all three approaches are latent variable models, i.e. they all model relationships between one or more unobservable, latent variables representing measured construct of interest with a set of observed measures, or instrument items. Thus, although it may be labeled differently, all three approaches have two main kinds of parameters, i.e. slope and intercept parameters.

The slope parameters indicate the strength of the effect of latent variable on observed variables, for each group (McDonald, 1999; Magidson & Vermunt, 2004). For the three methods they are conceptually similar. The intercept parameters, on the other hand, have only a similar interpretation in the IRT and MLCA approach, but are not directly comparable with those in the MCFA approach (Meade & Lautenschlager, 2004). This is because of the different treatment of the observed variables (continuous vs. discrete); the IRT and MLCA have as many intercept parameters as there are answer categories in observed variables while MCFA models have only one intercept parameter per item. Due to this difference the IRT and MLCA are better able to detect various forms of response biases than the MCFA approach. For example, it has been shown that when extreme response bias causes inequivalence in intercepts, the MCFA approach is not able to correctly identify inequivalent parameters (Kankaraš et al, in press).

Table 1 summarizes the comparison of relevant characteristics of the three approaches to measurement equivalence. It presents the assumptions related to the latent and the response variables, along with the conceptual similarities between the model parameters – intercepts and slopes – as well as between the two most important forms of inequivalence in these parameters.

Table 1 Characteristics of the CFA, IRT and MLCA models for measurement equivalence

	MCFA	IRT	MLCA
A. Model assumptions			
Distribution of latent variable	Continuous Normal	Continuous Normal	Discrete Multinomial
Distribution of response variables – $f(y_k\|\Theta)$	Continuous Normal	Discrete Multinomial	Discrete Multinomial
Regression model for response variables	Linear	Logit	Logit
B. Model parameters			
Intercept parameter	Item intercept	Function of difficulty parameter	Intercept
Slope parameter	Factor loading	Discrimination parameter	Beta loading
Inequivalence in intercepts	Scalar inequivalence	Uniform DIF	Direct effect
Inequivalence in slopes	Metric inequivalence	Non-uniform DIF	Interaction effect

2.3 Procedures in analyzing measurement equivalence

In all three approaches the analysis of measurement equivalence is based on the comparison of models with various degrees of inequivalence. In particular, models that allow more (intercept or slope) parameters to vary freely across groups (and thus to be inequivalent) are compared, in terms of their fit to the data, with more restricted models that constrains these parameters to be equal across groups, thus assuming their equivalence. The question now is what procedure to use in comparing models? The answer to this question is somewhat different in each approach.

In MCFA, a researcher starts from the unrestricted (heterogeneous) model in which all parameters are group specific and then compare it with the more restricted, nested models in a number of consecutive steps (Vandenberg & Lance, 2000; Steenkamp & Baumgartner, 1998). Models are first compared on a scale level, starting with the model with equal slope parameters in all items (metric equivalence), and followed by the model with equal slope and intercept parameters in all items (scalar equivalence). If inequivalence is found on a scale level, e.g. if the model with equal slope parameters fits worse than the initial, unrestricted model, a researcher can proceed with item-level analysis in search of partially equivalent models. Model fit statistics play an important role in

the procedure. The most commonly used model comparison test in MCFA is the chi-square difference test, which is in fact a likelihood-ratio (LR) test of nested models. Other popular fit indices are measures such as Root Mean Square Error of Approximation (RMSEA), Comparative Fit Index (CFI), and Akaike Information Criterion (AIC).

Similar procedures for model comparison using LR chi-square tests are used in the IRT based DIF analysis. However, differently form MCFA, DIF analysis starts from the most restricted, equivalent measurement model, which is then compared with models in which the parameters in a single item are allowed to vary freely across groups (Thissen et al., 1988; Meade & Lautenschlager, 2004). As for MCFA, the minimal requirement is that parameters of at least one item should be invariant across groups (Mead & Lautenschlager, 2004; Steenkamp & Baumgartner, 1998).

Researchers that use the third MLCA approach often ally with the strategy developed in the context of IRT in that they start with the restricted (homogeneous) model and compare this with models with increasing number of direct and interaction effects included. The MLCA typically relies on information criteria such as AIC, BIC and AIC3 that evaluate models both in terms of their fit and their parsimony, as well as on LR tests (Moors, 2004; Kankaraš & Moors, 2009).

CONCLUSION

The history of cross-cultural research has shown numerous examples of bold generalizations about differences between cultural groups that were based on flawed and sometimes even ill-intentioned measurement practices. In presented work we have argued that in cross-cultural studies it is not enough to address only the methodological challenges of monocultural research; a researcher also needs to investigate additional requirements for valid comparison of results across cultures that are specific for this type of research. Because even if a "perfect" translation of the original instrument into the language of other cultures is possible, it still does not ensure that each culture interprets questions in the same way. In fact, it can be expected that the validity of these instruments is increasingly questioned as cultures in which they are applied are more adrift from the original one (Hui & Triandis, 1985).

In this work we presented three quantitative methodological approaches that can be used to investigate the issue of measurement equivalence, namely the MCFA, designed for continuous latent and outcome variables, the LCFA for categorical latent and outcome variables, and DIF for continuous latent and categorical outcome variables. Although these three approaches come from different methodological realms and have a somewhat different focus, they share many of the fundamental assumptions and are conducted in a rather

similar manner. However, each one of them has comparative advantages and disadvantages compared to the other two, which prompts careful consideration on the part of a researcher in order to match particular research situations with the best suited method.

Cross-cultural studies enable us to investigate universality of psychological laws and cultural variations in people's characteristics, opinions and behaviors. As our world is increasingly becoming a "global village" it could be expected that the number and scope of cross-cultural comparisons rises as well. It is our hope that this increase will be paralleled with an increase in methodological vigor and quality of data analysis that will lead to more valid and trustworthy results.

REFERENCES

Arbuckle, J. L. (2003). *AMOS 5.0 update to the AMOS User's Guide*. Chicago, IL: Smallwaters.

Bentler, P.M. (1995). *EQS Structural Equations Program Manual.* Encino, CA: Multivariate Software, Inc.

Bock, R.D., & Zimovski, M.F. (1997). Multiple Group IRT. In W.J. van der Linden and R.K. Hambleton (Eds.), *Handbook of Modern Item Response Theory*. Springer: New-York.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the Equivalence of Factor Covariance and Mean Structures: The Issue of Partial Measurement Invariance. *Psychological Bulletin*, *105*(3), 456–466.

Cheung, G. W., & Rensvold, R. B. (2000). Assessing Extreme and Acquiescence Response Sets in Cross-Cultural Research using Structural Equations Modeling. *Journal of Cross-Cultural Psychology*, *31*, 187–212.

Clogg, C.C., & Goodman, L.A. (1985). Simultaneous Latent Structure Analysis in Several Groups. *Sociological Methodology, 15*, 81–110.

Church, A. T., & Lonner, W. J. (1998). The Cross-Cultural Perspective in the Study of Personality. *Journal of Cross-Cultural Psychology*, *29*(1), 32–62.

Fahrmeier E.D. (1975). The Effect of School Attendance on Intellectual Development in Northern Nigeria. *Child Development, 46*, 281–285.

Goddard, H. H. (1917). Mental tests and the immigrant. *Journal of Delinquency*, *2*, 243–277.

Gould, S. J. (1981). *The Mismeasure of Man*. New York: W.W. Norton & Co.

Greenleaf, E. A. (1992). Measuring Extreme Response Style. *Public Opinion Quarterly*, *56*, 323–351.

Hagenaars, J. A. (1990). *Categorical Longitudinal Data–Loglinear Analysis of Panel, Trend and Cohort Data.* Newbury Park, CA: Sage.

Horn, J. L., & McArdle, J. J. (1992). A Practical and Theoretical Guide to Measurement Invariance in Aging Research. *Experimental Aging Research, 18,* 117–144.

Hothersall, D. (1995). *History of Psychology* (3rd Ed.). McGraw Hill.

Hui, C.H., & Triandis, H.C. (1985). Measurement in Cross-Cultural Psychology: A Review and Comparison of Strategies. *Journal of Cross-cultural Psychology*, *16*(2), 131–152.

Hui, C.H. & Triandis, H.C. (1989). Effects of Culture and Response Format on Extreme Response Style. *Journal of Cross-Cultural Psychology*, *20*(3), 296–309.

Johnson, T. (1998). Approaches to Equivalence in Cross-Cultural and Cross-National Surveys. *ZUMA Nachrichten Spezial: Cross-Cultural Survey Equivalence*, *3,* 1–40.

Jöreskog, K.G. (1971). Simultaneous Factor Analysis in Several Populations. *Psychometrika, 6*, 409–426.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's Guide*. Chicago: Scientific Software.

Kankaraš, M., & Moors, G. (2009). Measurement Equivalence in Solidarity Attitudes in Europe. Insights from a Multiple Group Latent Class Factor Approach. *International Sociology*, *24*(4), 557–579.

Kankaraš, M., Moors, G., & Vermunt, J.K. (in press). Testing for Measurement Invariance with Latent Class Analysis. In E. Davidov, P. Schmidt and J. Billiet (Eds.), *Methods and Applications in Cross-Cultural Analysis*, Lawrence Erlbaum.

Lubke, G.H., & Muthén, B.O. (2004). Applying Multigroup Confirmatory Factor Models for Continuous Outcomes to Likert Scale Data Complicates Meaningful Group Comparisons, *Structural Equation Modeling*, *11* (4), 514–34.

Lynn R. (2006). *Race Differences in Intelligence: An Evolutionary Analysis*, Washington. Summit Books, Augusta, GA.

Magidson, J. & Vermunt, J.K. (2001). Latent Class Factor and Cluster Models, Bi-Plots and Related Graphical Displays. *Sociological Methodology,* 31, 223–264.

Magidson, J., & Vermunt, J.K. (2004). Latent Class Models, In D. Kaplan (Ed.), *Handbook of Quantitative Methods in Social Science Research*, Sage Publications, Newbury Park, CA.

Masters, G. N. (1982). A Rasch Model for Partial Credit Scoring. *Psychometrik, 47*, 149–174.

McCutcheon, A. (2002). Basic Concepts and Procedures in Singe– and Multiple-Group Latent Class Analysis. In J. Hagenaars and A. McCutcheon (Eds.), *Applied latent class analysis*, Cambridge University Press.

McDonald, R. P. (1999). *Test Theory: A Unified Treatment*. Mahwah, NJ: Erlbaum.

Meade, A.W., & Lautenschlager, G.J. (2004). A Comparison of Item Response Theory and Confirmatory Factor Analytic Methodologies for Establishing Measurement Equivalence/ Invariance. *Organizational Research Methods*, *7* (10), 361–88.

Meredith,W. (1993). Measurement Invariance, Factor Analysis and Factorial Invariance, *Psychometrika*, *58*, 525–543.

Moors, G. (2004). Facts and Artefacts in the Comparison of Attitudes among Ethnic Minorities. A Multi-Group Latent Class Structure Model with Adjustment for Response Style Behaviour. *European Sociological Review, 20*, 303–320.

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

Paulhus, D. L. (1991). Measures of Personality and Social Psychological Attitudes. In Robinson, J.P., and Shaver, R.P. (Eds.), *Measures of Social Psychological Attitudes Series* (Vol. 1, 17–59). San Diego: Academic.

Raju, N.S., Laffitte, L.J., & Byrne, B.M. (2002). Measurement Equivalence: A Comparison of Methods Based on Confirmatory Factor Analysis and Item Response Theory. *Journal of Applied Psychology*, *87* (3), 517–29.

Singh, J. (1995). Measurement Issues in Cross-Cultural Research. *Journal of International Business Studies*, *26*(3), 597–619.

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

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of Item Response Theory in the Study of Group Differences in Trace Lines. Pp. In H. Wainer and H. Braun (Eds.), *Test Validity,* 147–169. Hillsdale, NJ: Erlbaum.

Van de Vijver, F., & Leung, K. (1997). *Methods and Data Analysis of Cross-Cultural Research*. Thousand Oaks: Sage.

Van de Vijver, F. (1998). Towards a Theory of Bias and Equivalence. *Zuma Nachrichten*: *Cross-Cultural Survey Equivalence*, *3,* 41–65.

Vandenberg, R.J., & Lance, C.E. (2000). A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organizational Research Methods*, *2*, 4–69.

Vermunt, J.K., & Van Dijk, L. (2001). A nonparametric random-coefficients approach: the latent class regression model. *Multilevel Modelling Newsletter*, *13*, 6–13.

Vermunt, J. K. (2005). Mixed-effects logistic regression models for indirectly observed outcome variables. *Multivariate Behavioral Research*, *40*, 281–301.

Wicherts, J. M. (2007). Group Differences in Intelligence Test Performance. Unpublished doctoral dissertation, University of Amsterdam.