Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

ECONOMIA E STATISTICA AGROALIMENTARE

Ciclo XXV

Settore Concorsuale di afferenza: 13/D2

Settore Scientifico disciplinare: SECS – S/03

TITOLO TESI

In Defence of Modelling Simultaneity for a Correct Approximation of Cultural Aspects: Implications for Food Consumers Studies with Latent Variables

Presentata da: **dott. MARCO VASSALLO**

Coordinatore Dottorato
Prof. Roberto Fanfani

Relatore
Prof. Mario Mazzocchi

Esame finale anno 2013

IN DEFENCE OF MODELING SIMULTANEITY FOR A CORRECT APPROXIMATION OF CULTURAL ASPECTS: IMPLICATIONS FOR FOOD CONSUMER STUDIES WITH LATENT VARIABLES

by

MARCO VASSALLO

A DISSERTATION

SUBMITTED TO THE DEPARTMENT OF STATISTICS
IN TOTAL FULFILMENT OF THE REQUIREMENTS OF
THE DEGREE OF DOCTOR OF PHILOSOPHY IN
AGRI-FOOD ECONOMICS AND STATISTICS

UNIVERSITY OF BOLOGNA
Bologna, Italy

4ht July, 2013

Abstract

Dealing with latent constructs (loaded by reflective and congeneric measures) cross-culturally compared means studying how these unobserved variables vary, and/or covary each other, after controlling for possibly disturbing cultural forces. This yields to the so-called ‘measurement invariance’ matter that refers to the extent to which data collected by the same multi-item measurement instrument (i.e., self-reported questionnaire of items underlying common latent constructs) are comparable across different cultural environments. As a matter of fact, it would be unthinkable exploring latent variables heterogeneity (e.g., latent means; latent levels of deviations from the means (i.e., latent variances), latent levels of shared variation from the respective means (i.e., latent covariances), levels of magnitude of structural path coefficients with regard to causal relations among latent variables) across different populations without controlling for cultural bias in the underlying measures. Furthermore, it would be unrealistic to assess this latter correction without using a framework that is able to take into account all these potential cultural biases across populations simultaneously. Since the real world ‘acts’ in a simultaneous way as well. As a consequence, I, as researcher, may want to control for cultural forces hypothesizing they are all acting at the same time throughout groups of comparison and therefore examining if they are inflating or suppressing my new estimations with hierarchical nested constraints on the original estimated parameters. Multi Sample Structural Equation Modeling-based Confirmatory Factor Analysis (MS-SEM-based CFA) still represents a dominant and flexible statistical framework to work out this potential cultural bias in a simultaneous way. With this dissertation I wanted to make an attempt to introduce new viewpoints on measurement invariance handled under covariance-based SEM framework by means of a consumer behavior modeling application on functional food choices.

Publications

The PhD period from 2010 to 2012 partially overlapped with the second phase of dissemination activity concerning the Health-Grain European project (2005-2010) I was involved in and from which the data were kingly approved for being used in this thesis. As a consequence, a number of peer-reviewed publications, which I authored and co-authored, come out prior to and during the writing of this dissertation. However, only just one of them was particularly focused on the work presented here and it has been reported here:

Dean M., Lampila P., Shepherd R., Arvola A., Saba A., Vassallo M., Claupein E., Winkelmann M., Lähteenmäki L. (2012). Perceived relevance and foods with health-related claims. *Food Quality and Preference*, 24, 129-135.

Contents

List of figures

List of tables

.

Acknowledgements

First and foremost let me thank my boss and senior researcher at INRAN (Istituto Nazionale di Ricerca per gli Alimenti e la Nutrizione – National Research Institute on Food and Nutrition. At the present time, CRA – Centro per la Ricerca e la Sperimentazione in Agricoltura – Agricultural Research Council) Anna Saba. She was the first person who put me in contact with latent world and consumer behavior modeling from a psychosocial perspective. Without her I would never have discovered this so fascinating field of application and anything further would have been possible.

My supervisor professor Mario Mazzocchi who was the first person you believed in me and encouraged me to make my Ph.D. experience conceivable. I remember our first talk in a small bar in Rome during a break at a project meeting. He found my enthusiasm very catching and therefore suggested me to make an attempt in reaching this important academic goal.

Professor and chair Gregory R. Hancock who accepted my visiting scholar application for my Ph.D. program at his Department of Measurement, Statistics & Evaluation (EDMS) in the College of Education at the University of Maryland (USA) from 1 March 2011 through 30 June 2011. He literally was a 'guru' for me who taught me, and clarified, many SEM concepts, applications and beyond keeping up my enthusiasm and curiosity all the time.

Agri-Food Economics and Statistics Ph.D. coordinator prof. Roberto Fanfani and all committee members for having selected me as potential Ph.D. student. I hope I have met their expectations.

All my Ph.D. colleagues at University of Bologna with special thanks to Gianluca Parodi and Pier Maurice Reverberi who welcomed and introduced me in Bologna during my 4-mouth stay and always getting me involved in fantastic and funny going out.

Last but not least, let me thank a really special person who 'crashed me down' in my last three months of this thesis like a dance-rave hurricane. Without her I would never have completed this dissertation in time.

To all these people I am really grateful.

Dedication

There are only two people I may want to dedicate this effort of mine. They are my parents. My father Paolo and my mother Anna for loving me and supporting me especially when I was unlovable and unbearable being always there. They made me understand the importance of education and taught me the mentality to bring to an end whatsoever you begin with the best you can do. I am not able to thank you enough. I love you so much.

Preface

Readers of this thesis are assumed to be familiar with regression analysis, common factor analysis, measured and latent path analysis foundations at basic level, at least.

Chapter 1

Introduction

Let me begin by openly telling you the reason that made me tick to write this dissertation down. It comes from a sentence of Michael Pollan's about what he calls the American Paradox: "the more we worry about nutrition, the less healthy we seem to become". This sentence has been mostly paraphrased in the Pollan's book "In defense of food: an eater's manifesto" (2008) that has inspired me the title of this dissertation. He was talking about nutrition and healthy from a consumer perspective to spend time in selecting and preparing good food, but we can easily grab that paradox and thinking alike statisticians interested in studying psychological constructs across cultures and, as a result, substituting in that sentence the word 'nutrition' with 'cross-cultural-constructs' and the word 'healthy' with 'accurate'. If we try to do that we will come out with a new paradox regarding cross-cultural-constructs: "the more we (as researchers) worry about cross-cultural-constructs, the less accurate we seem to become". It looks like a sound paradox as it seems even more reasonable that we ought to be even more accurate if we want compare constructs across multi-cultural populations, since the same term 'cultural' reveals possible vast diversities. In contrast, it is even more common to come across research papers in which group-comparisons, or even pooling data, concerning measures that underlie psychological constructs have been made without controlling for cultural forces.

So then now, let me highlight two important keywords for this dissertation: 1) constructs and 2) multi-cultural population.

What is a construct? A construct is a psychological concept that has a latent nature and thus it may be conceptualized like a latent variable or latent construct. Everybody knows what 'latent variable' stands for, possibly having heard out such a customary 'singsong-like' definition: 'latent variable is a concept, construct that cannot be directly measured and needs of a stimuli (items weighed with measurement scales) in order to be quantified somehow'. But, let me boldly add something more here. A latent construct is that what outwardly surrounds us, and has been, in turn, quantified inside us. This is due to the fact that latent constructs like, for instance, 'attitudes', 'intentions', 'values', 'moods', 'beliefs', and so forth, have been made of interactions between ourselves and the real world or, better, between that variety of information each of us has got inside, in terms of culture of any kind, and the variety of information the world outside makes us known. These marvelous 'invisible' cultural interactions allow latent concept to be quantified inside us and this sort of quantification will be revealed throughout questions (items) that stimulate the constructs to come out. As a consequence, it is noteworthy how the word ‘culture’ is still strongly about since we answer to those construct-based questions basing on our own cultures and it is fair; even though these cultural forces might seriously influence our construct-based answers, make them biased and culturally-oriented, with the result of confounding the real meaning of the construct itself.

What do I mean with multi-cultural population?

A multi-cultural population is a group of people who differ for one (or many) characteristic(s) whichever nature the characteristic(s) is (are): physical or not physical. Only just one aspect from the most simple (like gender, social status or different language) to the most complex (like religion) makes one population dissimilar from another in respect to construct(s) of interest. But, independently from how complex the characteristic may be I, as researcher, have to be sure that it is related to that construct, and at the same time is not too much influencing that construct itself across populations. This latter may sound a bit awkward and thinking about gender, for instance, we might have doubts/queries like the following: which kind of typical forces belonging to male population and which one to female population have to be controlled for in making construct comparisons? What are the typical forces related to the constructs and those what are not?

It seems that there is no way out to this problem since typical aspects (not merely physical) of male and female population might be almost unlimited with all those nuances to which only God can give right and exhaustive answers. You might figure out if we wanted to compare a construct between two populations who differ in different, even controversial, religions. Here the so-called typical aspects might be really unlimited that would require the assistance of the two Divinities!!

In this respect, what we truly want to work out is not to discover what kind of cultural forces precisely are, since it would be impossible and useless, but if these forces are acting, or not, during the comparison of that construct across populations. And if these cultural aspects are really acting how much they are swaying the construct(s) object of comparison.

Granted that, how can we possibly collect or better quantify these cultural forces associated with common construct(s) of interest? And in case we are able to collect them, how an we control them for?

So then, to partially answer to those questions let me introduce the third keyword of this dissertation: 3) self-reported instruments, like questionnaires, since they are the most used way to collect those aforementioned construct-based answers from construct-based questions able to motivate us in bringing out constructs. These questions and answers have to be as much culturally invariant (i.e., invariant from both cultural forces related and not related to the construct object of the study) as possible when making subsequent multi-cultural comparisons across populations both at measurement (i.e., among observed measures) and at latent level (i.e, among constructs).

Hence, let me explain what happens with these self-reported questionnaires. The more a researcher deals with measuring, and so that quantifying, constructs of a multi-cultural world, and in turn studying differences among these quantified constructs across populations, the more he or she deals with self-reported instruments (i.e., multi-item questionnaires) that might be seriously fallible as they are unable to *weigh* all those aforementioned cultural differences the researcher wants to investigate (Gregorich, 2006). It is due to the fact that, despite of the best translation and back-translation a researcher may have in his/her own hands, despite of the best latest theory a researcher is able to set up, and despite of the best selection of common response-item-scales he/she is able to propose, self-reported instruments are affected by cultural forces that in no way can be extracted out and in no way remain constant, but instead they constantly evolve and change across time and situation.

As long as these cultural forces act, cross-cultural comparisons among self-reported instruments cannot be made and further discussions on cross-cultural differences among constructs, or even single measures, will inevitably lead to fully misleading conclusions although they seem consistent.

Cultural forces are particularly strong with self-instruments associated with latent constructs. That is because, as I already mentioned, a construct like an 'attitude towards something', for instance, involves in the respondent many inner statuses due to many cultural aspects related to his/her personal background when he/she tries to quantify that attitude through a score on an item-measurement-scale. Hence, the problem lies in *how much* or *how less* these cultural aspects contaminates the final response-scores and so that being considered respectively *non invariant* or *invariant* across different populations in which the self-reported instrument has been applied for.

Essentially and practically, different people from different cultures, languages, beliefs, races, religions, politics, or even different people from different groups within the same population, may not comprehend the meaning of multi-response-items, that assess common constructs, or the meaning of a common construct itself in the same way, but possibly in different ways because they belong to groups that may be culturally different. Although, this is merely the beginning of the story about cross-cultural invariance.

Thus now, before moving on with this story I am bound to stop here for a while and explain the main purposes of this dissertation I may want to define it as 'my personal challenge' in the covariance-based structural equation modeling (CB-SEM) field of application to food choices with latent variables cross-culturally compared. By doing so, even though this dissertation inevitably describes how to deal with cross-cultural measurement invariance at latent level from a statistical point of view, it will not be too computational-led, but conversely focused on keeping up reader's intuition and curiosity, hopefully. To this end, I may want to make an attempt to simplify foundations, assumptions as regards this topic with the challenge of 'pulling out the essence' as much plain and applied as possible in order to make this dissertation comprehensible, and with a bit of luck, useful to the widest audience possible. I do not know whether, or not, I am able to successfully reach this goal, but let me be a little bit bold in chasing it. Should I fail, the reader might get stuck with some good reference, at least ☺

By the way, let me apology in advance with all those proficient methodologists, who are possibly reading this dissertation, for bothering them with redundancies of well-known concepts. But, on the other hand, let me encourage the same methodologists to critically review my efforts in the hope that the aforementioned curiosity will be, even for them, stronger than their expected annoyance.

This dissertation is divided into five chapters: chapter 1, you are currently reading, tries to turn on reader's curiosity with providing some critical points on how to deal with cultural aspects that involve latent constructs conceptualized as reflective of the observed reality (i.e., observed measures); the second chapter draws attention to methodological backgrounds and anchors in coping with simultaneous approach to detect cultural aspects at latent level; chapter 3 is on some technical details and advice in applying multi-group structural equation modeling carried out from the literature and my personal experience; chapter 4 discusses a cross-cultural application to food choice providing results and implications; the last chapter 5 tries to gather all the things up with offering possibly suggestions.

1.1 Why simultaneity?

During my research studies on consumer decision making process throughout psychosocial models with latent variables I have been always fascinated by the Covariance-Based Structural Equation Modeling (CB-SEM) technique capacity of controlling for all relations in a simultaneous way, similarly to a whole picture enabling to depict what is happening at that precise instant. The more you are able to control for, the more you can understand what has happened and possibly influenced your research dynamics. This is particularly true when consumers make decisions to do something since tons of psychological motivations are simultaneously producing invisible effects before consumers make actual facts. All that is even more true when these psychological dynamics move across different cultural groups and so that the simultaneity is not only at model level but also at cultural level. As a consequence what it is desirable is looking for a technique that is able to model the hypothesized factorial structure and, at the same time, being able to test for cultural invariance of that structure itself detecting what dynamic has been culturally affected from. As a matter of fact, cross-cultural studies on latent constructs may concern different types of dynamics, also longitudinal, through comparisons among latent statistical moments, latent interrelations and/or structural relations among latent constructs themselves in terms of structural path coefficients. In all these situations of group-comparison, having a simultaneous way of estimation should be preferred, when possible, as it is the best way (as I am going to defend throughout this thesis) both to detect presence of invariance and making structural estimates (i.e., latent variances, covariances, correlations, un-standardized and standardized path coefficients) comparisons defensible.

As I am going to argue in the next chapters 2 and 3 simultaneous way of estimation works alike a hierarchical process of nested constraints on the un-standardized estimates across groups. The rationale of this constraints-chain process is to verify if the latent structure of interest may vary across groups/cultures, in terms of estimations, at each constraint-step simultaneously and not singularly in each group. This latter means that whenever each constraint is made at a time (i.e., a parameter is constrained to be equal across groups) the entire structure may entirely change both within and between groups at the same exact time the estimations are being provided. If I am able to do that I will control for possibly biases occurring in my latent structure due to different cultural aspects peculiar of each group-comparison. More technically speaking, if the final estimation process with constraints is not too much worse than the one without constraints in terms of estimates magnitude, their significant values, fit indices I may robustly claim that the latent structure of interest is cross-culturally invariant. This process of sequential constraints, fit indices and so forth will be described in chapter 3.

Presently, and from a statistical point of view, it is noteworthy that we are stepping into the field of common factor analysis and the Multi Sample Structural Equation Modeling-based Confirmatory Factor Analysis (MS-SEM-based CFA) that still represents the best framework to assess cross-cultural measurement invariance for true latent constructs with reflective/effect and congeneric measures[1] in a simultaneous way. Foundation of common factor-based analysis will be handle in chapter 2. Moreover, I may want to advise the reader that it will be out of the purpose of this dissertation talking about measurement invariance with regard to other kind of pseudo latent variables such as composite/emergent factors with formative/cause indicators[2] even though they can be analyzed using CB-SEM approach. The main reason lies in the fact that I personally share the opinion that composite factors are not

[1] Indicators as effects of the latent are termed reflective because they represent reflections, representations, manifestations of a construct. They are congeneric if they load on only one common factor (Jöreskog, 1971; Brown, 2006).

[2] Indicators as causes or formative of a composite factor as they form, induce, define characteristics of the construct itself and "omitting an indicator is omitting part of the construct" (Bollen & Lennox, 1991).

properly constructs with a true latent nature (although they are also unfairly associated with the word "latent factors"), but, on the contrary, something of "built-up" by researchers in order to summarize the total variance of a "bunch" of measures and, as a consequence, they are theoretical constructs using weighted composites of observed variables (Rigdon, 2013) as Principal Component Analysis (PCA) analysis does. Let me stop here with this latter provocation and with a nice definition of these principal component/composite/emergent or even formative constructs given by Cameron McIntosh in the SEMNET[3]: "I would disagree that formative constructs or principal components are latent variables. Synthetic, yes, but remember that we are 'building' them rather than 'tapping into' them. True latent variables have an existence independent of the observables and span a greater space. Components are simply a translation of exactly the same observed information". However, let me confess that it would be intriguing and challenging to discuss about possibly measurement invariance with models including both latent and composite factors together and therefore talking about *how well or how bad* covariance-based SEM and/or component-based SEM (i.e., Partial Least Square Path Modeling – PLS-PM) can deal with these two different approaches to measure theoretical constructs, but for the time being I may just provide a couple of recent good references in the case the reader cannot help waiting it out and want to put a new dissertation up before I presumably do ☺:

[3] Structural Equation Modeling Discussion Network
http://www2.gsu.edu/~mkteer/semnet.html

- Bollen, K.A. and Davis W.R., (2009). Causal Indicator Models: Identification, Estimation, and Testing. *Structural Equation Modeling*, 16, 498-522.
- Kline R.B., (2013). Reverse Arrow Dynamics: Feedback Loops and Formative Measurement. In G.R. Hancock & R.O. Mueller (Eds.), Structural Equation Modeling: a second course (pp. 41-79). Second Edition. Greenwich, CT: Information Age Publishing.
- Lee L., Petter S., Fayard D., Robinson S., (2011). On the use of partial least squares path modelling in accounting research. *International Journal of Accounting Information Systems*, 12, pp. 305-328.
- Rigdon E.E., (2013). Partial Least Squares Path Modeling. In G.R. Hancock & R.O. Mueller (Eds.), Structural Equation Modeling: a second course (pp. 81-116). Second Edition. Greenwich, CT: Information Age Publishing.
- Vinzi V.E., Chin W.W., Henseler J., Wang H. (Eds 2010). Handbook of Partial Least Squares: Concepts, Methods and Applications. Springer.

Chapter 2

Background: concepts and anchors

The main purpose of this chapter is to focalize as much clearly as possible what the study of the constructs comparisons does mean across populations in a simultaneous way. By doing so, it is chiefly interest of mine providing to reader a background on latent variable measurement invariance, as much practical and clear as possible. A background that cannot be too strictly computational (although there are some essential formulas from matrix algebra), as this dissertation does not want to be psychometrical-oriented, but even more practical-focused, above and beyond the fact that all the methodology has been well-explained and reported already by excellences in this field (e.g., Bentler, Bollen, Brown, Byrne, Hancock, Kaplan, Kline, Muthén and Muthén, Rigdon , and many others).

As I aforementioned in the preface I am assuming that readers are familiar with some basilar statistical concepts of regression models, factor analysis and path analysis so as to better follow this development on latent constructs measurement invariance.

2.1 – What does the invariance of a latent construct mean?

When we talk about latent constructs everybody knows that we are referring to concepts that, although cannot be directly measured, are around us and/or inside us such as: attitudes, values, beliefs… and so forth. These concepts play a role, make us decisions and so that they may have similar or different meanings from one person to another. If they have a similar meaning (across people/group) they are *invariant*, if they have a different meaning they are obviously *non-invariant*. From this perspective everything seems to be deadly easy, but since latent factors cannot be directly measured they need to be quantified/measured through observed measures carried out from self-reported measurement instruments such as questionnaires. So, granted that, the issue of invariance now seems turning into something even more complicated as we have introduced a new obstacle: self-reported measurement instrument. As a consequence, the question spontaneously raises up: which of the two 'guys' have to be invariant? The self-reported measurement instrument, the latent construct or both? The answer is not merely both, but it depends on what kind of invariance we want to detect and assume.

Hence, now the matter gets more and more tricky as it seems that there are two types of invariance across populations: one with regard to observed variables and another one to the latent variables.

Although these two aspects are all the time associated with the two words 'measurement equivalence' only the one concerning the observed variables is in truly associated with a proper analysis of measurement invariance as it refers to the measurement instrument used and so that to observed measures (items scores). The other aspect of invariance concerning latent variables is NOT a measurement testing indeed, but rather a study of similarities or dissimilarities (i.e., heterogeneity) of these error-free or true score variables (i.e., latent factors) of interest in terms of: a) latent dispersions (latent variances); b) latent covariances (latent interrelationships in presence of more than one factor across populations); c) latent levels (latent means). It is actually intuitive just from now that since latent factors are measured by a set of underneath observed measures any potential test of measurement invariance (1), as the same word 'measurement' is telling us, involves the observed measures and only just them. Besides, it is even more intuitive that once the assessment of the invariance of the measures is assumed, I can proceed with exploring latent variables heterogeneity across populations (2), otherwise not, and I cannot even proceed with the aspect (2) before having assessed the aspect (1) for the logical reason of measurement step I aforesaid. To this end, and getting back to the initial question (title of this subchapter) I may want to claim that invariance of a concept across populations is indeed a test on how this latent factor is statistically heterogeneous across groups of comparison under the assumption of measurement invariance of the observed variables in measuring that latent concept. It practically means that if I want to test if an 'attitude towards something' may differ, or not differ, across populations I have to test if the measures underneath that attitude are statistically invariant beforehand. If this latter is the case, I may want to proceed with cross-groups comparisons, and/or test of invariance, on all those statistical moments (i.e., means, variances and covariances) at latent level. Fortunately, this view has been conveyed to the literature from authors like Vandenberg and Lance who stated (2000; p.18), citing also Anderson and Gerbing's work (1988) the following brilliant words: "…we argue that tests of measurement invariance (associations of observed scores to the latent variable or variables) should precede tests of structural invariance (associations of latent variables with each other). Our logic is based on Anderson and Gerbing's (1988) argument that one needs to understand what one is measuring before testing associations among what is measured".

2.2 – The study of measurement and latent heterogeneity invariance across populations

Unfortunately, this way of conceptualizing those two up-titled aspects of the so-called 'measurement equivalence' analysis for latent variables has not been applied by all researchers and a tough conviction of mine is that not having this distinction clearly in mind is a reason why the issue of measurement equivalence seems so difficult to afford to, when it is not at all. In this respect, I am still struggling with myself why (as it happens in many times of our life and this is one of that) we like to complicate our living by ourselves. The only reasonable answer I found so far is that a latent ☺ component of pure masochism lives inside us and it is ready to bring out when we think that things are getting along too much well.

Backing to us and searching around the vast literature about measurement equivalence I found only just an author (without diminishing any other authors' contributions to this area of research) who clearly defined what measurement invariance steps and what population heterogeneity study at latent level respectively stand for. He is Timothy A. Brown, professor in the Department of Psychology at the University of Boston. In his respect, he precisely states on page 266: "The measurement model pertains to the measurement characteristics of the indicators (observed measures) and thus consists of the factor loadings, intercepts and residual variances. Hence, the evaluation of equivalence across groups of these parameters reflects tests of *measurement invariance*. The structural parameters of the CFA model involve evaluation of the latent variables themselves, and thus consist of the factor variances, covariances, and latent means. … Thus, the examination of the group concordance of structural parameters can be considered tests of *population heterogeneity*;

that is, do the dispersion, interrelationship, and levels of the latent factors vary across groups?" (Brown, 2006).

So now, please let me make an attempt to integrate what Brown brightly started to say in his book with adding a more complete sentence with regard to the issue of measurement equivalence for latent variables across populations, that may turn into *'the study of measurement and latent heterogeneity invariance across populations'* I already proposed as title for this subchapter. As a consequence the issue of 'measurement equivalence for latent variables' includes those two aspects that are fundamental to keep separate in our mind in order to understand the precise, mostly hierarchical, process occurring in testing equivalence of latent variables across groups. This hierarchical process will be successively described and I have trust that everything will be more and more clear with reading on.

2.3 – Omnibus test of invariance and compound symmetry

Thus now it seems that the story begins with testing how the observed measures are invariant across groups, and it is so as they are the only observed information a researcher have in his/her hands other than the hypotheses on possible latent constructs that should explain those observed relationships. Hence, and once again, let me stimulate your intuition and suggest that if all information we have in our hands is in the observed variables, this information includes, in turn, both potential latent factors and those cultural aspects I want to detect or, better, controlling for. As a consequence, we, as researchers, have to find out a way to use this information properly well to initially test whether, or not, my observed measures are culturally invariant across groups. In this respect, it is straightforward deducible that, since the measures encompass cultural aspects of each population, whether all the sources of covariation among these measures are not statistically different across groups (thus they can be attributable to the empirical finding that they came from just one population or parent population - Meredith, 1964 – taken from Jöreskog's 1971) the cultural forces are not acting and/or are so marginal that any kind of further group comparison on these measured variables can be possible and defensible.

Sources of covariation stand for variances and covariances of each observed variables within each population. On the other hand, whether all these sources of covariation are statistically different (as it often happens) the cultural forces are acting in a way or another, and the measures are culturally non-invariant and so are possibly latent traits (i.e., latent constructs). This latter means that a certain level of invariance in the observed variables exists across groups, but unfortunately I am not still able to isolate it since I am considering all the sources of covariation. Nevertheless, it seems a good starting point since I know if my data are affected or not by cultural aspects as a whole (related or not related to possibly latent factors). In the measurement invariance literature this test is termed as omnibus test of the equality of covariance matrices across groups (Vandenberg & Lance, 2000; Steenkamp & Baumgartener, 1998).

All that seems to recall the concept of sphericity and its more general form of compound symmetry that is well-known for repeated measures in panel studies. “Sphericity refers to the equality of variances of the *differences* between treatment levels. Whereas compound symmetry concerns the covariation between those treatments” (cit. from “A bluffer’s guide to Sphericity” prof. Andy Field – University of Sussex). In our case we do not have repeated measures, but potentially cultural different responses carried out from potentially different groups of respondents that we want to test whether they are culturally invariant. So that, the treatment levels here are the different populations in which each observed measure has been carried out as I may want to check if the cultural aspects are acting across groups as though they were related each other somehow. In this latter respect, it seems that I have to run a sort of Mauchly’s test (1940) in order to test if the covariance matrices are equal across groups, although here I do not have repeated measures, but just different responses from different groups collected at that same period.

So then, how can I work this matter out? The best intuitive answer I may give you is with having a method that is able to simultaneously estimate these covariance matrices Σ^i (i=1,.., n-group) at the same time and testing whether, or not, they are invariant across n groups of comparison with constraining them to be equal:

$$\Sigma^1 = \Sigma^2 = \ldots = \Sigma^n \tag{2.1}$$

Where Σ^n is a n-group pxp matrix with variances along the diagonal for each observed variable p and covariances off the diagonal for each pair of the same observed variables p:

$$\Sigma^n = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{bmatrix} \tag{2.2}$$

From this first step seems that the observed means μ_p (mean vector for each n-group) have been excluded:

$$\mu^n = [\mu_1 \quad \mu_2 \quad \cdots \quad \mu_p] \tag{2.3}$$

Furthermore, this exclusion looks logical as the using of observed variances and covariances (deviations and shared deviations from means) do not allow means to add further information. However, I will address next how this presumably logical exclusion is only temporary.

Eventually here, as first logical and still intuitive conclusion, if I am able to simultaneously know if all sources of covariation may be considered un-equal (non-invariant) across groups I am enable to detect how large is the level of measurement non-invariance across those groups I am comparing to. The more the sources of covariation are different the more the level of non-invariance in my observed variables is high.

But what about latent variables invariance or, as we have acknowledged to be, latent heterogeneity? How can we detect it? In the case that, for instance, all sources of covariation in the observed variables are invariant, possibly latent variables are still invariant? In order to answer to these questions we need to introduce theories on potential latent traits and making a step back to measurement modeling concern.

2.4 – Measurement modeling

As I aforementioned, everything seems to start from having a set of observed variables (carried out from a self-reported instrument in each population) that: a) vary and covary among them in each population; b) are function of another set of hypothesized latent variables that reflect, and so that explain, the manifest interrelation among these observed variables in terms of covariances; c) are presumably affected by cultural forces since they come from different populations and thereby needing of being tested for cultural invariance so as to proceed for a subsequent study of latent heterogeneity across populations.

Before moving on let me recall some fundamental concepts taken from the classic measurement process based on the Classical Test Theory (CTT; Lord & Novick, 1968) of true and error scores. In this theory it has been postulated that any measure x_i, even the one obtained with the most sophisticated procedures, is affected by a measurement error e_i (that is non-systematic, but normally distributed with zero mean and non-zero variance) and so that this measure is function/dependent of the true measure t_i that may be latent in nature (and thereby unknown) and the measurement error itself:

$$x_i = t_i + e_i \tag{2.4}$$

As logical computational consequence the true measure is indeed the expected value of the initial measures and is not related with the measurement error:

$$E\ (x_i) = t_i \tag{2.5}$$

$$Cov\ (t_i\ ,\ e_i) = 0 \tag{2.6}$$

Nevertheless, I need at least of two measures in order to model the measurement error and so that find a true measure from the equation (2.4). Hence, more measures I collect, more precise is the estimation of the measurement errors and more precise is the true measure I am looking for across the observed measures x_i. So then, accordingly with equation (2.4) and (2.6) I have a set of measures x_i with proper means and deviations from means (i.e., observed variances: σ_{xi}^2) that are function, and so that can be decomposed, of: a) another set of true measures with respective means and deviations from means (i.e., latent true-error free variable variances: σ_{ti}^2); b) a set of measurement errors with deviations from zero means (i.e., measurement error variances: σ_{ei}^2):

$$\sigma_{xi}^2 = \sigma_{ti}^2 + \sigma_{ei}^2 \tag{2.7}$$

$$\rho = \sigma_{ti}^2 \ / \ \sigma_{xi}^2 \tag{2.7.1}$$

Equation (2.7) reflects the famous definition of reliability[4] ρ (2.7.1) of the classic measurement process where a true value is a value free of measurement error. It means that it is a value that I do not know yet and I need of a set of observed measures to be able to partial out their measurement errors and therefore coming up to that true-still-unknown value as much precisely as possible.
Still, we have knowledge from the *common factor model* theory of Thurstone (1947), that constitutes the key of *factor analysis*, that each set of observed variables may be written, or better decomposed of, as a linear function of that part of common shared variance and that part that is unique in each observed itself. These two concepts of common shared variance and unique variance represent in truly what I tried to explain above formalized with the expression (2.7) where σ_{ti}^2 is indeed that common shared variance we need to reflect manifestation of a common latent factor (the true value we are looking for); whereas σ_{ei}^2 is indeed the unique variance, that stands for: a) the part of the observed variance we do not need to manifest the true value and b) the part of the observed variance that each observed variable does not share with the observed variances of the other observed variables and c) it represents the measurement error in finding out the true value.

Hence, combining the aforementioned classical test theory of measurement process with a typical Confirmatory Factor Analysis

[4] "Reliability is the ratio of true score's variance to the observed variable's variance" (Bollen,1989; p.208).

(CFA) model (Bollen, 1989), that it is a type[5] of common factor model where the relations between measures and factors are a priori specified, the equation (2.4) can be explicated in a system of simple linear regression equations[6] as follows:

$$x_i = \tau_i + \lambda_i \xi + \delta_i \quad (2.8)$$

where x_i is a set of observed variables (i=1,…, n), ξ is a hypothetical common latent factor, λ_i represent the factor loadings or regression slopes, τ_i the intercepts, δ_i the measurement errors. The difference between the equation (2.8) and an usual regression equation is that the independent variable is the latent factor and the criterion is constituted by multiple observed variables x_i. As a consequence, it means that the latent concept ξ is trying to explain, summarize, all those observed variables x_i and the magnitude of how the latent factor is able to do that is due to the regression slopes or factor loadings λ_i associated to each x_i, whereas the magnitude of what that was not captured by the latent factor is δ_i that represents an error in this sort of interpolation process. This error, has an expected value $E(\delta_i) = 0$ and $Cov(\xi;\delta_i) = 0$.

[5] The other type of common factor model is the most famous Explorative Factor Analysis (EFA) where the relations between measures and factors are not a priori specified. Both EFA and CFA are able to partial out common variance from unique variance, but the former assumes measurement error at random and so that it cannot be modeled whilst latter may assume measurement error at random or not and so that it can be modeled (Brown, 2006; Fabricar et al., 1999).

[6] Following and adapting from Jöreskog (1973), Keesling (1972), and Wiley (1973) notation (i.e., JKW).

Eventually, and in order to complete this interpolation process as linear, it is methodologically fundamental to consider the intercept τ_i that represents the expected value of x_i when the latent factor ξ is null. This latter definition deserves of more attention as follows.

So then, in order to find this true value ξ that, in our case, is latent in nature I need of a measure, or better a set of measures (quantitative or qualitative or count and so forth) that I may observe from a sample of respondents. These measures, as I have already stated, include also cultural characteristics of the respondents since they answer taking into account their cultures.

It means that the measures other having a metric for measuring the latent trait they should have also an origin, a location, from which they depart (for measuring the true latent trait itself). And so does the latent trait towards which we have to assign both a metric and a location as well.

In other plain and practical words the location of each observed measure basically represents its predicted value when the true value (i.e., latent construct) is not still present for the respondents. Nonetheless, even though the location is not directly related to true value it exists because the respondents give answer taking into account their culture and may play a role with implications in detecting cultural forces as I am going to explain in the next subchapter about structured means, but for the time being just keeping it in mind.

So now, backing again to the system of equations (2.8) we have to find a way to estimate the parameters τ_i , λ_i and δ_i since no observed measure is provided for the dependent latent variable ξ. So that, since the only information I have is due to the observed measures x_i I am going to use all sources of covariation of x_i as I stated in the subchapter 2.3. This leads to the main fundamental of the structural equation model as a whole both applied to measured variable and latent variables path analysis (Bollen, 1989): decomposition of observed variances and covariances into the model implied parameters

$$\Sigma = \Sigma[\theta] \qquad (2.9)$$

If a researcher is able to write the system of equations (2.9) he or she is able to identify all the necessary parameters of the model (2.8).

For making you an example of three measures x_1 , x_2 , x_3 and one latent factor ξ and following the system of equations (2.8) and the expression (2.9) we can re-write[7] the covariance matrix of the three measures as follows:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & & \\ \sigma_{21} & \sigma_2^2 & \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{bmatrix} = \begin{bmatrix} \lambda_1^2\sigma_\xi^2 + \sigma_{\delta 1}^2 & & \\ \lambda_2\lambda_1\sigma_\xi^2 & \lambda_2^2\sigma_\xi^2 + \sigma_{\delta 2}^2 & \\ \lambda_3\lambda_1\sigma_\xi^2 & \lambda_3\lambda_2\sigma_\xi^2 & \lambda_3^2\sigma_\xi^2 + \sigma_{\delta 3}^2 \end{bmatrix} = \Sigma[\theta] \quad (2.10)$$

[7] Using the variances and covariances algebra of linear composites: $y_i = \alpha_i + \beta_i x_i + e_i$; $\sigma_{yi}^2 = \beta_i^2\sigma_{xi}^2 + \sigma_{ei}^2$; $\sigma_{yiyj} = \beta_i\beta_j\sigma_{xi}^2$

To better visualize the system (2.10), along with what pieces of information (i.e., sources of observed variation and covariation) occur in estimating the unknown parameters, please consider the subsequent decomposition table 2.1 adapted from Hancock et al. (2009):

Table 2.1 – Decomposition table of structural parameters (adapted from Hancock et al., 2009).

information	decomposition	λ_1	λ_2	λ_3	σ_ξ^2	$\sigma_{\delta 1}^2$	$\sigma_{\delta 2}^2$	$\sigma_{\delta 3}^2$
σ_1^2	$\lambda_1^2\sigma_\xi^2 + \sigma_{\delta 1}^2$	√			√	√		
σ_2^2	$\lambda_2^2\sigma_\xi^2 + \sigma_{\delta 2}^2$		√		√		√	
σ_3^2	$\lambda_3^2\sigma_\xi^2 + \sigma_{\delta 3}^2$			√	√			√
σ_{21}	$\lambda_2\lambda_1\sigma_\xi^2$	√	√		√			
σ_{31}	$\lambda_3\lambda_1\sigma_\xi^2$	√		√	√			
σ_{32}	$\lambda_3\lambda_2\sigma_\xi^2$		√	√	√			

Reading the table horizontally we are aware of how many and which pieces of information we need to estimate the unknown parameters (Hancock et al., 2009). On the other hand, reading the table vertically we are aware of which decomposition expression is directly involved in the estimation of that particular parameter (Hancock et al., 2009). The checkmarks indicates the combinations.

It is noteworthy that in order to estimate the latent variance σ_ξ^2 we need of all the information available in the observed measures as expected. Furthermore, the latent variance σ_ξ^2 is also function of all the other parameters since it is involved in all the decomposition expressions whereas the other not, but unevenly.

All this should let you understand why testing for latent variances invariance across groups, or, better, studying for a latent homogeneity-heterogeneity across groups, requires of a well-defined hierarchical steps starting from an invariance testing of the observed measures as a whole and proceeding with possibly further steps of invariance of the other parameters λ_i and $\sigma^2_{\delta i}$ that respectively represents, as stated previously, the common variance and the error variance in measuring the latent factor ξ.

Let me conclude with stimulating your intuition once again. It would not make any sense testing for homogeneity of a latent construct if I did not know if the shared common variance (i.e., what I really need for measuring the latent concept) among the observed variables is invariant across groups. Still, it would not make any sense testing for differences in reliabilities of my measures (see equation (2.7.1)) if the precisions in measuring that latent concept (i.e., unique variances or measurement errors) along with the latent variances were both again cross-group invariant.

However, the new system of equations (2.10) it still not identified as we have 6 pieces of information in Σ and 7 parameters to be estimated in $\Sigma[\theta]$. This issue seems again easy to be solved out as it is again so much intuitive that since the latent factor ξ

cannot be directly measured it needs of metric and the most ideal solution is to assign the same metric of the observed variables. It practically means that one of the loadings λ_i has to be fixed to 1 and therefore that observed variable (i.e., indicator) becomes the so-called marker or reference indicator (Bollen, 1989; Brown, 2006).

But now, our intuition might make us a couple of questions: which observed variable (i.e., indicator) in the system (2.8) has to be the marker? Whichever I want? Once I selected the marker indicator, should it be the same in each group comparison? Or, in other words, once the marker indicator has been fixed, is that invariant across groups?

Let me openly admit that although these latter queries get the invariance issue even more complicated they make it so fascinating at the same time as it is deducible that once an indicator is fixed to a number, say 1, it cannot be tasted for invariance because a constant is indeed invariant since it does not vary. As a consequence, this strategy to fix a marker indicator does not seem to be good enough. In this respect, there is also another way to give a metric to the latent factor and it consists in fixing the variance of the latent factor ξ to 1. This strategy provides the same results of the one with the marker indicator when we deal with a latent factor within each group, but it is intuitive that we renounce to estimate the latent variance(s) and so we do with the latent heterogeneity study across groups as well, since the latent variances would be standardized to 1.

Thus, we are really bound to get back to the strategy of the marker indicator and try to give insightful answers to those previous questions in due course.

2.5 – Invariance steps

According to Brown (2006), Gregorich (2006), Steenkamp & Baumgartener (1998), Vandenberg & Lance (2000), the measurement invariance steps are four: 1) configural invariance; 2) metric invariance; 3) scalar invariance; 4) invariance of uniqueness or testing of equality of indicators residuals. The first three steps must follow an hierarchical sequence of assessment, whereas the fourth can be less restrictive as I am going to address next.

In order to understand each step, please refer again to the system of equations (2.8) that now it turns out to be into a multi-block system of equations for each group c (c stand for cluster; c =1,m) as follows:

$$x_i^c = \tau_i^c + \lambda_i^c \xi^c + \delta_i^c \qquad (2.11)$$

with the means μ_i^c of the observed variables x_i^c:

$$\mu_i^c = \tau_i^c + \lambda_i^c \kappa^c \qquad (2.12)$$

where κ^c is the mean of the latent variable ξ^c for each group c.

It is noteworthy that the systems (2.11) and (2.12) can be easily extended to more than one factor ξ_j^c (j= 1, q). Starting with the system (2.11) it is straightforward to notice that if I am able to exactly write the expression in the system (2.11) the subsequent testing for complete measurement invariance of those observed variables may concern at least as much steps as the parameters are: $\tau_i^c, \lambda_i^c, \delta_i^c$. So then, at least three steps of equal intercepts, equal factor loadings, equal measurement errors across groups; although this latter hierarchical order of the invariance steps will be different as I am going to address next.

On the other hand, since we know that all the information we need is provided by the observed variable variances and covariances matrix we do not need of the observed means as they do not add any further information. Hence, the system (2.12) seems to be useless to achieve measurement invariance in the observed variables across groups, whilst it conversely seems to play a role in the study of the heterogeneity of the latent factor ξ^c across groups in terms of its mean κ^c.

2.5.1 – Configural invariance

As I aforesaid, looking at the system (2.11), in order to achieve a proper measurement invariance across groups we should test for at least three parameters if I may properly write, assume, that a set of observed variables may be explained by a common latent factor equally well across groups. This latter is very intuitive and is the first and vital starting point for every measurement invariance regarding a set of measures that are loading a common factor as the covariation of the a potential set of observed variables must be univocally explained by the same theoretically-driven latent factor(s) across groups. It practically means that the same set of observed variables must significantly load, measure, the same latent factors across groups. The term “load” means relevant (different from zero) zero-order correlation between that observed variable(s) and the latent factor(s). Horn and McArdle (1992) define the test of ‘configural invariance’ the one in which the same salient (different from zero) and non-salient (zero or close to zero) pattern of each indicator in loading each factor has been specified and this specification must be equivalent across groups. Furthermore, this equivalence should be respected also for the sign of each loading that, again, must be the same across groups (Meredith, 1993). A consequence of this assessment is that the same “so-built” factorial configuration (in case of more than one factor) will have to hold across groups also in terms of factorial correlations that are expected to be below the unity to be able to discriminate the factors themselves. All this technically means that the so-performed CFAs hold in all groups both in terms of convergent and discriminant validity[8]. It is again intuitive that if the factorial structures are different across groups any further step of invariance will inevitably stop as we are not able to write the system (2.11) in the same way for all the groups, or, better, we are

[8] Convergent and discriminant validity are respectively achieved when the standardized factor loadings are moderate in magnitude (e.g., > .4 and <.95; Bagozzi and Yi, 1988) and the correlations among factors is not too high (e.g., <.85, Brown, 2006; Kline, 2005, 2011) .

trying to compare different configurations that it does not make any sense whatever. It is as though I may want to test if a cube and a sphere are able to roll along a surface even though they are made of the same material. In speculative way the configural invariance is assessing if theoretical hypotheses, initially made on a particular set of items, are effectively reflecting common manifestation(s) of latent construct(s) equally well across groups.

In order to assess configural invariance the formal expression from the equation (2.11) is:

$$\xi = \xi^c \tag{2.13}$$

If this first and basic step of measurement invariance holds I may argue that my measures are really congeneric and thereby the postulated theory behind holds equally well across groups. As a consequence, now I have got proper estimation of structural parameters in each group, but I cannot still make any comparison because I did not make any hypotheses on the invariance of the measures yet.

2.5.2 – Metric invariance

After having assessed configural invariance I may want to proceed with the first proper step of measurement invariance: the so-called metric invariance as it concerns the invariance of that part of metric in the observed measures useful for giving to latent factor a proper way to be measured. This latter is indeed the reverse meaning of each factor loading λ_i^c, or unstandardized regression weight, that links a latent factor to each measure and it is straightforward interpreted as "the expected number of unit changes in the observed variables for a one-unit change in the true level of ξ" (Bollen, 1989; p.182) and so that, may be reversely deduced as how much of the expected effect of the true value is apt to be passed in the measured variables. If these expected effects (i.e., the factor loadings λ_i^c) are equivalent, and thereby invariant, across groups I may say that the latent construct ξ has been understood in the same way across groups. In other words, the respondents, who belong to different groups, have attributed the same meaning at the construct ξ above and beyond possibly different cultural aspects.

In order to assess metric invariance the formal expression from the equation (2.11) is:

$$\lambda_i = \lambda_i^c \tag{2.14}$$

This step of measurement invariance is known as 'weak factorial invariance[9]' (Meredith, 1993; Brown, 2006). It is straightforward that assessing metric invariance without having assessed configural invariance early on does not make any sense, since I cannot test if a construct has been understood in the same way across groups if I am not certain that the same measures are being used to represent that factor across each group equally well.

---

[9] Vandenberg & Lance (2000) labeled metric invariance as a 'strong invariance', whereas configural with 'weak invariance' agreeing with Horn & McArdle (1992) position. Conversely, I would prefer to agree with Meredith's and Brown's position as metric invariance should be the first true step of measurement invariance testing.

For this latter reason the measurement invariance process is strictly hierarchical and so does (although partially) the population heterogeneity steps at latent level as we are going to address step by step.

In addition, it is intuitive from the decomposition of the true value and residual explained by the process (2.7) and the system (2.8) that if configural and metric invariance hold and so that the construct (factor) object of the study has the same meaning across groups I am able to defend if the respondents agree more, or less, to that construct meaningful well. Or more simply if there is more, or less, consensus around that construct in answering (scoring) to those questions/items associated with that construct itself. In technical words, it means that I am able to compare factor variances across groups above and beyond possibly different cultural aspects. These cultural aspects, although related to the construct, do not alter its meaning.

2.5.3 – Scalar invariance

If configural and metric invariance have been achieved I may want to go on with another step of measurement invariance. It concerns the intercepts in the system (2.8) and thereby (2.11) and (2.12) and it is termed as scalar invariance or strong invariance (Meredith, 1993; Brown, 2006; Gregorich, 2006).

I have already outlined about the meaning of the intercepts that stands for origin locations of the observed metrics and thus they represent those cultural aspects that are active in the respondents but are not directly related to the construct ξ. As a matter of fact, the intercepts constitute an additive term in every aforementioned systems, but they have really relevance only in the expectation systems (2.12) when both observed means and latent means are involved. This is due to the fact that, looking at the CFA model (2.11), although the locations formally exist they do not give any contribution since all the information is caught by observed variances and covariances, so that deviations from the intercepts themselves are clearly zero. As a consequence, this level of invariance may be evoked only if I want to test for means even though it is still hierarchical to configural and metric invariance. The reason is again straightforward as the configural is the essential condition and the loadings invariance (i.e., metric) is the necessary and sufficient condition to test for locations (i.e., scalar). In fact, looking at the system (2.12) if the slopes are different across groups (i.e., metric invariance has not been achieved) it would be useless testing for location invariance as both the observed and latent means will result biased of λ_i^c quantity in any case.

Even more philosophically speaking if metric invariance is not achieved the latent factor have, as claimed, different meaning across groups and thereby does not make any sense for further invariance testing. In order to assess scalar invariance the formal expression from the equation (2.11) is:

$$\tau_i = \tau_i^c \qquad (2.15)$$

Also here, we can easily notice that only if scalar invariance is assessed we can make comparisons about means at observed and latent level above and beyond possibly different cultural aspects.

2.5.4 – Uniqueness invariance

The last step of measurement invariance is the one regarding the measurement errors in the system (2.11) and it is labeled as 'strict invariance' (Meredith, 1993; Brown, 2006; Gregorich, 2006).

As we already know the uniqueness is that part of observed variance not in common with the latent factor, the so-called measurement error because we are indeed trying to measure something, that is the latent construct, and we can commit errors in catching this true latent value through the regression system (2.11). However, it seems useless testing for invariance of the measurement errors when we have already tested for what we really need to make the latent quantitatively represented somehow, that is the common variance (i.e., factor loadings).

On the other hand, I may want to test for precision of my measures in loading a common factor ξ. In other word, testing if the measurement errors have been of the same magnitude across groups. More formally, testing for homogeneity of the regression models (2.11) (Vandenberg & Lance, 2000; page 13).

In addition, still looking at the system (2.11) if I wanted to compare observed variances and covariances across groups above and beyond possibly cultural forces I have to test for equality of measurement error variances after having assessed for configural, metric and scalar invariance. As a matter of fact, always from the system (2.11) if the measurement error variances differ across groups they constitute an additive bias of σ_{δ}^{2} in making comparisons among observed variances and covariances across groups even though the configural, metric and scalar invariance would have been achieved. Because of this latter reason the uniqueness invariance preserves the hierarchy with the previous steps even though is not necessary for making comparisons at latent level. In order to assess uniqueness invariance the formal expression from the equation (2.11) is:

$$\sigma_{\delta}^{2} = \sigma_{\delta c}^{2} \qquad (2.16)$$

2.6 – The study of population heterogeneity at latent level

Once all the necessary steps of measurement invariance have been assessed the researcher may want to explore how much the latent constructs are heterogeneous across groups. The phases for studying the heterogeneity across different populations with regard to latent variables are basically three: a) factor variance invariance; b) factor covariance invariance; c) equality of factor means. Here I have used letters instead of numbers as all these phases are not so strict hierarchical, unlike the measurement invariance steps, but they depend on what kind of latent heterogeneity I want to assess.

2.6.1 – Factor variance invariance

Factor variance invariance concerns the test of latent factors variances equivalence across groups. As I have already stated, before testing if factor variances are the same across groups I am bound to assess two measurement tests early on: configural and metric.

In presence of two or more factors in a factorial design, if configural, metric and factor invariance hold I may make comparisons among standardized solutions at latent level (i.e., latent correlations across groups). The factor variance invariance is formally expressed as:

$$\sigma_{\xi}^{2} = \sigma_{\xi c}^{2} \qquad (2.17)$$

2.6.2 – Factor covariance invariance

This phase of latent invariance heterogeneity involves a factorial design with two or more latent constructs to which associations are being compared across groups. Hence, testing for factor covariance invariance means testing for the equality of all possible covariances in a CFA design in order to verify how much the factors are correlated each other. Also this phase requires that configural and metric invariance have been assessed early on. Besides, it is customary to test for factor covariance invariance together with factor variance invariance. This latter is logical and intuitive from the reason that in presence of more than two factors the interest will be obviously focused on both how similarly they vary and covary across groups.

The factor variance invariance is formally expressed as:

$$\sigma_{\xi i \xi j} = \sigma^{c}_{\xi i \xi j} \tag{2.18}$$

2.6.3 – Equality of factor means

This is the last phase of latent heterogeneity invariance and it regards the latent means. I am going to discuss more about structured means analysis in this dissertation, but presently I may want to complete this section just recalling that for testing equivalence in latent means across groups is necessary to have assessed three hierarchical steps of measurement invariance: configural, metric and scalar for the reasons that I have already outlined in the subchapter 2.5.

The equality of factor means is formally expressed as:

$$\kappa = \kappa^{c} \tag{2.19}$$

2.7 – Continuum of invariance

Thus now, we have seen that when we afford the issue of invariance with regard to latent variables we have to deal with two aspects of invariance: at observed measurement level early on and at latent level later on. We need of invariances at measurement level to make further comparisons and/or hypotheses of equivalence at latent level. Hence, it is straightforward noticing that a sort of continuum of this invariance exists. This continuum basically reflects how much a theoretical factorial design, or latent structure, is invariantly moving across possibly different groups and stopping when it cannot be considered invariant any longer until it reaches a cultural identity at latent level.

Granted that, we learn from the literature that a continuum of invariance is defined as a situation where: "… encompassing both covariance and mean structure models together" (cit. on page 138, from Hancock et al., 2009 who referred to Meredith's work in 1993). So that, I may want to introduce only now the concept of *complete measurement invariance* and therefore identity at latent level across groups comparison. The former had been already defined by Karl Jöreskog, one of the three fathers of the structural equation model era with latent variables in the seventies (i.e., Jöreskog (1973), Keesling (1972) and Wiley (1973)), who provided in the 1971 three important assumptions for a complete measurement invariance of measures that are intended to be reflected by latent constructs across groups: all the factor loadings (i.e., regression coefficients), all the error covariance matrices, all factor variances (and factor covariances for model with more than one factor) must be identical across groups of interest. This is, and was, an initial, and pioneering I daresay, logical definition of complete measurement invariance: if all the sources of measurement are equal across groups, our measurement instrument is logically reliable and independent from cultural aspects, although it includes also factor variance invariance that it is an invariance test at latent level.

However, If I had started with this definition I would have been too much demanding from beginners, let me be a bit conceited here, as Jöreskog's definition is indirectly referring to those two aspects of invariance we discussed: the proper measurement invariance of the self-reported instrument across groups and the proper invariance of the groups' heterogeneity. Hence, if we now integrate Meredith's concept of continuum of invariance with Jöreskog's complete measurement invariance definition we get to the point of having an identity at measurement and latent level when all the steps of measurement invariance and the heterogeneity study are indeed assessed. This latter reflects a perfect situation. In other words, if all seven steps are achieved across groups we can robustly affirm that neither cultural aspects nor differences in any moments at latent level are affecting groups with regards to construct(s) of interest and so that we can pool the observed data for further global analysis with regard to those latent constructs of interest working at enclaves levels without distinguishing groups.

Furthermore, according to Jöreskog's definition of complete measurement invariance, we might pooling data at measurement and latent level also stopping at factor variance/covariance invariance without considering the intercepts and locations if they are not of interest in our study.

In table 2.2 I have tried to summarize and retrace this important issue of how the continuum of invariance is associated with measurement invariance and heterogeneity steps in order to let you understand when comparisons at latent level are defensible, and when pooling data at latent level are possible.

Table 2.2 – Continuum of invariance steps and structural parameters comparison.

		Continuum of Invariance						
	Comparisons	Measurement Invariance →						Latent Heterogeneity
		Configural	Metric	Scalar	Uniqueness	L-Variances	L-Covariances	L-Means
CFA	L-Variances/Covariances	**1**	**2**	*3*	*4*	*5*	*6*	*7*
	L-Means	**1**	**2**	**3**	*4*	*5*	*6*	*7*
	L-Correlations	**1**	**2**	*3*	*4*	**5**	*6*	*7*
Structural Model	Un-standardized Paths	**1**	**2**	*3*	*4*	*5*	*6*	*7*
	Standardized Paths	**1**	**2**	*3*	*4*	**5**	*6*	*7*
	Identity at Latent Level	**1**	**2**	**3**	**4**	**5 Pooling data at latent level**	**6 Pooling data at latent level**	**7 Identity**
	Reliability*	**1**	**2**	**3**	**4**	**5**	*6*	*7*

* testing for reliabilities invariance requires a continuum of invariance till latent variances homogeneity as latent variance invariance and uniqueness assure that the ratio $\rho = \sigma_{ti}^2 / \sigma_{xi}^2$ (2.7.1) is meaningful comparable (adapted from Vandenberg & Lance, 2000; p. 34) where σ_{ti}^2 are latent variances and σ_{xi}^2 observed variances.

From table 2.2 we can notice that the ordered sequence of numbers represents the hierarchical steps needed to be achieved till reaching the identity or stopping early on when the previous step has not been assessed. On the other hand, numbers in bold together with yellow underlining are the compulsory hierarchical steps for making comparisons at latent level. Numbers in italic are the not necessary steps.

So now, it seems that we have answered to two questions placed at the end of subchapter 2.3: "what about latent variables invariance or, as we have acknowledged to be, latent heterogeneity? How can we detect it?

The third question is still left: "In the case that, for instance, all sources of covariation in the observed variables are invariant (i.e., omnibus tests both achieved), possibly latent variables are still invariant?

Answering to this third question is very straightforward now and intuitive. Since the observed variables include cultural aspects, if the two omnibus tests are both assessed, the cultural aspects are not acting in a significant way in the measures (loading common latent constructs) across groups and then we can both pool the data smoothly and making all kind of comparisons at latent level we desire.

Nevertheless, although those observed measures encompass cultural aspects they include also true latent values and unique values. Hence, in my opinion, it is not sufficient to achieve the two omnibus tests for granting also an identity at latent level in terms of structural estimates even though they are indeed function of those observed variances and covariances and observed means. This is due to the following two reasons: a) omnibus tests are rarely both perfectly achieved; b) at measurement level the observed variables have not been partial out yet through common factor model strategies since no latent structure has been hypothesized yet. As a consequence, differences at latent level may still exists, or, better, comparisons at latent level can be meaningfully defensible without assessing for measurement steps.

On the other hand, reaching an identity at latent level (achieving all the necessary hierarchical steps) assures me to have an invariance both at measurement and at latent level and so that I may pool the data with regard to those latent constructs of interest and claim that there are no differences across any kind of possible groups concerning those constructs in terms of statistical moments, or possibly causal path coefficients among constructs in structural models, above and beyond cultural aspects.

Eventually, in the rare situation when both the omnibus tests have been assessed we should not need to run the seven invariance steps either if our only objective is to pool data without concerning possibly latent traits.

All in all, in order to properly afford the issue of invariance we have to start proceeding with detecting how much invariant (or lack of invariant) our datasets of observed variables (that are presumably intended to reflectively measure common factors across groups) is as they include both cultural aspects and latent traits, throughout the two omnibus tests. If these two tests are achieved I may stop measurement invariance analysis and go ahead latent heterogeneity with all possible comparisons across groups at latent level. If they are not, as in most of the research cases, I may proceed with sequential steps of measurement invariance in order to find out where it is located in my data along with what kind of suitable further comparisons at latent level I will be able to defend across potential groups.

A way to make all this possible lies in the simultaneity ability of SEM to deal with multi-group analysis as I outlined in chapter 1 and am going to specify in chapter 3.

2.8 – Partial measurement invariance

Now, before going on let me introduce a very fascinating and important issue that has been made known to us by Byrne et al. (1989) for the first time. This concept basically starts from a very intuitive (again and again the intuition helps us) question about invariance: "all the items must invariant in order to make proper further comparisons at latent level?". As a matter of fact, we indirectly talked about a sort of full measurement invariance so far. We have established that once the measurement invariance has been achieved, it is a full invariance where all the measures are indeed invariant across groups. But, if you think over this latter statement it does not sound very well as in the real world may happen that not all the measures, the items we are drawing attention to, may result invariant across groups, but merely some of them. So then, another two questions raise up: "how many items have to be invariant in order to still make proper comparisons at latent level?" and "what happens to that latent with those non-invariant measures?". Byrne and al. (1989) try to give a proper answer to these queries. They intuitively claimed, but without any formal demonstration (as observed by Hancock et al., 2009), that: "…we believe that they are left with the impression that, given a non-invariant pattern of factor loadings, further testing of invariance and the testing for differences in factor mean scores are unwarranted. This conclusion, however, is unfounded when the model specification includes multiple indicators of a construct and at least one measure (other than the one that is fixed to 1.00 for identification purposes) is invariant (Muthén & Christoffersson, 1981)" (Byrne at al., 1989; page 458). Hence, let me openly say that I have appreciated a lot this conclusion as it is really proper and in line with the measurement invariance issue is supposed to be. It is obvious that we are talking about metric and scalar invariance and it seems reasonable having still metric and scalar invariance when at least one measure, other than the marker indicator, results at least invariant.

On the other hand and in other words, it means that the construct is still understood in the same way across groups (i.e., metric invariance) when two items are at least invariant (the fixed one is invariant for construction) and the cultural forces not related to the construct are still invariant for those two items (i.e., scalar invariance). Although with only two measures (or just the fixed one) seems very weak to defend these two invariance steps, even though I may have strong theory in supporting that true latent factor. As a consequence, another important issue comes up from Hancock et al. (2009): "…how to establish a proper initial minimum set" of measures? Basically, these latter authors work the matter out with having strong theoretical grounds on the construct of interest, and so have on the involved measures, that may assure construct invariance at theoretic level, even when it has failed with statistical evidence, until proclaiming a conditional minimal measurement invariance once only the marker indicator may be considered theoretically invariant and the others not.

The way I am viewing this matter is that having strong theoretical grounds set the baseline process up, but it cannot be considered a scapegoat at all times when empirical evidence does not go in parallel with that theoretical thought. I would prefer to proclaim those latent constructs with one or two invariant measures strongly affected by cultural aspects rather than leaving the things as they were culturally invariant just because of theoretical justifications. By the way, a researcher should look into what cultural problems the rest of items really have, and why, both in terms of measures and postulated theories.

2.9 – Rationale about structured latent means comparisons

So far I have briefly introduced the concept of latent mean in a way similar to the classical statistical moment definition of a mean, although at latent level, throughout the systems (2.11) and (2.12). Intuitively, since the latent variable is a measurement error-free variable the relative mean is still a measurement error-free moment retracing this fundamental aspect of the latent rationale. Thus, having a look at the systems (2.11) and (2.12), I may want to show you here again, it is noteworthy noticing that the expression (2.12.1) of computed latent mean(s) κ^c appears to be different from a common mean:

$$x_i^c = \tau_i^c + \lambda_i^c \xi^c + \delta_i^c \tag{2.11}$$

$$\mu_i^c = \tau_i^c + \lambda_i^c \kappa^c \tag{2.12}$$

$$\kappa^c = \frac{\mu_i^c - \tau_i^c}{\lambda_i^c} \tag{2.12.1}$$

As a matter of fact, (2.12.1) infers that the latent means are function of observed means μ_i^c, intercepts τ_i^c and regression slopes λ_i^c (i.e., factor loadings) and let you understand that error-free rationale at latent mean level. In this respect, the latent means κ^c derive from the observed means from which is necessary subtracting the intercepts of hypothesized CFA linear model (2.11) and dividing this amount for the regression slopes in order to have a mean that is still error-free, on average. The quantity $(\mu_i^c - \tau_i^c)$ indeed represents the difference from what has been observed and what has been partial out from a CFA model, whilst the intercepts τ_i^c represent, in turn, all those cultural causes not directly related to the common factor ξ that now we are able to control for (i.e., by subtracting them from observed means). Besides, since factor loadings embody the core of latent variable quantification (as they denote what the underlined measures have in common in defining a latent construct) they do not have to change, but have to be there as the meaning of the latent construct depends on them. For this latter reason the factor loadings have to be equal across groups (i.e., metric invariance).

Now backing to the systems (2.11) and (2.12) we can easily notice that it is impossible to calculate each single mean in each group for identification problems as the number of free parameters overcomes the number of the observations[10]. But, fortunately, we are interested in differences among latent means across groups and therefore we have to preserve this objective. As a consequence, the wording 'latent means differences' seems making even more sense in considering observed means and thereby latent means into a structural design in order to answer to the research question: "Do

---

[10] The number of observations is the number of observed variances and covariances (i.e., (v(v+1))/2; with v the number of observed variables) and the observed mean vector. Hence, if for instance i=3 the number of observations in the system (2.11) is nine (i.e., 6 variances, 6 covariances, 3 means) whilst the number of free parameters is ten (i.e., 1 latent variance, 1 latent mean, 3 error variances, 3 intercept terms, 2 factor loadings with fixing $\lambda_1^c = 1$ for each c group).

population differ with respect to the average amount of a particular latent construct?" (cit. from Hancock & Muller 2012). To which I may want to add: "above and beyond cultural aspects?".

The rationale of latent means is all here: I want to see if the latent constructs, object of the study, differ in average across groups above and beyond cultural forces. By doing so, and surprisingly, it is precisely this comparison/difference to give us keys to identify the systems (2.11) and (2.12) keeping up all the above mentioned rationale about error-free statistical moment and isolation from those cultural forces not related to the construct(s) of interest. Besides, the 'word' comparison suggests to fix a group as a reference in order to make proper comparison with it. And the easy way to do that is to fix one latent group mean to zero and so that the other means can be computed as deviations from the one as reference. This advice is the same used to solve an arithmetical problem when someone says that the number 5 is the difference between two numbers without giving you which numbers were involved in, with the consequence that there would be no unique answers if at least one number has revealed (adapted example from Hancock, 1997). Furthermore, in order to isolate the cultural aspects as a whole it is necessary that the construct has been both understood in the same way across groups and the other forces not related to that construct have resulted the same. In other words, it means that both metric and scalar invariance (at least both at partial level) have to be achieved first. For this latter reason I had previously claimed that these two types of invariance should have been assessed to make comparisons among latent means.

For making you an example with two groups c = a, b from the system (2.12) and (2.12.1) we have:

$$\mu_i^a = \tau_i^a + \lambda_i^a \kappa^a \quad (2.12.2)$$

$$\mu_i^b = \tau_i^b + \lambda_i^b \kappa^b \quad (2.12.3)$$

$$\kappa^a = \frac{\mu_i^a - \tau_i^a}{\lambda_i^a} \quad (2.12.4)$$

$$\kappa^b = \frac{\mu_i^b - \tau_i^b}{\lambda_i^b} \quad (2.12.5)$$

$$\kappa^d = \kappa^b - \kappa^a = \frac{\mu_i^b - \tau_i^b}{\lambda_i^b} - \frac{\mu_i^a - \tau_i^a}{\lambda_i^a} \quad (2.12.6)$$

Setting the group 'a' as reference group we fix $\kappa^a = 0$ the expression (2.12.2) and (2.12.6) are respectively solved as follows:

$$\mu_i^a = \tau_i^a \quad (2.12.7)$$

$$\kappa^d = \kappa^b = \frac{\mu_i^b - \tau_i^b}{\lambda_i^b} \quad (2.12.8)$$

Where κ^b becomes the difference between the two group means on the construct ξ. If metric and scalar are both assessed (i.e., $\lambda_i^a = \lambda_i^b = \lambda$ and $\tau_i^a = \tau_i^b$) substituting (2.12.7) in (2.12.8):

$$\kappa^d = \frac{\mu_i^b - \mu_i^a}{\lambda} \qquad (2.12.9)$$

As a result, the difference in latent means on the construct ξ between the two groups involves the observed means "standardized" with the equal factor loadings (that are bound to be equal since the factor ξ must be understood in the same way between the two groups) granting that the other forces (the intercepts τ_i^c) not related to the common factor ξ are still not influencing difference in latent means either.

In other words, only those items i that are metric and scalar invariant across groups are involved in the computation of the latent mean difference, the others not. Still more practically "… this implies the desired condition that any difference the groups may have on the observed variables is directly attributable to a difference in the underlying construct, and not to differences in the nature of the structural relationship" (cit. page 8, Hancock, 1997). The structural relationship stands for the CFA model (2.11).

On the other hand, it seems even more clear now that in case these other cultural aspects are acting the scalar invariance does not hold (i.e., $\tau_i^a \neq \tau_i^b$). And even though I fix the reference group (2.12.7) I am not able to wright down the expression (2.12.9) that it turns into:

$$\kappa^d = \kappa^b = \frac{\mu_i^b - \tau_i^b}{\lambda} \qquad (2.12.10)$$

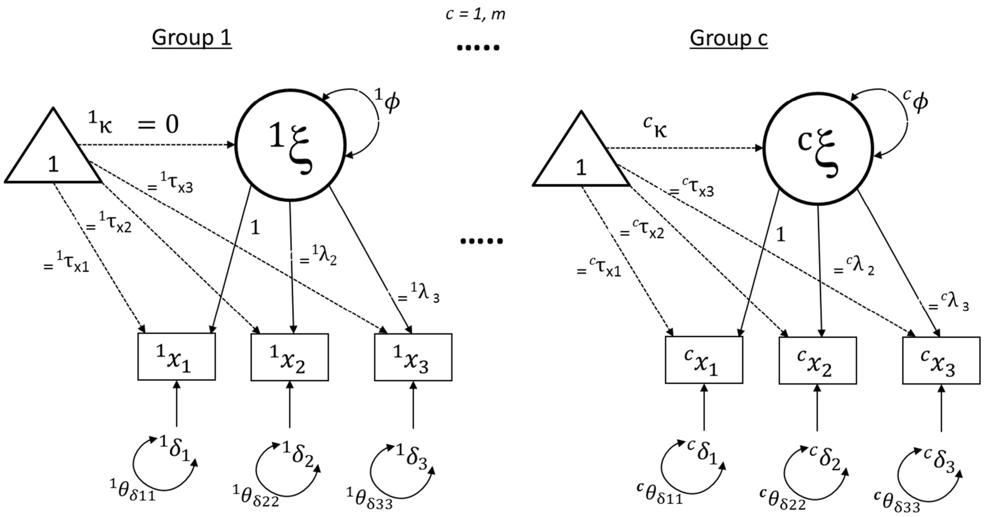
Consequently, the difference in latent means on the common construct ξ is affected by cultural forces (not directly related to the common factor) that are making this difference biased.

In details, the item i is culturally biased in locations since $\tau_i^a \neq \tau_i^b$, even though the factor loadings are the same. Hence, that item causes a $\Delta\tau = (\tau_i^a - \tau_i^b)$ adding bias to $(\mu_i^b - \mu_i^a)$ in the estimation of latent means difference. From the expression (2.12.6):

$$\kappa^d = \kappa^b - \kappa^a = \frac{\mu_i^b - \tau_i^b}{\lambda} - \frac{\mu_i^a - \tau_i^a}{\lambda} = \frac{(\mu_i^b - \mu_i^a) + (\tau_i^a - \tau_i^b)}{\lambda} \qquad (2.12.11)$$

The expression (2.12.10) is a special case of (2.12.11) with 'a' as reference group. In figure 2.1 it has depicted a structured means model (SMM) path diagram in order to better visualize the simultaneous process of estimation in the system (2.11) and (2.12) with a hypothetical common factor ξ loaded by three indicators x1, x2, x3 where: λs represent the factor loadings, τs the intercepts, δs the measurement errors, θδs the measurement error variances, ϕ the factor variance, the predictor variable depicted as a triangle defines a pseudo-variable with no variance which is equal to 1 for all the individuals because it represents the coefficient 1 of all intercept terms included the intercept κ of factor ξ. The intercept κ is also the factor mean since $\xi = 1\kappa + \phi$; $E(\xi) = 1\kappa$.

Figure 2.1. –Multi-group structured means model path diagram.



It is also noteworthy from figure 2.1 that the factor loadings and the intercept terms are respectively constrained to be equal for the required assumption of metric and scalar invariance. For more details about SMMs have a look at the book chapter of Thompson & Green (2013).

2.9.1 – Group code approach to latent means comparisons (differences): a special case of MIMIC models

I would like to conclude this chapter 2 with an important alternative approach to means comparisons (differences) at latent level that I did not want to mention so far both because it has not viewed as a proper simultaneous way of proceeding and it constitutes a special case of structured means models (SMMs) often applied when a sufficient sample size is not available.

This method is termed as MIMIC that stands for Multiple-Indicator Multiple-Cause (Jöreskog & Goldberger, 1975) and in the case of latent means difference it works like ANOVA with dummy variables that reflects the impact of different groups (e.g., contrasts, effects) on a dependent variable that now has a latent nature. To this end the latent construct is regressed on dummy variable(s) within a single structural model (Hancock, 1997) and the parameters of interest like γ are re-written as function of the latent indicators and group code variable as depicted in figure 2.2 (with figure 2.2 bis with k dummy) and table 2.3 for a dichotomous dummy variable.

Figure 2.2 – Group code approach to latent means model: MIMIC modeling with dichotomous dummy path diagram.

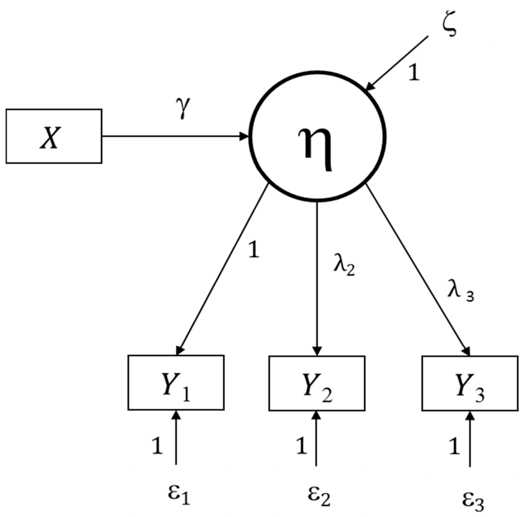


Figure 2.2 (bis) – Group code approach to latent means model: MIMIC modeling with k-dummy path diagram.

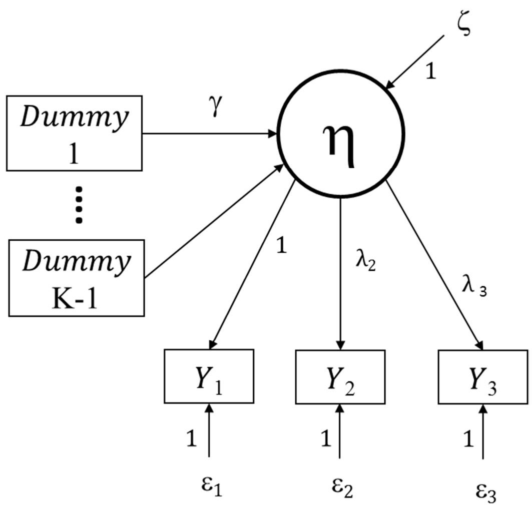


Table 2.3 - Equations and relationships for the Group Code Analysis (Hancock, 1997).

Structural equations	Model-implied relationships	
$Y_1 = 1\eta + 1\varepsilon_1$	$Var\ (Y_1)$	$= [\gamma^2 Var(X) + Var\ (\zeta)] + Var\ (\varepsilon_1)$
$Y_2 = 1\eta + 1\varepsilon_2$	$Var\ (Y_2)$	$= \lambda_2{}^2\ [\gamma^2 Var(X) + Var\ (\zeta)] + Var\ (\varepsilon_1)$
$Y_3 = 1\eta + 1\varepsilon_3$	$Var\ (Y_3)$	$= \lambda_3{}^2\ [\gamma^2 Var(X) + Var\ (\zeta)] + Var\ (\varepsilon_1)$
$\boldsymbol{\eta = \gamma X + 1\zeta}$	$Var\ (X)$	$= Var\ (X)$
	$Cov\ (Y_1,Y_2)$	$= \lambda_2\ [\gamma^2 Var(X) + Var\ (\zeta)]$
	$Cov\ (Y_1,Y_3)$	$= \lambda_3\ [\gamma^2 Var(X) + Var\ (\zeta)]$
	$Cov\ (Y_2,Y_3)$	$= \lambda_2 \lambda_3\ [\gamma^2 Var(X) + Var\ (\zeta)]$
	$Cov\ (X,Y_1)$	$= \gamma Var\ (X)$
	$Cov\ (X,Y_2)$	$= \gamma\lambda_2 Var\ (X)$
	$Cov\ (X,Y_3)$	$= \gamma\lambda_3 Var\ (X)$

It is again intuitive that the dummy coded variable X needs to be involved in a covariation with the factor indicators in order to provide a contribution to the latent factor itself through the parameter γ that indeed represents how the groups differ, on average, with respect to a latent construct. But what do the variance of the group code variable X and covariances between the group code variable and the indicators represent? And why the parameter γ represents the estimated difference in factor means?

Before answering to these questions a careful reader would have perceived that neither observed means/intercept terms nor separate groups data covariance matrices have been considered. As a consequence, since we are still in a covariance-based SEM situation all the observed variables (with the inclusion of the dummy-group code variable X) are deviated from their means and thus only variances and covariances are considered. These observed variances and covariances among factor indicators are the ones of the combined (pooled) sample without group-distinction and since our inference is at latent level the var (η) is the total variance that, in turn, is function of total variances and covariances among indicators. The var (X) is the between groups variance whereas the covariances between dummy X and each indicator actually represent the between groups covariances. As a matter of fact, covariances between the group code and each indicator embody an indication of how one group has more, or less, of that indicator with respect to the other group and do not represent a proper quantitative value.

For making you an example taken from Hancock & Muller (2012) the following correlation matrix (correlations are used instead of covariances for illustrative purposes) represents three indicators referring to the latent factor Math-Proficiency with the dummy group variable (fourth line from the top) stands for gender group (girls coded with 0 and Boys coded with 1):

$$\begin{bmatrix} 1 & & & \\ .747 & 1 & & \\ .736 & .666 & 1 & \\ .092 & .080 & .079 & 1 \end{bmatrix}$$

Since the last line represents the correlation between the dummy and each indicator, it is noteworthy that in all cases boys (coded with 1) have higher scores than girls (coded with 0).

So that, from the model-implied relationships in table 2.3 the presence of [Var (X) and Cov (X, Y_i)] along with [Var (Y_i) and Cov (Y_i, Y_j)] is respectively a MANOVA-like situation of between variances (covariances) and total variances (covariances), but under the covariance-based SEM whereas the within group (WG) variance var (WG) = var (η) - var (X) = ζ that precisely represents the model disturbance in figure 2.2, or, better, that part of within group variance that was not explained by dummy X.

Alike for (M)ANOVA we want to test if the difference among 'between' and 'within' group variance (covariances) is significant. If yes, it is due to the dependent variable mean of interest that, in our case, has a latent nature. So now, looking back at table 2.3, since the latent factor has zero mean ($E(\eta_0) = 0$; because of considering data deviated from means) it is computationally easy to derive from the structural equation $\eta = \gamma X + \zeta$ with dichotomous dummy X (X_0=0; X_1=1; with $E(\zeta) = 0$ for definition) that $E(\eta_1) = \gamma$ and then $E(\eta_1) - E(\eta_0) = \gamma$.

Granted that, and at a first glance, it seems that MIMIC approach to latent means comparison is easier than SMM as I do not need to separate group or matrices, but only running a complete model with setting the dummy(ies). But, the "HUGE BUT" still lies in the measurement invariance testing.

As a matter of fact, a careful reader would have asked again: "Would it be proper to combine the two groups together without any testing on that?".

This is the crucial point about MIMIC approach to latent means comparison. The main weakness of this approach is that we are unable to test for invariance before making such a comparison since MIMIC assumes a complete measurement invariance across groups. That is, as we know, assuming configural, metric, uniqueness invariance along with factor variance invariance without any formal test and this is a tough assumption that might not hold. If this latter is the case the subsequent estimation of γ will result biased since based on group constrains that does not hold. Clearly, as we have acknowledged, when complete measurement invariance hold we are able to pool the data with no worries about cultural bias and so that results from MIMIC and SMM will be identical. However, in my opinion and how it is pretty evident, whenever the sample size permits, it is always recommended to apply SMM with simultaneous way of estimation as it is the only way to detect which items and locations are, or not, cross-culturally invariant for making subsequent comparisons at latent level.

2.10 – Best "sellers" about measurement invariance issue

A list of current literature about the topic is likely what someone would have expected by a background chapter in a dissertation and therefore I could not tear myself away from that, but do hope not to be too much 'outlier' ☺ in presenting this part as a sort of open-shelf selection of 'best sellers' suggesting readings about the gigantic amount of documents as regards measurement invariance. This selection cannot be obviously exhaustive, and it does not want to be like that at all, but it is aimed at emphasizing some chosen manuscripts, selected both from those which you can find cited spread around this thesis and not, so as to offer in a 'nutshell' a hopefully good orientation towards measurement invariance topic for both beginners and experts.

Hence, let me start with a very illuminating book chapter by Hancock, Stapleton and Arnold-Berkovits (2009) entitled "The Tenuousness of Invariance Tests within Multi-Sample Covariance and Mean Structure Models" in which the authors brilliantly argue how instable can be assessing measurement invariance with following just statistical evidence and how are important theoretical grounds instead, untangling also the constraints mechanism of invariance in a very exhaustive way.

Let me continue with another book chapter by Brown's (2006) entitled "CFA with Equality Constraints, Multiple Groups, and Mean Structures" that yields a remarkable overview on the measurement invariance and CFA arena as well.

After these two suggestions I may want to highlight two possible milestones in this field: "A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research" by Vandenberg & Lance (2000); "Assessing Measurement Invariance in Cross-National Consumer Research" by Steenkamp & Baumgartener (1998). These two publications are still very often cited nearly everywhere in applied research journals and they actually put many keystones on measurement invariance application providing also loads of further references.

Successively, let me suggest a very clever article, possibly not too popular, entitled: "Do Self-Report Instruments Allow Meaningful Comparisons Across Diverse Population Groups? Testing Measurement Invariance Using the Confirmatory Factor Analysis Framework" in which the author Steven Gregorich (2006) gives us a revisited overview of the topic simplifying many measurement invariance matters.

Eventually, I may want to recommend a classic, but still hands-on, document by Karl Jöreskog (2005) "Structural Equation Modeling with Ordinal Variables using LISREL". This downloading doc from the Scientific Software International (SSI) website constitutes a nice handbook on how to deal with multiple groups comparisons at latent level providing clear explanation of the theory and its empirical application together with several programming steps.

At the very end, I highlight a challenging book by Thanh V. Tran (2009): "Developing Cross-Cultural Measurement" edited by Oxford University Press. As it is subtitled, this book concerns a sort of pocket guide to social work research methods alike cross-cultural measurement actually is. This publication provides an overview on cross-cultural assessment from different disciplines and not only from the statistical point of view. Nevertheless, many applications using SEM have been presented as well.

Chapter 3

Beyond technicality and fit indices

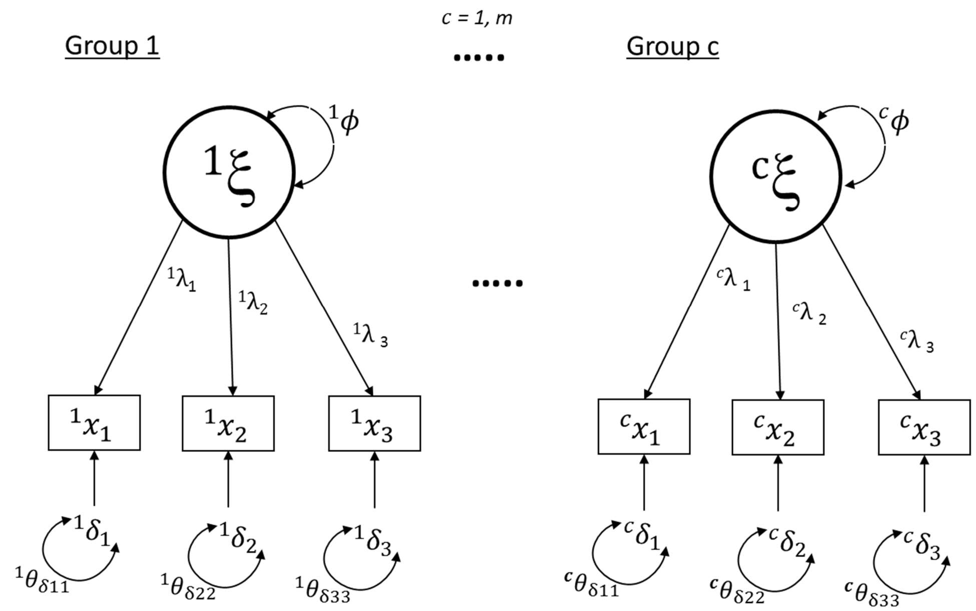
I may want to set about writing this chapter with quoting Glymour et al. (1987, pages 32-33) (reference found in Bollen's book (1989) on page 72) about the sense of approximation with regard to theories postulated by scientists: "In the natural sciences, nearly every exact, quantitative law ever proposed is known to be literally false. Kepler's law are false, Ohm's law is false, …, and on and on. These theories are still used in physics and in chemistry and in engineering, even though they are known to be false. They are used because, although false, they are approximately correct. Approximation is the soul of science".

3.1 – The process of constraints

In this subchapter I may want to briefly explain how the measurement invariance process of constraints works out. This is due to the reason that having an idea on how the 'mechanism's in running order' might be particularly helpful in understanding (and having "trust" of) thresholds and boundaries in achieving invariance. It is straightforwardly intuitive just from the word 'constraint' that we try to compel something and so that yielding to possibly tenuousness of the entire process itself or, better saying, checking the tenability of the constraints. But, it is properly what I am looking for. I may want to verify how the entire process is simultaneously tenuousness, or conversely robust, when I make such constraints to the parameters of interest.

Looking at the figure 3.1 we acknowledged from the Chapter 2 that I am able to specify, in each c group, variances and covariances of the observed variables (three in this case) as function of the structural parameters related to a hypothetical common factor ${}^{c}\xi$, loaded by three indicators, where ${}^{c}\lambda_{i}$ represents the factor loadings, ${}^{c}\delta_{i}$ the measurement errors, ${}^{c}\theta_{\delta i}$ the measurement error variances, ${}^{c}\phi$ the factor variances.

Figure 3.1 – Multi-group model path diagram for the latent ξ.



So that, with regard to the observed variables X_1, X_2 and X_3 I have a decomposition of variances and covariances for the groups 1 and c (with c = 2, m) as follows:

$^{1}Var\ (^{1}X_1) = (^{1}\lambda_1)^2\ ^{1}\phi + {}^{1}\theta_{\delta 1}$

$^{1}Var\ (^{1}X_2) = (^{1}\lambda_2)^2\ ^{1}\phi + {}^{1}\theta_{\delta 2}$

$^{1}Var\ (^{1}X_3) = (^{1}\lambda_3)^2\ ^{1}\phi + {}^{1}\theta_{\delta 3}$

$^{c}Var\ (^{c}X_1) = (^{c}\lambda_1)^2\ ^{c}\phi + {}^{c}\theta_{\delta 1}$

$^{c}Var\ (^{c}X_2) = (^{c}\lambda_2)^2\ ^{c}\phi + {}^{c}\theta_{\delta 2}$

$^{c}Var\ (^{c}X_3) = (^{c}\lambda_3)^2\ ^{c}\phi + {}^{c}\theta_{\delta 3}$ (3.1)

$^{1}Cov\ (^{1}X_1, {}^{1}X_2) = {}^{1}\lambda_1\ ^{1}\lambda_2\ ^{1}\phi$

$^{1}Cov\ (^{1}X_1, {}^{1}X_3) = {}^{1}\lambda_1\ ^{1}\lambda_3\ ^{1}\phi$

$^{1}Cov\ (^{1}X_2, {}^{1}X_3) = {}^{1}\lambda_2\ ^{1}\lambda_3\ ^{1}\phi$

$^{c}Cov\ (^{c}X_1, {}^{c}X_2) = {}^{c}\lambda_1\ ^{c}\lambda_2\ ^{c}\phi$

$^{c}Cov\ (^{c}X_1, {}^{c}X_3) = {}^{c}\lambda_1\ ^{c}\lambda_3\ ^{c}\phi$

$^{c}Cov\ (^{c}X_2, {}^{c}X_3) = {}^{c}\lambda_2\ ^{c}\lambda_3\ ^{c}\phi$ (3.2)

If I want to make constraints on factor loadings $^{c}\lambda_i$ in a context of metric invariance I will start from the system (3.2) as it directly involves each loading along with factor variances and I do not need of error variances information $^{c}\theta_{\delta i}$ at this step. Should I want to constrain, for example, $^{1}\lambda_1 = {}^{c}\lambda_1$ with initial numerical values of $^{1}\lambda_1 = z_1$ and $^{c}\lambda_1 = z_c$, I have to adjust the quantity $^{c}\lambda_2{}^{c}\phi$ and $^{c}\lambda_3{}^{c}\phi$ with a multiplicative factor of z_1/z_c in response to the changing in $^{c}\lambda_1$ in the $^{c}Cov\ (^{c}X_1, {}^{c}X_2)$ and $^{c}Cov\ (^{c}X_1, {}^{c}X_3)$ expressions respectively. Likewise, the quantity $^{1}\lambda_2{}^{1}\phi$ and $^{1}\lambda_3{}^{1}\phi$ will be adjusted with a multiplicative factor of zc/z_1 in response to the changing in ${}^{1}\lambda_1$ in the ^{1}Cov (${}^{1}X_1$,${}^{1}X_2$) and ^{1}Cov (${}^{1}X_1$,${}^{1}X_3$) expressions, respectively. Essentially, these multiplicative factors are going to adjust each loading and so that rescaling the factor variances ${}^{c}\phi$ and ${}^{1}\phi$ as well solving the system (3.2) with the multiplicative factors dividing ${}^{c}\phi$ and ${}^{1}\phi$ by $(z_c/z_1)^2$ and $z_1/z_c)^2$, respectively. As consequence, new values will be substituted to the expressions in the system (3.1) in order to calculate new values for measurement error variances ${}^{c}\theta_{\delta i}$ as well. In the case of the marker indicator if ${}^{1}\lambda_1 = {}^{c}\lambda_1 = 1$ with initial numerical values of ${}^{1}\lambda_1 = z_1$ and ${}^{c}\lambda_1 = z_c$, we have to adjust the quantity ${}^{c}\lambda_2{}^{c}\phi$ and ${}^{c}\lambda_3{}^{c}\phi$ with a multiplicative factor of $1/z_c$ in response to the changing in ${}^{c}\lambda_1$ in the ^{c}Cov (${}^{c}X_1$, ${}^{c}X_2$) and ^{c}Cov (${}^{c}X_1$, ${}^{c}X_3$) expressions respectively. Similarly, the quantity ${}^{1}\lambda_2{}^{1}\phi$ and ${}^{1}\lambda_3{}^{1}\phi$ will be adjusted with a multiplicative factor of $1/z_1$ in response to the changing in ${}^{1}\lambda_1$ in the ^{1}Cov (${}^{1}X_1$,${}^{1}X_2$) and ^{1}Cov (${}^{1}X_1$,${}^{1}X_3$) expressions respectively. As a consequence the factor variances ${}^{c}\phi$ and ${}^{1}\phi$ will be adjusted by $(1/z_1)^2$ and $(1/z_c)^2$, respectively.

Setting up a numerical example:

${}^{1}\lambda_1 = 0.5$; ${}^{1}\lambda_2 = 0.6$; ${}^{1}\lambda_3 = 0.7$

${}^{c}\lambda_1 = 0.8$; ${}^{c}\lambda_2 = 0.7$; ${}^{c}\lambda_3 = 0.9$

${}^{1}\phi = 1.10$

${}^{c}\phi = 2.10$

${}^{1}Cov\ ({}^{1}X_1, {}^{1}X_2) = {}^{1}\lambda_1\, {}^{1}\lambda_2\, {}^{1}\phi = 0.33$

${}^{1}Cov\ ({}^{1}X_1, {}^{1}X_3) = {}^{1}\lambda_1\, {}^{1}\lambda_3\, {}^{1}\phi = 0.385$

${}^{1}Cov\ ({}^{1}X_2, {}^{1}X_3) = {}^{1}\lambda_2\, {}^{1}\lambda_3\, {}^{1}\phi = 0.462$

${}^{c}Cov\ ({}^{c}X_1, {}^{c}X_2) = {}^{c}\lambda_1\, {}^{c}\lambda_2\, {}^{c}\phi = 1.176$

${}^{c}Cov\ ({}^{c}X_1, {}^{c}X_3) = {}^{c}\lambda_1\, {}^{c}\lambda_3\, {}^{c}\phi = 1.512$

${}^{c}Cov\ ({}^{c}X_2, {}^{c}X_3) = {}^{c}\lambda_2\, {}^{c}\lambda_3\, {}^{c}\phi = 1.323$

a) with ${}^{1}\lambda_1 = {}^{c}\lambda_1$ the following loadings ${}^{1}\lambda_1$, ${}^{1}\lambda_2$, and ${}^{1}\lambda_3$ will be multiplied for (0.8/0.5) = 1.6 in order to adjust the ${}^{1}\lambda_1$ (i.e., former 0.5) to be equal to ${}^{c}\lambda_1$ (i.e., 0.8) in the system (3.2) as follows: ${}^{1}\lambda_1 = 0.8$; ${}^{1}\lambda_2 = 0.96$; ${}^{1}\lambda_3 = 1.12$. Therefore the factor variance ${}^{1}\phi = 1.10$ will decrease to the value of 0.429 and that is expected as the loadings ${}^{1}\lambda_1$ and ${}^{c}\lambda_1$ are quite different in magnitude (i.e., 0.5 vs 0.8). This new value of ${}^{1}\phi^{*} = 0.429$ can be also obtained scaling the original value of $1.10/(0.8/0.5)^2$. Similarly for the loadings ${}^{c}\lambda_1$, ${}^{c}\lambda_2$, and ${}^{c}\lambda_3$ they will be multiplied for (0.5/0.8) = 0.625 in order to adjust the ${}^{c}\lambda_1$ (i.e., former 0.8) to be equal to ${}^{1}\lambda_1$ (i.e., 0.5) in the system (3.2) as follows: ${}^{c}\lambda_1 = 0.5$; ${}^{c}\lambda_2 = 0.437$; ${}^{c}\lambda_3 = 0.562$. Therefore the factor variance ${}^{1}\phi = 2.10$ will increase to value of 5.38 and that is expected as the loadings ${}^{1}\lambda_1$ and ${}^{c}\lambda_1$ are quite different in magnitude (i.e., 0.5 vs 0.8). Also this new value of ${}^{c}\phi^{*} = 5.34$ can be also obtained scaling the original value of $2.10/(0.5/0.8)^2$.

b) with ${}^{1}\lambda_1 = {}^{c}\lambda_1 = 1$ the following loadings ${}^{1}\lambda_{1,}$ ${}^{1}\lambda_{2,}$ and ${}^{1}\lambda_3$ and will be multiplied for $(1/0.5) = 2$ in order to adjust the ${}^{1}\lambda_1$ (i.e., former 0.5) to be equal to 1 in the system (3.2) as follows: ${}^{1}\lambda_1 = 1$; ${}^{1}\lambda_2 = 1.2$; ${}^{1}\lambda_3 = 1.4$. Therefore the factor variance ${}^{1}\phi = 1.10$ will decrease to the value of 0.275 and that is expected as the loadings ${}^{1}\lambda_1$ and ${}^{c}\lambda_1$ are quite different in magnitude from 1 (i.e., 0.5 vs 0.8). This new value of ${}^{1}\phi^{*} = 0.275$ can be also obtained scaling the original value of $1.10/(1/0.5)^2$. Similarly for the loadings ${}^{c}\lambda_{1,}$ ${}^{c}\lambda_{2,}$ and ${}^{c}\lambda_3$ they will be multiplied for $(1/0.8) = 1.25$ in order to adjust the ${}^{c}\lambda_1$ (i.e., former 0.8) to be equal to 1 in the system (3.2) as follows: ${}^{c}\lambda_1 = 1$; ${}^{c}\lambda_2 = 0.875$; ${}^{c}\lambda_3 = 1.125$. Therefore the factor variance ${}^{1}\phi = 2.10$ will decrease to value of 1.344 and that is still expected as the loadings ${}^{1}\lambda_1$ and ${}^{c}\lambda_1$ are quite different in magnitude from 1 (i.e., 0.5 vs 0.8). Also this new value of ${}^{c}\phi^{*} = 1.344$ can be also obtained scaling the original value of $2.10/(1/0.8)^2$.

From this simple example you can see how the whole process can be strongly altered because of only just a single constraint. In this case I have constrained the factor loadings, but it would have been the same with constraining intercept terms ${}^{c}\tau_i$ or uniqueness ${}^{c}\theta_{\delta i}$ and so forth (see Hancock et al., 2009 for more details). In this respect, what it is really important is that the whole process of testing for invariance must hold. For ‘whole process’ I do not mean only having good model fit diagnostics, but not having too much different estimations after those constraints have been made. In other words, it means that the initial estimations cannot be too much modified by the constraining-process even if the global model diagnostics are still reasonably unchanged in terms of cut-off boundaries.

The same situation happens when the marker indicator is selected as we constrain the loadings to 1 and so that all the process of adjustment in reference to 1 re-starts again.

Furthermore, the marker indicator, as I aforementioned in the previous chapter 2, is the only indicator that is not tested for invariance, but we declared that is invariant by default for identification purposes (i.e., defining the metric) with regard to each latent factor. Hence, it is intuitive that a researcher, in making this hypothesis of marker-indicator invariance, should have strong theoretical grounds of invariance itself in advance, above and beyond subsequent statistical verifications. These latter are important in terms of psychometrical properties in any case. To this end the marker indicator should be selected also among the most reliable items in loading a common factor (Brown, 2006; Kline, 2012). In my opinion, a simple way to deal with this is looking at the strongest item-total correlation for each item (i.e., psychometric property from Cronbach’s coefficient alpha[11] solution that tests how consistent each item may be with the averaged behavior/scores of other items in loading a common factor and so that how much shared-variance may possibly pass onto the latent factor of interest with regard to each item). However, since ”The statistical is conditional upon the theoretical” (Hancock et al., 2009, page 171) psychometric properties should never overcome the theoretical, but going in parallel. So that, a researcher should work in advance for making up a good theory and good reliable items afterwards. It is obvious that when a marker indicator is selected, it should be the same used for all latent variables in each group as it strongly influences both the dynamic of adjustment constraining process and the estimation of the latent variable statistical moments (i.e., variances, covariances, means).

3.2 – Fit diagnostics

In applying SEM framework, and so that MS-SEM, tons of data-model fit indices have been developed across ages. Here, I am going to make a selection (without entering in computational details of these indices as it would be beyond the purposes of this dissertation) of the most used, although, as I am going to address at the end of this subchapter, these fit

---

[11] Cronbach's alpha solution is a merely descriptive value of the consistency of the measures but it is not assessing convergent validity of the measures themselves in loading a common factor (Bollen, 1989) and therefore it is only an initial indication of how reliable a set of measures is. Subsequent CFA is able to assess both convergent and discriminant validity (in case of more than one factor).

indices should not be taken as being the "absolute truth", but just as good indication of it.

The SEM fit indices are divided in three classes (Muller & Hancock, 2010):

a) Absolute fit indices that evaluate the overall discrepancy between observed matrix and the model-implied matrix (i.e., $\Sigma - \Sigma[\theta]$). They clearly improve as more parameters are added to the model; Examples of most popular ones are the Chi-square and the Standardized Root Mean Squared Residual (SRMR).
b) Parsimonious fit indices that still evaluate the overall discrepancy between observed matrix and model-implied matrix, but they take into account the model complexity as well. It means that they test for useful contribution of those more added parameters. Examples of most popular ones are the Root Mean Squared Error of Approximation (RMSEA) and the Akaike Information Criterion (AIC);
c) Incremental fit indices that evaluate our hypothesized model in relation to a baseline model named 'null model' where the correlations among factors are independent (i.e., close to zero). Examples of most popular ones are the Comparative Fit Index (CFI) and the Non-normed Fit Index (NNFI; as known as Tucker-Lewis Index – TLI).

Methodologists and statistical "gurus" the world over have calculated cut-off values of these indices for having boundaries criteria about assessing bad or good data-model fit as follows:

a) Chi-square values should be low and not-significant for assessing a good data-model fit, although this index has many methodological drawbacks (i.e., sensitive to violation of multi-normality assumptions, model complexity, sample size, etc.; Browne & Cudeck, (1993); Schermelleh-Engel & Moonsbrugger, 2003)) it is commonly reported as an indication of how our model-implied matrix is approaching to the observed one. SRMR values below 0.09 are considered good data-model fit (Hu & Bentler, 1999).
b) RMSEA values equal or less than .05 were considered a good fit (Hu & Bentler, 1999), in the range between .05 to .08 marginal, and greater than .10 a poor fit (Browne & Cudeck, 1993; Steiger, 1989). Small values of AIC in non-nested model comparisons are considered good and most parsimonious model (Rigdon, 1999).
c) Values greater than .90 for CFI and NNFI-TLI are considered adequate for a good model fit (Bentler, 1990) although values approaching and over .95 are preferred (Hu & Bentler, 1999).

Selection of model-fit indices is even more harsh when we deal with complex models, and nested, as the ones related to measurement invariance testing usually are.

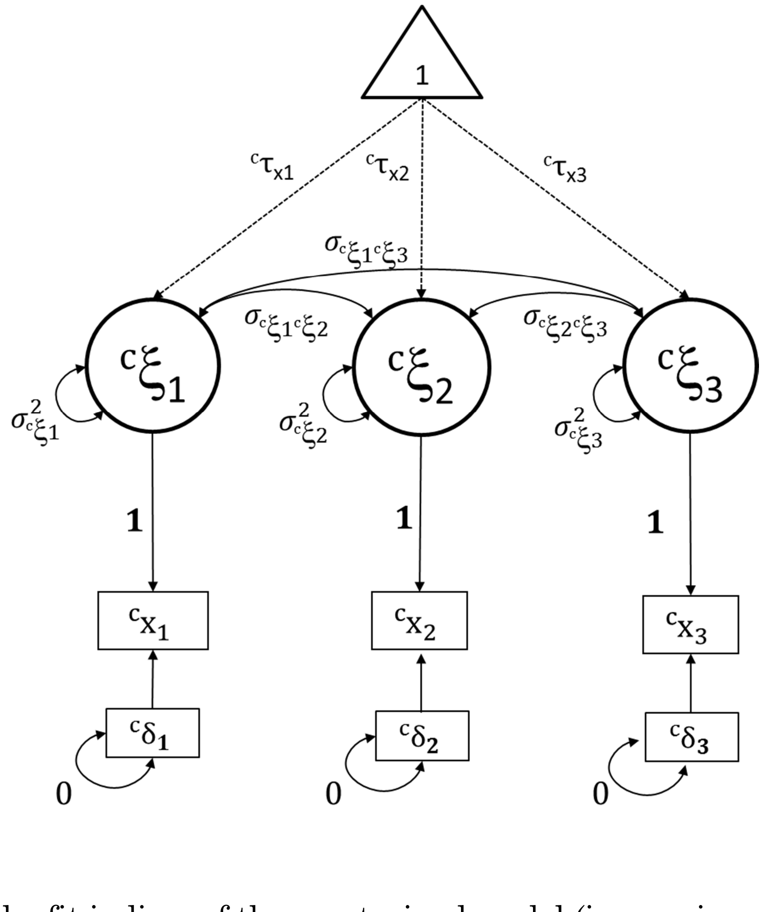
Since measurement invariance models are actually all nested constrained-models they have to be examined always as a difference between the most and the least restricted models in respect of the hierarchy of constraints. Specifically, in a context of nested-model comparison the chi-square difference test is usually applied (Steiger et al., 1985). Besides, Sörbom (1989) introduces the Modification Index (MI) computed for fixed or constrained parameters as a reflection of "…how much the overall model Chi-Square would decrease if that fixed/constrainted parameter is freely estimated" (Brown, 2006; page 119) and therefore a further useful indication of lack of invariance. But, since the Chi-square has the aforementioned limitations, although robust improvement have been made by Satorra & Benlter (2001 – they proposed a scaled chi-square correction for non-normality incorporating kurtosis of the variables), many methodologists have suggested to use the other above proposed indices in conjunction with Chi-Square difference in order to assess the hierarchical steps of invariance (Vandenberg & Lance, 2000). In addition, Cheung & Rensvold (1999) found that also differences in CFI (ΔCFI) between -0.1 and -0.2 are indicative of lack of invariance.

Hence, generally, if the fit indices (and differences) of the constrained model result 'much worse' than the ones of un-constrained model, the constrained model is invariant to those restrictions and therefore that level of invariance is achieved. However, and as you can foresee, the big deal lies in that 'much worse'. How that difference has to be 'much worse'? One quite safe answer seems to be when the aforementioned cut-off criteria remain within their acceptable boundaries with regard to the constrained model. But, on the other hand, all those fit indices are more or less, and so are the differences, affected by methodological upsides and downsides that are continuously object of simulation studies (Vandenberg & Lance, 2000) so that cannot be taken for granted as "absolute truth", as I outlined at the beginning. As a consequence, it seems that there is no way out. I do think that there is a way out and it can be shortened with this conclusion: "Ultimately, a researcher must combine these statistical measures with the human judgment when reaching a decision about a model fit" (Chen, Curran, Bollen, Kirby & Paxton, 2008). This 'human judgment' recalls what I have already outlined about having an overview on the model results as much coherent as possible with the following hierarchical points: 1) postulated theory on the constructs, measures and their possibly invariance; 2) significance and magnitude of the estimations found after the constraints process; 3) having model fits reasonably good in terms of indices and Modification Indices, but not looking for the best model fits, and so that model specifications, you can computationally handle from your data. These three reasons because of what we need is "…to assess whether a model has a reasonable correspondence to reality" (Bollen, 1989) and I may want to add that if the reality is "bad", difficult to explain and so forth, the model has to retrace this inconsistency without being 'scared' if it happens since “approximation is the soul of science” both from a theoretical and empirical point of view. Now you can get the right sense of this quoting I made at the beginning of this chapter 3. In this respect and eventually, let me conclude that if a researcher had thought in advance over those 3 points he/she could have controlled for also possibly annoying “Reviewer B” (as there is always a “Reviewer B” - quoting a comments made by Greg Hancock during the ‘Three-day Workshop on Structural Equation Modeling and Latent Variable Models’ organized by myself and held at the Department of Statistical Sciences of the University of Bologna, 12-14 September 2012) who every so often disagrees on that fit index, proposing his/her own view and making you lose time for your publication.

3.3 – Detecting measurement invariance

As we have seen in the previous chapter 2 the first step to detect presence of invariance is to assess omnibus tests of equally observed variances/covariances and observed means across groups. By doing so through MS-SEM we need to program each observed variable as it was a single latent variable. In other words, as if these single latent variables were perfectly measured by each observed variable. This strategy takes a more general name from piecewise identification (Bollen & Davis, 2009). It computationally consists in fixing each factor loading λ_i to 1 and each associated error variance δ_i to 0, since the latent variables programmed as such have no measurement error. Figure 4 depicts a complete omnibus test both for variances/covariances and mean vectors (τ_i) using the same representation for latent means model showed in chapter 2. In appendix A are shown SIMPLIS programs for running these omnibus tests with LISREL.

Figure 3.2 – Omnibus tests of observed variances/covariances and means vector un-constrained path diagram for three measures.



If the fit indices of the constrained model (i.e., variances and covariances, mean vectors constrained to be equal) are not much worse than the un-constrained ones (i.e., variances and covariances, mean vectors free to vary) we can robustly claim that the omnibus tests are assessed and our data are not affected by cultural forces. On the contrary, when the constrained model fit indices are worse than the un-constrained ones (the omnibus tests are rejected, the most common situation found, and so that we can proceed with singular steps of measurement invariance and latent heterogeneity following a nested sequence of comparison between the less un-constrained model fit indices and the much constrained model (i.e., sequential constrained model) fit indices (i.e., metric vs configural; scalar vs metric; and so forth) stopping when that step-test has been achieved. This sequential strategy is applied also for a context of partial invariance where the comparisons will be made starting with the less partially constrained model and so on. Regarding this latter situation of partial invariance (i.e, metric and/or scalar, basically) it is a common strategy of detecting which item is not invariant through checking also for Modification Indices values associated to each item (Hancock & Muller, 2012). Since MI can be defined as a Chi-square with 1 degree of freedom (df), scores associated to fixed parameters of 3.84 or greater (i.e., 6.63; Schermelleh-Engel et al., 2003) reveal critical values of Chi-square (1) at $p < .05$, so that the model may improve if that parameter is freely estimated (Brown, 2006).

In sum, a four-step strategy is recommended for detecting and dealing with measurement invariance:

1) Run CFAs for each group separately for assessing if the postulated theory on the construct(s) of interest hold, just fixing the same marker indicators for each latent in each

group. These marker indicators will be successively considered invariant because of theoretical grounds and possibly psychometric properties like item-total correlation. Since modifying the marker indicator will modify latent parameters you cannot run CFAs within each group and afterwards changing those marker indicators when running multi-sample CFAs based SEM (MS-CFA-based SEM) across groups.

2) Run MS-CFA-based SEM simultaneously on observed variables variances/covariances matrix and mean vectors in order to check for presence of invariance (omnibus tests of invariance).
3) If the omnibus test of invariance is rejected, you can proceed with further steps of invariance. By doing so, run MS-CFA-based SEM simultaneously across groups without any constraint other than the marker indicators. This is the configural model, the starting model of comparison for further steps of invariance at measure and latent level.
4) Run MS-CFA-based SEM simultaneously across groups with sequential constraints on loadings, intercepts, (uniqueness), factor variances, factor covariances where occur and make comparisons with earlier constrained model until test of that step of invariance is assessed.

Chapter 4

Implications for food choice

In this chapter will be discussed a measurement invariance application using a construct applied to functional grain products consumer choice research area and object of investigation across four European countries (Italy, United Kingdom, Germany and Finland). The data used for the analyses were carried out from HEALTH-GRAIN project (http://www.healthgrain.eu/pub/) - 6th Framework Food Research Program - I was involved in over the period 2005-2010. Precisely, I have worked through the research module on "Consumers Expectations and Attitudes on Healthy Cereal Foods" headed by prof. Richard Shepherd of the University of Surrey.

4.1 – Attitude towards using food as a medicine

The construct object of this cross-cultural study was initially defined as an attitude towards functional cereal foods 'as tools to repair flaws in healthiness of the diet' (Dean et al., 2012) and therefore as these functional cereal foods were seen like medicines. This latter reason was one of the aims of the aforementioned module project so as to study whether attitudes towards functional foods may have an influence on consumer perception towards products with health claims and thus indirectly towards diseases.

Hence, the latent construct was named 'attitude towards using food as a medicine' (AFM) and measured by four items rated on a 7-point scale (1 = 'strongly disagree'; 7 = 'strongly agree') (Dean et al., 2012). Three of the four items were selected adapting them from past works on 'reward from using functional foods' in Urala and Lähteenmäki (2007): 'I can prevent diseases by regularly eating foods with health claims', 'Foods with health claims can repair the damage caused by unhealthy diet', 'Foods with health claims make it easier to follow a healthy lifestyle'. On the other hand, the fourth item was thought to emphasize, alike the first item, possibly prevention of certain diseases with the eating of functional products, but with 'help me' instead of 'prevent': 'Eating foods with health claims will *help me* no to get some diseases'. This item has been particularly discussed during international meetings, so as to have as much cross-culturally consensus as possible with regard to its meaning in capturing the sense of an 'attitude towards food as a medicine'. These four items were also pre-tested (on 114 respondents belong to each country) within a set of other 20-items in order to verify their ability to discriminate individuals in all countries. Furthermore, although the measures used to conceptualize this AFM were already used in the already cited publication by Dean et al.'s (2012) they have been never cross-culturally analyzed at latent level. In this respect, this dissertation may constitute also an opportunity to try and provide possibly further insightful implications for considering this construct in a consumer-decision making process modeling context at cross-cultural level.

4.2 – Data and preliminary results

Data were professionally collected, using a self-reported questionnaire, by sub-contractor agencies in each country between April and May 2008. The questionnaire was put up in English, translated into the other three languages and so that back-translated into English again. The target of subjects was of consumers over 35 year old, with the same quota for men versus women and solely or jointly responsible for family's grocery shopping. The total sample size in the four countries was of 2395 respondents distributed as follows: 662 in Italy, 504 in Germany, 547 in UK, 682 in Finland.

Tables 4.1 and 4.2 respectively report descriptive statistics about moments and reliabilities as regards the observed measures. Looking at the third and fourth moments (i.e., skewness and kurtosis) in table 4.2 we may note that they are both not-so-distant from zero and so that the observed variables have slightly offended multi-normally assumptions (values of univariate skewness and kurtosis respectively over 2 and 7 might violate multi-normality assumptions – Chou & Bentler, 1995; Curran, West & Finch, 1996; Muthén&Kaplan,1985).

Table 4.1 - Descriptive statistics and reliabilities of the Attitude towards using Food as a Medicine (AFM) items by four country. AM stands for the short named 'Attitude Medicine'.

Items	ITA		GER		UK		FIN	
	Mean (sd)	**r[a]**	**Mean (sd)**	**r[a]**	**Mean (sd)**	**r[a]**	**Mean (sd)**	**r[a]**
AFM[b]								
AM1: I can prevent diseases by regularly eating foods with health claims.	4.47 (1.73)	.665	3.85 (1.84)	.719	4.29 (1.66)	.620	4.64 (1.61)	.595
AM2 :Foods with health claims can repair the damage caused by an unhealthy diet.	4.38 (1.73)	.599	3.99 (1.36)	.806	3.66 (1.70)	.605	3.95 (1.74)	.555
AM3: Foods with health claims make it easier to follow a healthy lifestyle.	4.73 (1.77)	.678	4.26 (1.32)	.793	4.57 (1.64)	.650	4.80 (1.57)	.631
AM4: Eating foods with health claims will help me not to get some diseases.	4.13 (1.77)	.690	3.84 (1.38)	.848	3.66 (1.67)	.688	4.41 (1.67)	.710
Cronbach a	.830		.899		.820		.806	
Effective Sample Size	654		504		547		671	

Note: a: item-total correlation; b: 1=strongly disagree-7=strongly agree

Table 4.2 – Skewness (Sk) and Kurtosis (Ku) of the Attitude towards using Food as a Medicine (AFM) items by four country. AM stands for the short named 'Attitude Medicine'.

Items	ITA		GER		UK		FIN	
	Sk	**Ku**	**Sk**	**Ku**	**Sk**	**Ku**	**Sk**	**Ku**
AFM								
AM1: I can prevent diseases by regularly eating foods with health claims.	-.351	-.688	.011	-1.108	-.221	-.611	-.326	-.643
AM2: Foods with health claims can repair the damage caused by an unhealthy diet.	-.320	-.739	-.103	.375	.094	-.740	-.048	-.876
AM3: Foods with health claims make it easier to follow a healthy lifestyle.	-.500	-.608	.333	.098	-.342	-.535	-.597	-.237
AM4: Eating foods with health claims will help me not to get some diseases.	-.194	-.789	.264	.173	.065	-.733	-.262	-.705
Effective Sample Size	654		504		547		671	

Table 4.3 – CFAs for Attitude towards using Food as a Medicine (AFM) - Un(standardized) factor loadings and fit indices by four countries. AM stands for the short named 'Attitude Medicine'.

Items	ITA	GER	UK	FIN	ITA	GER	UK	FIN
AFM – factor variance	**1.96**	**1.59**	**1.77**	**1.98**	**1.00**	**1.00**	**1.00**	**1.00**
	Un(standardized) factor loadings				Un(standardized) factor loadings			
AM1: I can prevent diseases by regularly eating foods with health claims.	.93 (.75)	1.11 (.76)	.88 (.70)	.77 (.67)	1.30 (.75)	1.40 (.76)	1.16 (.70)	1.09 (.67)
AM2: Foods with health claims can repair the damage caused by an unhealthy diet.	.82 (.66)	.94 (.87)	.87 (.68)	.77 (.63)	1.14 (.66)	1.19 (.87)	1.16 (.68)	1.09 (.63)
AM3: Foods with health claims make it easier to follow a healthy lifestyle.	.97 (.77)	.90 (.86)	.91 (.74)	.81 (.73)	1.36 (.77)	1.13 (.86)	1.21 (.74)	1.14 (.73)
AM4: Eating foods with health claims will help me not to get some diseases.	1.00 (.79)	1.00 (.91)	1.00 (.79)	1.00 (.84)	1.40 (.79)	1.26 (.91)	1.33 (.79)	1.41 (.84)
Goodness-of-fit indices								
Effective sample size	654	504	547	671	654	504	547	671
NT Chi-Square (df)	2.54	5.06	.28	.86	2.54	5.06	.28	.86
p-value	.28	.079	.87	.65	.28	.079	.87	.65
CFI	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
NNFI (TLI)	1.00	.99	1.00	1.00	1.00	.99	1.00	1.00
RMSEA	.020	.055	.00	.00	.020	.055	.00	.00
90% CI for RMSEA	(.000; .083)	(.000; .12)	(.000; .043)	(.000; .060)	(.000; .083)	(.000; .12)	(.000; .043)	(.000; .060)
SRMR	.0089	.0090	.0034	.0056	.0089	.0090	.0034	.0056

Tables 4.3 illustrates the CFAs singularly computed in each country. In order to define a metric for the latent AFM the loading of the AM4 item was selected to be fixed to 1 as it is the item theoretically hypothesized as cross-cultural invariant and with the strongest item-total correlation found in all countries. As we know, with fixing one loading to 1 we are assuming, by default, that it is cross-culturally invariant and it is also the indicator that drives all the process of nested constraints and structural parameter estimations (see chapter 3).

Thus, it is straightforward noticing from the left side of table 4.3, where the AM4 is the marker indicator, that: a) the factor loadings (standardized values between brackets) are all >.50 and so that the convergent validity of the measures is well assessed; b) the goodness of fit indices are again all very satisfactory. Looking at the right side of table 4.3, where factor variance has been put to 1 and all the loading were freely estimated, we can again straightforwardly noticing that the unstandardized loadings associated with AM4 are the strongest in magnitude (followed by AM3). That is was expected since AM4 was already found the most promising reliable item (see table 4.1) other than the most theoretical invariant. It means that I am "walking on a pretty safe path" in terms of invariance as in all country the item AM4 really resulted the one in which the latent trait AFM is most reflected. In practical words, respondents coming from all four countries believe that 'Eating foods with health claims will help me no to get some diseases' can better represent their AFM. As a consequence, I would bet on that item AM4 as the most cross-culturally reliable for subsequent structural estimations.

Hence we can easily conclude that covariation among the 4 measures is well represented by the common factor AFM. In other words, a latent construct really exists and it is really representing relationships among those measures. This is a first crucial step assuring that in each country the latent factor AFM hold and therefore the four items are defining it. In practical words, it means that the respondents 'mirrored' their attitude towards as a medicine in those items that are indeed its reflection, manifestation. Looking at the unstandardized factor loadings, with the exception of the one of AM4, we notice that they are fairly different across countries especially the ones of AM1 and AM2. It means that although AFM factor has been understood pretty well in all countries (standardized factor loadings >.50) there are some items that resulted as a better manifestation of AFM than others. This implies that AFM may not have been effectively understood in the same way across countries and the estimation of loadings may have been affected by cultural forces that have more conveyed answers towards certain items than other. This may generally constitutes a problem in self-reported questionnaires and it is due to many reasons, not only to back-translations mistakes, but to important cultural aspects that may affect the construct of interest. Furthermore, fixing AM4 to 1, the most reliable item, it has modified the other items-loading, as expected, making them less reliable in reflecting AFM, and possibly less invariant. But, as I aforementioned I still stand by AM4 since more shared variance has passed onto the AFM latent factor in 'quantifying' it. As a consequence, the same item AM4 can be a further 'leader' for checking invariance of the other items. By the way, you can here understand how is important the selection of the marker indicator as it really leads all the cross-cultural process that may dramatically change if the marker indicator in turn changes. And if a researcher does not have any theoretical grounds and strong reliable items either, he/she cannot work out this matter properly well ending up to misleading conclusions.

Furthermore, the error variances are indicative of which item has more variance concerning specific aspects not related to the construct of interest and variance of a random error in scoring that item. This random error still refers to the possibility that an item has not been understood perfectly well. This dichotomy (i.e., specific and random) is what unique variance stands for in which there are many things we really do not know, but we can partial out and so that control them for. Technically speaking, with squaring the standardized factor loadings we obtain the proportion of variance in the measure that has been explained by the latent factor whereas the error variance is the proportion of variance in the measure that has NOT been explained by the latent factor (Brown, 2006). The reader can easily compute, as a hands-on exercise from table 4.3, the error variances associated to each item with subtracting 1 from each squared standardized factor loading and so that finding that the less error variances are the one related to AM4 item.

Another important result we pick up from table 4.3 is the AFM factor variance. That is what we are looking for and want to speculate on. By the way, I may want to recall the meaning of factor variance at latent level that is telling us how much disperse is the error-free concept object of the study (i.e., AFM) within each group/country. In other words, how much consensus (i.e., homogeneity) exists around this concept within each group/country. In speculative way it means that the more disperse is this consensus the more people are uncertain with regard to the construct of interest. In our case, Italian and Finnish consumers, followed by British and German consumers, have an attitude much more uncertain with regards to using food as a medicine. On the other hand, German consumers seem having more consensus instead. It practically means that Italians and Finnish have an attitude in using food as a medicine that requires much more attention as it is composed of different point of views. In order to find a reason why Italian and Finnish consumers have such a common result, even though they are culturally different people, we might draw attention to exploring how AFM differs, or does not, in a subsample of respondents with relevance and not-relevance towards a particular diseases within Italy and Finland.

However, before proceeding with this further analysis, let me stop for a while and make comments about this first and interesting result on similar AFM variance found in two so-different countries. First of all, I may want to tell you how much practical may be this result even though it apparently does not seem so. As a matter of fact, should I do not have further information available, like a theory of decision making process in which I am able to introduce this factor as possibly determinant of an intention to buy functional products, I might provide some opening deduction only just from this AFM variance as follows: if I were, for instance, a business man involved in the market of functional food-products, and I wanted to sell these products in Italy and Finland, I had to pay attention to promote them like medicine in helping diseases as I might risk of not being consensually understood by most Italian and Finnish consumers. That would not be the case of British and German consumers who might fairly accept this 'food as a medicine' promotion for functional foods. But now the main concern comes up: 'am I able to make this AFM factor variance comparison among countries thoroughly well?'. The answer lies, as we know, in assessing measurement invariance at least at metric level. In other words, if the AFM concept has been understood across all group-countries equally well I am able to make cross-group-country comparisons among AFM latent variances (covariances in case of two or more factors) meaningfully well. As a consequence we are bound to proceed in the measurement invariance testing.

4.3 – Measurement Invariance results at country-level

What is important now is to run the hierarchical necessary steps to make comparisons among the AFM statistical moments at latent level: factor variances and hopefully factor means. As I had discussed in the chapter 2 the very first steps of invariance are the omnibus tests of equality covariance matrices and means of the observed measures. These omnibus tests are important for having an initial indication of the existence of a possible degree of lack of invariance in our datasets across groups (countries here).

Table 4.4 – Test of partial measurement invariance for Attitude towards using Food as a Medicine.

	Model	df	χ^2	p-value	SRMR	RMSEA	90% CI for RMSEA	TLI	CFI	AIC	Δdf	$\Delta\chi^2$	ΔCFI
Omnibus tests	1.Invariance of covariance matrices	30	332.12	.000	.068	.130	(.120; .140)	.92	.90	352.12	-	-	-
	Invariance of observed means	*12*	*213.95*	*.000*	*.000*	*.168*	*(.150; .190)*	*.91*	*.95*	*301.95*	-	-	-
Measurement Invariance	2. Configural invariance	8	8.74	.364	.015	.013	(.000; .051)	1	1	72.74	-	-	
	3. Full metric invariance (λ_{AM4}=1) ; 3vs2	17	39.98	.001	.044	.048	(.029; .067)	.99	.99	85.98	9	31.24	-.01
	3.1 Partial metric invariance (λ_{AM1} free; λ_{AM4}=1)	14	22.96	.061	.031	.033	(.000; .056)	1	1	74.96	3	(-)17.02	.01
	3.2 Partial metric invariance (λ_{AM1} and λ_{AM2} free; λ_{AM4}=1)	11	13.92	.24	.024	.021	(.000; .051)	1	1	71.92	3	(-) 9.04	0
	3.2. Partial metric invariance (λ_{AM3} and λ_{AM4} free; λ_{AM1}=1)	*11*	14.03	*.23*	*.021*	*.022*	*(.000; .051)*	*1*	*1*	*72.02*	*0*	*.10*	*0*
	4. Full scalar invariance; 4vs3	26	209.72	.000	.041	.110	(.096; .120)	.96	.96	296.72	9	169.74	-.03
	4.1 Partial scalar invariance (τ_{AM4} and τ_{AM3} fixed; λ_{AM4}=1)	14	59.86	.000	.023	.074	(.056; .094)	.98	.99	143.86	3	(-)149.86	.03
	4.2 Partial scalar invariance (τ_{AM4} fixed; λ_{AM4}=1)	8	8.74	.364	.015	.013	(.000; .051)	1	1	104.74	6	(-)51.12	.01
	5. Full Uniqueness Invariance (λ_{AM4}=1); 5vs4	38	481.81	.000	.066	.140	(.130; .150)	.92	.87	517.81	12	272.09	-.09
Heterogeneity of populations	5. Factor Variance invariance* 5vs3	20	43.91	.001	.057	.045	(.027; .063)	.99	.99	83.91	3	3.93	0

*testing for equality of variances after full metric invariance

From table 4.4 it is unequivocally evident that these tests are both rejected due to bad fit indices. It means that a degree of country-wise lack of invariance exists across the measures both in terms of means (i.e., scale locations, observed means), deviations from these means (i.e., observed variances) and the shared variation from these means as well (i.e., observed covariances). The problem is to find out where and to what extent the measurement invariance has been "spread" across different cultures as the countries, object of this study comparison, doubtless represent.

Thus, we start with configural invariance step that should be easily achieved as it is the obvious consequence of the singular CFA model reported in table 4.3. As a matter of fact, I would expect that hypothesizing the same configuration across countries the subsequent simultaneous estimation holds and it would, actually, as the goodness of fit indices associated to the configural invariance hypothesis are all very good (see table 4.4). Hence, we can proceed with metric invariance testing at full level, that is with all factor loadings fixed to be equal across groups with Italy as reference group. The new diagnostics of this metrically constrained model get worse in comparison to the ones of configural invariance (i.e., the unconstrained model) although within acceptable cut-off criteria boundaries (see chapter 3).

However, the parsimonious fit indices like RMSEA and AIC along with the chi-square difference are really increasing instead of decreasing or keeping them stable. It means that constraining all the loadings to be equal may not be the best strategy, even though the full metric invariance diagnostics are not so bad. In this respect, looking at table 4.5 where the MIs have been reported we notice that at full metric invariance level they are particularly high in correspondence of all the four items, although the greatest is in correspondence of AM1 in Germany.

Table 4.5 – Modification Indices in metric invariance for Attitude towards using Food as a Medicine - four countries. AM stands for the short named 'Attitude Medicine'. In bold when the item is constrained to be equal across countries.

	ITA	**GER**	**UK**	**FIN**		**ITA**	**GER**	**UK**	**FIN**		**ITA**	**GER**	**UK**	**FIN**
Full Metric Invariance	Modification Indices				*Partial Metric Invariance*	Modification Indices				*Partial Metric Invariance*	Modification Indices			
***AM1**: I can prevent diseases by regularly eating foods with health claims.*	.01	15.33	1.65	6.90	*AM1*	.00	.00	.00	.00	*AM1*	.00	.00	.00	.00
***AM2**: Foods with health claims can repair the damage caused by an unhealthy diet.*	3.95	4.63	.01	.59	***AM2***	4.28	8.57	.05	2.13	*AM2*	.00	.00	.00	.00
***AM3**: Foods with health claims make it easier to follow a healthy lifestyle.*	4.52	4.23	.43	.06	***AM3***	5.21	1.25	.11	1.47	***AM3***	2.55	.01	.08	3.93
***AM4**: Eating foods with health claims will help me not to get some diseases.*	*.02*	*7.21*	*.21*	*10.02*	***AM4***	*.09*	*2.37*	*.01*	*5.80*	***AM4***	2.55	.01	.08	3.87
										AM1	.30	2.97	1.80	2.10
										AM2	.30	3.14	1.82	2.13
										AM3	.00	.00	.00	.00
										AM4	.00	.00	.00	.00

Table 4.6 –Attitude towards using Food as a Medicine factor variances by level of measurement invariance – four countries.

Level of Measurement Invariance	ITA	GER	UK	FIN
Configural Invariance	**1.96**	**1.59**	**1.77**	**1.98**
Full metric Invariance	**1.97**	1.69	1.72	1.65
Partial metric invariance (λ_{AM1} and λ_{AM2} free; λ_{AM3} fixed; λ_{AM4}=1)	2.09	**1.59**	**1.79**	**1.81**

In addition, looking at the table 4.6 the new AFM factor variance computed when the loadings have been constrained to be equal across groups switch up, or down, in comparison to the ones at configural level. This is another warning that not all the items are properly metrically invariant as they are changing the estimation of the associated factor variance. In our case, since the diagnostics associated to full metric invariance are not so bad the difference in factor variance estimation are small, although exists and therefore it is worthwhile looking into a partial metric invariance testing. Thus, we proceed with relaxing one parameter at a time (since at any releasing or constraining parameter the process of adjustments occurs in changing all the other estimations as we have acknowledged from chapter 3) and so that the first one to freely estimated is the loading associated with AM1, given that it has got the greatest MI. As expected there is an improvement of the fit indices (see table 4.4), although high MIs still remains especially for the loading of the item AM2 still in Germany (see table 4.5). To keep going on this process of freeing the loadings with the highest associated MI (i.e., AM2) we get to a very nice situation both in terms of fit indices, MIs themselves, and AFM factor variance estimation that gets back to values found at the configural level (see table 4.6) with the exception of Italy, although within the same magnitude.

Hence and at the end, partial metric invariance of two items (i.e., AM3 and AM4) out of four seems to be a better result than the full metric invariance one, even though the full metric diagnostics were acceptable. It practically means that items AM1 and AM2 are not so metrically invariant as they seemed and have to be taken with caution when we want to consider them into a cross-cultural questionnaire for future research. All in all, we may conclude that the factor AFM has been partially understood in the same way across the four countries and constraining all the loading to be culturally invariant has made factor variances attenuated.

In addition and for illustrative purposes, I made a double check on the invariance of AM3 and AM4. This latter was hypothesized as marker indicator but, as we know, for this reason was not formally checked for invariance because of being fixed to 1. It is noteworthy from table 4.5 that when relaxing AM3 and AM4 the MIs increase for Germany and Uk, attenuate for Finland and decrease just for Italy (i.e., one country out of four). Hence, we might assume that AM3 and AM4 are more country-wise invariant than the other two items are since MIs resulted really good in two countries out of four (i.e., Germany and UK).

Thus now, after having assessed (partial) metric invariance and so that being able to make comparisons among factor variances we can proceed with the scalar invariance for making further comparisons at latent means level. Still looking back at table 4.4 we can easily see that full scalar invariance (most constrained model) has not been assessed as the fit indices are worse than the ones concerning metric invariance (less constrained model). It means that cultural biases in the four items exist when making comparisons among means at latent level above and beyond the true latent factor AFM. These cultural forces act even when they are not straight related to the factor (i.e., when the latent variable is zero). As a consequence, a partial scalar invariance testing is necessary in order to check which item is group-invariant and so that be taken for computing latent means differences.

As we have acknowledged from chapter 2, it would not make any sense to constrain intercepts without having equal factor loadings. Hence, since partial metric invariance has been already achieved with having AM3 and AM4 invariant, the subsequent partial scalar invariance will be tested with constraining the corresponding intercept terms of AM3 and AM4 measures (i.e.,

τ_{AM3} and τ_{AM4}) to be equal across groups, respectively. Looking at the fit indices in table 4.4 they seem fairly acceptable in comparison to full scalar invariance ones, although the parsimonious index of RMSEA is still not very nice and it means that constraining τ_{AM3} and τ_{AM4} to be equal across groups are not still giving an useful (the same) contribution (i.e., parsimony) to the whole model (i.e., the reality, as the phenomenon really is) as they were unconstrained. If we release τ_{AM3} we get back to a configural invariance situation where the latent factor AFM cannot be really tested either for metric or properly for scalar invariance since the only two constrained parameters are λ_{AM4} to 1 and τ_{AM4} to be equal across groups. But since λ_{AM4} is the marker indicator is not properly tested for being invariant and since τ_{AM4} is the only constrained intercept the subsequent only observed mean involved in the structured mean difference is the one of the item AM4. As a consequence, it seems to me pretty useless testing for latent means differences basing on a single item as it might be too much optimistic assuring that the mean of AFM is culturally invariant from those forces not directly related to it, above and beyond the strongest theoretical ground I may postulate. So that, in this case I would discourage making such a latent means comparison basing on the single item AM4 and on both AM4 and AM3 items either. Although the location of this latter item AM3 might be taken into consideration, but with caution.

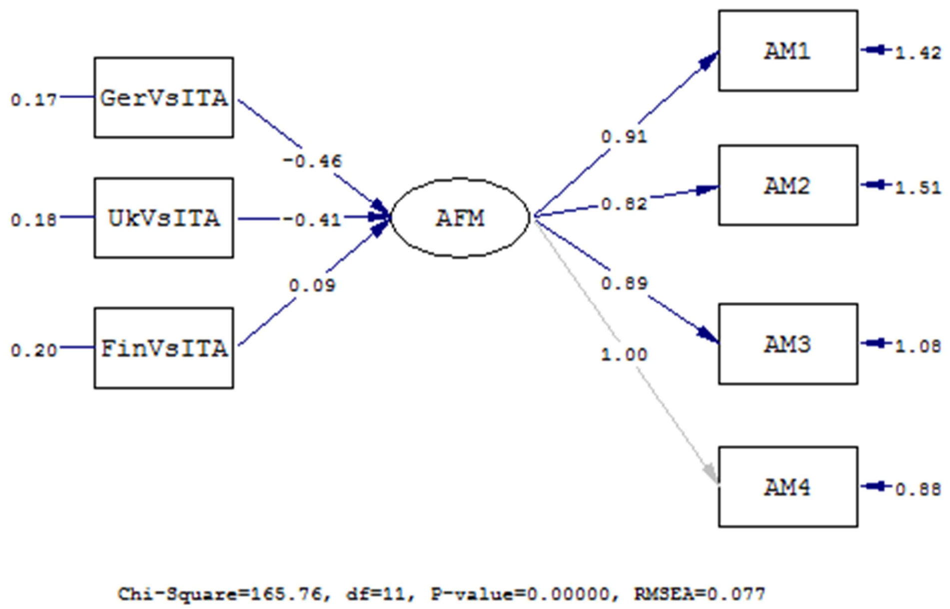
Now, if we looking at table 4.7 and, for completeness of MIMIC model, at figure 4.1 also, we interestingly notice that testing for latent means difference using a full scalar invariance assumption or MIMIC modeling will lead to pretty equal results as expected (see chapter 2), although they are both biased in comparison to the ones of partial scalar invariance. But, if to one side it is pretty evident from the bad fit indices that full scalar invariance cannot be achieved and so latent means differences cannot be made, from another side the same situation is not so equally evident from the MIMIC modeling. Someone might accept those fairly satisfactory fit indices with regard to MIMIC models and therefore taking for granted the following latent means differences showed in table 4.7 when they are indeed biased, especially for Finland, since it has assumed that all four AFM items are strong culturally invariant (i.e., metric and scalar) when they are merely weak culturally invariant (i.e., metrically invariant).

Table 4.7- Attitude towards using Food as a Medicine estimated structured means differences (with Italy as reference group) and MIMIC group model.

	Italy	Germany	UK	Finland	Diagnostics
Full scalar invariance	.00	-.41	-.40	.09*	$\chi^2(26) = 209.72$ RMSEA = .11 CFI = .96
Partial metric and scalar invariance (λ_{AM1} λ_{AM2} τ_{AM1} τ_{AM2} free; λ_{AM3} τ_{AM3} τ_{AM4} fixed; λ_{AM4}=1)	.00	-.37	-.35	.20	$\chi^2(14) = 59.86$ RMSEA = .074 CFI = .99
Partial metric and scalar invariance (λ_{AM1} λ_{AM2} λ_{AM3} τ_{AM1} τ_{AM2} τ_{AM3} free; τ_{AM4} fixed; λ_{AM4}=1)	.00	-.29	-.47	.28	$\chi^2(8) = 8.74$ RMSEA = .013 CFI = 1
MIMIC	.00	-.46	-.41	.09*	$\chi^2(11) = 165.76$ RMSEA = .077 CFI = .97

*not significant at the 95% confidence level

Figure 4.1 - MIMIC model with 3 (k-1; with k = 4) countries as dummy-coded variables and with Italy as reference group – Unstandardized solutions from LISREL output.



4.4 – Measurement Invariance results at relevance-level of type 2 diabetes in Italy and Finland

As I aforementioned in the previous sub-chapter the result of having equal AFM factor variances in Italy and Finland makes the matter very intriguing since the countries do not seem to be very culturally closed each other. To this end, it might be of interest exploring a 2x2 design with the relevance towards focused diseases[12] as a grouping dummy variable in both Italy and Finland. This approach might provide reasons about motivating this heterogeneous consensus around attitude towards food as a medicine in the two countries. By doing so, we have to split each country sample in two subgroups of relevance and not relevance and running CFAs both separately and across these two new groups. From table 4.8 we notice that AFM is well represented in both groups of not relevance/relevance within each country both in terms of factor loadings magnitude and fit diagnostics.

Besides, cultural invariance testing depicted in table 4.9 shows a complete measurement invariance for both countries, thus there are no cultural forces between the two groups affecting the observed measures, and a complete homogeneity at latent level only in Finland. As a consequence, we can make conclusions with regard to AFM between the two groups of relevance within both countries. In this respect, the core of results is the following latent statistical moments depicted on the top of table 4.8: factor variances, and factor means difference.

[12] Relevance of type 2 diabetes risk was measured on two questions with dichotomous answers (i.e., Yes, No): "Do you suffer from diabetes or do you consider yourself as having a high risk for developing diabetes?" "Do you have a relative or close acquaintance who has diabetes or difficulties in balancing their blood glucose levels?" Respondents who answered 'yes' to both relevance questions were classified as the 'relevant' group. Conversely, those respondents who answered 'no' or 'do not know' were classified as the 'not relevant' group (Dean et al., 2012).

Table 4.8 – CFAs for Attitude towards using Food as a Medicine (AFM) Un(standardized) factor loadings and fit indices by relevance in ITALY and FINLAND.

	ITALY		FINLAND	
Items	**N-REL**	**REL**	**N-REL**	**REL**
AFM – factor variance	**2.25**	**1.66**	**1.86**	**2.09**
AFM – factor mean difference	**0**	**0.40**	**0**	**0.04**[ns]
	Un(standardized) factor loadings from each CFA within each country and relevance group			
AM1: I can prevent diseases by regularly eating foods with health claims.	.87 (.68)	.97 (.80)	.74 (.67)	.81 (.68)
AM2: Foods with health claims can repair the damage caused by an unhealthy diet.	.79 (.63)	.81 (.67)	.81 (.67)	.73 (.58)
AM3: Foods with health claims make it easier to follow a healthy lifestyle.	1.02 (.78)	.91 (.74)	.79 (.74)	.84 (.71)
AM4: Eating foods with health claims will help me not to get some diseases.	1.00 (.77)	1.00 (.80)	1.00 (.88)	1.00 (.80)
Goodness-of-fit indices				
Effective sample size	329	314	338	332
NT Chi-Square (df=2)	1.69	3.87	.27	1.16
p-value	.43	.14	.87	.56
CFI	1.00	1.00	1.00	1.00
NNFI (TLI)	1.00	.99	1.00	1.00
RMSEA	.020	.055	.00	.00
90% CI for RMSEA	(.00; .10)	(.00; .14)	(.00; .054)	(.00; .093)
SRMR	.0120	.0160	.0045	.0098

Table 4.9 – Test of measurement invariance for Attitude towards using Food as a Medicine in ITALY by relevance.

	Model	df	χ^2	p-value	SRMR	RMSEA	90% CI for RMSEA	TLI	CFI	AIC	Δdf	$\Delta\chi^2$	ΔCFI
Omnibus tests	1.Invariance of covariance matrices	10	16.56	.084	.096	.045	(.00; .083)	.99	.99	36.56	-	-	-
	Invariance of means vector	*4*	*12.63*	*.013*	*.000*	*.082*	*(.034; .14)*	*.98*	*.99*	*60.63*	-	-	-
Measurement Invariance	2. Configural invariance	4	5.56	.234	.016	.035	(.00; .097)	1	1	37.56	-	-	-
	3. Full Metric invariance	7	9.23	.236	.029	.032	(.00; .080)	1	1	35.23	3	3.67	0
	4. Full Scalar invariance	10	11.31	.334	.030	.020	(.00; .060)	1	1	47.31	3	2.08	0
	5. Uniqueness invariance	14	15.67	.334	.027	.019	(.00; .059)	1	1	43.67	4	4.36	0
Heterogeneity of populations	6. Factor variance invariance	15	21.31	.127	.057	.036	(.00; .069)	.99	.99	84.22	3	4.24	0
	7. Factor covariance invariance	-	-	-	-	-	-	-	-	-	-	-	-
	8. Factor means invariance	16	32.03	.001	.091	.056	(.027; .084)	.99	.99	56.03	1	10.72	0

Table 4.10 – Test of measurement invariance for Attitude towards using Food as a Medicine in FINLAND by relevance.

	Model	df	χ^2	p-value	SRMR	RMSEA	90% CI for RMSEA	TLI	CFI	AIC	Δdf	$\Delta\chi^2$	ΔCFI
Omnibus tests	1.Invariance of covariance matrices	10	13.34	.205	.031	.032	(.00; .071)	1	1	33.34	-	-	-
	Invariance of means vector	*4*	*1.08*	*.90*	*.000*	*.000*	*(.00; .036)*	*1*	*1*	*49.08*	-	-	-
Measurement Invariance	2. Configural invariance	4	1.43	.840	.009	.000	(.00; .047)	1	1	33.43	-	-	-
	3. Full Metric invariance	7	3.27	.860	.026	.000	(.00; .036)	1	1	29.27	3	1.84	0
	4. Full Scalar invariance	10	4.20	.940	.020	.000	(.00; .014)	1	1	40.20	3	.93	0
	5. Uniqueness invariance	14	15.18	.370	.029	.016	(.00; .056)	1	1	43.18	4	10.98	0
Heterogeneity of populations	6. Factor variance invariance	15	15.23	.435	.030	.007	(.00; .069)	1	1	41.23	1	.05	0
	7. Factor covariance invariance	-	-	-	-	-	-	-	-	-	-	-	-
	8. Factor means invariance	16	15.43	.493	.030	.000	(.00; .049)	1	1	39.43	1	.20	0

Interestingly, there is an opposite factor variance situation in Italy and Finland with regard to AFM across the relevance groups. Italian respondents who have relevance towards diseases have more consensus on their attitude towards food seen as a medicine than the ones with having no relevance, and it was expected as seems coherent considering foods, and so that potential functional foods, as medicine when you have some disease. This trend is also confirmed by the factor mean difference as the relevance group of Italian consumers have more, on average, of this attitude towards food as a medicine than the one concerning no relevance towards diseases. Precisely, the relevance group of Italian consumers has significantly, on average, 0.40 units more of AFM than those Italians with no relevance. In terms of magnitude, and thereby in terms of standardized effects sizes or, better, in standard deviations units, the Italian consumers with relevance towards diseases are 0.28[13] standard deviation higher than Italian consumers with no relevance on AFM construct.

All this means that only those Italians consumers who have a potential disease may have a significant attitude towards food as a medicine and they meant it. It may be seen also as the widespread

[13] Hancock (2001) computed the estimated standardized effect size $\hat{d}$ of the structured means difference with the following formulas:

$$\hat{d} = \frac{|\kappa^1 - \kappa^k|}{\sqrt{\hat{\phi}}} = \frac{|\kappa^k|}{\sqrt{\hat{\phi}}} \, ; \, \hat{\phi} = \frac{\sum_{k=1}^{j} n_k \, (\hat{\phi}_k)}{\sum_{k=1}^{j} n_k}$$

with $|\kappa^k|$ the estimated latent means difference between each j group of comparison and the one of reference, $\hat{\phi}$ is the pooled variance estimate of the latent factor of interest ξ, n_k and $\hat{\phi}_k$ are respectively the sample size and the variance estimate of ξ in each group.

Italian vision of food as a pleasure itself and therefore seeing food as a medicine only just an additional cure of such diseases.

On the other hand, this is not the case of Finnish respondents but rather the opposite. Those Finnish consumers who have more relevance towards disease are also more disperse in a consensus on attitude towards food as a medicine although this difference between factor variances is not so high and not-so-strongly-significant (i.e., see table 4.10 where factor variance invariance has been well-achieved). So does the factor means difference that it is not significant at all. Hence, it means that both those Finnish consumers who might have potential disease and those who have not, care and worry (as the AFM factor variance is slightly higher in the relevance group than in the one of not relevance; see table 4.8) about the issue of food seen as a medicine and so that this attitude towards this topic is really heterogeneous in Finland independently of having or not relevance concerning a particular disease. As a conclusion, we have just given an answer to the reason why the factor AFM was found equally heterogeneous both in Italy and Finland. In the former country the heterogeneity is due to no relevance towards diseases, in the latter country it is due to more wide concern about having functional food as a medicine not so much delimited to consumers with diseases such as type II diabetes. Interestingly this result for Finnish consumers is in line with previous findings on functional foods in Finland where the general skepticism towards these products has decreased (Urala & Lähteenmäki, 2007) since Finnish consumers seem to be even more confident with using functional food products and thus associating these products to the concept of conventionally healthy foods (Urala & Lähteenmäki, 2007) rather than medicine.

Eventually and for completeness sake have a look at table 4.11 and figures 4.2 and 4.3 in which you can easily verify as the results from SMMs and MIMIC are absolutely identical here since complete invariance is assessed.

Table 4.11 - Attitude towards using Food as a Medicine estimated structured means differences with Not-Relevance as reference group by Italy and Finland and MIMIC group model.

		Not-Relevance	Relevance	Diagnostics
ITALY	Full scalar invariance	.00	.40	$\chi^2(10) = 11.31$ RMSEA = .020 CFI = 1
	MIMIC	.00	.40	$\chi^2(5) = 4.38$ RMSEA = .00 CFI = 1
FINLAND	Full scalar invariance	.00	.05*	$\chi^2(10) = 4.20$ RMSEA = .00 CFI = 1
	MIMIC	.00	.05*	$\chi^2(5) = 1.72$ RMSEA = .00 CFI = 1

*not significant at the 95% confidence level

Figure 4.2 - MIMIC model with relevance as dichotomous dummy (0 = not relevance; 1 = relevance) in ITALY - Unstandardized solutions from LISREL output.

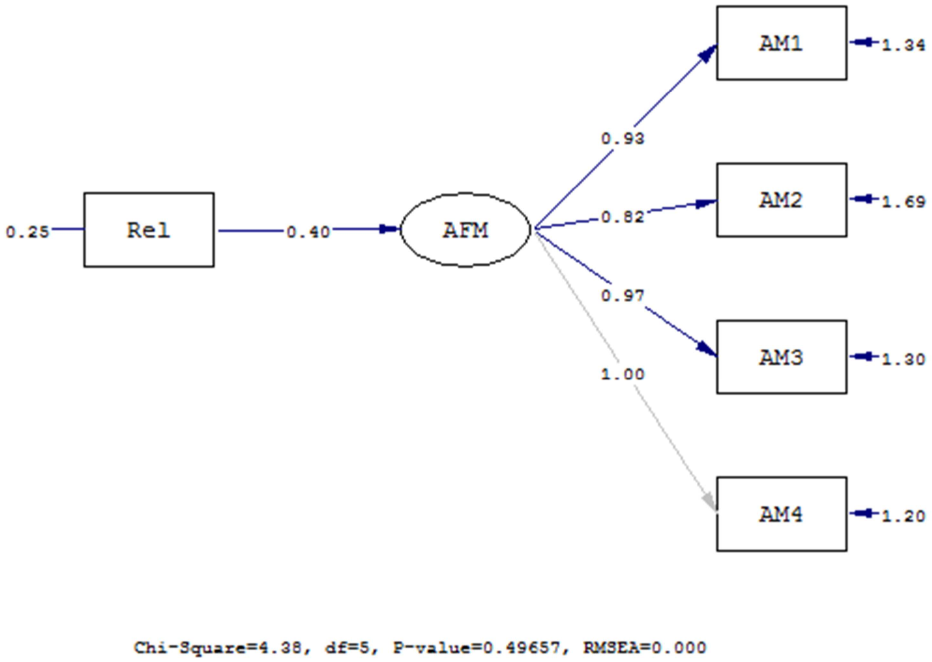
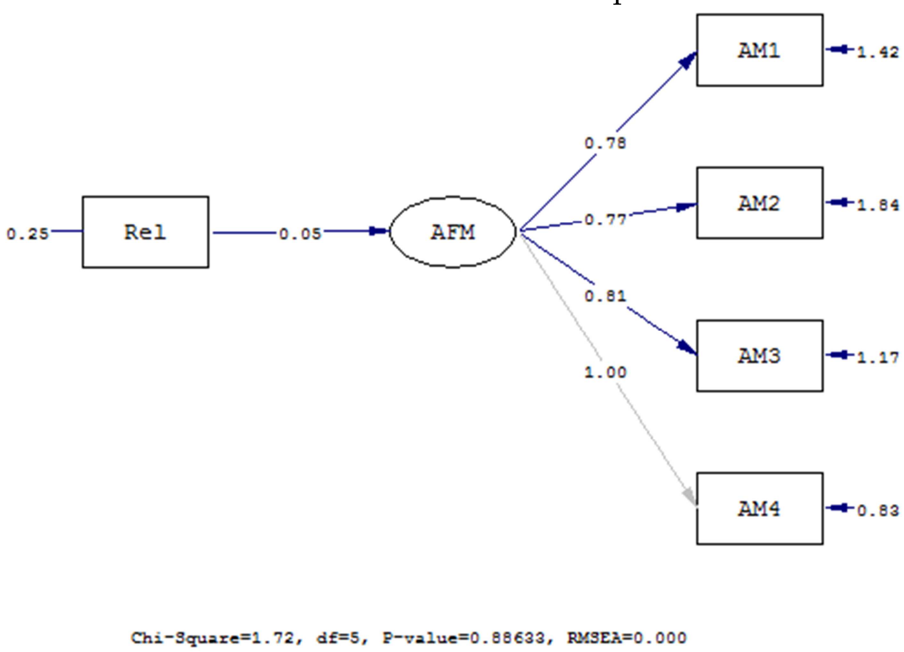


Figure 4.3 - MIMIC model by relevance as dichotomous dummy (0 = not relevance; 1 = relevance) in FINLAND - Unstandardized solutions from LISREL output.



Chapter 5

Wrap-up

Let me conclude this dissertation with my personal experience in handling the issue of detecting cultural aspects in a context of congeneric measures, carried out from self-reported instruments, reflectively loading common latent traits across groups of comparison.

First of all, I may want to clarify that cultural aspects are not ”Enemies at Gates” (title of a war movie directed by Jean-Jacques Annaud in 2001) ready to be defeated somehow, but something to control for, instead and keeping them up, I dare say. As a matter of fact, what it is speculative in common latent traits compared across different groups is to preserve both what is peculiar of each group and what that each group ‘has got’ from the latent trait itself, object of group-contrast. In practical words, I do not want to standardize, remove somehow, groups’ peculiarities, but I want to control them for, making groups comparable each other at latent level in a way that those characteristics, although in the groups, let the groups themselves distinctive with regard to that construct of interest without too much inflating, or attenuating, its meaning and therefore the true comparison, either.

From a statistical point of view, what drives latent traits (that is the estimated structural parameters such as factor loadings and intercepts, essentially), have to result, more or less, tenable across groups if I may want to successively declare that those latent traits are not too much affected by cultural aspects in terms of their own meaning, average, dispersion, interrelations and causal relationships (path coefficients) among them. In other words, I am interested in how differ a latent factor (i.e., getting to study heterogeneity at latent level) after having controlled for cultural aspects that, at measurement level, may seriously influence the meaning of that factor (i.e., measurement invariance steps).

By doing so, the aforementioned way I need, in order to make latent comparisons possible, is a simultaneous way of proceeding since cultural aspects are always evolving with times (and so are the latent traits) and therefore may be only simultaneously controlled for. In this respect, throughout a sequence of nested equal hypothesized constraints, estimated at the same time across groups of comparison, would be possible to verify whether, or not, the cultural aspects are influencing those latent traits object of the study. This simultaneous process of nested constrains initially acts alike a test of 'compound symmetry across cultures' (i.e., omnibus tests of invariance across groups) of the observed sources (i.e., measures) of covariation, to successively continue at latent level. It looks as if these group-of-measures were culturally related each other, somehow, and therefore enabling to verify whether, or not, they might be considered invariant, roughly equal, with regard to the latent traits of interest when they simultaneously 'meet' each other.

Furthermore, a simultaneous strategy is able to find out which measures are invariant and which are not throughout a partial measurement invariance approach. As a matter of fact, we do not need of having all the measures to be invariant to make proper construct comparisons at latent level, but at least one of them other than the marker indicator (Byrne et al., 1998). Even though it is intuitive that a latent trait with only one invariant measure, out of three measures for instance, is very culturally affected and, as consequence, it should be treated with caution when making these comparisons at latent level. By the way, in this latter extreme context of partial measurement invariance, theoretical ground on how that latent trait has been initially conceptualized, and so have been the measures, ought to be very substantial for supporting an invariance assumption. This is even more true in the case of having only the marker indicator as invariant. In light of this, Hancock et al. (2009) argue that measurement invariance can be still proclaimed, as minimal, precisely when the scale indicator is the only proper constraint to be invariant across groups in presence of strong theoretical cross-cultural invariant hypotheses on that single measure and thus on that latent trait.

Besides, this latter sort of theoretical conclusion is both in line on what structural equation modeling as a whole is applied for: "…structural equation modeling cannot proceed in the absence of theory" (Hancock et al., 2009, page 171) and it indirectly yields to a further defense of simultaneity way of cross-cultural empirical verification in detecting levels and causes of invariance and back to the theory.

5.1 – How to handle measurement invariance: practical advice and future suggestions

The first important step to be assessed by a researcher when he/she deals with observed measures whose covariation underneath common latent factors is the omnibus tests of the invariance of covariance matrices and means vectors of the observed measures. Above and beyond the fact that there is a wide, and obvious, agreement from methodologists about assessing these omnibus tests (tons of references reported in Vandenberg & Lance, 2000), it is extremely intuitive, since the observed measures gather all the aspects of interest both the ones strictly cultural and the ones more specifically focused on the construct object of the study as I explained in chapter 2, that before starting with working out what type of invariance my observed data may be affected, I am bound to know if lack of invariance really exists. If it does not, and my observed data are invariant, I can smoothly proceed with pooling them and/or exploring possible latent traits in terms of statistical moments and path coefficients in possibly structural models within a pooled sample and between different samples above and beyond cultural forces. It is far more intuitive that if my observed data are perfectly invariant (almost perfect fit indices associated to the simultaneous omnibus tests) the subsequent potential differences at latent level will be presumably invariant as well. Although it is worthwhile exploring them across groups in any case, especially for complex designs (e.g., gender x age class x educational level) or complex structural relationships, if the sample size permits.

Nevertheless, these two omnibus tests are both rarely seen in scientific journals concerning issues of measurement invariance at latent level and hardly they have been perfectly assessed and so that successively measurement invariance and latent heterogeneity steps are necessary.

Hence, what a researcher have to do in case that datasets have been found so strongly affected by cultural forces not to allow any group comparisons (i.e., metric invariance does not hold)? The first direct answer is that any comparison neither at latent nor at measurement levels can be defensible, so that, in my opinion a good solution would be to stay with the original data and make speculations on the reasons why this lack of invariance is occurring. By doing so, the way outs would be hierarchical and threefold: 1) getting back to the theoretical postulates that need to be revised as they have been revealed too much culturally affected; 2) doing further qualitative research on the phenomenon object of the study focused on those latent traits found non-invariant in terms of measures; 3) as a consequence of the point (2) putting up a self-reported questionnaire with items/measures as much as reliable as possible across cultures.

Frankly, I would prefer such a heuristic approach with a final result that needs to be confirmed by a simultaneous mechanism of estimation process, like MS-SEM does, for a correct approximation of the cultural aspects rather than applying pragmatic solutions like centering (with subtracting means or other values) and/or standardization of the observed variables where they are not respectively achieving scalar invariance and metric and scalar invariance together. I am arguing that as you should have really strong theoretical reasons to make such “manual” adjustment to your data renouncing to detect cultural forces and trying to understand what is really going on in your survey. On the contrary, with leaving the things as they are the researcher is able to make proper conclusions on the collected data, and so that on the real state of the phenomenon, along with providing future perspective and suggestions in controlling for cultural forces involved in a reflective-led latent trait cross-culturally studied.

In this latter respect, as I have demonstrated with the application to food choice, if the researcher collect a good enough sample size he/she will able to detect hidden cultural differences that make the measures not comparable and therefore making speculative conclusions rather than discard them. As I pointed out, a cross-cultural measure is not a mistake in my data, but a source of knowledge on which speculating on, instead. A speculation that is “mirror”, even if a “dirty mirror”, of the reality since “…We can only reject a model – we can never prove a model to be valid. A good model-to-data fit does not mean that we have the true model. We need to examine other plausible specifications that fit; we need to explore various avenues to assess whether a model has a reasonable correspondence to reality” (Bollen, 1989; page 72) and “If a model is not a reasonable approximation of reality, then the results regarding the parameters contained therein are largely without meaning” (Hancock et al, 2009) above and beyond the best pragmatic adjustment procedures I may find out.

So then, we are at the end of this dissertation. If you have arrived until here it would have meant that I have been able to get you not too much bored at long last. This would be a sound result for me already. By the way, and in the hope that this dissertation may have highlighted some interesting points, I would like to encourage your critical sense in making comments and remarks so as to share our experience and knowledge on dealing with cross-cultural studies concerning latent aspects of decision-making processes.

References

Anderson, J.C., & Gerbing, D.W. (1988). Structural equation modeling in practice: a review and recommended two-step approach. *Psychological Bulletin*, 103(3), 411-423.

Bagozzi, R.P., & Yi, Y. (1988). On the evaluation of structural equation models. *Journal of the Academy of Marketing Science*, 16(1), 74– 94.

Bentler, P.M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107, 238-246.

Bollen, K.A. (1989). Structural equations with latent variables. NY: Wiley.

Bollen, K.A., and Davis, W.R. (2009). Causal indicator models: identification, estimation, and testing. *Structural Equation Modeling: A Multidisciplinary Journal*, 16, 498-522.

Bollen, KA., and Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*. 110, 305–14.

Brown, T.A. (2006). Confirmatory factor analysis for applied research. New York: The Guilford Press.

Browne, M.W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen, & J. S. Long (Eds.), *Testing Structural Equation Models* (pp. 136– 162). Newbury Park, CA7 Sage.

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.

Chen, F., Curran, P.J., Bollen, K.A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods and Research*, 36, 462-94.

Cheung, W.G., Rensvold, R.B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 9(2), 233-255.

Chou, C.P., & Bentler, P.M. (1995). Estimates and tests in structural equation modeling. In R. H. Hoyle (Ed.), *Structural Equation Modeling: Concepts, Issues, and Applications* (pp. 37–54). Thousand Oaks, CA: Sage.

Curran, P.J., West, S.G., & Finch, G.F. (1996). The robustness of test statistics to non-normality and specification error in confirmatory factor analysis. *Psychological Methods, 1,* 16–29.

Dean, M., Lampila, P., Shepherd, R., Arvola, A., Saba A., Vassallo, M., Claupein, E., Winkelmann, M., Lähteenmäki, L. (2012). Perceived rlevance and foods with health-related claims. *Food Quality and Preference*, 24, 129-135

Fabricar, L.R, Wegener, D.T., MacCallum, R.C., Strahan, E.J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4, 272-299.

Field, A. "A Bluffer's guide to sphericity" - University of Sussex http://www.statisticshell.com/docs/sphericity.pdf

Glymour, C., Scheines, R., Spirtes P., and Kelly, K. (1987). Discovering causal structure. Orlando, FL: Academic Press.

Gregorich, S.E., (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care*, 44, 11(3), S78-94.

Hancock, G.R. (1997). Structural equation modeling methods of hypothesis testing of latent variable means. *Measurement and Evaluation in Counseling and Development*, 30, 91-105.

Hancock, G.R. (2001). Effect size, power, and sample size determination for structured means modeling and MIMIC approaches to between-groups hypothesis testing of means on a single latent construct. *Psychometrika*, 66, 373-388.

Hancock & Muller (2012) – Latent Means Models- http://www.education.umd.edu/EDMS/fac/Hancock/Course_Materials/EDMS722materials.html

Hancock, G.R., Stapleton, L.M., Arnold-Berkovits, I., (2009). The tenuousness of invariance tests within multi-sample covariance and mean structure models. In T. Teo and M.S. Khine (Eds.), *Structural Equation Modeling in Educational Research: Concepts and Applications* (pp. 137-174). Rotterdam, Netherlands: Sense Publishers.

Horn, J.L., & McArdle, J.J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, 18, 117-144.

Hu, L.T., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6, 1-55.

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

Jöreskog, K.G. (1973). A general method for estimating a linear structural equation system. In A.S. Goldberger and O.D. Duncan (Eds.). *Structural Equation Models in the Social Sciences*. NY: Academic Press, pp. 85-112.

Jöreskog, K.G, & Goldberger, A.S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, 70, 631-639.

Keesling, J.W. (1972). Maximum Likelihood approaches to causal analysis. Ph.D. dissertation. Department of Education: University of Chicago.

Kline, B.R. (2005; 2011). Principles and practice of structural equation modeling. (Second Edition; Third Edition). The Guilford Press. NY.

Lord, F.M., & Novick, M.R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Mauchly,J.W. (1940). Significance test for sphericity of a normal n-variate distribution. *The Annals of Mathematical Statistics*, 11, 204-209.

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika*, 29, 177-185.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58, 525-543.

Muller, R.O. & Hancock, G.R. (2010). Structural equation modeling. In G.R. Hancock and R.O. Muller (Eds.). *The Reviewer's Guide to Quantitative Methods in the Social Sciences* NY: Routledge, pp. 371-383.

Muthén, B., & Christoffersson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psycometrika*, 46, 4, 407-419.

Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology*, 38, 171-189.

Pollan, M. (2008). In Defense of food: an eater's manifesto. Penguin Press: NY.

Rigdon, E.E. (1999). Using the Friedman method of ranks for model comparison in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 6, 219-232.

Rigdon, E.E. (2013). Partial least squares path modeling. In G.R. Hancock, & R.O. Mueller (Eds.), *Structural Equation Modeling: A Second Course* (Second edition, pp. 81-116). Greenwich, CT: Information Age Publishing.

Satorra, A., & Bentler, P.M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, 66, 507-514.

Schermelleh-Engel, K., Moosbruger, H., Müller, H., (2003). Evaluating the fit of structural equation models: tests of significance and descriptive-of-fit measures. *Methods of Psychological Research Online*, 8, 2, 23-74.

Sörbom, D. (1989). Model modification. *Psychometrika*, 54, 371-384.

Steiger, J.H. (1989). EzPATH: A supplementary module for SYSTAT and SYGRAPH. Evanston, IL: SYSTAT.

Steiger, J.H., Shapiro, A., & Browne, M.W. (1985). On the multivariate asymptotic distribution of sequential chi-square statistics. *Psychometrika*, 50, 253-264.

Steenkamp, Jan-Benedict E.M., & Baumgartener, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, 25, 78-90.

Thompson, MS., and Green, SB. (2013). Evaluating between-group differences in latent variable means. In G.R. Hancock, & R.O. Mueller (Eds.), *Structural Equation Modeling: A Second Course* (Second edition, pp.163-218). Greenwich, CT: Information Age Publishing.

Thurstone, L.L. (1947). Multiple-factor analysis. Chicago: University of Chicago Press.

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,* 3 (1), 4-69.

Wiley, D.E. (1973). The identification problem for structural equation models with unmeasured variables. In A.S. Goldberger and O.D. Duncan (Eds.). *Structural Equation Models in the Social Sciences*. NY: Academic Press, pp. 69-83.

Urala, N. and Lähteenmäki, L. (2007). Consumers' changing attitudes towards functional foods. *Food Quality and Preference*, 18, 1-12.

Appendix – SIMPLIS SYNTAX

Omnibus test of variances and covariances

('!' is a SIMPLIS command that stands for comments)

```
!LISREL data system file (.dsf) is the file including all observed matrices and means

Group1: Italy
Observed Variables
AM1 AM2 AM3 AM4
system file from file Ita_AM.dsf
Latent Variables AFM1 AFM2 AFM3 AFM4
Sample Size is 654
Relationships
AM1 = 1*AFM1
AM2 = 1*AFM2
AM3 = 1*AFM3
AM4 = 1*AFM4
Set error variances of AM1-AM4 to 0

Group 2: Germany
system file from file Ger_AM.dsf
Sample size is 504
Relationships
AM1 = 1*AFM1
AM2 = 1*AFM2
AM3 = 1*AFM3
AM4 = 1*AFM4
Set error variances of AM1-AM4 to 0

Group 3: Uk
system file from file Uk_AM.dsf
Sample size is 547
Relationships
AM1 = 1*AFM1
AM2 = 1*AFM2
AM3 = 1*AFM3
AM4 = 1*AFM4
Set error variances of AM1-AM4 to 0

Group 4: Finland
system file from file Fin_AM.dsf
Sample size is 671
Relationships
AM1 = 1*AFM1
AM2 = 1*AFM2
AM3 = 1*AFM3
AM4 = 1*AFM4
Set error variances of AM1-AM4 to 0

Print Residuals
Path diagram
Options: MI                              !Modification Indices
End of Problem
```

Omnibus test of means vectors

```
Group1: Italy
Observed Variables
AM1 AM2 AM3 AM4
system file from file Ita_AM.dsf
Latent Variables AFM1 AFM2 AFM3 AFM4
Sample Size is 654
Relationships
AM1 = 1*AFM1
AM2 = 1*AFM2
AM3 = 1*AFM3
AM4 = 1*AFM4
AFM1 AFM2 AFM3 AFM4 = const          ! ‘const’ stands for constant, intercepts terms. This command line will be not
Set error variances of AM1-AM4 to 0  !repeated in the next groups in order to assume means to be equal

Group 2: Germany
system file from file Ger_AM.dsf
Sample size is 504
Relationships
AM1 = 1*AFM1
AM2 = 1*AFM2
AM3 = 1*AFM3
AM4 = 1*AFM4
Set Variance of AFM1-AFM4 free       ! variances and covariances are freely estimated
Set Covariance between AFM1 and AFM2 free
Set Covariance between AFM1 and AFM3 free
Set Covariance between AFM1 and AFM4 free
Set Covariance between AFM2 and AFM3 free
Set Covariance between AFM2 and AFM4 free
Set Covariance between AFM3 and AFM4 free
Set error variances of AM1-AM4 to 0

Group 3: Uk
system file from file Uk_AM.dsf
Sample size is 547
Relationships
AM1 = 1*AFM1
AM2 = 1*AFM2
AM3 = 1*AFM3
AM4 = 1*AFM4
Set Variance of AFM1-AFM4 free
Set Covariance between AFM1 and AFM2 free
Set Covariance between AFM1 and AFM3 free
Set Covariance between AFM1 and AFM4 free
Set Covariance between AFM2 and AFM3 free
Set Covariance between AFM2 and AFM4 free
Set Covariance between AFM3 and AFM4 free
Set error variances of AM1-AM4 to 0

Group 4: Finland
system file from file Fin_AM.dsf
Sample size is 671
Relationships
AM1 = 1*AFM1
AM2 = 1*AFM2
AM3 = 1*AFM3
AM4 = 1*AFM4
Set Variance of AFM1-AFM4 free
Set Covariance between AFM1 and AFM2 free
Set Covariance between AFM1 and AFM3 free
Set Covariance between AFM1 and AFM4 free
Set Covariance between AFM2 and AFM3 free
Set Covariance between AFM2 and AFM4 free
Set Covariance between AFM3 and AFM4 free
Set error variances of AM1-AM4 to 0

Print Residuals
Path diagram
Options: MI
End of Problem
```

Configural invariance

```
Group1: Italy
Observed Variables
AM1 AM2 AM3 AM4
system file from file Ita_AM.dsf
Latent Variables AFM
Sample Size is 654
Relationships
AM1 = AFM
AM2 = AFM
AM3 = AFM
AM4 = 1*AFM

Group 2: Germany
system file from file Ger_AM.dsf
Sample size is 504
Relationships
AM1 = AFM
AM2 = AFM
AM3 = AFM
AM4 = 1*AFM
Set Variance of AFM free                !with more than one factor also covariances must freely
Set Error Variances of AM1-AM4 free     !estimated and so that specified

Group 3: Uk
system file from file Uk_AM.dsf
Sample size is 547
Relationships
AM1 = AFM
AM2 = AFM
AM3 = AFM
AM4 = 1*AFM
Set Variance of AFM free
Set Error Variances of AM1-AM4 free

Group 4: Finland
system file from file Fin_AM.dsf
Sample size is 671
Relationships
AM1 = AFM
AM2 = AFM
AM3 = AFM
AM4 = 1*AFM
Set Variance of AFM free
Set Error Variances of AM1-AM4 free

Print Residuals
Path diagram
Options: MI
End of Problem
```

Full Metric invariance

```
Group1: Italy
Observed Variables
AM1 AM2 AM3 AM4
system file from file Ita_AM.dsf
Latent Variables AFM
Sample Size is 654
Relationships
AM1 = AFM
AM2 = AFM
AM3 = AFM
AM4 = 1*AFM

Group 2: Germany
system file from file Ger_AM.dsf
Sample size is 504
Relationships
!AM1 = AFM           !you can put '!' or even leave out those command lines
!AM2 = AFM           !in order to fix the loading to be equal across groups. In
!AM3 = AFM           !case of partial invariance, just leaving out the line in
!AM4 = 1*AFM         !correspondence to the loading to be fixed.
Set Variance of AFM free
Set Error Variances of AM1-AM4 free

Group 3: Uk
system file from file Uk_AM.dsf
Sample size is 547
Relationships
!AM1 = AFM
!AM2 = AFM
!AM3 = AFM
!AM4 = 1*AFM
Set Variance of AFM free
Set Error Variances of AM1-AM4 free

Group 4: Finland
system file from file Fin_AM.dsf
Sample size is 671
Relationships
!AM1 = AFM
!AM2 = AFM
!AM3 = AFM
!AM4 = 1*AFM
Set Variance of AFM free
Set Error Variances of AM1-AM4 free

Print Residuals
Path diagram
Options: MI
End of Problem
```

Full Scalar invariance

```
Group1: Italy
Observed Variables
AM1 AM2 AM3 AM4
system file from file Ita_AM.dsf
Latent Variables AFM
Sample Size is 654
Relationships
AM1 = const AFM                      !const is the intercept term
AM2 = const AFM
AM3 = const AFM
AM4 = const 1*AFM
AFM = 0*const                        !latent mean for Italy has been set to 0
                                     !as group reference
Group 2: Germany
system file from file Ger_AM.dsf
Sample size is 504
Relationships
!AM1 = const AFM
!AM2 = const AFM
!AM3 = const AFM
!AM4 = const 1*AFM
AFM = const
Set Variance of AFM free
Set Error Variances of AM1-AM4 free

Group 3: Uk
system file from file Uk_AM.dsf
Sample size is 547
Relationships
!AM1 = const AFM
!AM2 = const AFM
!AM3 = const AFM
!AM4 = const 1*AFM
AFM = const
Set Variance of AFM free
Set Error Variances of AM1-AM4 free

Group 4: Finland
system file from file Fin_AM.dsf
Sample size is 671
Relationships
!AM1 = const AFM
!AM2 = const AFM
!AM3 = const AFM
!AM4 = const 1*AFM
AFM = const
Set Variance of AFM free
Set Error Variances of AM1-AM4 free

Print Residuals
Path diagram
Options: MI
End of Problem
```

Full Uniqueness invariance
Exactly the same program of scalar invariance, but with leaving out lines of observed error variances (i.e., ! Set Error Variances of AM1-AM4 free)

Full Variance (Covariance) invariance
Exactly the same program of metric invariance, but with leaving out lines of latent variances (i.e., ! Set Variances of AFM free)

Full Covariance invariance (in case of more than one latent)
Exactly the same program of metric invariance, but with leaving out lines of latent variances (i.e., ! Set Variances of AFM free) and j latent covariances (i.e., !Set covariances of $Latent_1$ and $Latent_j$ free)

Equal of latent means - Identity
Exactly the same program of scalar invariance, but with leaving out lines of latent means differences (i.e., ! Set AFM = const)

MIMIC model approach to latent means comparison across 4 countries

```
system file from file Mimic.dsf          !Mimic.dsf includes 3 dummy variables from the
Latent Variables AFM                     !original data file with Italy as reference.
Relationships
AM1 = AFM
AM2 = AFM
AM3 = AFM
AM4 = 1*AFM

AFM = GerVsITA  UkVsITA FinVsITA    !GerVsITA UkVsITA FinVsITA are the labels I have
                                    !selected for the 3 dummies in the original data file
Print Residuals
Path diagram
Options: MI
End of Problem
```