University of Milano-Bicocca

PhD thesis
in Statistics
XXVI cycle

PhD Coordinator Professor Fulvia Mecatti

Testing cross-national construct equivalence in international surveys

Candidate:
Rossana Rosati

Supervisor:
Prof. Fulvia Mecatti
Dr. Daniele Vidoni

University of Milano-Bicocca

PhD thesis
in Statistics
XXVI cycle

Testing cross-national construct equivalence in international surveys

Applications on international civic and citizenship education survey data

to obtain the degree of PhD
this thesis will be defended in public on

Thursday 17 December, 2015 at 10.00 hours

by

Rossana Rosati

PhD Coordinator:

Prof. Fulvia Mecatti

Supervisors:

Prof. Fulvia Mecatti

Dr. Daniele Vidoni

Special thanks for her support and precious advice to Dr. Maria Magdalena Isac – University of Groningen.

Acknowledgements

I would have liked to write these words some years ago, but unforeseen and well-accepted family events – together with less welcome working priorities – contributed to delaying this moment. Nevertheless, I am now at the end of a long and challenging journey that cost me a lot, but from which I have collected much more. I am not just referring to the education and training received; it alone would have been sufficient to repay any sacrifice, but I think and cherish about the great support, friendship, and encouragement that I have received and that I will never forget.

First and foremost, I would like to express my gratefulness to my PhD Coordinator Fulvia Mecatti for providing me with an outstanding professional model and for being generous with her time, advice, and helpful support whenever I needed them.

I greatly thank my supervisor Daniele Vidoni who was for me an irreplaceable guide. He actively contributed to shape me into a researcher, and continuously encouraged and supported me, giving an example of enthusiastic and critical approach to the scientific domain. Without him this thesis would have never been completed.

I am also very grateful to my Head of Unit at the European Commission Joint Research Centre, Andrea Saltelli. A unique *beautiful mind* that I had the chance to meet.

The Department of Statistics of the University of Milan–Bicocca provided me with a challenging learning environment; I must thank all Professors, researchers, colleagues, and other people who contributed to my development.

Thinking about these last years, I really have a long list of mentors, colleagues, friends, and unexpected people to be thankful. I could not mention all of them individually.

However, in the economy of these lines, I must really thank full-heartedly the colleagues and friends Dorota, Paola A., Daniel, Giuseppe, and Luca as well as Nicoletta who were essential at different times during my PhD career.

I express my great thankfulness to Paola G. and Ivano, they are a precious constant in my personal and professional life.

Words cannot express my gratitude to Magda. Without her I would not be where I am today. It is quite unbelievable that such a brilliant researcher could also be such a good, incomparable, and admirable friend. The day I met her, my life took an unexpected and welcome turn.

To M. & D.

Table of Contents

PART I

Chapter 1

GENERAL INTRODUCTION

Over the last decades, phenomena such as globalization and a more extensive integration of the policies of some nations have ignited a growing interest in comparisons among countries to understand their performance in different policy areas. The reach of official statistics has been increasingly widened and, in many cases, it is currently possible to obtain standardized and reliable information beyond demographics or basic economic indicators. Official statistics offer a good picture of the situation in areas such as productive investment or employment, but the same is not true for other equally relevant policy objectives, such as education outcomes, social integration, or attitudes towards migrants. Yet, the relevance of the topics has opened the way to joint efforts resulting in large-scale studies across countries and cultures.

The data collected in these studies are often used for building country level indicators (e.g. country averages), which serve for international comparisons. In the field of education, the focus is frequently on measuring student achievement in basic subjects (e.g. mathematics, reading); still, international large-scale assessments (ILSA) also enable researchers, educators, and policy makers to compare educational systems regarding several other aspects such as students' values, attitudes, behavioral intentions, and beliefs. Such findings are frequently included in international reports in the form of league tables presenting country averages on different measures. These data, the rankings, and the further developed secondary analyses often become the tools used for important country comparisons and subsequent decisions.

For these comparisons to be done in a valid way, it is very important that the concepts are measured in a sufficiently equivalent way in all countries involved in the survey. Yet, statistical information on construct comparability that will justify valid comparisons of country factor means is not readily available in all cases.

In LSAs, considerable efforts are spent to ensure measurement equivalence of international test instruments (e.g. measuring student achievement), but not the same attention is devoted to the issue of equivalence of questionnaire data measuring values and attitudes. Hence, the cross-cultural generalizability of attitudinal measures and the possibility of country comparisons cannot always be reached; statistical tests of measurement invariance (MI) should be carried out to ensure meaningful country comparisons and related conclusions.

This dissertation aims to address the issue of MI of attitudinal measures and the statistical tests to be carried out to verify equivalence in ILSAs. A case is made for valid country comparisons of measures collected in cross-national surveys. Making use of the International Civic and Citizenship Education Study – ICCS conducted by the

International Association for the Evaluation of Educational Achievement – IEA in 2009, we illustrate the issue with a practical example.

The reminder of this introduction is structured as followed. The first section briefly presents the main issues of the study and introduces the research questions on which this dissertation is based. In the second section, we describe the data used in the research. In the last sections an overview is provided and the chapters of this dissertation are presented.

1.1 The current research

The current study intends to contribute to the issue of MI of attitudinal measures in ILSA's.

Awareness of the measurement invariance issue has progressively increased as proven by studies concerning equivalence, recommended practices, and applications of tests (Vandenberg & Lance, 2000; Schmitt & Kuljanin, 2008; Steenkamp & Baumgartner, 1998; Byrne & Stewart., 2006). Nevertheless, different aspects of measurement equivalence are still rarely evaluated and data are used without the due concerns and cautions in country rakings, leagues tables, and secondary analysis.

The dissertation takes as an example the data collected in the International Civic and Citizenship Education Study – ICCS conducted by the International Association for the Evaluation of Educational Achievement – IEA in 2009 and its further latent variables analyses and reporting.

Cross-country validity of the defined constructs has been a priority for the ICCS team since the trial stage (Schulz et al., 2011), but the actual invariance of all the measures involved could not been tested. '*The implication is that most scales in ICCS are still to be validated in order to compare constructs with some confidence across countries*' (Weziak-Bialowolska & Isac, 2014).

This work investigates the non-cognitive outcomes concerning students' attitudes towards immigration, collected through the ICCS 2009 questionnaire (ICCS 2009 International Report; Schulz et al., 2008). Apart from the data structure, the subject has attracted our attention because of the higher and higher mobility at European level and the more recent migration phenomena.

The analysis has been conducted at European level with regard to two formats of the instrument: the six-items battery of the original ICCS 2009 study and the five items battery used by the ICCS 2009 team to construct the *students' attitudes toward equal rights for immigrants* scale as reported in the ICCS 2009 International Report and the ICCS 2009 European Report (ICCS 2009 European Report, 2011, p. 92).

Moreover, in assessing the measurement invariance of these measures we took further cues from the mentioned league tables. More specifically, we take note that different scales are distinguished for native and immigrant background students, and we have operationalized the topic addressing four main research questions as follows:

a. Can country average levels of student attitudes toward equal rights for immigrants be compared with confidence among all European countries and/or relevant sub-groups of countries?

b. Can such comparisons be carried out also for sub-groups of students such as the non-immigrant/native students in these countries?

c. Can country average levels of student attitudes toward equal rights for immigrants also be compared when we consider only the group of students with an immigrant background in these countries?

d. Is it possible to identify reference country/variables sub-groups for which measurement invariance holds at higher levels?

1.2 Data

For the purpose of this dissertation, we use the information collected in the International Civic and Citizenship Education Study – ICCS conducted by the International Association for the Evaluation of Educational Achievement – IEA in 2009.

The ICCS provides data about civic knowledge, citizenship competences, values, and attitudes of Grade 8 students (14-year-olds) in 38 countries in Europe, Asia and Latin American. The ICCS rules concerning target population implied that if the average age of students in Grade 8 was below 13.5 years in a country, as in the case students started formal schooling at age five, the target grade became Grade 9 (ICCS 2009 International Report, 2011).

The survey provides data on the measurement of both cognitive and non-cognitive student outcomes, as well as data concerning the background of students and the context (i.e. school and family) in which the student civic competences are developed.

We approached the research topic taking as example the measure of students' attitudes toward equal rights for immigrants collected at European level (European Union Countries participating in European Module of the survey and Switzerland).The choice of this particular grouping of countries is motivated by the practical example considered as a starting point for this empirical exercise (data reported for the ICCS 2009 European module) as well as by presumed cultural similarity of the European countries as opposed to the entire international sample of countries surveyed in ICCS 2009 (including Latin American and Asian countries) which may, in principle, increase the possibility of accurate country comparisons.

1.3 Overview of the dissertation

The current dissertation is composed of four main parts and a final overview relating to research conclusions, limitations of the work, and suggestions for future studies in the field.

Chapter 2 introduces the main issue of the comparative use of data collected in large-scale surveys across countries and cultures, in particular with regard to questionnaire data. A brief history of the international large-scale assessment – ILSA is drawn and the growing interest for this kind of studies both for research and policy-making objectives is illustrated. In order to establish whether country scores on a scale are comparable, we apply to the notion of measurement invariance (MI). As documented in Chapter 2, measurement invariance implies that scale scores from different countries measure the same construct with the same measurement unit and reference point.

In Chapter 3, we present a comprehensive literature review concerning measurement invariance and measurement invariance testing in a multi-group confirmatory factor analysis - MGCFA framework. As detailed in the chapter, in the factor analysis framework three levels of measurement invariance can be distinguished and will be tested: a) configural invariance - common factors are associated with the same items across compared groups; b) metric invariance - the factor loadings across groups are invariant, that is the common factors have the same meaning across groups and the same measurement unit; c) scalar invariance - factor intercepts are identical across groups. This later level of equivalence enables meaningful comparisons of the group means, as the factors have both the same measurement unit and the same reference point. Only meeting the criteria of scalar invariance will justify country comparisons. In the event, the criteria is not met, alternative strategies (e.g. partial measurement invariance) could be investigated and tested.

Chapter 4 provides an overview of the International Civic and Citizenship Education Study – ICCS conducted by the International Association for the Evaluation of Educational Achievement – IEA in 2009, which is the source of the data used for our research. Data on Grade 8 (approximately 14 years of age) students' citizenship competences from 38 countries were collected. In particular, we describe in detail research concerns and results relating to the *students' attitudes toward equal rights for immigrants* at European countries level. Referring to the available data and the earlier defined research questions the chapter describes the method of our empirical study.

In Chapter 5 research results are illustrated. The methodology and statistical analyses presented in the previous chapters are applied to multiple sets of data according to the two batteries of items considered and all the research questions: all students in all European countries and sub-groups of countries; sub-groups of students such as non-immigrant/native students; and, students with an immigrant background. The estimation takes into account the specific properties of the data and a detailed account of the data analysis strategy is provided. The results are discussed for both instrument formats, for the entire sample and, the sub-samples.

Finally, in Chapter 6 the main findings of the research work are summarized. The core conclusions concerning the research questions are provided and critically discussed. Some limitations of our current study are indicated and some suggestions are made with regard to possible further research avenues in the field of measurement invariance of questionnaire data collected across groups and cultures.

Chapter 2

INTERNATIONAL LARGE-SCALE ASSESSMENT AND MEASUREMENT INVARIANCE

This chapter introduces the main issue of the comparative use of data collected in international large-scale survey – ILSA across countries and cultures, in particular with regard to questionnaire data. In the first section, a brief history of ILSAs is drawn while in the following one, the growing interest for this kind of studies both for research and policy-making objectives is illustrated. In order to establish whether country scores on a scale are comparable, we apply to the notion of measurement invariance. An introduction to the issue is given in section three, where we point out that different levels of measurement invariance – MI can be achieved – i.e. configural invariance, metric invariance, and scalar invariance – but for a meaningful comparisons, scale scores from different countries must measure the same construct with the same measurement unit and reference point.

2.1 International large-scale assessment (ILSA) – brief history

International large-scale assessments (ILSA) are '*large-scale survey of knowledge, skills, or behaviors in a given domain*' (Kirsch et al., 2013, p. 1) generally standardized across countries and/or different populations and cultures. These assessments take into account 'group scores' and comparisons between groups/countries and differ from large-scale testing programs mainly focused on measuring individuals.

Over the last decades, globalization and a more extensive integration of the policies of some nations called for a growing interest in large-scale comparative studies. Progressively but rapidly, the analyses and domains of investigation of these international studies at system level have broadened to include a high number of student learning areas, skills, knowledge, and attitudes and have reached several groups of countries (Kamens & McNeely, 2010; Kamens, 2013).

The origins of ILSAs date back to the early 1960s. Following a pioneering idea arisen during a scholars' meeting in Hamburg at the UNESCO Institute for Education (1958), between 1959 and 1962 a pilot Twelve-Country Study focused on five domains was conducted to investigate the feasibility of undertaking more extensive assessments of educational achievements[1]. The very first line of its final report is quite symptomatic '*The present study may well be described as* ***an unusual addition*** *to the literature of education*' (Foshay et al., 1962, p. 5).

On the basis of the positive results of this preliminary study, the International Association for the Evaluation of Education Achievement (IEA) organized the First International Mathematics Study (FIMS) on 13-year-old and pre-university students. Data were collected in 1964.

[1] http://www.iea.nl/pilot_twelve-country_study.html

At the same time, in the United States an advisory group was constituted (chaired by John Tukey head of the Department of Statistics at Princeton University) and its work led to the National Assessment of Educational Progress (NAEP). The NAEP was, and remains, a large national representative assessment of US students' achievements in various subject areas. The NAEP researchers firstly introduced methodologies such as Item Response Theory (IRT) or the balanced incomplete block spiraling (BIB) and developed the use of marginal estimation procedures and covariance information. In the 1980s, these innovative methodologies allowed to progress towards more complex questions concerning construct domains, population generalizations, and scale comparisons across multiple test forms (Kirsch et al., 2013), which allowed for overcoming the simple item analyses and the 'descriptive' assessments.

These methodologies were adopted by both the International Association for the Evaluation of Educational Achievement (IEA) and the Organization for Economic Co-operation and Development (OECD), which are currently running some of the main ILSAs at worldwide level.

Since its first large-scale assessment, IEA has conducted more than 25 studies in different domains of student achievements. Today, the IEA continuous cycles for the Trends in Mathematics and Science Study – TIMSS (started in 1995) and the Progress in Reading Literacy Study – PIRLS (launched in 2001) attract country participants all over the world and representing most of the worldwide GDP.

For instance, 40 systems participated in TIMSS 1999, *'66 systems for TIMSS 2007, and 79 participants for the TIMSS 2011 assessment, which includes a number of benchmarking US states and other subnational systems such as Dubai*' (Wagemaker, 2013, p. 18).

Table 2 - 1 History of ILSAs

Research Organisation	Assessment/Study	Year(s)
IEA	Pilot Twelve-Country Study	1960
IEA	First International Mathematics Study (FIMS)	1964
IEA	First International Science Study (FISS) / Six Subject Survey: Science	1970-971
IEA	Six Subject Survey	1970-1971
IEA	Second International Mathematics Study (SIMS)	1980-1982
IEA	Classroom Environment Study	1981-1983
IEA	Second International Science Study (SISS)	1983-984
IEA	Written Composition Study	1984-1985
IEA	Preprimary Project (PPP)	1987-1989, 1992, 1995-1997
IEA	Computers in Education Study (COMPED)	1989, 1992
IEA	Reading Literacy Study	1990-1991
OECD	International Adult Literacy Survey (IALS)	1994, 1996, 1998
IEA	Language Education Study	1995
IEA	Trends in International Mathematics and Science Study (TIMSS)	1995,1999, 2003, 2007, 2011
SACMEQ/International Institute for Educational Planning (IIEP)	The Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ I, II, III)	1995-1998, 1999-2004, 2005-2009
IEA	Second Information Technology in Education Study (SITES-M1)	1998-1999
IEA	TIMSS 1999 Video Study	1998-2000
Latin American Laboratory of Assessment of Quality of Education (LLECE)/UNESCO	First International Comparative Study (PERCE)	1998
IEA	Civic Education Study (CIVED)	1999
OECD	Program for International Student Assessment (PISA)	2000, 2003, 2006, 2009, 2012
IEA	Progress in International Reading Literacy Study (PIRLS)	2001, 2006, 2011
IEA	Second Information Technology in Education Study Module 2 (SITES-M2)	2001
NCES/OECD	Adult Literacy and Lifeskills (ALL) Survey	2003, 2006-2008
IEA	Second Information Technology in Education Study (SITES)	2006
IEA	Teacher Education and Development Study in Mathematics (TEDS-M)	2007-2008
OECD	Teaching and Learning International Survey (TALIS)	2007, 2008, 2013
IEA	TIMSS Advanced 2008	2008
IEA	International Civic and Citizenship Education Study (ICCS)	2009
IEA	YES "Youth in Europe"	2010, 2011, 2012
OECD	Program for the International Assessment of Adult Competencies (PIAAC)	2011
OECD	Assessment of Higher Education Learning Outcomes (AHELO) feasibility study	2010-2013
IEA	International Computer and Information Literacy Study (ICILS)	2013
LLECE/UNESCO	Latin America Second Regional Comparative and Explanatory Study (SERCE)	2005-2006
LLECE/UNESCO	Third Regional Comparative and Explanatory Study (TERCE)	2013
IEA	ECES (Early Childhood Education Study)	2014, 2015-2016 (projected)

Source: William & Engel, 2013, p. 217

IEA recently organized other major assessments such as the Teacher Education Development Study-Mathematics – TEDS-M, about Mathematics teachers' competences in primary and lower-secondary schools in 17 countries, the Civic Education Study - CIVED, and its further cycle the International Civic and Citizenship Education Study – ICCS in 2009.

ICCS 2009 is a comparative assessment of students' knowledge and conceptual understanding, dispositions, beliefs, attitudes, and behaviors related to civics and citizenship. Additional questionnaires collect data and context information from different stakeholders (i.e. teachers, school principals). IEA has recently launched a new round of the program, the ICCS 2016 (http://iccs.iea.nl/).

In the late 1990s, the OECD launched the Programme for International Student Assessment – PISA for the assessment of 15-year-old students in Mathematics, Science, and Reading in over 30 countries. Around 510,000 students in 65 economies took part in the PISA 2012 representing about 28 million 15-year-olds globally. PISA 2012 also grew in terms of '*range of domains assessed, with cross-curricular areas such as problem solving and financial literacy being added to the assessment*' (Kirsch et al., 2013, p. 4). More than 70 economies have signed up to take part in the assessment in 2015, which focuses on Science.

In 2012, the OECD also assessed adult competencies through the first cycle of the Programme for the International Assessment of Adult Competencies – PIAAC across 25 OECD countries in 33 languages. The assessment regarded literacy and basic numeracy skills, and it was the first computer-based household survey of adults (aged 15-65).

The OECD surveys also include the Teaching and Learning International survey (TALIS) concerning teaching and learning environment in school and teachers' working conditions. The target population is teachers at the secondary school level, and the study intends '*to measure study participants on latent variables that deal with attitudes,*

perceptions, and experiences […] summarized in terms of measurement model-based scale scores' (Rutkowsky & Svetina, 2013, p. 2).

In these pages, we specifically focused on educational assessments because of their undeniable relevance and particular interest for the purpose of this dissertation, but the reach of international comparative researches is much wider. For example, UNICEF has already carried out three cycles of the Multiple Indicator Cluster Survey, with the aim of observing women and children conditions at international level. Likewise, the World Health Organization – WHO has carried out a World Health Survey on over 70 countries in 2002-04.

2.2 Growing importance of ILSA

International large-scale surveys - ILSAs contribute to describing populations with regard to a specific field, and they offer unique opportunities for comparing a comprehensive range of achievements, values, behaviors, abilities, and opinions of large groups of people within and across countries.

Their development responds to the challenging questions posed by researchers, policy-makers, and general public all over the world (Kirsch et al., 2013; Stanat & Lüdtke, 2010). In fact, they provide valuable benchmarking information for researchers and policy-makers in different fields and across countries and cultures (Rutkowski et al., 2013). In various cases, ILSA's results have reached the large public and stimulated media debate (i.e. 'TIMSS shock' and 'PISA shock' in Germany[2]).

Going beyond the mere aim of measuring educational outcomes (Robitaille & Garden, 1989; Postlethwaite & Ross, 1992; Postlethwaite & Wiley, 1992), ILSAs currently contribute to the development of evidence-based policies and stimulate countries to progress or mark their unexpected achievements (Lockheed & Wagemaker, 2013; Hanushek & Woessmann, 2009; Beatty & Pritchett, 2012).

Today, large-scale surveys, including ILSAs, are recognized as a prime way for learning about system quality and understand '*the contexts in which the achievements of a country's economic competitors take place*' (Wagemaker, 2014, p. 19) and consequently improve through the sharing of best practices. '*We have become an 'assessment society*' […] *developed in previously almost unimaginable ways'* (Broadfoot & Black, 2004, p. 19).

[2] 'TIMSS shock' dates winter 1996/1997 when very poor results were highlighted for German fourth and eighth-graders level (Lehmann, 2011). In December 2001, followed the 'PISA shock' when Germany ranked at the lower end of the comparative scale. German 15-year-old students did poorly in all of the three tested subjects.

As stated by William & Engel (2013) 'borrowing' for improving is not new (Sadler, 1900; Noah & Eckstein, 1969; Postlethwaite, 1999), but globalization has amplified opportunities for external referencing, frequently presented as benchmarking across countries (Phillips & Ochs, 2003).

Following the Board on International and Comparative Studies in Education of the National Academy of Sciences (BICSE) as reported by Heyneman and Lee (2014), among the main contributions of ILSA, it is possible to identify:

- A wider laboratory on which to observe the consequences of different domestic policies and practices;
- International test information, which helps define what is realistic in terms of domestic education policy;
- The identification of concepts overlooked at local level;
- The highlighting of important questions and challenge long-held assumptions.

As independent references, ILSAs and, generally speaking, the connected international rankings have also generated animated public debates and have attracted extensive attention from the media, with the creation of new categories of winners - high scoring countries - and losers - countries lower averages (Williams & Engel, 2013).

In fact, the use of individual indicators, synthesizing scores representing overall country performance and commonly used to scale nations in comparative league tables, has showed an incredible potential for policy-makers and media and for the academics. Still, the reliability of these comparative league tables and rankings depends on the underlying assumptions and the rigorous analysis of the data used.

Therefore, a necessary prerequisite to meaningfully compare cross-country and cross-cultural survey results is the effective measurement of same constructs of interest and the use of the same instruments for collecting data across nations, and '*this is especially true for subjective attributes such as values, attitudes, opinions, or behavior*' (Davidov et al., 2014, p. 55).

However, international studies have not always paid the necessary attention to verifying actual cross-country comparability (invariance), which is often assumed as a sort of implicit capacity of the data collected by using the same instruments (i.e. questionnaires or tests).

2.3 The comparability issue - Measurement and invariance

Measurement refers to the act of '*ascertaining the size, amount, or degree of (something) by using an appropriate instrument or device marked in standard units*' (adapted from *Oxford Advanced Learner's Dictionary*). Researchers try to profile and translate into a '*common language*' the traits of interests, and this '*common language*', which should justify valid comparison, is in quantitative terms. Still, the comparability only holds if concepts are measured in a sufficiently invariant way, which means that the same constructs are measured with the same equivalent measurement instrument in all the countries involved in a survey (see Meredith, 1993; Chen, 2008; Weziak-Bialowolska & Isac, 2014).

'*The crux is that cross-group comparisons require pre-requisite assumptions of invariant measurement operations across the groups being compared*' (Vandenberg & Lance, 2000, p. 9), thus the equivalence (or invariance – the two terms are used as synonyms) in measurement should be a critical and major concern in comparative researches (i.e. Hui & Triandis, 1985; Byrne & van de Vijver, 2010).

Horn and McArdle (1992) authoritatively argue about *measurement invariance or measurement equivalence* as follows:

> *'The general question of invariance of measurement is one of whether or not, under different conditions of observing and studying phenomena, measurements yield measures of the same attributes. If there is no evidence indicating presence or absence of measurement invariance […] findings of differences between individuals and groups cannot be unambiguously interpreted'* (p. 117).

Intuitively, measurement equivalence is valid when '*members of different populations who have the same standing on the construct being measured receive the same observed*

score' while ' a test violates invariance when two individuals from different populations who are identical on the construct score differently on it' (Schmitt & Kuljanin, 2008, p. 210).

Less intuitive, and more sensitive, is the operational translation of this goal, which should ensure that the discovered differences and scalar positions in cross-national researches are entirely depending on country and cultural differences in the measured construct, and they are not due to other disturbing causes. In other words, the respondents' results on the measurement instrument (in case of survey, questionnaires) are not affected by other noise factors, specific at country level, which would lead to non-equivalence in measurement.

Non-invariance can depend on different reasons (Byrne & van de Vijer, 2010, Schulz, 2008; Rutkowsk & Svetina, 2013, .Weziak-Bialowolska, 2014). Traditionally, language differences and the related need of translating the instruments are recognized as possible sources of non-equivalence. Strict verifications are normally planned and a great attention is given to the correct translation of the instruments, but even small differences in the meaning of a term can have a significant impact on the item responses (Schulz, 2008; Mohler et al., 1998; Harkness et al., 2004). At local level, diverse modalities of administration i.e. of questionnaires or dissimilar implementing procedures could result in causes of non-equivalence. Apart from these issues, which most international studies identify as possible sources of non-invariance and tend to prevent by rigorous reviews and strict implementing criteria, non-equivalence can arise from cultural diversities across surveyed countries.

Different cultural behaviors and habits at country level can lead to different approaches to an item statement; dissimilar characteristics of the educational system or the national context may condition the way in which answers are understood and interpreted (Schulz, 2008; Schulz, 2003; Kankaras et al., 2010). Therefore, survey results could be affected by

bias in measurement, which is a systematically biased score on the measured construct, independent of the fact that the instruments are correctly employed (van de Vijver & Leung, 1997; van de Vijver & Tanzer, 1997).

Three main kind of bias may affect cross-country studies (Byrne, 2003):

- *Item bias* occurs at specific item level because of different cultural habits and does not necessarily have an impact on the general measurement of the constructs.

- *Construct bias* refers to an actual dissimilarity of the investigated construct. Thus, the construct meaning is not shared – or there is only a partial overlap – across countries, which leads to evident limitations to cross-cultural comparisons.

- *Method bias* is connected to the methodological aspects of a large-scale research. Key examples of method bias are extreme response bias – ERB, which implies a systematic positioning at the limits of the rating scale (i.e. very good / very bad) and the tendency to acquiescence (also called agreement tendency or yea-saying), which is the tendency to systematically agree with the item statements (Schulz, 2008; Kankaras & Moors, 2010).

Concerning measurement invariance in a factor analysis framework, which is relevant for this research work, three main levels of non-equivalence can be distinguished:

- configural invariance - common factors are associated with the same items across compared groups;
- metric invariance - the factor loadings across groups are invariant that is the common factors have the same meaning across groups and the same measurement unit;
- scalar invariance - factor intercepts are identical across groups. This level of equivalence enables meaningful comparisons of the group means as the factors have both the same measurement unit and the same reference point.

As showed in the following of this dissertation, only meeting the criteria of scalar invariance will justify full country comparisons.

Chapter 3

MEASUREMENT INVARIANCE

Chapter 3 gives an overview of the invariance issue and the assessment of measurement equivalence. In the first section a comprehensive literature review concerning measurement invariance and measurement invariance testing in a multi-group confirmatory factor analysis - MGCFA - framework is provided. The second section presents, still in the factor analysis framework, three main levels of measurement invariance and the connected statistical tests: a) configural invariance - common factors are associated with the same items across compared groups; b) metric invariance - the factor loadings across groups are invariant, that is the common factors have the same meaning across groups and the same measurement unit; c) scalar invariance - factor intercepts are identical across groups. This later level of equivalence enables meaningful comparisons of the group means. The next section distinguishes between measurement invariance and structural invariance. Last section offers possible alternative strategies

(e.g. partial measurement invariance) in case the measurement invariance requirements are not met.

3.1 Measurement invariance

As previously discussed, cross-group comparisons are meaningless without assuming measurement invariance, for this reason adequate equivalence tests and procedures should be applied to avoid ambiguous interpretation of data and improper conclusions.

Awareness of the measurement invariance issue has grown as proven by studies concerning equivalence, recommended practices, and applications of tests (Steenkamp & Baumgartner, 1998; Byrne & Stewart., 2006; Davidov et al. 2014). Articles have been published in quite diverse research fields (i.e. educational research, organizational research, and medical care), but researchers in social and behavioral sciences show the most interest in the topic (see for a review Vandenberg & Lance, 2000; Schmitt & Kuljanin, 2008).

Nevertheless, various aspects of measurement equivalence are still rarely evaluated. In addition, a common definition about MI has not been agreed upon yet. The nomenclature can vary '*considerably across studies for all ME/I tests and usually reflected the authors' particular substantive concerns*' (see Vandenberg & Lance, 2000, p. 36).

One of the most influential works in the field was carried out by Meredith (1993). He proposes four hierarchical levels of measurement equivalence: configural equivalence, weak equivalence, strong equivalence, and strict equivalence. Similarly, Steenkamp & Baumgartner (1998) distinguish increasingly levels of measurement invariance, but the authors refer to weak invariance as metric one, while strong equivalence is called scalar.

Van de Vijver & Leung (1997) define three planes of invariance: construct equivalence, measurement unit, and scalar equivalence. Construct equivalence should be considered as the basic condition to proceed in any kind of comparison.

Gregorich (2006) suggests a preliminary stage, the dimensional invariance, which should be tested before any other.

Measurement invariance establishing is conceived as a hierarchical process by most researchers. Higher invariance levels are characterized by more severe constrains on measurement parameters, and define more restricted models to be compared. Thus, equivalence across groups is progressively more demanding, but any further validation step allows more extended cross-group comparisons

Configural invariance (first step) '*implies that the concept has the same cross-group meaning but is not sufficient for meaningful statistical comparison*' (Weziak-Bialowolska, 2014, p. 56). The weak invariance[3] (second step), or its variations called also pattern invariance (Meredith & Teresi, 2006) or metric factorial invariance (de Jong et al., 2007; Davidov, 2008), assures that the measurement unit is analogous across the studied countries, and it implies the same one-unit difference. This level of equality may be sufficient for researchers interested only in construct validity. The scalar invariance (third step) (de Jong et al., 2007; Davidov, 2008), or strong factorial invariance (Meredith & Teresi, 2006), allows valid cross-group comparisons of the scores (i.e. country rankings based on mean scale scores), because it does guarantee the same origin of the scale.

Following Byrne & van de Vijver (2010) - see also Byrne et al., (1989); Byrne (2012) - two different issues must been distinguished: measurement equivalence and structural equivalence. The former is related to the observed variables and the extent of their

[3] Horn & McArdle (1992) define configural invariance as a weak one.

relation to the latent factors (generally speaking the CFA model), while the latter concerns the relations among the latent factors (unobserved variables).

The structural model '*specifies the manner by which particular latent variables directly or indirectly influence (i.e 'cause') changes in the values of certain other latent variables in the model*' (Byrne, 2012 p. 14). Therefore, in principle, tests for measurement invariance should be planned before assessing structural equivalence, i.e. the analysis of the constructs should precede the check of the their possible relations (Anderson & Gerbing's, 1988; Byrne & van de Vijver, 2010)

Statistical assessment of measurement invariance (level of accepted equivalence) strictly depends on the comparison purposes and on the research objectives of the practitioners. Although these theoretical and applied measurement works can vary, generally, three major testing approaches are traditionally implemented to test measurement equivalence (Davidov, 2008). These approaches are: 1. the Item functioning approach (i.e. in Jansen, 2011), 2. the Item Response Theory Models (i.e. in de Jong et al, 2007) and, 3. the factor analysis framework (Davidov et al., 2008; Gregorich, 2006; Wu et al., 2007).

However, the multi-group confirmatory factor analysis (MG-CFA) is today the most frequently followed, and it will be used in this study. The extensive literature on measurement invariance in a multi-group confirmatory factor analysis framework includes theoretical and didactic papers (Vandeberg & Lance, 2000; Byrne & Stewart, 2006, Schmitt & Kuljanin, 2008; Davidov et al., 2014), two-group cases, or fewer ones, and small samples sizes (Chen, 2007; French & Finch, 2006), large-scale analysis (Gregorich, 2006; Byrne & van, de Vijver, 2010; Rutkowski & Svetina, 2013 among others), and new approach proposals (Asparouhov & Muthén, 2014; Weziak-Bialowolska, 2014).

3.2 Testing for measurement invariance

In Chapter 2, we have already introduced different levels of measurement equivalence. In this section we intend to discuss these concepts focusing on tests used to validate invariance, and their statistical facets.

In the context of the MG-CFA model, literature reviews and applied studies recommend some ME/I tests to be usually applied and satisfied as a precondition for valid cross-group comparisons. Nevertheless, according to the literature, we stress that these tests do not represent a compulsory list to be used as well as it may not be considered as an exhaustive one. Far from it, practitioners and researchers are supposed to evaluate the test opportunity case by case and to focus on the measurement equivalence hypotheses to be tested (i.e. factor loadings, factor covariances, latent factor means…) depending on the research objectives and the kind of analyses undertaken.

As explained before, testing for invariance of a measuring instrument and/or for equivalence of a theoretical construct is a hierarchical process, where sets of parameters are increasingly constrained from the least to the most restrictive model.

The testing of the model, or rather the level reached in the progression of nested tests, necessarily refers to the research questions and study interests, and it should be carefully designed by the researcher prior to testing the hypotheses. This avoids conducting demanding and time consuming tests (i.e. a strict invariance test when construct validity of an assessment scale is investigated) without any usefulness for the carried out analysis or even undermining the results.

With regard to the pivotal work of Jöreskog (1971), traditionally, the recommended practice begins with an omnibus test of the equality of covariance structure across groups (Begozzi & Edwards, 1998; Horn & McArdle, 1992; Steenkamp & Baumgartner, 1998).

This first step tests the null hypothesis concerning difference of variance-covariance matrices:

$$H_0 : \sum^1 = \sum^2 = \ldots = \sum^G \qquad (3.1)$$

where $\sum$ is the population variance –covariance matrix, and for each *G*-group observed, it is given by:

$$\sum^G = \Lambda_X^G \, \Phi^G \Lambda_X^{G\prime} + \Theta_\delta^G \qquad (3.2)$$

Being $\sum^G$ the covariance matrix among the items (observed variables) in the G-th groups, Λ_X^G is the matrix of items' factor loadings relating to the latent variable vector ξ^G (unobserved variable), with associated covariance matrix Φ^G, and Θ_δ^G [4] is usually the diagonal matrix of unique variances. (Rutkowski & Svetina, 2013; Vandenberg & Lance, 2000; Schmitt & Kuljanin, 2008).

If the null hypothesis is verified then the lack of difference is confirmed. This normally leads to considering the measure as invariant and no further tests are needed (Alwin &

[4] Θ_δ^G is typically assumed to be diagonal, this implies no correlated measurement errors – However, this is not strictly necessary. The equation is in the framework of factor analysis, where observed item covariance is defined as a function of common and unique factors, and it can be extended to mean structure including a vector of intercepts.
Most applications of covariance structure analysis assume the intercepts to be zero, so their estimations is not conducted (Vanderber & Lance, 2000; Joreskog & Sorbom, 1996, p 297).

Jackson, 1981; Begozzi & Edwards, 1998; Jöreskog, 1971; Steenkamp & Baumgartner, 1998). If the condition is not met, a set of nested tests are undertaken and the sources of invariance are specifically investigated.

Nevertheless, Byrne (2012, p. 195) argues that this overall test can lead to contradictory findings, i.e. the null hypothesis is not verified yet further tests for measurement or structural invariance hold. The author stresses that '*such inconsistencies in the global test for equivalence stem from the fact that there is no baseline model for the test of invariant variance-covariance matrices'.* Therefore, she strongly suggests starting with a test for invariance in terms of configural model.

The reduced interest for a prior investigation of the differences in the variance-covariance matrices seems also proved by the fact that recent studies and articles (as reviewed in Schmitt & Kuljanin, 2008) do not report these tests.

On the contrary, there is full consensus in considering the configural invariance test as the further indispensable step in the equivalence assessment process (or as the first necessary test to be conducted if the analysis of the covariance matrices is omitted).

3.2.1 Configural invariance

The configural invariance test aims at demonstrating that the observed measures represent the same construct across groups, that is the studied concept is actually shared and thus meaningfully discussed (Davidov et al., 2014).

Clearly, there is no sense at all in comparing measurement results if the underlying construct is differently considered by respondents. For this reason, the configural invariance should be viewed as a sort of pre-requirement to be established before testing for further aspects of measurement equivalence.

Configural invariance implies that an equal number of factors and the factor-loading pattern be the same within countries/groups, in other word '*it ensures that common factors are associated with the same items*' (Weziak-Bialowolska, 2014, p. 56). The model proved to be valid by the test is the least restrictive one, and it purely implies a common factorial structure (or configuration), without any constraints regarding factor loadings or other specific parameters (Figure 3 -1 exemplifies a latent variable and its observed variables).

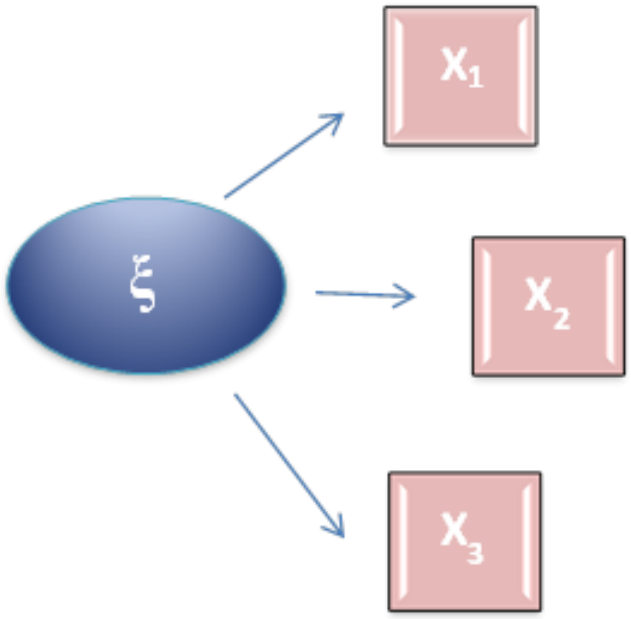
Practically, the way in which the tested model is hypothesized can vary significantly across research studies. For example, it can be based on theory, prior studies, researcher intuition, or established specifically referring to data.

Byrne (2012), suggests that prior to any further investigation, a baseline model should be estimated for each group/country. This specific model is the one that best fits data in terms of both parsimony and significance[5]. This estimation does not imply any between-group constraints.

[5] '*It ideally represents one for which fit to the data and minimal parameter specification are optimal*'

Once these similar group-base models are established, the multigroup baseline model is obtained by repeating again the process with all the data at the same time. This step is essential because a well-fitting multigroup baseline model implies that parameters are estimated for all groups simultaneously, and only by testing this overall model we have the baseline value for further model comparison.

Figure 3-1 Configural Invariance: each group has the same factor structure. The latent variable ξ has the same factor pattern (observed variables X_1-X_3) across groups



Normally, the test of configural invariance[6] verifies the null hypothesis that the defined baseline model (a model with invariant pattern of factors and determinate but freely estimated factor loadings) fits the measures' components (observed variables) across groups (Horn & McArdle, 1992).

This configuration reflects the underlying concept and its *configural factor structure* (Vandenberg & Self, 1993). Consequently, the evidence for a common factor structure implies no conceptual difference between groups (Vandeberg & Lance, 2000), or at least an acceptable similarity.

Configural equivalence must be tenable in order to proceed with any other more constraining test, and this model also serves as '*the baseline against which all subsequent tests for equivalence are compared*' (Byrne & van de Vijver, 2010, p. 109).

If the hypothesis holds (i.e. via a chi-square test of model fit with appropriate degrees of freedom), further tests allow to evaluate if – in addition to the same number of factors (latent variables ξ) – the same associated loadings (Λ_x), scale intercepts (υ_x), and measurement errors (δ) underlie the set of indicators.

Alternatively, if the null hypothesis is rejected, additional tests are not justified because different constructs are being measured. Therefore, it makes no sense to compare group results (Vandenberg & Lance, 2000).

As outlined, configural invariance represents the first indispensable step in the invariance assessment process, but it is not sufficient for meaningful statistical comparisons, such as the relationships between factor scores and items or the country rakings based on mean scale score.

[6] In not recent articles, we find also different definitions of the test, i.e. a test of 'same form' (Bollen, 1989) or the 'practical scientist's' invariance (Horn & McArdle, 1992).

3.2.2 Metric invariance

A further level of analysis is a test of metric invariance (Horn & McArdle, 1992; Steenkamp & Baumgartner, 1998), also defined as weak measurement equivalence (Meredith, 1993) or measurement unit equivalence (van de Vijver & Leung, 1997).

The metric invariance test is a more constraining test of equivalence than the test of configural invariance. In addition to an equal number of factors and the same factor-loading pattern, factor loadings are also constrained to be invariant (Figure 3 -2).

The equivalence of factor loadings (Λ_x) means that the regression slopes linking observed variables to latent variables (ξ) are the same within the compared groups, thus '*the expected change in the observed score on the item per unit change on the latent variable*' are forced to be same (Vandenberg & Lance, 2000, p. 37).

The null hypothesis to be tested for metric invariance confirmation is:

$$H_0 : \Lambda_X^1 = \Lambda_X^2 = \ldots = \Lambda_X^G \qquad \forall\ G\text{-group observed} \qquad (3.3)$$

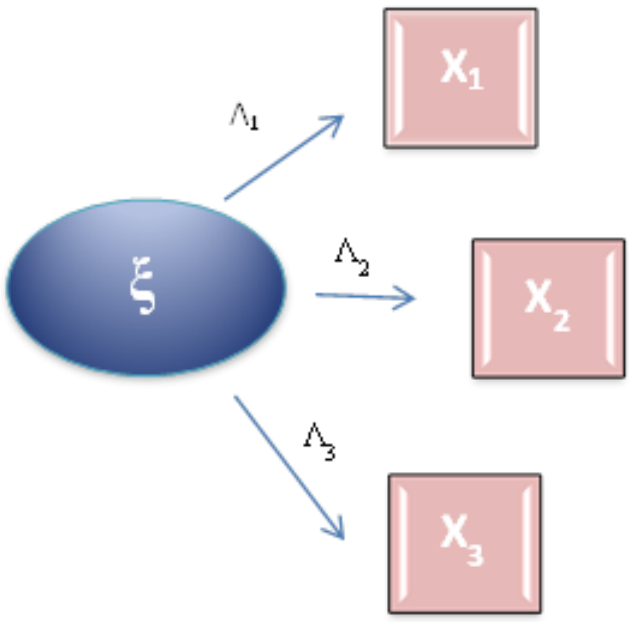
where, as afore mentioned, Λ_X^G is the matrix of items' factor loadings on the latent variable vector.

Thus, metric invariance ensures the cross-group equality of scaling units underlying the latent variables assessment (Jöreskog, 1969; Vandenberg & Self, 1993), '*an increase of one unit on the measurement scale has the same meaning in population A as in population B*' (Davidov et al., 2014, p. 63).

If metric invariance holds, then comparison of different population scores are allowed (Steenkamp & Baumgartner, 1998), and researches involving only construct validity questions or relationships between latent factors (i.e. factor scores/scales and /or other observable variables or test on invariance of factor variances or covariances) are fully validated (Weziak-Bialowolska, 2014; Schmitt & Kuljanin, 2008).

The metric invariance test is conducted by fixing the factor loadings for all the involved groups at the same level. Practically, these parameters can be freely set only for the first group (or generally speaking for a chosen group), which serves as the reference one (Bryne & van de Vijver, 2010). For all other groups/countries, the factor loadings are forced to be equal to those of the reference group, and these parameters will remain fixed in case of further analysis of invariance taking the fitting metric model as a baseline one.

Figure 3-2 Metric Invariance: factors loadings (λ_1-λ_3) between the observed variables (X_1-X_3) and the latent variable are the same across groups



Vandenberg & Lance (2000) indicate that almost every paper they reviewed reported tests of factor loadings (similarly in 2008 Schmitt & Kuljanin), however the mentioned studies were not unanimously agreeing on the consequences of the null hypothesis rejection.

Following a strict line, some researchers suggest that metric invariance must be considered as a requisite for any further measurement invariance analysis; thus if the null hypothesis about factor loading matrices does not hold, any additional test should be considered meaningless (Millsap & Hartog, 1988; Bollen, 1989). On the other hand, some authors (Byrne & van de Vijver, 2010, Byrne, 2013; Steenkamp & Baumgartner, 1998) propose to relax metric invariance constraints and carry out analysis at a partial level when the overall test of metric equivalence must be rejected.

Albeit in the literature partial measurement equivalence has never produced a vast discussion and there is not a consensus about the statistical criteria for relaxing metric equivalence constraints, the approach of limiting the subset of invariant measurement parameters is quite common in cross-group studies (Byrne et al, 1989; Byrne & van de Vijver, 2010).

Vandenberg & Lance (2000, p. 38) recommend a conservative approach and restrict the use of relaxed metric invariance constraints *'(a) only for a minority of indicators, (b) on as strong a theoretical basis as is possible, and (c) when cross-validation evidence points to their viability. Alternately, indicators that do not meet metric invariance restrictions may be removed from analysis*'.

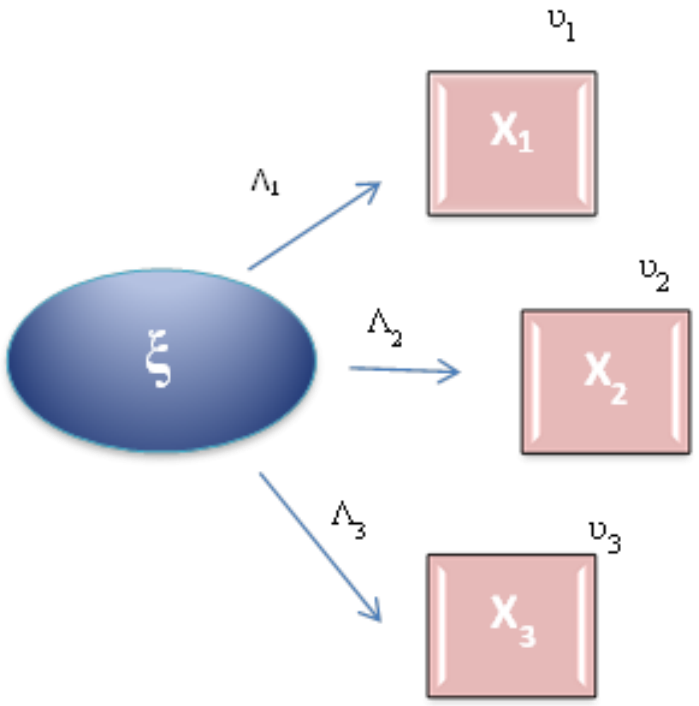
Other applied studies suggest replacing in the scale the items causing invariance when possible (Byrne & van de Vijver, 2010; Welkenhuysen-Gybels, 2003).

Steenkamp & Baumgartner (1998) argue that metric invariance constraints can be relaxed up to the limit that at least the parameters of two indicators per latent variables (the reference indicator and an additional one at least) result invariant across groups in a MG-CFA model.

3.2.3 Scalar and strict invariance

Metric invariance guarantees that the strength of the link between items and latent factors is the same for all the analyzed groups, but this level of equivalence still results insufficient for valid means comparisons and meaningful scaling. For this it is necessary to move on to an increased level of restrictiveness, testing for scalar equivalence (Steenkamp & Baumgartner, 1998, Byrne & van de Vijver, 2010) or strong factorial equivalence (Meredith, 1993).

Figure 3-3 Scalar Invariance: not only the factor loadings (λ_1-λ_3) but also the regression intercepts (υ_1-υ_3) are equal across groups



Scalar invariance implies that also the origin of the scale is the same across groups, i.e. the intercepts (υ_X^G) of the regression equations of the observed items on the latent variables are constrained to be equivalent (Figure 3 -3).

The null hypothesis to be proved for scalar invariance is:

$$H_0 : \upsilon_X^1 = \upsilon_X^2 = \ldots = \upsilon_X^G \qquad \forall\ G\text{-group observed} \qquad (3.4)$$

where υ_X^G is a vector of observed variable intercepts.

The metric equivalence test as well as the configural invariance one found on the analysis of the covariance matrices, consequently all observed indicators (i.e. item scores) could be computed as deviations from their means (i.e. fixed to zero). On the contrary, constraining item intercepts, the scalar equivalence test implies analysis of both mean and covariance structures (moment matrix analysis), consequently element such as item means cannot be longer fixed to zero (Byrne & van de Vijver, 2010).

If the null hypothesis holds, it is statistically verified that equal latent variable scores are related to the same expected scores on observed variables across groups (Rutkowski & Svetina, 2013), '*concretely, this means that all observed mean differences in the items must be conveyed through mean differences in the latent factor*' (Davidov et al., 2014, p. 64).

Given the above, scalar invariance supports meaningful comparison of latent factor means (Marsh et al., 2009; Weziak-Bialowolska, 2014) as well as valid country level analysis in a multilevel regression analysis framework.

In addition, some authors suggest a test for invariance of item intercepts as a standard step and contend that it should always be conducted (Little, 1997; Meredith, 1993; Selig et al., 2008), but this approach is not totally shared. Actually, other researchers relax the approach (Marsh et al., 2006; Byrne, 1993, Byrne & van de Vijver, 2010) and claim that less 'strong' level of equivalence, such as factor loadings invariance, could be sufficient, and more appropriated, on the basis of the kind of the conducted researches (i.e. construct validity studies).

Nevertheless, if until recently the scalar invariance was no widely investigated (see Vandenberg & Lance, 2000), awareness of the relevance of such analyses is growing among researchers (about 12% of field studies in 2000 vs 54% in 2008), maybe also because they give the possibility to testing means and covariances (Chan, 1998; Schmitt & Kuljanin, 2008).

The last test for measurement invariance is a test of the invariance of the unique variances related to each observed variable across groups (Schmitt & Kuljanin, 2008). Thus, the residuals of the regression equations are fixed for each item, and it may make sense only if (at least partial) metric and scalar invariance have already been proven (Vandenberg & Lance, 2000).

The null hypothesis to be tested is:

$$H_0 : \Theta_\delta^1 = \Theta_\delta^2 = \ldots = \Theta_\delta^G \qquad \forall\ G\text{-group observed} \qquad (3.5)$$

where Θ_{δ}^{G} is a covariance matrix of the measurement errors (δ) for the observed variables. Normally, this matrix is assumed to be diagonal, so that measurement errors are uncorrelated.

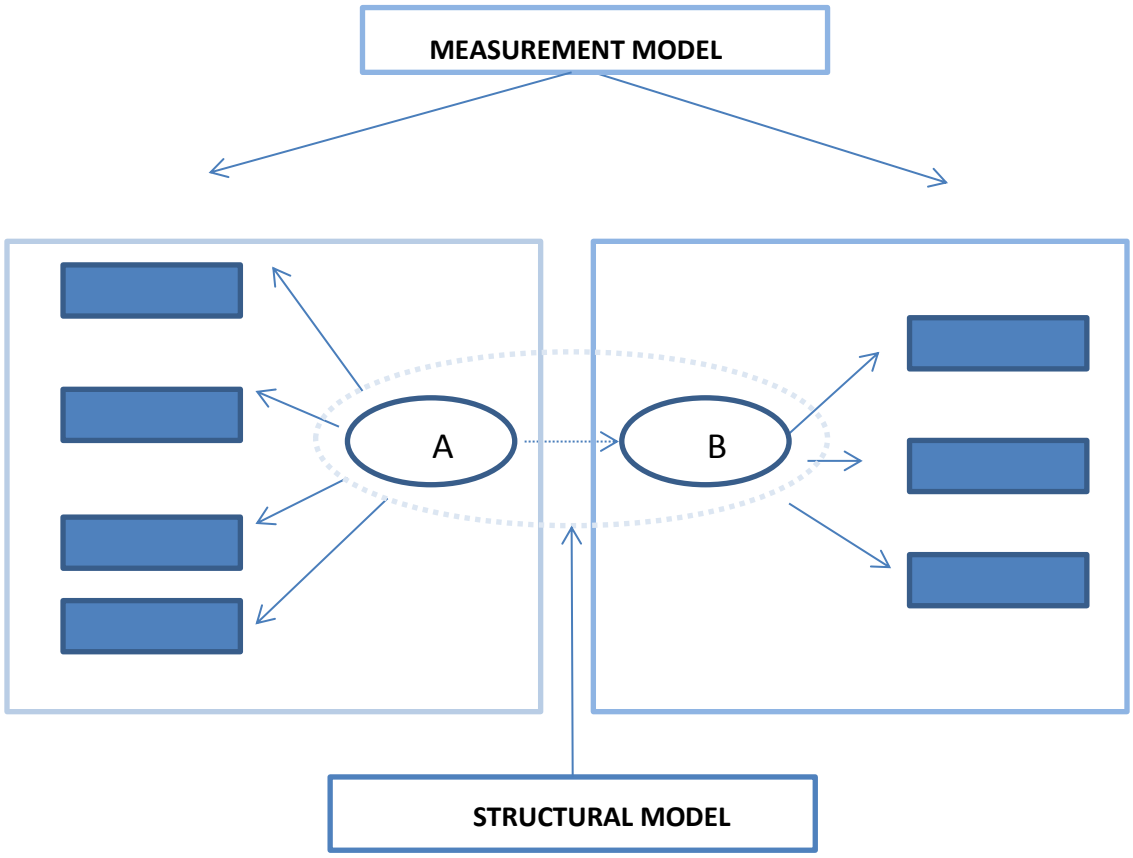
Due to its high strictness and difficulty to be achieved, this test has been termed strict invariance (Meredith, 1993) and most researches often consider it of little concern (Bentler, 2005; Widaman & Reise, 1997), unnecessary or not recommended (Little et al., 2007; Selig et al., 2008; Byrne & Stewart, 2006). These authors argue that the strict invariance test cannot provide further improved information regarding the most common questions of interest in the field such as investigating differences in factor structure or latent means or construct validity for scale assessment.

Conversely, some authors acknowledge the value of the strict invariance test in testing for multigroup equivalence of item reliability (i.e. Byrne, 1988) Yet, as stressed by Vandenberg & Lance (2000) strict invariance only holds when the equality of the factors variances has also been established (Schmitt & Kuljanin, 2008).

3.3 Structural invariance

The tests illustrated in the previous section are concerned with the relationships between observed variables and latent factors (i.e. how measured indicators load on latent variables) and are often referred to as tests for measurement invariance. In contrast, the tests presented below concern the relationships between the unobserved variables themselves and are termed tests for structural invariance (i.e. Byrne et al., 1989; Vandeberg & Lance, 2000; Schmitt & Kuljanin, 2008).

Figure 3-4 Measurement model and Structural model



Source: Adapted from Byrne, 2012, p. 15

In the same way, the tested models are distinct in measurement models (see section 3.1) and structural models. The later depict relations among latent factors and specify their direct or indirect influence on the model (Byrne, 2012).

Typically, aspects of structural invariance are investigated by three different tests concerning the invariance of factor variances, covariances, and means respectively. They assess the independent (or not dependent) issues concerning invariance. For this reason the tests for structural invariance do not need to be carried out in a hierarchical or sequential order,[7] as it is the case for the measurement testing (where each test is *de facto* nested in the previous one).

Further, they are not necessarily looked at in the equivalence assessment process, but they are conducted only on the basis of the specific research objectives. In particular, construct validity researches related to dimensionality and assessment scale embody such studies (Byrne & Shavelson, 1987; Schmitt & Kuljanin, 2008), where factor covariances are the major concern or as a prerequisite in item reliability invariance test (see above).

The factor variance-invariance test assesses whether the variances of the latent variables are equal across groups, thus in the tested model factor variances are constrained to equality (i.e. the diagonal element of Φ).

[7] Even if tests for invariance of factor covariances and variances are often conducted before a test of the difference of latent factor means (Schmitt & Kuljanin, 2008).

The null hypothesis to be tested is:

$$H_0 : \Phi_j^1 = \Phi_j^2 = \ldots = \Phi_j^G \qquad \forall\ G\text{-group observed} \quad (3.6)$$

where Φ_j^G is the variance matrix of the latent variable ξ_j.

If the null hypothesis holds, the groups work in an equal way, while if it is rejected they show a different use of 'the range of the construct continuum' (Vandenberg & Lance, 2000): smaller factor variance denoting a closer range. This test is sometimes considered together with the metric invariance test, being the detected non-invariance in factor variances linked to the group non-invariance in score setting (Schmitt, 1982; Vandenberg & Self, 1993).

Analogously, the invariance of the factor covariances is tested by constraining the covariances of latent variable pairs to be equal across groups.

The test has been considered by some authors as a test of stability (with a test of configural invariance) of the factor relations (Schmitt, 1982; Vandenberg & Self, 1993): actually, accepting the null hypothesis means that the relationships among unobserved factors are statistically the same in all the groups (Milfont & Fischer, 2010).

In other words, if the conceptual domain is invariant for all the groups then the relationships among latent variables should not substantially vary. Conversely, if the conceptual domain differs, the invariance of the covariances cannot hold. Nevertheless, the additional value of the factor covariance test in this kind of analysis with respect to the configural invariance test has been questioned (Vendenberg & Lance, 2000), the latter being more stringent than the former.

The last test for structural invariance is a test for equal factor means. It evaluates the way groups '*differ in level on the underlying construct(s)* ξ_j *that are operationalized (and approximated) by the composite of the* X_{jk}*s*' (Vandeberg & Lance, 2000 p. 40). Normally, measurement invariance tests would be carried out priory to validate testing for group differences (Vandeberg & Lance, 2000).

The null hypothesis to be tested is:

$$H_0 : \kappa_j^1 = \kappa_j^2 = \ldots = \kappa_j^G \qquad \forall \text{ } G\text{-group observed} \qquad (3.7)$$

where κ_j^G is the mean of the latent variable ξ_j.

If the null hypothesis does not hold, further analyses are necessary to identify specific causes for the differences among groups (Schaubroeck & Green, 1989; Vandenberg & Self, 1993).

3.4 Addressing non-invariance

Despite the interesting developments in the field, empirical studies on measurement equivalence show that obtaining invariance across groups (countries or cultures) is a quite difficult issue, in particular when the mere configural invariance level is not sufficient to support the research objectives (Davidov et al, 2012; Vandenberg & Lance, 2000; Millsap, 2011; van de Schoot et al., 2013).

If measurement equivalence does not hold, a meaningful comparison of the data across countries is not justified (i.e. scores on a latent factor when the underlying model fails to be proved equivalent); generally speaking, the researchers should desist from comparisons across groups.

Actually, in case of any such '*impasse*', some alternative 'strategies' (Davidov et al., 2014) can be undertaken (or tried) by researchers to overpass the hard obstacle of the missing data fit. As reported in Davidov et al. (2014), researchers could:

1 Identify sub-groups of countries where measurement invariance is tenable, and continue limiting the comparison to this set (or independent sets) of countries. This approach is recommended in the case of cross-cultural research (a) when the underlying construct is found to be inappropriate (structurally and psychometrically) or (b) when cluster analyses increase both within-cluster homogeneity and between-cluster heterogeneity (Byrne & van de Vijver, 2010; Weziak-Bialowolska & Isac, 2014).

2 Conduct further studies to better understand and detect invariance sources, and evaluate the possible removal of some of the items causing invariance (Meuleman, 2012; Gregorich, 2006). '*However, this can be done only if a few invariant items*

remain to measure the latent variable after the unusable items have been dropped' (*infra*, p. 66)

3 Accept and justify measurement invariance on a specific, historical, and/or societal level or control for sources of bias such as acquiescence or extreme responding (Welkenhuysen-Gybels et al., 2003; Weijters et al, 2008).

Strictly linked to point (2) is a partial treatment of the data, in the sense that researchers are ready to relax some parameters to solve the non-equivalence problem, at the cost of losing information.

Implementing a condition of partial measurement invariance – that is '*some but not all measurement parameters are constrained equal across groups in testing*' (Byrne, 2012, p. 198) – implies to give up the plain consistency of the models described above. However, if some parameters are held constant, whereas others are freely estimated, in some cases, it is possible to recover a model where measurement invariance (at partial level) still satisfactory holds when full measurement equivalence is not given.

Most of the studies exploring partial invariance tests show an empirical approach more than a theoretical one, as stated by Schmitt & Kuljanin (2008) in their review, '*when researchers found evidence for a lack of invariance […] allowed parameters to be freed across groups until they were satisfied that the remainder of the parameters were invariant across groups'*.

As stressed by Barbara Byrne (2012, p. 255), one of the first authors to discuss in depth the subject of partial invariance (Byrne et al., 1989), partial measurement equivalence has been a highly controversial issue in the technical literature (Marsh & Grayson, 1994; Widaman & Reise, 1997; Kaplan & George, 1995).

In large-scale cross-country studies, where it is often problematic to reach an acceptable level of invariance (Rutkowski & Svetina, 2013; Davidov et al. 2008), partial invariance can also be unsatisfactory. Partial measurement equivalence works efficiently when few items are the source of large differences (van de Schoot et al., 2013) and these items can easily be identified. In large-scale studies characterized by a large number of countries, the identification of the parameters to be relaxed is a quite difficult aim, '*due to many possible violations of invariance and many possible modifications*' (Weziak-Bialowolska, 2014, p. 57) of the model.

PART II

Chapter 4

DATA AND METHOD

Chapter 4 provides an overview of the data used and describes the methods applied to the empirical study. The first section offers a brief outline of the data source, the International Civic and Citizenship Education Study – ICCS conducted by the International Association for the Evaluation of Educational Achievement – IEA in 2009. In particular, we describe the research focus of the study and the results related to *students' attitudes toward equal rights for immigrants* in European Countries. Building on these data, we formulated the research questions presented in section 2. Section 3 provides a technical description of the input data sample downloaded from the IEA's website, while section 4 describes the datasets generated from the original IEA data and used for the analyses. The last section refers to the data analysis strategy developed for the study.

4.1 ICCS 2009 - Students' perceptions of equal rights for immigrants

This empirical study investigates data from the International Civic and Citizenship Education Study (ICCS), a large scale survey organized by the International Association for the Evaluation of Educational Achievement – IEA. The first cycle of the study, which is the object of the current research, took place in 2009; the data collection for the second cycle of the study (ICCS 2016) is scheduled for 2016. The final data set of ICCS 2009 includes data on citizenship competences of Grade 8 (approximately 14 years of age) students from 38 countries[8]. The ICCS rules concerning target population implied that if the average age of students in Grade 8 was below 13.5 years then Grade 9 students were used as target population instead of Grade 8.

The International Association for the Evaluation of Educational Achievement, known as IEA, is an independent, international consortium of national research institutions and governmental research agencies, with headquarters in Amsterdam.

Its primary purpose is to conduct large-scale comparative studies of educational achievement with the aim of gaining more in-depth understanding of the effects of policies and practices within and across systems of education.

[8] Austria, Belgium (Flemish), Bulgaria, Chile, Chinese Taipei, Colombia, Cyprus, Czech Republic, Denmark, Dominican Republic, England, Estonia, Finland, Greece, Guatemala, Hong Kong SAR, Indonesia, Ireland, Italy, Republic of Korea, Latvia, Liechtenstein, Lithuania, Luxembourg, Malta, Mexico, The Netherlands, New Zealand, Norway, Paraguay, Poland, Russian Federation, Slovak Republic, Slovenia, Spain, Sweden, Switzerland, and Thailand.

The ICCS 2009 Research Question 3 investigated '*what is the extent of interest and disposition to engage in public and political life among adolescents and which factors within or across countries are related to it* ' (ICCS 2009 International Report, p. 87).

With the aim of investigating this broad subject, more specific sub-issues were defined to cover all its facets. Among the various aspects investigated, they identified students' perceptions of democracy and citizenship, students' perceptions of equal rights in society, students' perceptions of their country, and students' engagement with religion. Each of these matters was further developed into sets of sub-questions, which finally were operationalized in questionnaire items.

More specifically, the students' perceptions of equal rights in society subject was translated into three main research questions connected to students' attitudes toward gender equality, equal rights for all ethic/racial group in society, and equal rights and opportunities for immigrants, which is the specific field of interest for this dissertation.

In this latter research area, various dimensions were considered for the analysis. Specifically, the survey items referred to students' perceptions of equal rights in society, students' attitudes toward intercultural relations as well as students' attitudes toward race, migration, immigration and cohesion.

Figure 4 -1 Supplement 1 – International Version of the ICCS 2009 Questionnaires

Q26 People are increasingly moving from one country to another. **How much do you agree or disagree with the following statements about <immigrants>?**

(Please tick only one box in each row)

		Strongly agree	*Agree*	*Disagree*	*Strongly disagree*
IS2P26A	a) <Immigrants> should have the opportunity to continue speaking their own language	☐ 1	☐ 2	☐ 3	☐ 4
IS2P26B	b) <Immigrant> children should have the same opportunities for education that other children in the country have	☐ 1	☐ 2	☐ 3	☐ 4
IS2P26C	c) <Immigrants> who live in a country for several years should have the opportunity to vote in elections	☐ 1	☐ 2	☐ 3	☐ 4
IS2P26D	d) <Immigrants> should have the opportunity to continue their own customs and lifestyle	☐ 1	☐ 2	☐ 3	☐ 4
IS2P26E	e) <Immigrants> should have all the same rights that everyone else in the country has	☐ 1	☐ 2	☐ 3	☐ 4
IS2P26F	f) When there are not many jobs available, <immigration> should be restricted	☐ 1	☐ 2	☐ 3	☐ 4

Source: ICCS 2009 User Guide for the International Database Supplement 1, 2014 – p. 75

In particular, with regard to students' attitudes toward equal rights for immigrants (Schulz at al., 2010) the ICCS 2009 student questionnaire (Figure 4 -1) included the following six Likert-type items (with possible answer categories 'strongly agree', 'agree', 'disagree', 'strongly disagree'):

- Immigrants should have the opportunity to continue speaking their own language;
- Immigrant children should have the same opportunities for education that other children in the country have;
- Immigrants who live in a country for several years should have the opportunity to vote in elections;
- Immigrants should have the opportunity to continue their own customs and lifestyle;
- Immigrants should have all the same rights that everyone else in the country has;
- When there are not many jobs available, immigration should be restricted.

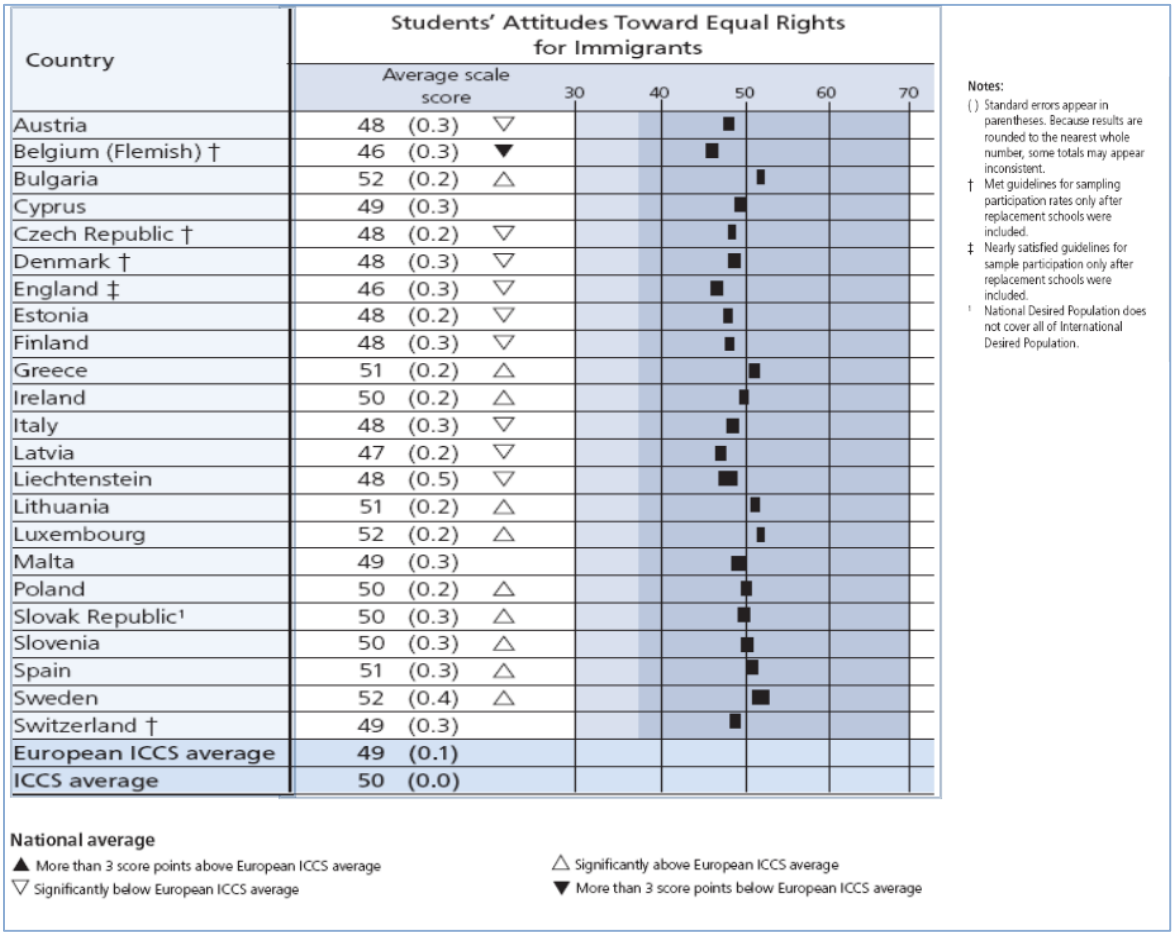
It must be pointed out that the ICCS 2009 research aim was to capture students' attitudes toward the principle of equality in rights and opportunities for immigrants, for this reason a stem question introducing the above items related to immigration to any country, and was formulated as follows:

People are increasingly moving from one country to another How much do you agree or disagree with the following statements about <immigrants>?

This approach allowed including also countries with very low levels of immigration.

Table 4-1 Comparison of national averages for students' attitudes toward rights for immigrants[9]

Country	Students' Attitudes Toward Equal Rights for Immigrants: Average scale score		
Austria	48	(0.3)	▽
Belgium (Flemish) †	46	(0.3)	▼
Bulgaria	52	(0.2)	△
Cyprus	49	(0.3)	
Czech Republic †	48	(0.2)	▽
Denmark †	48	(0.3)	▽
England ‡	46	(0.3)	▽
Estonia	48	(0.2)	▽
Finland	48	(0.3)	▽
Greece	51	(0.2)	△
Ireland	50	(0.2)	△
Italy	48	(0.3)	▽
Latvia	47	(0.2)	▽
Liechtenstein	48	(0.5)	▽
Lithuania	51	(0.2)	△
Luxembourg	52	(0.2)	△
Malta	49	(0.3)	
Poland	50	(0.2)	△
Slovak Republic[1]	50	(0.3)	△
Slovenia	50	(0.3)	△
Spain	51	(0.3)	△
Sweden	52	(0.4)	△
Switzerland †	49	(0.3)	
European ICCS average	**49**	**(0.1)**	
ICCS average	**50**	**(0.0)**	



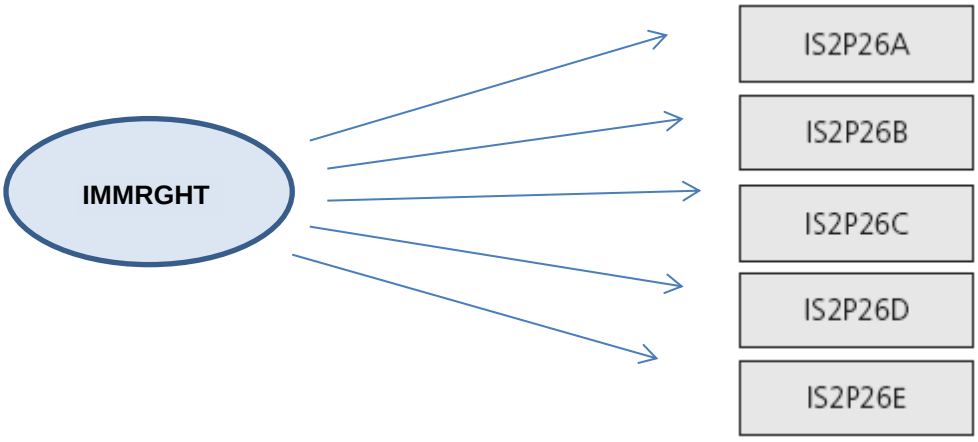
Source: adapted from ICCS 2009 European Report p. 92

The collected data were used for elaborating the ICCS 2009 *students' attitudes toward equal rights for immigrants* scale (Table 4 -1). The scale includes five items[10]. The sixth

[9] A similar scale with regard to all 38 countries is reported in the ICCS 2009 International Report (2010, p.102)

item =-*When there are not many jobs available, immigration should be restricted*' was not used for scaling.

Figure 4-2 Students 'attitudes toward equal rights for immigrants (IMMRGHT).



Source: ICCS 2009 Technical Report, 2011

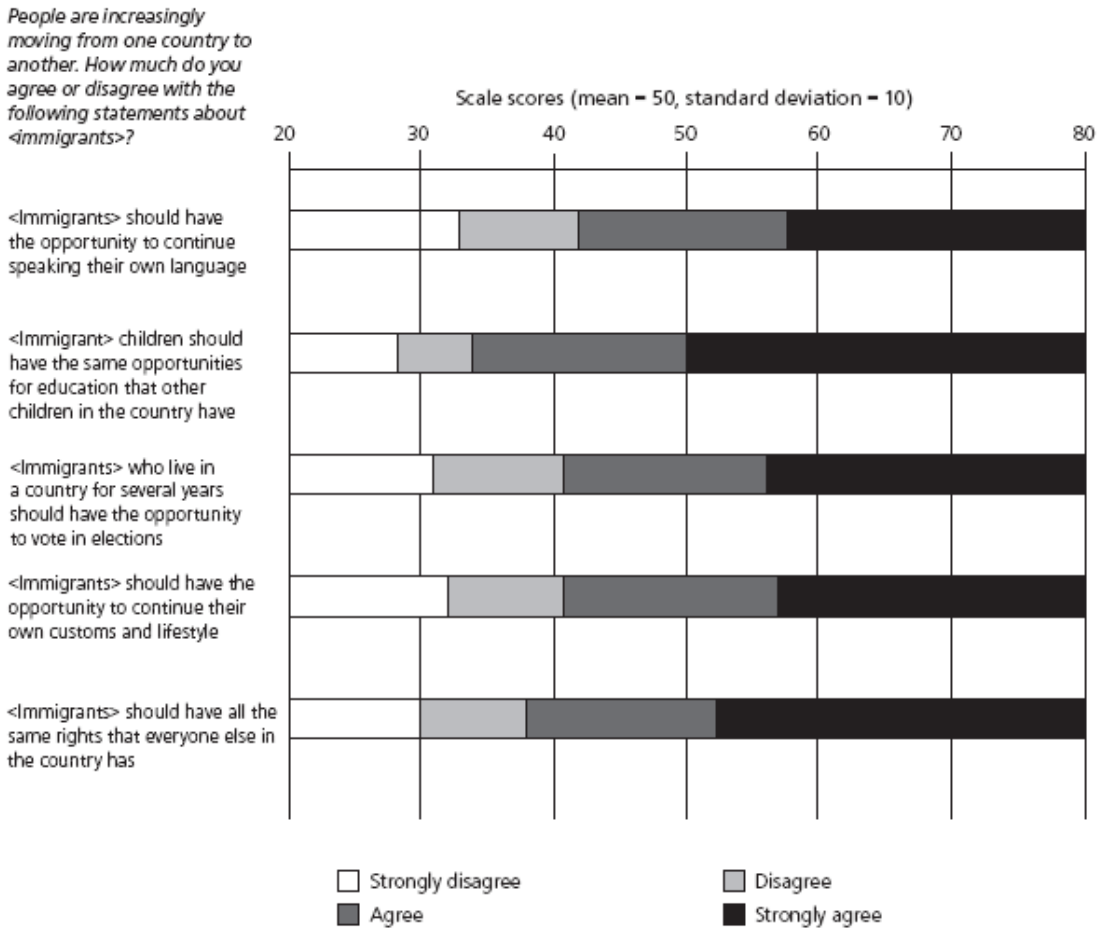
The ICCS 2009 International Report (p. 100) states that the five discussed items form a highly reliable scale, with a Cronbach's alpha of 0.90 for the whole international dataset (38 countries).

The higher scale scores indicate higher levels of support for the rights of immigrants. On the basis of these data, albeit the important differences between countries, it could be assumed that a student with an ICCS average score of 50 had more than 50 percent

[10] Analogously, the CIVED survey in 1999 (a predecessor of ICCS 2009) considered a set of eight items to capture students' attitudes toward immigrants, but only five of these were included in the scale (Schulz, 2004).

likelihood of agreeing with all five items. Figure 4 -3a (from the Appendix E of the ICCS 2009 International Report, p. 275) illustrates the item-by-score map for the scale.

Figure 4-3a - Item-by-score map for students' attitudes toward equal rights for immigrants

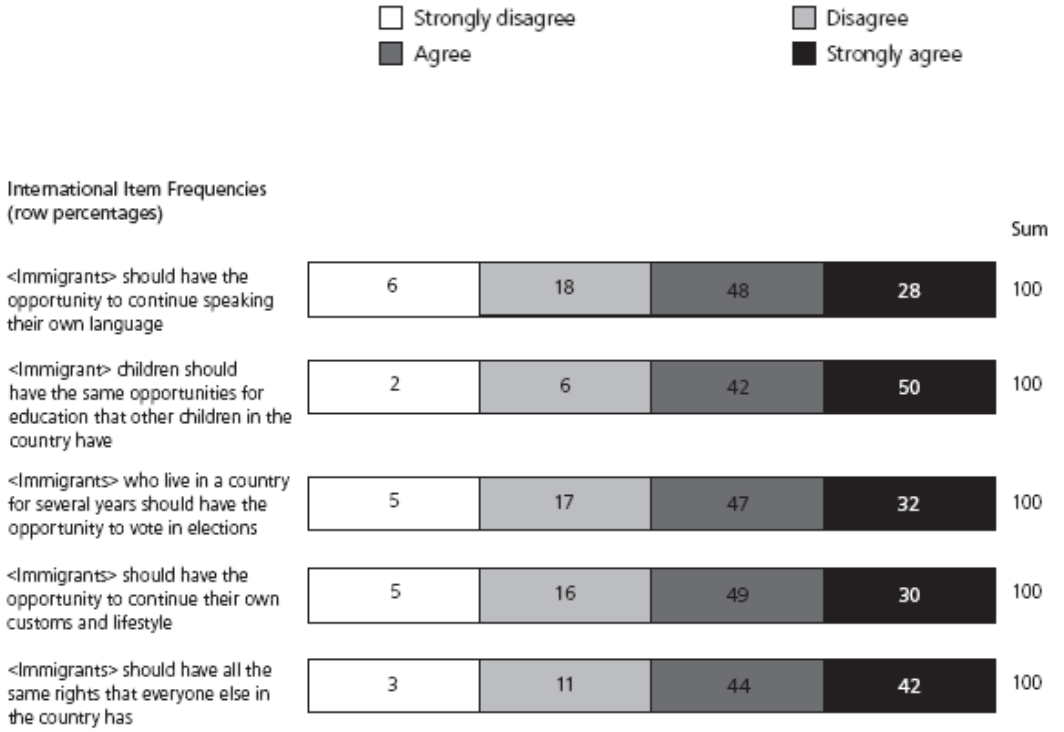


Source: adapted from ICCS 2009 International Report

The agreement ranged from 76 percent with the first statement '*immigrants should have the opportunity to continue speaking their language*' to 92 percent referred to the

statement '*immigrant children should have the same opportunities for education'* (Figure 4- 3b).

Figure 4-3b - Item-by-score map for students' attitudes toward equal rights for immigrants (Item frequencies)



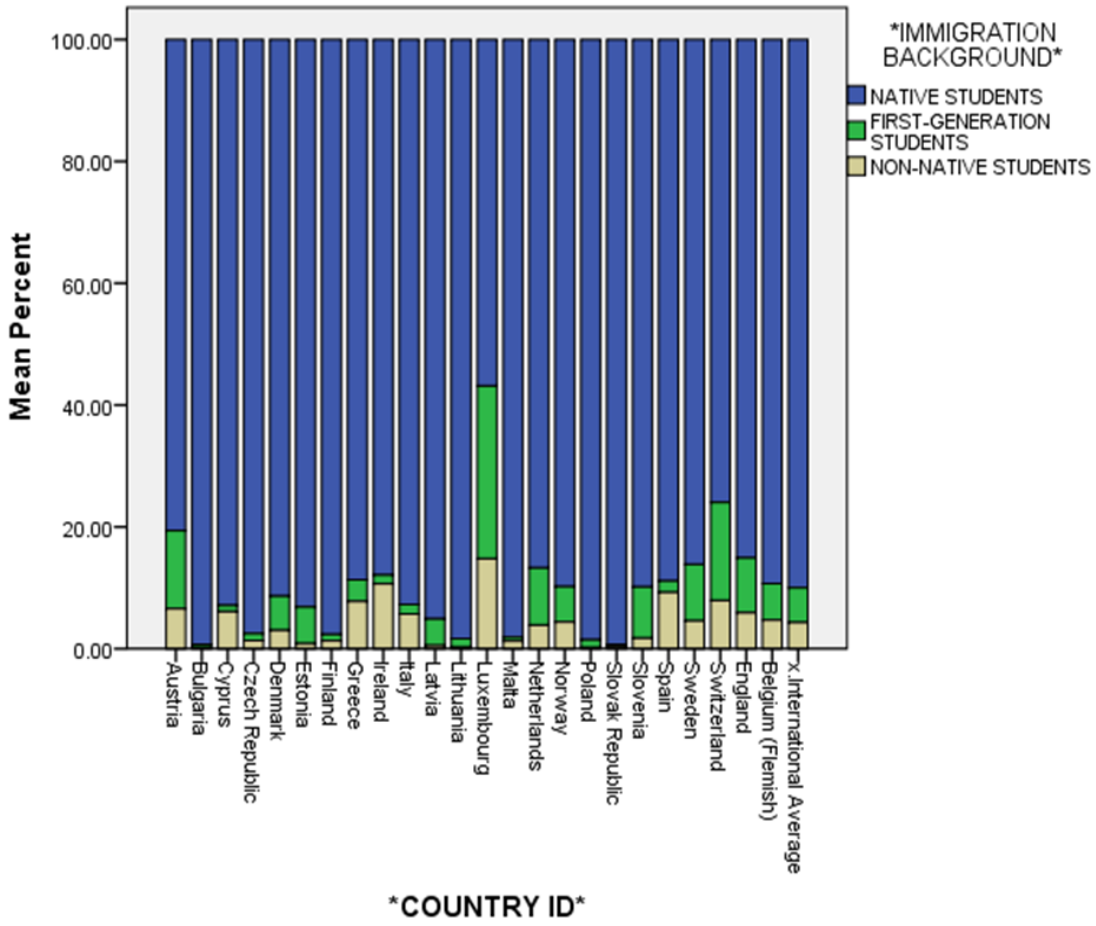
Note:
Average percentages for 36 equally weighted participating countries that met sample participation requirements. Because results are rounded to the nearest whole number, some totals may appear inconsistent.

Source: adapted from ICCS 2009 International Report

At European level, on average, these percentages were some score points lower (ICCS 2009 European Report, 2010). The agreement ranged from 72 percent with the first statement '*immigrants should have the opportunity to continue speaking their language*'

to 91 percent referred to the statement '*immigrant children should have the same opportunities for education'*. The European average score was 49 points (the ICCS international average was 50 points), and the scores for the European countries ranged from 46 to 52 points. Belgium (Flemish), England, and Latvia showed the lowest national averages, while Bulgaria, Luxembourg, and Sweden the highest levels of attitude toward equal rights for immigrants.

Figure 4-4 Student participation in ICCS 2009 survey – European Countries for immigration background



Source: ICCS 2009 International Dataset – Data elaborated using IEA IDB analyzer and SPSS software

The ICCS 2009 European Report stresses that the differences across European national scores of students' attitudes toward equal rights for immigrants may be influenced by the different immigration context and history of the participating countries.

It is argued that the levels and the origin of immigrant populations vary greatly in Europe. This difference is also highlighted by the variance in the surveyed student rates with respect to the immigrant background (see Figure 4 -4 and Table 4 -3). The government and policy actions concerning immigration and the perception of immigrants in society are dissimilar within and across European countries.

As stated in the same report, studies confirm the complexity and different impact of immigration in Europe (Penninx, 2005; Penninx at al., 2006), for example:

- Some Western European countries (such as England, France, and the Netherlands) display a lasting and quite complex immigration histories, in same case strongly intertwined with colonialism;
- Some Southern and Northern European countries (such Finland, Greece, Italy, Norway and Spain) have been facing new significant flows of migrants;
- Finally, some Central and Eastern European countries have experienced immigration only in recent years.

Further, the ICCS 2009 research team considered also that cultural factors (such as family background) may effectively influence students' attitudes toward minorities and immigrants at an European level (papers as Dejaeghere & Quintelier, 2008; Torney at al., 2008 were reviewed).

Table 4-2 Student participation in ICCS 2009 survey for immigration background – Descriptive Statistics

COUNTRY ID	*IMMIGRATION BACKGROUND*	Sum of Cases	Sum of TOTWGTS	Percent
Austria	NATIVE STUDENTS	2646	68656	80.62
	FIRST-GENERATION STUDENTS	408	10919	12.82
	NON-NATIVE STUDENTS	207	5588	6.56
Bulgaria	NATIVE STUDENTS	3185	62101	99.27
	FIRST-GENERATION STUDENTS	15	311	0.5
	NON-NATIVE STUDENTS	7	148	0.24
Cyprus	NATIVE STUDENTS	2880	8033	92.88
	FIRST-GENERATION STUDENTS	32	88	1.02
	NON-NATIVE STUDENTS	198	528	6.1
Czech Republic	NATIVE STUDENTS	4484	92700	97.53
	FIRST-GENERATION STUDENTS	50	1060	1.11
	NON-NATIVE STUDENTS	62	1293	1.36
Denmark	NATIVE STUDENTS	3901	54571	91.35
	FIRST-GENERATION STUDENTS	273	3359	5.62
	NON-NATIVE STUDENTS	142	1806	3.02
Estonia	NATIVE STUDENTS	2510	10770	93.14
	FIRST-GENERATION STUDENTS	169	696	6.02
	NON-NATIVE STUDENTS	25	97	0.84
Finland	NATIVE STUDENTS	3196	60711	97.64
	FIRST-GENERATION STUDENTS	35	651	1.05
	NON-NATIVE STUDENTS	43	817	1.31
Greece	NATIVE STUDENTS	2790	88596	88.68
	FIRST-GENERATION STUDENTS	112	3555	3.56
	NON-NATIVE STUDENTS	223	7757	7.76
Ireland	NATIVE STUDENTS	2882	47388	87.92
	FIRST-GENERATION STUDENTS	47	763	1.42
	NON-NATIVE STUDENTS	339	5750	10.67
Italy	NATIVE STUDENTS	3070	491014	92.74
	FIRST-GENERATION STUDENTS	49	8171	1.54
	NON-NATIVE STUDENTS	184	30280	5.72
Latvia	NATIVE STUDENTS	2568	19532	95.09
	FIRST-GENERATION STUDENTS	137	901	4.38
	NON-NATIVE STUDENTS	14	108	0.53
Lithuania	NATIVE STUDENTS	3671	36577	98.32
	FIRST-GENERATION STUDENTS	153	546	1.47
	NON-NATIVE STUDENTS	26	79	0.21

Source: ICCS 2009 International Dataset – Data elaborated using IEA analyzer and SPSS software

Table 4-2 Student participation in ICCS 2009 survey for immigration background – Descriptive Statistics (continued)

COUNTRY ID	*IMMIGRATION BACKGROUND*	Sum of Cases	Sum of TOTWGTS	Percent
Luxembourg	NATIVE STUDENTS	2860	3206	56.86
	FIRST-GENERATION STUDENTS	1272	1599	28.37
	NON-NATIVE STUDENTS	595	833	14.77
Malta	NATIVE STUDENTS	2053	4653	98.13
	FIRST-GENERATION STUDENTS	12	28	0.59
	NON-NATIVE STUDENTS	24	61	1.29
Netherlands	NATIVE STUDENTS	1682	135028	86.73
	FIRST-GENERATION STUDENTS	150	14638	9.4
	NON-NATIVE STUDENTS	69	6022	3.87
Norway	NATIVE STUDENTS	2602	51534	89.8
	FIRST-GENERATION STUDENTS	172	3336	5.81
	NON-NATIVE STUDENTS	127	2520	4.39
Poland	NATIVE STUDENTS	3167	428414	98.55
	FIRST-GENERATION STUDENTS	42	5481	1.26
	NON-NATIVE STUDENTS	5	839	0.19
Slovak Republic	NATIVE STUDENTS	2909	50601	99.27
	FIRST-GENERATION STUDENTS	11	200	0.39
	NON-NATIVE STUDENTS	8	171	0.34
Slovenia	NATIVE STUDENTS	2708	15330	89.84
	FIRST-GENERATION STUDENTS	266	1437	8.42
	NON-NATIVE STUDENTS	56	297	1.74
Spain	NATIVE STUDENTS	2918	387701	88.87
	FIRST-GENERATION STUDENTS	63	8171	1.87
	NON-NATIVE STUDENTS	293	40380	9.26
Sweden	NATIVE STUDENTS	2715	88703	86.14
	FIRST-GENERATION STUDENTS	419	9562	9.29
	NON-NATIVE STUDENTS	199	4709	4.57
Switzerland	NATIVE STUDENTS	2109	61298	75.99
	FIRST-GENERATION STUDENTS	495	12961	16.07
	NON-NATIVE STUDENTS	233	6409	7.94
England	NATIVE STUDENTS	2401	451871	85.09
	FIRST-GENERATION STUDENTS	247	47691	8.98
	NON-NATIVE STUDENTS	161	31463	5.92
Belgium (Flemish)	NATIVE STUDENTS	2585	59396	89.28
	FIRST-GENERATION STUDENTS	180	4010	6.03
	NON-NATIVE STUDENTS	147	3120	4.69

Source: ICCS 2009 International Dataset – Data elaborated using IEA IDB analyzer and SPSS software

Starting from this assumption, the European research team decided to explore whether attitudes toward rights for immigrants varied significantly among students from non-immigrant and immigrant families. Therefore, scale scores were produced and compared for these two groups of students. The associated table (see Table 4 -3) of the ICCS 2009 European Report showed the average score for each European country on the construct (in a similar fashion with the ICCS 2009 International Report) and added two different columns to compare 'student from non-immigrant families' and 'student with immigrant-background' scores. These enabled average comparisons within countries (e.g. native versus immigrant students in country X) and mean comparisons across countries (e.g. average scores of native students in country X compared with country Y).

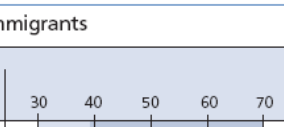
Following the distinction already adopted in the ICCS 2009 International Report, only two categories of students were compared referred to 'students from non-immigrant families', including students who were born in another country but whose parents had been born in the country of the test, and 'students with immigrant background', including non-native students and first-generation students.

As previously mentioned, the European picture is fairly mixed (see data relating to participants' immigration background as illustrated below), and not all the ICCS 2009 European countries presented sufficient large sub-samples of students with an immigrant background to be included in the analysis. The ICCS researchers fixed the minimum sub-sample size at 50 students from immigrant families.

As showed in Table 4 -2, for some European countries the number (and percentage) of participants in the 'first-generation students' and/or the 'non-native students' categories were very poor, consequently, despite the aggregation explained above, Slovak Republic, Poland, Malta, and Bulgaria immigrant background students' data were not investigated.

Table 4-3 National averages for students' attitudes toward equal rights for immigrants by immigrant background

Country	Students' Attitudes Toward Equal Rights for Immigrants								
	All students		Students from non-immigrant families		Students with immigrant background		Differences (non-immigrant)*		
Austria	48	(0.3)	46	(0.3)	54	(0.5)	8	(0.5)	
Belgium (Flemish) †	46	(0.3)	45	(0.3)	52	(0.6)	7	(0.7)	
Cyprus	49	(0.3)	49	(0.3)	52	(0.6)	3	(0.7)	
Czech Republic †	48	(0.2)	48	(0.2)	53	(1.0)	5	(1.0)	
Denmark †	48	(0.3)	48	(0.3)	55	(0.5)	7	(0.5)	
England ‡	46	(0.3)	45	(0.3)	53	(0.6)	8	(0.6)	
Estonia	48	(0.2)	47	(0.2)	52	(0.8)	4	(0.8)	
Finland	48	(0.3)	48	(0.3)	57	(1.0)	9	(1.0)	
Greece	51	(0.2)	51	(0.2)	54	(0.8)	3	(0.7)	
Ireland	50	(0.2)	49	(0.2)	55	(0.7)	6	(0.7)	
Italy	48	(0.3)	48	(0.3)	55	(0.7)	7	(0.7)	
Latvia	47	(0.2)	47	(0.2)	50	(1.1)	3	(1.1)	
Liechtenstein	48	(0.5)	46	(0.7)	50	(1.0)	4	(1.2)	
Lithuania	51	(0.2)	51	(0.2)	52	(0.9)	1	(0.9)	
Luxembourg	52	(0.2)	49	(0.2)	55	(0.3)	6	(0.4)	
Slovenia	50	(0.3)	50	(0.3)	53	(0.7)	3	(0.8)	
Spain	51	(0.3)	50	(0.3)	56	(0.6)	6	(0.7)	
Sweden	52	(0.4)	50	(0.4)	60	(0.5)	10	(0.7)	
Switzerland †	49	(0.3)	47	(0.3)	54	(0.5)	7	(0.6)	
European ICCS average	49	(0.1)	48	(0.1)	54	(0.2)	6	(0.2)	
ICCS average	50	(0.0)	50	(0.1)	53	(0.2)	3	(0.2)	
Country not meeting sampling requirements									
Netherlands	46	(0.4)	45	(0.3)	53	(1.2)	8	(1.3)	



Native students' score +/– confidence interval
Immigrant students' score +/– confidence interval

On average, students with a score in the range indicated by this color have more than a 50% probability of responding statements regarding equal rights for immigrants with:

	Disagree or strongly disagree
	Agree or strongly agree

Notes:

* Statistically significant ($p < 0.05$) coefficients in **bold**.
( ) Standard errors appear in parentheses. Because results are rounded to the nearest whole number, some totals may appear inconsistent.
† Met guidelines for sampling participation rates only after replacement schools were included.
‡ Nearly satisfied guidelines for sample participation only after replacement schools were included.
[1] National Desired Population does not cover all of International Desired Population.

Source: ICCS 2009 European Report, p. 92

The ICCS 2009 International Report (Schulz et al., 2010), the ICCS 2009 European Report (Kerr et al., 2010), and the ICCS 2009 Technical Report (Schulz et al., 2011) can be consulted for an exhaustive description of the ICCS 2009 methodologies and factor/scale properties.

4.2 Research Questions

The data concerning students' attitudes towards immigration, collected through the ICCS 2009 questionnaire (ICCS 2009 International Report; Schulz et al., 2008) and the league tables built and reported on such information (ICCS 2009 European Report, 2010), provided the premises for our empirical investigation of assessing measurement invariance.

Apart from the data structure and richness, the topic of immigrant rights caught our attention also from a theoretical point of view due to the increased mobility at European level and the recent migration phenomena. The relevance of such topic makes the issues of measurement invariance quite relevant and justified country comparisons that are very important for any further meaningful societal, scientific and policy discourses on the topic.

Therefore, building on the information described in this chapter, we elaborated our plan of testing for measurement invariance. To address the legitimacy of all possible comparisons at the international scale (mean comparisons across countries) we have operationalized the topic addressing four main research questions as follows:

a) Can country average levels of student attitudes toward equal rights for immigrants be compared with confidence among all European countries and/or relevant sub-groups of countries?

b) Can such comparisons be carried out also for sub-groups of students such as the non-immigrant/native students in these countries?

c) Can country average levels of student attitudes toward equal rights for immigrants also be compared when we consider only the group of students with an immigrant background in these countries?

d) Is it possible to identify reference country sub-groups for which measurement invariance holds at higher levels?

The plan of research took into account the data collected on the full battery of six items (see Figure 4 -1). Therefore, the analysis has been conducted at European level with regard to two formats of the instrument: the six-items battery of the original ICCS 2009 study and the five items battery used by the ICCS 2009 team to construct the *students' attitudes toward equal rights for immigrants* scale as reported in the ICCS 2009 International Report and the ICCS 2009 European Report (ICCS 2009 European Report, 2011, p. 92).

Finally, the scaling procedures referred to in the ICCS 2009 European Report significantly contributed to formulate the research questions set above. Nevertheless, it must be noted that the European country scaling presented in the aforementioned Report ranks European countries on the basis of the analyses done on the full international dataset (European scale scores have been estimated for the full international sample as reported in the ICCS 2009 International Report table). This approach is understandable especially if the aim is to keep the European Countries anchored to the overall results presented in the main study; while as clearly stated in the research questions, in this dissertation we have chosen to conduct all the analyses strictly on the sample (and subsamples) of European countries of interest.

4.3 Sample structure

The study has been conducted using data from the International Civic and Citizenship Education Study (ICCS) study carried out by the International Association for the Evaluation of Educational Achievement - IEA in 2009. The ICCS 2009 final data set includes data on Grade 8 (approximately 14 years of age) students' citizenship competences from 38 countries.

A two-stage cluster sample procedure was applied to identify the samples. In the first step to sampling, probability proportional to size procedures (referring to the number of students enrolled in a school) were employed obtaining a sample of about 150 schools in each country (ICCS 2009 International Report – Schulz et al., 2010).

During the second phase, within each school the students of a unique (*intact*) class were randomly selected to participate in the survey. This has resulted in country student samples varied from between 3.000 and 4.500 elements.

At country level the needed participation rate was 85 percent of the designed schools and 85 percent of the selected students for each of the participating schools, or '*a weighted overall participation rate of 75 percent*' (for more details see ICCS 2009 International Report, 2010).

With regard to the research questions of this dissertation, only European countries were of interest, therefore we initially considered 25 of the 38 countries participating in the ICCS 2009 survey. After a preliminary descripted statistical analysis, we excluded Liechtenstein due to its very small population (less than 40.000).

Table 4 - 4 ICCS 2009 European participation – Descriptive Statistics

Country	N Students (All)	Only immigrant background	Percentage
Austria	3261	615	19.38
Belgium (Flemish)	2912	327	10.72
Cyprus	3110	230	7.12
Czech Republic	4596	112	2.47
Estonia	2704	194	6.86
Denmark	4316	415	8.65
England	2809	408	14.91
Finland	3274	78	2.36
Greece	3125	335	11.32
Ireland	3268	386	12.08
Italy	3303	233	7.26
Latvia	2719	151	4.91
Lithuania	3850	179	1.68
Luxembourg	4727	1867	43.14
Neherlands	1901	219	13.27
Slovenia	3030	322	10.16
Spain	3274	356	11.13
Sweden	3333	618	13.86
Switzerland	2837	728	24.01
Total	62349	7773	

Note: Due to the necessary data cleaning procedures the numbers reported here can be different from those published in the ICCS report (Schulz et al., 2010)

Furthermore, concerning the subsample of students with immigrant background, the ICCS 2009 European Report did not investigate data from sub-samples with fewer than 50 students. Accordingly, Bulgaria, Malta, Poland, and Slovak Republic were not included in the ICCS 2009 European table due to the small size of the samples of students with immigrant background.

Even though Norway participated in the ICCS 2009 survey, after the preliminary stage the national research coordinators (NRCs) decided not to be part of the European module, therefore this country was also excluded from the dataset.

The Netherlands score appeared in a distinct section of the table due to Dutch sample characteristics (lower participation rates), which imposed a separate treatment of its results by the ICCS 2009 research team[11].

Moreover, following the distinction already adopted in the ICCS 2009 European Report, only two categories of students were considered: 'students from non-immigrant families', including students who were born in another country but whose parents had been born in the country of the test, and 'students with immigrant background', including non-native students and first-generation students.

As a result, we created a further subsample of 18 countries with a relevant immigrant subgroup of at least 50 students, as follows: Austria, Belgium (Flanders), Cyprus, Czech Republic, Denmark, England, Estonia, Finland, Greece, Ireland, Italy, Latvia, Lithuania, Luxembourg, Slovenia, Spain, Sweden, and Switzerland (all the countries reported in the ICCS 2009 European Report but Liechtenstein).

Figures are reported in Table 4 -4 (see Appendix for further details on sample characteristics concerning single items) concerning the student data sample. The number of students selected across these 18 countries was 60,448, of which 7,773 students with an immigrant background. These numbers are estimated after the data cleaning procedures. These referred to the categorical variable indicating the background-status of the student (native, first or second generation immigrant). If was answer categories on

[11] With respect to student sampling participation rates countries were distinguished into three Categories. The Netherlands were placed into Category 3: Unacceptable sampling response rate even when replacement schools were included.

this variable were omitted or invalid, the student record was not further examined. The missing data represented about 1.7% of the data included in the analysis for all countries.

After a careful examination of country data, and referring to the recent literature on measurement equivalence testing with multivariate and large-scale sample size, we decided to adopt a more conservative approach to sample size selection. We therefore carried out also an investigation of optimal sample sizes by country and identified countries with an immigrant sub-sample of at least 200 students (Boomsma & Hoogland, 2001; Byrne & van de Vijver, 2010). In spite of our prior aggregation of the two categories of the first-generation and non-native students into the unique group of students with immigrant background, Czech Republic, Estonia, Finland, Latvia, and Lithuania did not reach the required threshold of 200 elements for the immigrant sub-sample, and consequently they were excluded from this research step.

Following this further cut, we obtained a sub-sample of 13 European countries (12 European Union member states and Switzerland): Austria, Belgium (Flanders), Cyprus, Denmark, England, Greece, Ireland, Italy, Luxembourg, Slovenia, Spain, Sweden, and Switzerland. The number of students selected for the 13 countries resulted 43,305, of which 6,840 students with immigrant background.

4.4 Data sources and characteristics

The data were downloaded in SPSS format from the ICCS dataset available on the IEA site (http://rms.iea-dpc.org/#) selecting all the original 24 countries of interest (Liechtenstein was not included). IEA provides single raw data files at country level.

Table 4-5 ICCS 2009 International database – Relevant variables

Variable name	Type	Width	Decimals	Label	Values/Comments
IDCNTRY	Numeric	5	0	*COUNTRY ID*	Code of the country
IS2P26A	Numeric	1	0	IMMIGRANTS-SPEAKING OWN LANGUAGE	1 - STRONGLY AGREE 2 - AGREE 3 - DISAGREE 4 - STRONGLY DISAGREE
IS2P26B	Numeric	1	0	IMMIGRANTS-SAME OPPORTUNITIES EDUCATION	1 - STRONGLY AGREE 2 - AGREE 3 - DISAGREE 4 - STRONGLY DISAGREE
IS2P26C	Numeric	1	0	IMMIGRANTS-OPPORTUNITY VOTE IN ELECTIONS	1 - STRONGLY AGREE 2 - AGREE 3 - DISAGREE 4 - STRONGLY DISAGREE
IS2P26D	Numeric	1	0	IMMIGRANTS-CONTINUE OWN CUSTOMS	1 - STRONGLY AGREE 2 - AGREE 3 - DISAGREE 4 - STRONGLY DISAGREE
IS2P26E	Numeric	1	0	IMMIGRANTS-SAME RIGHTS AS EVERYONE	1 - STRONGLY AGREE 2 - AGREE 3 - DISAGREE 4 - STRONGLY DISAGREE
IS2P26F	Numeric	1	0	IMMIGRANTS-NOT MANY JOBS RESTRICT IMMIG.	1 - STRONGLY AGREE 2 - AGREE 3 - DISAGREE 4 - STRONGLY DISAGREE
TOTWGTS	Numeric	8	3	*FINAL STUDENT WEIGHT*	The final student weight of each student k in class j of school i in stratum h is the product of the five student-weight components: $TOTWGTS_{hijk} = WGTFAC1_{hi} \times WGTADJ1S_{hi} \times WGTFAC2Shij \times WGTADJ2S_{hi} \times WGTADJ3S_{hijk}$
IDGRADE	Numeric	2	0	*GRADE ID*	7 - GRADE 7 8 - GRADE 8 9 - GRADE 9 10 - GRADE 10 99 - OMITTED
IMMIG	Numeric	1	0	*IMMIGRATION BACKGROUND*	1 - NATIVE STUDENTS 2 - FIRST-GENERATION STUDENTS 3 - NON-NATIVE STUDENTS 7 - INVALID 9 - OMITTED

Source: ICCS 2009 International Database (http://rms.iea-dpc.org/#)

From the same site, we obtained the IDB Analyzer developed by IEA. As stated in the relevant page '*the IDB Analyzer is software used to combine and analyze data from IEA studies such as TIMSS*[12]*, TIMSS Advanced*[13]*, PIRLS*[14]*, SITES*[15]*, TEDS*[16]*, CivED*[17]*, ICCS and other large-scale assessments. It creates SPSS syntax that can be used to perform analysis with the aforementioned international databases*' (http://www.iea.nl/eula.html) and to merge the files.

Table 4–6 Students data – Descriptive Statistics

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	NATIVE STUDENTS	68492	87.4	89.3	89.3
	FIRST-GENERATION STUDENTS	4809	6.1	6.3	95.6
	NON-NATIVE STUDENTS	3387	4.3	4.4	100.0
	Total	76688	97.8	100.0	
Missing	INVALID	60	.1		
	OMITTED	1237	1.6		
	System	418	.5		
	Total	1715	2.2		
Total		78403	100.0		

[12] Trends in International Mathematics and Science Study
[13] Trends in International Mathematics and Science Study Advanced
[14] Progress in International Reading Literacy Study
[15] Second Information Technology in Education Study
[16] Teacher Education and Development Study
[17] Civic Education Study

Using the Merge Module of the IDB Analyzer, a unique data set was created by combining data files from different countries. Information was selected at student level, including among other variables relating to country identifier, the six dependent variables referring to the six surveyed items, the final student weight, and the immigrant background status variable (Table 4 -5). Subsequently, the data set - combined using the Merge Module - was elaborated through the Analysis Module – IDB Analyzer and SPSS codes were created to obtain an initial dataset (please refer to Appendix). This constituted of about 78,400 records (Table 4 -6). Based on these records, some basic descriptive statistics were produced to better illustrate the data (i.e. Table 4-7).

Table 4-7 Students data – Descriptive Statistics

		IS2P26A	IS2P26B	IS2P26C	IS2P26D	IS2P26E	IS2P26F
N	Valid	76518	76607	76407	76209	76396	76120
	Missing	1885	1796	1996	2194	2007	2283

In the ICCS 2009 European Report table (ICCS 2009 European Report, 2010 - p. 92), the higher scale scores indicate more positive attitudes toward the rights of immigrant in society, accordingly item values were re-coded in our dataset:

1 - STRONGLY DISAGREE

2 - DISAGREE

3 - AGREE

4 - STRONGLY AGREE.

Moreover, the immigrant status variable was re-coded to 0 (zero) for first-generation students (prior value 2) and non-native students (prior value 3) to obtain only two categories of students:

0 - IMMIGRANT
1 - NATIVE.

Finally, missing data was recoded to '-9' to be used for further analysis in Mplus 7.3.

Table 4-8 ICCS 2009 Students data – IS2P26A item

IS2P26A - Immigrants should have the opportunity to continue speaking their own language						
COUNTRY - immigrant sub-sample less than 50	**Status**	**STRONGLY DISAGREE**	**DISAGREE**	**AGREE**	**STRONGLY AGREE**	**Sum of cases**
Bulgaria	immigrant		3	5	12	20
	native	117	374	1366	1203	3060
Malta	immigrant	4	4	15	12	35
	native	86	233	975	714	2008
Poland	immigrant	2	5	27	13	47
	native	72	364	1866	842	3144
Slovak Republic	immigrant	2	1	9	6	18
	native	157	796	1501	451	2905
COUNTRY - immigrant sub-sample less than 200	**Status**	**STRONGLY DISAGREE**	**DISAGREE**	**AGREE**	**STRONGLY AGREE**	**Sum of cases**
Czech Republic	immigrant	1	17	51	41	110
	native	373	1382	2170	532	4457
Estonia	immigrant	3	22	93	72	190
	native	194	840	1128	322	2484
Finland	immigrant	1	2	25	50	78
	native	213	837	1666	420	3136
Latvia	immigrant	2	19	85	43	149
	native	138	732	1296	381	2547
Lithuania	immigrant	1	11	97	68	177
	native	44	311	1993	1311	3659

Source: ICCS 2009 International Database (http://rms.iea-dpc.org/#)

Even though the two categories of the first-generation and non-native students were aggregated into the unique group of students with immigrant background, for some countries the sub-sample of students with an immigrant background was very small. These were Bulgaria, Malta, Poland, and Slovak Republic with immigrant sub-sample of less than 50 elements, and Czech Republic, Estonia, Finland, Latvia, and Lithuania with less than 200 students. Table 4 –8 gives an example of the small rate of response in in these countries for the immigrant subsamples with respect to the first questionnaire item.

Summary descriptive statistics for the five items used by IEA researchers to measure the *Student's attitudes towards equal right for immigrants* scale as well as for the 6th relevant item of the ICCS 2009 questionnaire are reported in Appendix.

4.5 Main data analysis strategy

Building upon the data previously detailed, we proceeded with testing for measurement invariance. The data were analyzed in a multi-group factor analysis framework using Mplus 7.3. software (Muthén & Muthén, 2012; Muthén & Muthén, 2014).

Prior to performing the main analysis, the conversion of data was required due to the specific format input accepted by Mplus. Thus, the dataset produced in SPSS was converted into a database file (.dat) and adequately structured to be used with Mplus.

As extensively explained in Part I, in a first step it was necessary to test the least restrictive model – configural invariance (Horn & McArdle, 1992), and only if this level of equivalence was established (Byrne & van de Vijver, 2010; Byrne, 2012) we could verified more restrictive models, that is metric (weak) invariance and scalar (strong) invariance.

Moreover, we considered that data had three discriminant variables to be taken into account for our research objectives: countries for the multi-group analysis, students' codified answers for factor analysis (six or five item model), and the background status (native/immigrant) relevant for the two separated analyses (as indicated in the previous chapter).

Based on these preliminary considerations, our plan of testing was articulated into two analogue but parallel analyses respectively referring to the six-item model and the five-item model, in this last case with specific reference to the ICCS 2009 *Students' attitudes towards immigrant* scale. Therefore, the steps described hereafter were common for both the models.

Initially, adopting a bottom-up approach and with the aim of confirming the possibility of a common baseline model for the measured countries, we started with a CFA of a one-factor model with respect to each of 24 countries considered individually.

After testing for the fit of the one-factor model in each country, we could proceed with the check of the measurement invariance properties.

As said, the first step involved testing for configural measurement invariance. The configural measurement equivalence analysis was conducted for the whole group of 24 European countries. Furthermore, following the ICCS 2009 European research team's choice of analyzing only countries with an immigrant subgroup of at least 50 students, a second run for testing configural invariance involved the 18 countries which reached the immigrant subgroup limit. Additionally, assuming a more restrictive approach concerning sample dimensionality in large-scale analysis, a third test was run for only the 13 countries with an immigrant subgroup of at least 200 students (Boomsma & Hoogland, 2001; Byrne & van de Vijver, 2010). In case configural equivalence held, further tests for metric and scalar invariance were run and results evaluated.

In the second step, measurement invariance properties were analyzed with respect to both the two subgroups of non-immigrant/native and immigrant background students. Starting from the original dataset, two different input files were prepared in SPSS with the aim of distinguishing the students' records on the background status variable. As previously described for the whole sample dataset, the two SPSS datasets were converted into database files (.dat files) for their use with Mplus software. As usual, we started with a configural invariance assessment. When configural equivalence was verified, further tests for metric and scalar invariance were run and results evaluated.

Next, we proceeded with the examination of the immigrant students subgroup. Due to the small subgroup size and in particular for the too much reduced number of cases surveyed

for some European countries, we were obliged to restrain our analysis at the 13 countries level.

Still, our prior concern was to check for configural measurement invariance, and further investigating both the weak invariance and strong invariance properties of the immigrant background student data.

Moreover, on the basis of the information provided by the Mplus output files during the previously analyses, and in particular evaluating the misfit contributions of the different countries, we tried to identify a possible subgroup of countries better fitting a baseline model and for which measurement invariance resulted at the highest level.

Finally, equality constraints regarding variables (full invariant analysis) were relaxed and tentative partial versions of invariance were investigated with regard to both the preliminary item models.

Concerning the model fit statistics used to evaluate measurement invariance in our study, based on the latest available literature, we decided to refer to the root-mean-square error of approximation – RMSEA (Steiger & Lind, 1980), the comparative fit index – CFI (Bentler, 1990), and the Tucker-Lewis fit index – TLI (Tucker & Lewis, 1973).

Actually, Mplus outputs offer several goodness-of-fit values, all of which relate to a model as a whole (Byrne, 2012, p. 66). Nevertheless, on the basis of our sample/data structures we were obliged to avoid more common fit statistic like a chi-square (χ^2) statistic.

As an 'absolute misfit index' (Browne et al. 2002), the RMSEA is correlated in a negative way to model fit, that is it increases as goodness of fit decreases. Commonly, RMSEA values below 0.05 indicate good fit, and values until 0.08 can be considered as a signal of an acceptable level of errors of approximation, thus a reasonable low level of

noise in the model (Browne & Cudek, 1993; MacCallum et al., 1996; Hu & Bentler, 1999).

Therefore, for the RMSEA range we initially referred to a value as high as 0.08 both for the index and its upper boundary of 90% confidence interval. When this limit was exceed we decided to adopt a less restrictive rule and refer the model fit assessment to RMSEA values less than 0.10 (Byrne & van de Vijver, 2010; Kline, 2011, Rutkowski & Svetina, 2013), but clearly stating this decision.

The CFI and the TLI are incremental indices which measure the improvement in model fit comparing the constrained model with the less restricted nested one. These indices are positively correlated to model fit, meaning that they increase as goodness of fit increases.

For both these indices, values higher than 0.90 were normally considered acceptable. More recently, a revised cutoff value close to 0.95 has been suggested, but its strength was questioned (Bentler, 1992; Hu & Bentler, 1999; Marsh et al., 2012; Byrne, 2012; Kline, 2011).

Consequently, we referred to the cutoff value of 0.95, but when this lower limit was not satisfied we decided to adopt a less strict approach and refer the model fit assessment to CFI/TLI values over 0.90. When this approach was assumed we clearly stated the decision.

In the estimation procedure the categorical character of our item variables has been taken into account and the robust weighted least square estimator – WLSMV was selected in CFA analyses run with Mplus software (Muthén & Muthén, 2012)[18].

[18] When at least one factor indicator or other observed dependent variable is binary or ordered categorical, Mplus has seven estimator choices: weighted least squares (WLS), robust weighted least squares (WLSM, WLSMV), maximum likelihood (ML), maximum likelihood with robust standard errors and chi-square (MLR, MLF), and unweighted least squares (ULS) When at least one factor indicator or other observed dependent variable is censored, unordered categorical, or a count, Mplus has six estimator choices: weighted least squares (WLS), robust weighted least squares (WLSM, WLSMV), maximum likelihood (ML), and maximum likelihood with robust standard errors and chi-square (MLR, MLF). (http://www.statmodel.com). Our choice was the default one.

Chapter 5

RESULTS

In the current chapter research results are showed. More precisely, results are illustrated in answer to the 4 research questions and two separate sets of analyses: one conducted with regard to the six items of the original ICCS 2009 questionnaire and the other for the only five items used for the building of the *students' attitudes toward equal rights for immigrants* scale as reported in the ICCS 2009 International Report and the ICCS 2009 European Report (ICCS 2009 European Report, 2010, p. 92). Having as reference these two analogous but different constructs based respectively on six items (observed variables for questionnaire) and five items (observed variables used for scale), each research question was investigated twice. In the first section of this chapter, results for the six item model are presented. Investigations are applied both to all European countries and subgroups (18 and 13 countries), and with regard to all students sample (first research question) as well as to subgroups of students, that is the non-immigrant/native students (second question) and students with an immigrant background (third research question). Similarly, in the next section, results are shown for the five item construct,

following each of the prior analysis steps (all European countries and subgroups, all students sample and native/immigrant background subgroups). Finally, the last section is dedicated to explorative results about country groupings with better fit and partial invariance (fourth research question).

5.1 Six item model

All students

The first model tested for measurement invariance involved all the six items included in the ICCS questionnaire (Brese et al., ICCS 2009 User Guide for the International Data base Supplement 1, 2014).

The results indicated a modest fit of the one-factor model when 24 European countries[19] were analyzed one by one. In fact, as shown in Table 5-1, the one-factor model was not well fitted in all the countries included in the study.

[19] Austria, Belgium (Flemish), Bulgaria, Cyprus, Czech Republic, Denmark, England, Estonia, Finland, Greece, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, The Netherlands, Norway, Poland, Slovak Republic, Slovenia, Spain, Sweden, and Switzerland. Liechtenstein was excluded due to its very small population (less than 40.000)

Table 5-1 Six item Model - fit statistics

Country	RMSEA	90 Percent C.I.	CFI	TLI	Country	RMSEA	90 Percent C.I.	CFI	TLI
Austria	0.086	0.076 0.096	0.979	0.966	Latvia	0.058	0.048 0.069	0.982	0.970
Belgium (Flemish)	0.134	0.124 0.145	0.933	0.889	Lithuania	0.071	0.062 0.080	0.970	0.950
Bulgaria	0.061	0.051 0.071	0.984	0.974	Luxembourg	0.064	0.056 0.072	0.988	0.981
Cyprus	0.054	0.044 0.065	0.991	0.985	Malta	0.046	0.034 0.059	0.992	0.986
Czech Republic	0.078	0.070 0.086	0.981	0.969	Netherlands	0.134	0.121 0.147	0.918	0.864
Denmark	0.133	0.124 0.141	0.965	0.942	Norway	0.112	0.102 0.122	0.980	0.966
England	0.103	0.093 0.113	0.987	0.979	Poland	0.053	0.044 0.064	0.990	0.984
Estonia	0.111	0.100 0.121	0.954	0.924	Slovak Republic	0.081	0.071 0.091	0.975	0.958
Finland	0.124	0.114 0.134	0.980	0.967	Slovenia	0.100	0.090 0.110	0.974	0.957
Greece	0.065	0.055 0.075	0.986	0.976	Spain	0.130	0.120 0.139	0.960	0.934
Ireland	0.082	0.072 0.092	0.986	0.977	Sweden	0.117	0.108 0.127	0.988	0.980
Italy	0.129	0.120 0.139	0.957	0.928	Switzerland	0.087	0.077 0.097	0.977	0.962

Note: RMSEA = root mean square error of approximation; 90 Percent C.I. = 90% Confidence Interval for RMSEA; CFI = comparative fit index; TLI = Tucker-Lewis index

More specifically, only 9 countries fitted the model perfectly, RMSEA < 0.08 and CFI/TLI > 0.95: Bulgaria, Cyprus, Czech Republic, Greece, Latvia, Lithuania, Luxembourg, Malta, and Poland. Adopting the proposed less strict approach for the fit assessment, that was to reduce the limit for the RMSEA to RMSEA < 0.10 (CFI and TLI are confirmed over 0.95), further 5 countries could be considered as fitting the model: Austria, Ireland, Slovak Republic, Slovenia, and Switzerland. The remaining 10 countries did not fit the model.

Nevertheless, when the data for all 24 countries were analyzed simultaneously (pulled dataset without applying a multi-group approach), the results showed that the fit of the one-factor model was very good: RMSEA = 0.069 (90 Percent C.I. 0.067 - 0.071), CFI = 0.969, and TLI = 0.948.

With the aim of answering the first research question, we proceed with the measurement invariance testing in a multiple-group confirmatory factor analysis framework. The results for the six-item model revealed a quite questionable level of configural invariance when 24 countries were considered (Table 5-2).

Table 5 -2 Six items - fit statistics

Country	RMSEA	90 Percent C.I.	CFI	TLI
Configural Invariance				
All European countries	0.095	0.093 - 0.097	0.979	0.964
18 countries	0.098	0.096 - 0.100	0.978	0.964
13 countries	0.102	0.100 - 0.105	0.978	0.964
Metric Invariance				
All European countries	0.110	0.108 - 0.112	0.956	0.952
18 countries	0.116	0.114 - 0.117	0.954	0.950
Scalar Invariance				
All European countries	0.109	0.108 - 0.111	0.923	0.953
18 countries	0.112	0.111 - 0.114	0.925	0.953

Note: RMSEA = root mean square error of approximation; 90 Percent C.I. = 90% Confidence Interval for RMSEA; CFI = comparative fit index; TLI = Tucker-Lewis index

As a matter of fact, our results for the configural invariance test might have been considered as borderline, with the RMSEA = 0.095 (90 Percent C.I. 0.093 - 0.097), CFI = 0.979 and TLI=0.964. The required value of RMSEA < 0.08 was not achieved, yet we had to consider the peculiar large-scale structure of our data and the cross-group analysis conducted. Therefore, following Byrne and Van de Vijver (2010), Kline (2011), and Rutkowski and Svetina (2013) we accepted a less strict value for the RMSEA, that is RMSEA < 0.10, and consequently we considered the configural invariance verified for all 24 European countries.

Having established the configural invariance for 24 countries, the metric and scalar invariance were investigated.

The results showed that neither metric invariance nor scalar equivalence held: the RMSEA value resulted unacceptable in both analyses (RMSEA = 0.110 for weak invariance and RMSEA = 0.109 for the strong invariance).

In the second step, following the lead of the procedures applied in the ICCS 2009 European Report (which investigated data only from subsamples bigger than 50 students), we reduced the country sample to the countries that fitted this criterion. Consequently, only 18 out of the 24 European countries[20] were considered for the analyses.

Adopting the prior discussed less strict approach for the RMSEA, the configural invariance could be verified for this subgroup with RMSEA = 0.098 (90 Percent C.I. 0.096 - 0.100), CFI=0.978 and, TLI = 0.964.

[20] Bulgaria, Malta, Poland, and Slovak Republic were not included in the study due to the small immigrant background student samples. In addition, The Netherlands and Norway were not included because of their exclusion from the European Report analysis.

Having established the configural invariance for 18 countries, the metric and scalar invariance were analyzed for this subgroup.

Further results for the weak measurement invariance and strong invariance attested a poor fit: namely RMSEA = 0.116 and CFI = 0.954 for metric equivalence, RMSEA = 0.112 and CFI=0.925 for scalar equivalence.

Thirdly in a subsequent step we further reduced the number of countries to those with an immigrant subsample of at least 200 students[21]. This procedure resulted in the selection of a subgroup of 13 countries with an immigrant subsample of at least 200 students[22]. Analyses applied to this subsample showed that the configural invariance did not hold with a poor level of RMSEA = 0.102 (90 Percent C.I. 0.100 - 0.105), CFI = 0.978 and TLI = 0.964.

Thus, given that configural invariance was not achieved higher levels of equivalence were not investigated further for this subgroup of countries.

Overall, the results obtained for the six-item model showed a modest level of measurement invariance. In fact, for all 24 European countries and the two subgroups of countries considered, both metric invariance and scalar invariance were not reached. Only a common structure in the factor held for 24 and 18 countries cases, when a more lenient approach was taken to model fit evaluation (RMSEA).

[21] On the basis of documented literature about sample dimensionality in large-scale data analysis (Boomsma & Hoogland, 2001; Byrne & van de Vijver, 2010), it was decided to adopt a conservative approach in terms of sample dimension. Consequently, only countries with an immigrant subsample of at least 200 students have been investigated for the third research question: '*Can country average levels of student attitudes toward equal rights for immigrants also be compared when we consider only the group of students with an immigrant background in these countries?*'

[22] In the final sample were included: Austria, Belgium (Flemish), Cyprus, Denmark, England, Greece, Ireland, Italy, Luxembourg, Slovenia, Spain, Sweden, and Switzerland. For the completeness of the dissertation also other sample students (all students and native ones) were investigated with regard to this 13 country sample

Native students

With regard to the second research question, similar analyses were applied to the subsample of native students.

When 24 countries were considered, the results for the six-item model showed that configural invariance held only if we accepted a less strict value for the RMSEA, namely RMSEA < 0.10, while both metric and scalar invariance were not verified (Table 5-3).

For the subgroup of 18 countries, configural invariance could not be assumed with RMSEA=0.103 (90 Percent C.I. 0.101 - 0.106), CFI = 0.975 and TLI = 0.959. Therefore, higher levels of equivalence were not tested for this subgroup of countries.

Similarly, for the subgroup of 13 countries, configural invariance was not demonstrated and no other higher levels could be investigated as illustrated in Table 5 -3.

Table 5 -3 Six items - fit statistics

Country	RMSEA	90 Percent C.I.	CFI	TLI
		Configural Invariance		
All European countries	0.100	0.097 0.102	0.976	0.959
18 countries	0.103	0.101 0.106	0.975	0.959
13 countries	0.109	0.106 0.112	0.974	0.957
		Metric Invariance		
All European countries	0.111	0.110 0.113	0.953	0.949
		Scalar Invariance		
All European countries	0.113	0.112 0.114	0.915	0.947

Note: RMSEA = root mean square error of approximation; 90 Percent C.I. = 90% Confidence Interval for RMSEA; CFI = comparative fit index; TLI = Tucker-Lewis index

Immigrant background students

The third research question regarded the possibility of valid data comparisons for the immigrant background student subsample.

Following relevant literature (Boomsma & Hoogland, 2001; Byrne & van de Vijver, 2010), a conservative approach has been adopted. Consequently, only countries with an immigrant subsample of at least 200 students have been investigated. Moreover, given the reduced number of response for some countries[23] involved in the ICCS 2009 survey[24], measurement invariance tests with respect to the immigrant sub-sample could be possible only referring to a subsample of 13 countries.

The results for configural invariance test showed a good level of equivalence: RMSEA=0.071 (90 Percent C.I. 0.064 - 0.079), CFI = 0.986 and TLI = 0.977. Being the RMSEA < .08 and CFI/TLI > 0.95, consequently we could assume the needed cross-country configural measurement equivalence (Table 5 -4).

Furthermore, higher levels of equivalence were investigated.

Table 5 -4 Six items - fit statistics

Country	RMSEA	90 Percent C.I.		CFI	TLI
Immigrant background students - 13 countries					
Configural invariance	0.071	0.064	0.079	0.986	0.977
Metric invariance	0.055	0.049	0.062	0.987	0.986
Scalar invariance	0.056	0.051	0.061	0.978	0.986

Note: RMSEA = root mean square error of approximation; 90 Percent C.I. = 90% Confidence Interval for RMSEA; CFI = comparative fit index; TLI = Tucker-Lewis index

[23] Bulgaria, Cyprus, Czech Republic, Estonia, Finland, Latvia, Lithuania, Malta, and Poland

[24] The Netherlands and Norway were not included because of their exclusion in the European Report table.

The results obtained for the metric invariance test denoted a very good fit: RMSEA=0.055 (90 Percent C.I. 0.049 - 0.062), CFI = 0.987 and TLI = 0.986.

Moreover, the strong (scalar) measurement invariance was found tenable. These last results confirmed that all the examined measurement invariance tests held for the subsample of the immigrant background students.

Interpreting these results on the basis of the theory illustrated in the previous chapter allows to state that data from immigrant background students (included in the considered subsample of 13 countries) can be validly compared.

The configural equivalence assures a common factor model across the country-groups. The metric equivalence guarantees the same 'strength of the relation' between the independent factor and other variables. Finally, the reached scalar invariance would allow investigations in a comparative perspective (i.e. comparing of average performances) and a wide range of secondary data analysis.

5.2 Students' attitudes towards equal rights for immigrants - Five item model

All students

In a second stage, we analyzed the one-factor model with regard to the 5 items answers on which the *Students' attitudes towards equal rights for immigrants* concept had been scaled in the ICCS 2009 study (ICCS 2009 International Report – Schulz et al., 2010**;** ICCS 2009 European Report – Kerr et al., 2010).

The results for 24 European countries analyzed on a one by one basis showed a worse fit compared to the prior analysis with six items (in Table 5 -5 fitting countries are highlighted).

More specifically, only 5 countries fitted the model perfectly complying with the standard values of RMSEA < 0.08 and CFI/TLI > 0.95: Bulgaria, Cyprus, Latvia, Lithuania, Malta, and Poland. Adopting a less strict approach to model fit assessment (that is reducing the limit for the RMSEA to RMSEA< 0.10 and accepting CFI and TLI values of over 0.95), 5 additional countries might be considered as fitting the model: Austria, Czech Republic, Greece, Lithuania, and Luxembourg. The last 14 countries did not fit the model, with very poor fit particularly for Belgium (Flemish), Denmark, Italy, The Netherlands, and Spain.

With regard to the fit of the one-factor model for all 24 countries simultaneously analyzed (pulled data, no multiple group approach), the results showed a questionable level of RMSEA = 0.092 (90 Percent C.I. 0.089 0.094), a CFI=0.968, and TLI = 0.936 Still, we took the decision of adopting a less strict approach referring to the following limits for the evaluated indices: RMSEA < 0.10, CFI > 0.90, and TLI>0.90. After this, we could consider the one-factor model (5 items) fitting for all countries.

Table 5 -5 Five item Model - fit statistics

Country	RMSEA	90 Percent C.I.	CFI	TLI	Country	RMSEA	90 Percent C.I.	CFI	TLI
Austria	0.100	0.088 0.113	0.984	0.968	Latvia	0.067	0.053 0.082	0.977	0.954
Belgium (Flemish)	0.179	0.166 0.193	0.932	0.863	Lithuania	0.099	0.088 0.111	0.967	0.933
Bulgaria	0.076	0.063 0.090	0.986	0.972	Luxembourg	0.081	0.071 0.092	0.989	0.979
Cyprus	0.042	0.029 0.057	0.997	0.994	Malta	0.055	0.039 0.072	0.993	0.986
Czech Republic	0.099	0.088 0.110	0.983	0.966	Netherlands	0.176	0.159 0.193	0.922	0.844
Denmark	0.172	0.161 0.183	0.967	0.933	Norway	0.146	0.132 0.160	0.981	0.961
England	0.126	0.113 0.140	0.989	0.978	Poland	0.074	0.062 0.088	0.990	0.990
Estonia	0.142	0.128 0.156	0.956	0.912	Slovak Republic	0.113	0.099 0.126	0.972	0.944
Finland	0.154	0.141 0.167	0.983	0.965	Slovenia	0.135	0.122 0.149	0.973	0.946
Greece	0.081	0.068 0.095	0.987	0.974	Spain	0.170	0.157 0.183	0.961	0.921
Ireland	0.108	0.095 0.121	0.986	0.972	Sweden	0.140	0.127 0.153	0.990	0.981
Italy	0.168	0.155 0.181	0.959	0.917	Switzerland	0.110	0.097 0.124	0.979	0.958

Note: RMSEA = root mean square error of approximation; 90 Percent C.I. = 90% Confidence Interval for RMSEA; CFI = comparative fit index; TLI = Tucker-Lewis index

Regarding the first research question, we tested for measurement invariance. The results were not better (Table 5-6). The configural measurement invariance did not hold with RMSEA = 0.124 (90 Percent C.I. 0.121 - 0.126), CFI=0.979, and TLI = 0.959. Thus, higher levels of equivalence were not investigated.

Table 5 -6 Five items - fit statistics

Country	RMSEA	90 Percent C.I.		CFI	TLI
Configural Invariance - All students					
All European countries	0.124	0.121	0.126	0.979	0.959
18 countries	0.127	0.124	0.130	0.979	0.959
13 countries	0.132	0.129	0.136	0.979	0.959

Note: RMSEA = root mean square error of approximation; 90 Percent C.I. = 90% Confidence Interval for RMSEA; CFI = comparative fit index; TLI = Tucker-Lewis index

Furthermore, the test for invariance regarding the 18 countries subsamples returned the same misfit values (see Table 5-6). Therefore, no level of measurement equivalence could be identified.

Finally, also the results for 13 countries revealed no configural measurement invariance: RMSEA= 0.132 (90 Percent C.I. 0.129 0.136), CFI = 0.979, and TLI = 0.959 (see Table 5-6) .

Native students

Relating to native students, under the scope of the second research question of this dissertation, results showed that measurement invariance did not hold for all countries as well as for the two analyzed subsamples.

When 24 countries were considered, configural invariance tests showed a result of RMSEA=0.130. Similar results were reached with tests for measurement equivalence for the subgroup of 18 countries (RMSEA = 0.134) and for the subgroup of 13 countries (RMSEA = 0.142). Due to the lack of configural measurement invariance, higher levels of equivalence were not further investigated.

Table 5-7 presents the results of the configural invariance tests:

Table 5 -7 Five items - fit statistics

Country	RMSEA	90 Percent C.I.		CFI	TLI
Configural Invariance - Native students					
All European countries	0.130	0.127	0.133	0.976	0.953
18 countries	0.134	0.130	0.137	0.976	0.953
13 countries	0.142	0.138	0.146	0.975	0.951

Note: RMSEA = root mean square error of approximation; 90 Percent C.I. = 90% Confidence Interval for RMSEA; CFI = comparative fit index; TLI = Tucker-Lewis index

Immigrant background students

Analyses carried out for the subsample of the immigrant background students (third research question) with regard to the *Students' attitudes towards equal rights for immigrants* concept showed that configural equivalence could be assumed if a less strict approach was applied for the thresholds of the goodness-of-fit indices evaluating the model. As for the prior case, we decided to consider the new boundaries RMSEA < 0.10 and CFI/TLI > 0.90 as acceptable.

After this choice, referring to the subsample of 13 countries, the configural invariance was proved on the basis of the following results: RMSEA = 0.087 (90 Percent C.I. 0.078 0.097), CFI = 0.988 and TLI = 0.977.

Moreover, both weak and strong measurement invariance were confirmed. Table 5-8 shows the relevant results of the measurement invariance tests.

Table 5-8 Five items - fit statistics

Country	RMSEA	90 Percent C.I.		CFI	TLI
Immigrant background students - 13 countries					
Configural invariance	0.087	0.078	0.097	0.988	0.977
Metric invariance	0.058	0.050	0.066	0.991	0.990
Scalar invariance	0.060	0.054	0.065	0.981	0.989

Note: RMSEA = root mean square error of approximation; 90 Percent C.I. = 90% Confidence Interval for RMSEA; CFI = comparative fit index; TLI = Tucker-Lewis index

The results for strong invariance were very satisfactory: RMSEA = 0.060 (90 Percent C.I. 0.054 - 0.065), CFI = 0.981, and definitely authorize researchers to data comparison at cross-country level.

5.3 Further findings

As shown in the previous paragraphs, there are evident limitations in guaranteeing equivalence and thus ample cross-country comparability.

As posited in Chapter 3, these limitations can be overcome either by relaxing the requirements for configural invariance, or by exploring alternative strategies with more limited or specific scope.

Hence, with regard to the last research question relating to the possibility of identifying sub-groups for which measurement invariance held at higher levels, we conducted further tests. This final step of investigation was carried out along two different strands as follows:

- Discover possible country subgroups with better test performances referring to the prior batteries of six-items and five-items (Byrne & van de Vijver, 2010; Weziak-Bialowolska, 2013);

- Isolate a subset of variables (items) with improved results – partial invariance (Byrne, 2012, Byrne & van de Vijver, 2010).

This last step of study was conducted only with respect to the whole student sample.

On the basis of the misfit contributions values and proceeding with meticulous analyses, a subsample of 8 countries (Austria, Cyprus, England, Greece, Ireland, Luxembourg, Sweden, and Switzerland) was identified[25].

[25] Only countries with an immigrant subsample of at least 200 students were taken into account in this phase.

Table 5 -9 Six items - fit statistics

Country	RMSEA	90 Percent C.I.	CFI	TLI
All students - 8 countries				
Configural invariance	0.081	0.077 0.084	0.987	0.979
Metric invariance	0.083	0.080 0.086	0.980	0.987
Scalar invariance	0.078	0.076 0.080	0.970	0.980

Note: RMSEA = root mean square error of approximation; 90 Percent C.I. = 90% Confidence Interval for RMSEA; CFI = comparative fit index; TLI = Tucker-Lewis index

The results for the configural measurement invariance tests definitely improved respect both the six-item (Table 5 -9) and the five-item (Table 5 -10) cases conducted with other country subgroups.

Table 5 -10 Five items - fit statistics

Country	RMSEA	90 Percent C.I.	CFI	TLI
All students - 8 countries				
Configural invariance	0.099	0.095 0.104	0.989	0.978
Metric invariance	0.076	0.072 0.079	0.989	0.987
Scalar invariance	0.079	0.077 0.082	0.977	0.986

Note: RMSEA = root mean square error of approximation; 90 Percent C.I. = 90% Confidence Interval for RMSEA; CFI = comparative fit index; TLI = Tucker-Lewis index

The configural invariance test showed a result of RMSEA = 0.081 when six observed variables were considered, and a border value of RMSEA = 0.099 in the case of the five-item model. Still, a less strict approach was applied for the thresholds of the goodness-of-fit indices evaluating the models, and consequently the new boundaries of RMSEA < 0.10 and CFI/TLI > 0.90 were accepted.

After this decision, higher levels of equivalence were investigated.

For the six-item case, metric invariance was assumed on the basis of the prior choice of relaxing the RMSEA limit, being RMSEA = 0.083 (90 Percent C.I. 0.080 - 0.086), CFI = 0.980 and TLI = 0.987. The scalar invariance test gave a value of RMSEA = 0.078 (Table 5 -9).

The configural invariance test showed heavily borderline results for the alternative case of only 5 variables. However, the first level of equivalence was accepted applying a less strict rule for the RMSEA value.

Moreover, both weak and strong measurement invariance were confirmed with good values of RMSEA = 0.076 and RMSEA = 0.079 respectively. Table 5 -10 shows the relevant results for the measurement invariance tests in case of five observed variables.

Finally, we were able to detect a five-item battery non-including the first observed item[26], for which measurement invariance testing showed tenable and better fits for all levels of equivalence (Table 5 -11). This last case referred to all students and the 13 countries subsample.

Table 5 -11 Five items - fit statistics

Country	RMSEA	90 Percent C.I.	CFI	TLI
All students - without IP2P26A				
Configural invariance	0.052	0.048 0.055	0.996	0.992
Metric invariance	0.053	0.050 0.056	0.993	0.992
Scalar invariance	0.075	0.073 0.077	0.973	0.984

Note: RMSEA = root mean square error of approximation; 90 Percent C.I. = 90% Confidence Interval for RMSEA; CFI = comparative fit index; TLI = Tucker-Lewis index

[26] The first ICCS 2009 questionnaire question relating students' attitudes toward equal rights for immigrants was: IS2P26A- *Immigrants should have the opportunity to continue speaking their own language*?

Chapter 6

DISCUSSION AND CONCLUSIONS

This chapter concludes the dissertation. In section one, the main findings of the research work are summarized. Research questions are critically linked to the results of the study and discussed. Next, core conclusions concerning the research work are provided. Moreover, the second section is dedicated to pinpoint some limitations of the current study. Furthermore, some suggestions are made with regard to possible avenues for further research in the field of measurement invariance of instruments and constructs based on data collected across groups and cultures.

6.1 Summary of the findings and conclusions

This dissertation aimed to address the issue of MI of attitudinal measures and the need of statistical tests to be carried out in order to verify the comparability of data collected in International Large Scale Assessments - ILSAs via questionnaire instruments. In particular, the levels of invariance necessary to guarantee the significant comparison of the observed items as well as the meaningful definition of country scales referring to a common construct (i.e. built on country averages) have been examined.

The empirical study conducted for this dissertation is based on data from the International Civic and Citizenship Education Study (ICCS). The ICCS research was carried out by the International Association for the Evaluation of Educational Achievement - IEA in 2009 and its final data set includes data on Grade 8 (approximately 14 years of age) students' civic competences from 38 countries.

The assessment of measuring instruments and the validity of cross-country comparisons have been a priority for the ICCS team since the trial stage (Schulz et al., 2011), but a comprehensive invariance testing has not been drawn-out. '*The implication is that most scales in ICCS are still to be validated in order to compare constructs with some confidence across countries*' (Weziak-Bialowolska & Isac, 2014, p. 3) Starting from these views, we drew a research plan for assessing the measurement invariance of non-cognitive outcomes concerning European students' attitudes towards immigration. This data was collected through the ICCS survey via the student questionnaire (ICCS 2009 International Report, 2010; Schulz et al., 2008). More precisely, two different analyses have been conducted with regard to both the six-item battery of the original ICCS 2009 questionnaire and the only five-item battery used by the ICCS team for building the *students' attitudes toward equal rights for immigrants* scale as reported in the ICCS 2009 International Report and the ICCS 2009 European Report (ICCS 2009 European Report, 2010, p. 92).

Specifically, we addressed the following research questions:

a) Can country average levels of student attitudes toward equal rights for immigrants be compared with confidence among all European countries and/or relevant sub-groups of countries?

b) Can such comparisons be carried out also for sub-groups of students such as the non-immigrant/native students in these countries?

c) Can country average levels of student attitudes toward equal rights for immigrants also be compared when we consider only the group of students with an immigrant background in these countries?

d) Is it possible to identify reference country sub-groups for which measurement invariance holds at higher levels?

Having as reference two analogous but different models based respectively on six items (observed variables for questionnaire) and five items (observed variables used for scale), each research question was investigated twice. Namely, two similar parallel analyses were carried out referring to both the six-item model and the five-item model.

Measurement invariance was tested under a factor analytical framework and starting from a perspective of full measurement equivalence of the instrument used to measure students' attitudes toward equal rights for immigrants.

With respect to the first research question, we initially tested for measurement invariance using data on the full sample of students (regardless of immigrant status) for all 24 European countries[27] involved in the ICCS 2009 survey.

Furthermore, following the ICCS 2009 European research team's choice of analyzing only countries with an immigrant subgroup of at least 50 students, the measurement invariance properties were investigated regarding the only 18 countries which satisfied this subgroup limit (Austria, Belgium- Flanders, Cyprus, Czech Republic, Denmark, England, Estonia, Finland, Greece, Ireland, Italy, Latvia, Lithuania, Luxembourg, Slovenia, Spain, Sweden, and Switzerland).

Finally, based on related literature (i.e. Boomsma & Hoogland, 2001; Byrne & van de Vijver, 2010), a third set of tests was run for only the 13 countries (Czech Republic, Estonia, Finland, Latvia, and Lithuania were excluded) with an immigrant subgroup of at least 200 students.

With regard to the six-item model, results for all 24 countries and the subgroup of 18 countries revealed a quite questionable level of configural invariance when the whole sample of students was taken into account. As a matter of fact, results might be considered borderline. Nevertheless, supported by relevant literature (Byrne & van de Vijver, 2010; Kline, 2011) and given the large-scale structure of our data, we decided to adopt a less strict approach in evaluating the model fit. Consequently, we considered the configural invariance verified for these two cases, and we could proceed with further tests

[27] Austria, Belgium (Flemish), Bulgaria, Cyprus, Czech Republic, Denmark, England, Estonia, Finland, Greece, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands (The), Norway, Poland, Slovak Republic, Slovenia, Spain, Sweden, and Switzerland. Liechtenstein was excluded due to its very small population (less than 40.000)

of metric (weak) and scalar (strong) invariance. The results showed that neither metric invariance nor scalar equivalence held.

Next, we analyzed the subgroup of only 13 countries. The first level of equivalence did not hold. Due to the fact that configural invariance was not achieved, higher levels of equivalence were not investigated further for this subgroup of countries.

For all previous cases (including all students and only native students) the detected misfits implied that a broad measurement invariance of the instrument across countries (or subgroups) is not verified. In all 24 countries and the subgroup of 18 countries, where at least certain levels of configural equivalence were showed, a common understanding of the investigated concept could be assumed. Nevertheless due to the missing scalar invariance, the valid comparison of average performance on a scale was not supported.

With regard to five-item model of *students' attitudes toward equal rights for immigrants* scale, first, we tested measurement invariance respect to the whole sample of students. In all cases considered, that is all 24 countries, 18 countries, and 13 countries, the configural measurement invariance did not hold. Consequently, higher levels of equivalence were not investigated.

Next, with the aim of answering the second research question, our analyses were replicated considering the non-immigrant/native subsample, firstly referring to six observed variables and then to the case of only five items.

With respect to the six-item model, as for the whole student sample, in case of 24 countries, the results showed that configural invariance held only if we assumed a less strict rule in evaluating the model fit. Nevertheless, further exams for testing both metric and scalar invariance did not lead to a good model fit. Furthermore, for both the subgroups of 18 countries and the subgroup of 13 countries, configural invariance could

not be assumed. Therefore, higher levels of equivalence were not analyzed for these subgroups of countries.

For the parallel case of five-items, our investigation applied to the native students sample showed the similar results as for all students. More specifically no level of equivalence was reached.

Once more, the detected misfit for all levels of measurement equivalence illustrated the limitations of the data with regard to possible comparisons across countries. We can conclude that, given such findings, researchers must be cautious when using these data for secondary analyses and for comparing studies.

In response to the third research question, the last subgroup to be investigated was the subgroup of students with an immigrant background. The test was conducted with regard to the only 13 countries with an immigrant subgroup of at least 200 students.

In this last case, results were satisfactory in terms of verified levels of invariance both for the six and the five-item models. Tests for configural invariance showed a good level of fit, consequently we could assume the needed configural measurement equivalence and proceed to further tests. The results obtained for the metric invariance tests denoted a very good equivalence. Finally, the strong (scalar) measurement invariance was found tenable in this case.

On the basis of these results, we assumed that data relating to these subgroups of immigrant students can be validly compared across the analyzed countries. As argued in this dissertation, the configural equivalence assures a common factor model across the country-groups. The metric equivalence guarantees the same ‘strength of the relation’ between the independent factor and other variables. And finally, the reached scalar invariance allows investigations in a comparative perspective (i.e. comparisons of country average performance) and a wide range of other types of secondary data analysis.

In other words, it means that for this student subsample not only the loading configuration is the same across countries, but also the associated loading values, and intercepts. Therefore, the measurement instrument can be used for assessment including comparisons across groups involving mean values.

Having controlled for measurement invariance of the available data on a six-items and five items basis on all potential groupings, in a last step, we investigated explorative scenarios addressing our fourth research question. We proceeded to identify within or across the original groupings, smaller subgroups of countries or subsets of items that might be in fact comparable.

This additional step, conducted regarding the all students sample, resulted in the detection of two possible configurations with high level of measurement invariance.

First, it was possible to recognize a smaller subgroup of 8 countries (Austria, Cyprus, England, Greece, Ireland, Luxembourg, Sweden, and Switzerland) for which also scalar invariance held.

Second, a preliminary analysis for invariance of the whole student samples referring to subsets of observed variables showed the possibility of identifying models with higher levels of equivalence. A new subgroup of variables, where the first questionnaire item was excluded, was detected and investigated (partial invariance) referring to the 13 countries subgroup (Austria, Belgium -Flanders, Cyprus, Czech-Republic, Denmark, England, Greece, Ireland, Italy, Luxembourg, Spain, Sweden, and Switzerland). This last analysis showed very good results at all levels of equivalence.

Although these findings open up promising avenues, this scenario needs to be further investigated.

6.2 Conclusion, limitations and avenues for future studies

In this dissertation we stressed that ILSAs data provide a unique opportunity of information and their development responds to the '*increasingly challenging questions posed by researchers and policymaker around the world*' (Kirsch et al., 2013, p. 1).

However we also made the point that such studies have not always received the due attention in guaranteeing comparability (measurement invariance), nevertheless the collected data are used for cross-country comparisons.

In this respect, our current findings provide some valuable information on measurement invariance tests for non-cognitive LSA data illustrated with an empirical example. This example is based on ICCS 2009 and the items measuring students' attitudes towards immigrants. We stress however that the ICCS study and data are only one random example chosen for illustrative purposes and that the main conclusions outlined here are equally applicable to all other LSA's that valuably feed scientific and policy discussions.

Taken together, our results confirm an increasing body of literature that indicates that measurement invariance of questionnaire data in LSA's cannot be achieved easily. With our empirical example we illustrated that country averages comparisons are not always possible even if they are based on a sample of countries that share cultural similarity (in this case European countries). As our tests applied to several groupings of countries and individuals show, country mean comparisons may be defensible at times but, in many instances, their validity is not guaranteed. This has important implications for researchers and policy makers that may draw conclusions and take decisions based on country rankings.

Therefore, the most important conclusion of this research is that measurement invariance tests are a useful tool in assessing the validly of country comparisons and they should be employed and presented in official reports and empirical research papers in order to

enable readers to arrive to a correct interpretation of the data. Only when measurement invariance is empirically demonstrated, country rankings may lead to meaningful debate.

Nevertheless, as showed here, in some cases, lower levels of invariance such as configural invariance can be achieved or invariance may be reached for sub-groups of countries or individuals. This leads to the conclusion that researchers should investigate more in-depth possibilities to improve measurement invariance and that although country rankings may not be possible at times, LSA data remains useful for answering other research questions. Regarding the latter, when only tests of configural invariance are met, the researcher cannot compare means but may proceed with other types of analyses such as studying associations between the latent construct and other constructs of interest across countries.

Nonetheless, next to these conclusions, here we must also highlight some limitations that characterize this research and may provide some relating leads to be further tested in the future.

First, given the aims of the current study, we conducted an extensive exam of the data on different country subgroups and with respect to different subsets of variables. Yet our grouping remains substantially based on the geographical location (our analyses were always orientated to group comparisons). Still, following Byrne and van de Vijver (2010), in large-scale research other ways to group countries may be found and investigated (i.e. other groupings or sub-groupings based on a more refined cultural or political similarity).

Second, we could identify some cases in which measurement invariance did not hold and, in particular, showed that strong MI was often absent. Yet this kind of analysis does not help in detecting the reasons for non-invariance. Such dissimilarities could be explained also on the basis of a comprehensive framework involving cultural country characteristics and historical/political country background (i.e. historical correspondences or common

political experiences). Consequently, further research could be integrated with the examination of contextual country level variables that may explain non-invariance (Davidov et al., 2012; Weziak-Bialowolska & Isac, 2014). In addition, specific complementary cognitive tests could contribute to examine the reasons for measurement non-invariance and improve measurement quality (Davidov, 2008). Therefore, an interesting topic to investigate in the future could be the way in which measurement invariance analysis can be instrumental in detecting country difference explanations.

Moreover, we could not explore alternative techniques to assess MI or even just if other methods were possible. We worked under a full measurement invariance framework and our partial invariance test (subsets of variable) was only explored. In particular, further research could involve in a comparative way other procedures like Multilevel Structural Equations Modelling - SEM techniques (Davidov et al., 2012, Byrne, 2013), exploring partial invariance more extensively, or alignment method as very recently introduced (Muthén & Asparouhov, 2013).

Furthermore, our first requirement was to refer to a factor analytical framework, but as for the ICCS 2009 study, analyses could be done under an Item Response Theory context.

As a final point, in model fit evaluation, we accepted a less strict approach of the relevant statistics. In this respect, we have been fully supported by the relevant literature (Byrne & van de Vijver, 2010; Kline, 2011; Rutkowski & Svetina, 2013), but clearly without this first decision and taking a more conservative approach (Hu & Bentler, 1999) some of the findings and conclusions of this dissertation would have been different (i.e. configural invariance could not be accepted regarding the whole student sample).

Appendix

Table A -1 : IS2P26A* -Descriptive statistics

COUNTRY	Status	STRONGLY DISAGREE	DISAGREE	AGREE	STRONGLY AGREE	Sum of cases	STRONGLY DISAGREE	DISAGREE	AGREE	STRONGLY AGREE
Austria	immigrant	20	38	239	310	607	2.95	6.06	40.13	50.86
	native	331	581	1125	584	2621	13.09	22.71	42.31	21.88
Belgium (Flemish)	immigrant	15	57	144	110	326	3.76	17.95	45.04	33.25
	native	419	963	981	201	2564	17.23	37.23	37.62	7.92
Bulgaria	immigrant		3	5	12	20		13.82	21.44	64.74
	native	117	374	1366	1203	3060	3.95	12.11	45.10	38.85
Cyprus	immigrant	10	14	100	81	205	4.97	7.06	48.44	39.53
	native	215	335	1344	852	2746	7.93	12.05	48.98	31.04
Czech Republic	immigrant	1	17	51	41	110	0.87	15.70	46.35	37.07
	native	373	1382	2170	532	4457	8.41	31.04	48.59	11.96
Denmark	immigrant	11	41	192	160	404	3.26	10.93	46.78	39.03
	native	387	1210	1856	364	3817	10.32	32.02	48.34	9.32
England	immigrant	17	44	182	151	394	4.54	11.36	45.94	38.17
	native	298	683	1086	290	2357	12.32	29.09	46.00	12.58
Estonia	immigrant	3	22	93	72	190	1.79	12.99	49.16	36.06
	native	194	840	1128	322	2484	7.30	33.98	45.81	12.91
Finland	immigrant	1	2	25	50	78	1.27	3.27	33.17	62.29
	native	213	837	1666	420	3136	6.99	27.01	52.65	13.35
Greece	immigrant	8	24	110	180	322	2.75	8.16	33.67	55.41
	native	133	287	1318	973	2711	5.05	10.54	48.50	35.91
Ireland	immigrant	11	36	178	148	373	3.04	9.84	48.02	39.11
	native	207	573	1459	579	2818	7.56	20.27	51.53	20.64
Italy	immigrant	7	24	111	84	226	3.24	10.53	49.12	37.11
	native	263	848	1503	425	3039	8.72	27.67	49.38	14.23
Latvia	immigrant	2	19	85	43	149	1.96	12.06	56.63	29.34
	native	138	732	1296	381	2547	5.16	29.00	50.80	15.03
Lithuania	immigrant	1	11	97	68	177	2.67	9.85	51.02	36.46
	native	44	311	1993	1311	3659	1.23	8.21	53.45	37.10
Luxembourg	immigrant	38	155	700	923	1816	2.18	8.18	37.59	52.05
	native	313	620	1276	607	2816	10.95	21.73	44.93	22.39
Malta	immigrant	4	4	15	12	35	8.94	8.81	43.12	39.13
	native	86	233	975	714	2008	4.52	11.35	47.25	36.88
Netherlands	immigrant	8	27	103	76	214	3.08	12.64	48.74	35.54
	native	344	656	539	123	1662	21.43	37.92	33.01	7.64
Norway	immigrant	4	17	118	149	288	1.87	5.46	41.04	51.63
	native	217	509	1270	480	2476	9.07	20.95	50.51	19.47
Poland	immigrant	2	5	27	13	47	5.54	11.04	56.63	26.79
	native	72	364	1866	842	3144	2.37	11.48	59.78	26.36
Slovak Republic	immigrant	2	1	9	6	18	12.91	4.14	50.78	32.17
	native	157	796	1501	451	2905	5.28	27.40	51.87	15.45
Slovenia	immigrant	11	49	135	125	320	3.40	15.67	43.02	37.90
	native	252	544	1256	630	2682	9.47	20.35	46.52	23.67
Spain	immigrant	15	30	147	157	349	4.33	8.80	40.72	46.14
	native	235	616	1367	650	2868	8.15	20.96	47.56	23.32
Sweden	immigrant	8	23	147	422	600	1.83	3.35	24.42	70.41
	native	241	510	1284	644	2679	8.97	19.45	48.36	23.22
Switzerland	immigrant	19	68	317	304	708	2.33	7.99	43.11	46.57
	native	172	505	994	409	2080	8.25	25.71	46.49	19.56
Grand Total		**5639**	**16040**	**35949**	**17684**	**75312**				

*Immigrants should have the opportunity to continue speaking their own language

Note: Due to the necessary data cleaning procedures the numbers reported here can be different from those published in the ICCS 2009 International Report (Schulz et al., 2010)

Table A -2: IS2P26B* - Descriptive statistics

COUNTRY	Status	STRONGLY DISAGREE	DISAGREE	AGREE	STRONGLY AGREE	Sum of cases	STRONGLY DISAGREE	DISAGREE	AGREE	STRONGLY AGREE
Austria	immigrant	10	52	193	351	606	1.56	9.06	32.66	56.72
	native	123	270	1110	1114	2617	4.67	10.54	42.77	42.03
Belgium (Flemish)	immigrant	6	8	116	197	327	1.50	2.11	35.98	60.41
	native	66	137	1350	1023	2576	2.71	5.41	51.66	40.23
Bulgaria	immigrant			6	14	20			27.84	72.16
	native	64	168	1118	1712	3062	2.02	5.66	37.71	54.61
Cyprus	immigrant	8	10	76	110	204	3.80	5.02	36.80	54.38
	native	139	238	1078	1274	2729	5.18	8.83	38.98	47.02
Czech Republic	immigrant	1	4	37	68	110	0.83	3.37	34.30	61.50
	native	77	214	2209	1960	4460	1.73	4.79	49.53	43.96
Denmark	immigrant	4	12	113	276	405	0.79	3.00	28.77	67.45
	native	53	189	1935	1666	3843	1.40	4.72	50.43	43.45
England	immigrant	7	18	149	221	395	1.81	4.87	37.07	56.25
	native	129	288	1280	665	2362	5.59	11.78	54.02	28.61
Estonia	immigrant		8	79	103	190		3.96	43.29	52.75
	native	32	153	1205	1095	2485	1.10	6.10	48.08	44.72
Finland	immigrant	1	1	17	58	77	1.28	1.52	22.02	75.17
	native	93	206	1500	1340	3139	3.10	6.47	47.79	42.64
Greece	immigrant	5	20	84	211	320	2.67	6.41	26.09	64.83
	native	94	200	996	1423	2713	3.39	7.36	36.61	52.64
Ireland	immigrant	3	17	111	242	373	0.80	4.39	29.86	64.95
	native	73	167	1366	1214	2820	2.64	6.01	48.28	43.08
Italy	immigrant	4	3	43	176	226	1.84	1.55	18.68	77.92
	native	77	130	1400	1433	3040	2.58	4.34	45.94	47.14
Latvia	immigrant	14	55	58	22	149	9.58	36.31	40.51	13.60
	native	163	687	1132	569	2551	6.40	26.61	44.76	22.22
Lithuania	immigrant	4	12	66	95	177	0.98	7.27	36.26	55.48
	native	23	150	1518	1967	3658	0.66	3.55	40.60	55.19
Luxembourg	immigrant	19	86	524	1184	1813	0.95	5.99	29.95	63.12
	native	74	194	1266	1285	2819	2.72	6.95	44.78	45.55
Malta	immigrant		2	14	20	36		4.05	30.99	64.96
	native	86	168	854	895	2003	4.51	8.43	41.54	45.51
Netherlands	immigrant	8	7	65	134	214	2.73	3.36	30.13	63.78
	native	44	129	915	578	1666	2.46	6.80	55.14	35.60
Norway	immigrant	3	5	73	208	289	1.71	1.75	25.55	70.99
	native	66	106	1006	1320	2498	2.80	4.41	40.80	51.99
Poland	immigrant	1	3	19	24	47	2.97	6.58	36.86	53.59
	native	30	142	1520	1457	3149	0.83	4.69	48.76	45.72
Slovak Republic	immigrant		1	5	12	18		3.05	27.01	69.94
	native	26	78	1404	1397	2905	0.88	2.96	48.01	48.16
Slovenia	immigrant	7	13	107	193	320	2.05	4.38	36.15	57.42
	native	64	151	1024	1444	2683	2.32	5.78	38.85	53.05
Spain	immigrant	7	15	73	252	347	2.32	3.98	19.97	73.73
	native	72	137	1010	1651	2870	2.49	4.84	35.38	57.29
Sweden	immigrant	4	13	81	503	601	0.96	2.53	14.91	81.61
	native	94	156	1032	1405	2687	3.72	5.92	38.57	51.79
Switzerland	immigrant	9	35	224	443	711	1.32	4.90	31.80	61.98
	native	60	121	1004	906	2091	3.26	6.08	48.23	42.43
Grand Total		**1947**	**4979**	**32565**	**35910**	**75401**				

* Immigrant children should have the same opportunities for education that other children in the country have

Note: Due to the necessary data cleaning procedures the numbers reported here can be different from those published in the ICCS 2009 International Report (Schulz et al., 2010)

Table A -3 : IS2P26C* -Descriptive statistics

COUNTRY	Status	STRONGLY DISAGREE	DISAGREE	AGREE	STRONGLY AGREE	Sum of cases	STRONGLY DISAGREE	DISAGREE	AGREE	STRONGLY AGREE
Austria	immigrant	18	61	215	307	601	2.97	10.24	36.94	49.86
	native	231	491	1109	785	2616	8.99	19.89	41.88	29.24
Belgium (Flemish)	immigrant	6	33	159	127	325	1.58	10.25	50.63	37.54
	native	174	531	1350	508	2563	6.88	21.19	51.35	20.59
Bulgaria	immigrant	1	0	9	10	20	5.36		42.24	52.41
	native	156	566	1394	936	3052	5.26	18.58	46.45	29.70
Cyprus	immigrant	16	25	73	90	204	8.08	12.68	34.98	44.26
	native	233	512	1206	767	2718	8.43	18.87	44.37	28.34
Czech Republic	immigrant	3	11	62	32	108	2.57	10.46	57.53	29.44
	native	178	729	2429	1119	4455	4.07	16.43	54.35	25.16
Denmark	immigrant	11	30	155	207	403	2.71	6.25	37.82	53.22
	native	131	504	2018	1180	3833	3.26	13.19	52.55	31.00
England	immigrant	12	39	185	157	393	3.10	9.87	47.20	39.84
	native	158	518	1188	489	2353	6.90	21.82	50.27	21.01
Estonia	immigrant	6	35	91	58	190	2.49	17.04	52.54	27.92
	native	83	530	1290	581	2484	3.13	21.34	52.53	23.01
Finland	immigrant	5	5	30	38	78	6.93	5.54	38.34	49.19
	native	143	553	1591	848	3135	4.63	17.79	50.83	26.75
Greece	immigrant	4	57	133	129	323	1.31	19.27	41.17	38.25
	native	170	452	1232	844	2698	6.22	16.87	45.56	31.35
Ireland	immigrant	3	23	161	182	369	0.85	6.62	43.85	48.68
	native	138	467	1279	926	2810	5.06	16.77	45.30	32.88
Italy	immigrant	6	21	102	97	226	2.75	9.57	45.68	42.00
	native	168	632	1498	737	3035	5.50	21.08	49.03	24.39
Latvia	immigrant	2	20	65	63	150	1.03	17.25	40.77	40.96
	native	67	447	1320	711	2545	2.41	18.90	51.65	27.03
Lithuania	immigrant	1	33	94	49	177	0.18	20.92	48.41	30.49
	native	101	900	1845	805	3651	3.21	25.36	50.63	20.80
Luxembourg	immigrant	33	143	734	898	1808	1.87	8.61	42.10	47.42
	native	119	354	1397	944	2814	4.48	12.70	49.28	33.54
Malta	immigrant	1	6	19	10	36	2.43	13.57	53.14	30.86
	native	208	442	785	567	2002	11.16	20.85	39.06	28.92
Netherlands	immigrant	5	24	86	99	214	1.14	8.88	44.11	45.86
	native	93	295	922	354	1664	5.38	17.66	54.06	22.90
Norway	immigrant	5	28	95	158	286	2.30	10.00	32.13	55.57
	native	119	354	1112	906	2491	5.05	14.54	44.41	35.99
Poland	immigrant	1	8	25	13	47	2.97	15.45	54.42	27.16
	native	89	588	1662	807	3146	2.88	18.75	53.07	25.30
Slovak Republic	immigrant	0	3	7	7	17		19.22	36.37	44.40
	native	56	389	1610	845	2900	2.01	13.60	55.71	28.67
Slovenia	immigrant	12	35	137	135	319	3.59	11.30	44.20	40.91
	native	131	402	1233	913	2679	4.77	15.25	46.09	33.90
Spain	immigrant	13	26	124	184	347	4.21	7.93	34.34	53.51
	native	161	411	1221	1070	2863	5.73	14.12	42.83	37.32
Sweden	immigrant	5	36	131	426	598	0.58	7.02	22.50	69.90
	native	127	346	1199	1008	2680	4.91	12.64	45.07	37.38
Switzerland	immigrant	20	106	286	295	707	2.32	15.01	39.25	43.41
	native	161	467	958	496	2082	8.12	24.03	44.96	22.89
Grand Total		**3584**	**12688**	**36026**	**22917**	**75215**				

*Immigrants who live in a country for several years should have the opportunity to vote in elections

Note: Due to the necessary data cleaning procedures the numbers reported here can be different from those published in the ICCS 2009 International Report (Schulz et al., 2010)

Table A -4 : IS2P26D* - Descriptive statistics

COUNTRY	Status	STRONGLY DISAGREE	DISAGREE	AGREE	STRONGLY AGREE	Sum of cases	STRONGLY DISAGREE	DISAGREE	AGREE	STRONGLY AGREE
Austria	immigrant	17	62	172	352	603	2.37	10.85	28.16	58.62
	native	232	447	1036	896	2611	9.48	17.66	39.51	33.34
Belgium (Flemish)	immigrant	4	22	121	179	326	1.19	7.21	36.69	54.92
	native	104	286	1302	872	2564	4.10	10.97	50.27	34.66
Bulgaria	immigrant	2	2	5	10	19	9.96	10.06	21.51	58.46
	native	102	315	1231	1403	3051	3.07	10.34	41.07	45.52
Cyprus	immigrant	13	22	62	107	204	6.10	11.30	30.09	52.50
	native	200	355	1154	1011	2720	7.37	12.77	42.43	37.43
Czech Republic	immigrant	0	9	38	64	111		8.33	34.15	57.52
	native	118	346	2168	1822	4454	2.68	7.78	48.60	40.94
Denmark	immigrant	9	17	125	250	401	2.43	3.41	31.57	62.59
	native	101	381	1934	1407	3823	2.47	10.23	50.59	36.71
England	immigrant	9	29	153	204	395	2.20	7.64	39.03	51.14
	native	172	405	1138	630	2345	7.38	17.26	48.05	27.31
Estonia	immigrant	2	21	63	104	190	0.91	9.43	34.85	54.81
	native	67	341	1188	887	2483	2.50	13.79	47.84	35.87
Finland	immigrant	1	3	21	53	78	1.27	4.49	26.74	67.50
	native	95	330	1513	1198	3136	3.14	10.54	47.87	38.44
Greece	immigrant	8	35	95	182	320	2.60	11.02	30.74	55.64
	native	155	324	1105	1121	2705	5.64	11.81	40.94	41.61
Ireland	immigrant	5	16	119	231	371	1.43	4.30	32.01	62.25
	native	90	235	1290	1200	2815	3.35	8.49	45.44	42.72
Italy	immigrant	5	10	65	146	226	2.32	4.48	28.50	64.70
	native	98	221	1396	1320	3035	3.25	7.31	46.09	43.34
Latvia	immigrant	2	30	71	46	149	2.07	22.52	43.92	31.48
	native	95	639	1242	562	2538	3.65	25.86	48.58	21.90
Lithuania	immigrant	3	13	74	86	176	0.69	7.46	38.35	53.49
	native	46	483	1749	1377	3655	1.46	13.39	47.61	37.54
Luxembourg	immigrant	36	88	528	1162	1814	1.77	5.41	31.02	61.80
	native	103	277	1242	1190	2812	3.89	9.93	44.30	41.89
Malta	immigrant	1	7	10	18	36	2.50	14.69	24.37	58.44
	native	114	286	829	772	2001	6.13	14.02	40.31	39.54
Netherlands	immigrant	8	16	69	121	214	2.33	6.50	36.02	55.15
	native	85	207	881	485	1658	4.60	12.19	53.57	29.64
Norway	immigrant	7	11	70	194	282	3.22	4.39	24.05	68.34
	native	100	243	1003	1134	2480	4.15	10.25	40.07	45.53
Poland	immigrant	2	7	21	17	47	4.78	12.98	44.51	37.73
	native	58	364	1668	1054	3144	1.74	11.56	53.71	32.98
Slovak Republic	immigrant	1	0	7	10	18	3.05		41.00	55.94
	native	39	185	1407	1271	2902	1.49	6.25	48.13	44.13
Slovenia	immigrant	8	20	108	182	318	2.59	6.34	35.67	55.40
	native	85	260	1144	1193	2682	3.15	10.02	42.96	43.88
Spain	immigrant	8	15	98	227	348	2.45	3.80	27.80	65.95
	native	106	245	1133	1380	2864	3.92	8.74	39.61	47.74
Sweden	immigrant	5	16	108	471	600	1.07	3.24	19.75	75.94
	native	115	175	1057	1342	2689	4.31	6.59	39.63	49.47
Switzerland	immigrant	11	52	241	405	709	1.83	7.98	31.58	58.61
	native	103	299	976	696	2074	6.03	14.97	46.09	32.90
Grand Total		**2750**	**8172**	**33230**	**31044**	**75196**				

*** Immigrants should have the opportunity to continue their own customs and lifestyle**

Note: Due to the necessary data cleaning procedures the numbers reported here can be different from those published in the ICCS 2009 International Report (Schulz et al., 2010)

Table A -5 : IS2P26E* - Descriptive statistics

COUNTRY	Status	STRONGLY DISAGREE	DISAGREE	AGREE	STRONGLY AGREE	Sum of cases	STRONGLY DISAGREE	DISAGREE	AGREE	STRONGLY AGREE
Austria	immigrant	17	62	172	352	603	2.37	10.85	28.16	58.62
	native	232	447	1036	896	2611	9.48	17.66	39.51	33.34
Belgium (Flemish)	immigrant	4	22	121	179	326	1.19	7.21	36.69	54.92
	native	104	286	1302	872	2564	4.10	10.97	50.27	34.66
Bulgaria	immigrant	2	2	5	10	19	9.96	10.06	21.51	58.46
	native	102	315	1231	1403	3051	3.07	10.34	41.07	45.52
Cyprus	immigrant	13	22	62	107	204	6.10	11.30	30.09	52.50
	native	200	355	1154	1011	2720	7.37	12.77	42.43	37.43
Czech Republic	immigrant	0	9	38	64	111		8.33	34.15	57.52
	native	118	346	2168	1822	4454	2.68	7.78	48.60	40.94
Denmark	immigrant	9	17	125	250	401	2.43	3.41	31.57	62.59
	native	101	381	1934	1407	3823	2.47	10.23	50.59	36.71
England	immigrant	9	29	153	204	395	2.20	7.64	39.03	51.14
	native	172	405	1138	630	2345	7.38	17.26	48.05	27.31
Estonia	immigrant	2	21	63	104	190	0.91	9.43	34.85	54.81
	native	67	341	1188	887	2483	2.50	13.79	47.84	35.87
Finland	immigrant	1	3	21	53	78	1.27	4.49	26.74	67.50
	native	95	330	1513	1198	3136	3.14	10.54	47.87	38.44
Greece	immigrant	8	35	95	182	320	2.60	11.02	30.74	55.64
	native	155	324	1105	1121	2705	5.64	11.81	40.94	41.61
Ireland	immigrant	5	16	119	231	371	1.43	4.30	32.01	62.25
	native	90	235	1290	1200	2815	3.35	8.49	45.44	42.72
Italy	immigrant	5	10	65	146	226	2.32	4.48	28.50	64.70
	native	98	221	1396	1320	3035	3.25	7.31	46.09	43.34
Latvia	immigrant	2	30	71	46	149	2.07	22.52	43.92	31.48
	native	95	639	1242	562	2538	3.65	25.86	48.58	21.90
Lithuania	immigrant	3	13	74	86	176	0.69	7.46	38.35	53.49
	native	46	483	1749	1377	3655	1.46	13.39	47.61	37.54
Luxembourg	immigrant	36	88	528	1162	1814	1.77	5.41	31.02	61.80
	native	103	277	1242	1190	2812	3.89	9.93	44.30	41.89
Malta	immigrant	1	7	10	18	36	2.50	14.69	24.37	58.44
	native	114	286	829	772	2001	6.13	14.02	40.31	39.54
Netherlands	immigrant	8	16	69	121	214	2.33	6.50	36.02	55.15
	native	85	207	881	485	1658	4.60	12.19	53.57	29.64
Norway	immigrant	7	11	70	194	282	3.22	4.39	24.05	68.34
	native	100	243	1003	1134	2480	4.15	10.25	40.07	45.53
Poland	immigrant	2	7	21	17	47	4.78	12.98	44.51	37.73
	native	58	364	1668	1054	3144	1.74	11.56	53.71	32.98
Slovak Republic	immigrant	1	0	7	10	18	3.05		41.00	55.94
	native	39	185	1407	1271	2902	1.49	6.25	48.13	44.13
Slovenia	immigrant	8	20	108	182	318	2.59	6.34	35.67	55.40
	native	85	260	1144	1193	2682	3.15	10.02	42.96	43.88
Spain	immigrant	8	15	98	227	348	2.45	3.80	27.80	65.95
	native	106	245	1133	1380	2864	3.92	8.74	39.61	47.74
Sweden	immigrant	5	16	108	471	600	1.07	3.24	19.75	75.94
	native	115	175	1057	1342	2689	4.31	6.59	39.63	49.47
Switzerland	immigrant	11	52	241	405	709	1.83	7.98	31.58	58.61
	native	103	299	976	696	2074	6.03	14.97	46.09	32.90
Grand Total		**2750**	**8172**	**33230**	**31044**	**75196**				

*** Immigrants should have all the same rights that everyone else in the country has**

Note: Due to the necessary data cleaning procedures the numbers reported here can be different from those published in the ICCS 2009 International Report (Schulz et al., 2010)

Table A -6 : IS2P26F* - Descriptive statistics

COUNTRY	Status	STRONGLY DISAGREE	DISAGREE	AGREE	STRONGLY AGREE	Sum of cases	STRONGLY DISAGREE	DISAGREE	AGREE	STRONGLY AGREE
Austria	immigrant	117	169	217	95	598	19.43	27.98	37.07	15.53
	native	254	511	1053	789	2607	9.95	18.59	40.61	30.85
Belgium (Flemish)	immigrant	75	91	120	35	321	23.72	26.46	38.46	11.36
	native	261	779	1122	393	2555	10.12	30.09	43.97	15.81
Bulgaria	immigrant	1	8	8	4	21	4.98	36.24	34.34	24.43
	native	351	880	1247	576	3054	11.30	28.64	41.50	18.56
Cyprus	immigrant	58	53	44	47	202	27.41	27.46	21.43	23.70
	native	452	959	813	505	2729	16.40	35.13	29.77	18.69
Czech Republic	immigrant	10	33	50	16	109	8.88	31.02	46.05	14.06
	native	260	969	2328	879	4436	5.89	21.91	52.41	19.79
Denmark	immigrant	120	128	116	33	397	28.56	32.55	30.74	8.15
	native	497	1340	1558	394	3789	12.96	35.79	40.99	10.26
England	immigrant	63	142	128	51	384	16.50	36.37	33.15	13.98
	native	227	604	996	513	2340	9.81	25.65	42.68	21.86
Estonia	immigrant	33	64	60	32	189	17.03	34.18	32.36	16.43
	native	218	831	1038	387	2474	9.06	34.34	41.44	15.16
Finland	immigrant	21	17	30	9	77	27.89	21.33	39.10	11.68
	native	164	770	1538	654	3126	5.25	24.43	49.21	21.11
Greece	immigrant	85	103	88	45	321	24.87	33.63	28.30	13.20
	native	400	857	952	495	2704	14.62	31.55	35.56	18.28
Ireland	immigrant	71	117	121	56	365	19.76	31.33	33.24	15.67
	native	266	620	1172	745	2803	9.57	22.39	41.64	26.39
Italy	immigrant	34	63	95	34	226	15.69	27.42	42.03	14.86
	native	215	630	1335	851	3031	7.14	20.70	44.25	27.92
Latvia	immigrant	3	11	65	70	149	1.96	9.99	42.82	45.23
	native	42	244	1388	875	2549	1.61	10.30	54.74	33.35
Lithuania	immigrant	16	48	79	34	177	10.11	27.61	47.73	14.55
	native	283	1155	1696	520	3654	8.28	31.43	46.50	13.80
Luxembourg	immigrant	447	497	558	298	1800	23.26	25.87	31.67	19.21
	native	375	847	1020	557	2799	13.30	29.61	37.04	20.05
Malta	immigrant	6	7	16	7	36	15.83	22.24	42.26	19.67
	native	213	454	877	447	1991	11.19	21.57	43.84	23.41
Netherlands	immigrant	36	64	78	35	213	16.36	31.85	34.49	17.30
	native	164	471	773	249	1657	9.31	27.78	46.20	16.71
Norway	immigrant	46	79	91	63	279	17.00	27.83	33.26	21.91
	native	226	603	1096	515	2440	9.31	24.36	44.70	21.63
Poland	immigrant	8	13	17	9	47	18.78	26.80	35.79	18.63
	native	290	961	1427	464	3142	9.01	30.87	45.33	14.78
Slovak Republic	immigrant	5	5	4	4	18	25.26	26.92	23.45	24.37
	native	318	1051	1157	376	2902	11.05	36.34	40.10	12.52
Slovenia	immigrant	62	96	106	53	317	18.62	30.73	34.42	16.23
	native	341	793	1027	516	2677	12.63	29.99	37.98	19.41
Spain	immigrant	78	107	108	48	341	22.55	31.97	31.43	14.05
	native	440	798	1001	616	2855	15.54	27.48	35.48	21.50
Sweden	immigrant	196	157	151	89	593	32.54	25.89	26.25	15.32
	native	307	752	1044	558	2661	11.49	27.96	39.00	21.55
Switzerland	immigrant	117	201	280	106	704	17.36	29.92	38.95	13.77
	native	158	434	928	555	2075	8.53	20.09	43.91	27.47
Grand Total		**8430**	**20586**	**31216**	**14702**	**74934**				

*When there are not many jobs available, immigration should be restricted

Note: Due to the necessary data cleaning procedures the numbers reported here can be different from those published in the ICCS 2009 International Report (Schulz et al., 2010)

References

Alwin, D. F., & Jackson, D. J. (1981). Applications of simultaneous factor analysis to issues of factorial invariance. In D. J. Jackson&E. F. Borgatta (Eds.), Factor analysis and measurement in sociological research (pp. 249-279). Beverly Hills, CA: Sage.

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

Asparouhov, T., & Muthén, B. (2006). Multilevel modelling of complex survey data. Proceedings of the Survey Research Methods Section, American Statistical Association 2006, 2718–2726.

Bagozzi, R. P., & Edwards, J. R. (1998). A general approach for representing constructs in organizational research. Organizational Research Methods, 1, 45-87.

Beatty, A. & Pritchett, L. (2012) From Schooling Goal to Learning Goals: how fast can student learning improve? CGD Policy Paper 012. Washington, DC: Center for Global Development

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

Bentler, P. M. (1992) On the fit of models to covariances and methodology to the Bulletin. Psychological Bulletin, Vol 112(3), Nov 1992, 400-404.

Bentler, P. M. (2005). EQS 6 structural equations program manual. Encino, CA: Multivariate Software.

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

Bollen, K. A. (1989). Structural equations with latent variables. New York: John Wiley.

Boomsma, A, & Hoogland, J.J. (2011). The robustness of LISREL modeling revisited. In R. Cudeck, S. DuToit, & D Sorbom (Eds.), Structural equations modeling: Present and future. Lincolnwood, IL; Scientific Software International.

Brese, F, Jung, M., Mirazchiyski, P., Schulz, W., & Zuehlke,O. (2014). ICCS 2009 User Guide for the International Database Supplement 1. Amsterdam, The Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

Broadfoot, P., & Black, P. (2004). Redefining assessment? The first ten years of Assessment in Education. Assessment in Education 11(1): 7-26.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. Sociological Methods and Research, 21, 230-258

Browne, M. W., MacCallum, R.C., Kim, C-T, Andersen, B.L., & Glaser, R.. (2002). When fit indices and residuals are incompatible. Psychological Methods, 7, 403-421.

Byrne, B. M. (2008). Testing for multigroup equivalence of a measuring instrument: a walk through the process. Psicothema, 20(4), 872–882.

Byrne, B. M., & Shavelson, R. J. (1987). Adolescent self-concept: Testing the assumption of equivalent structure across gender. American Educational Research Journal, 24, 365-385.

Byrne, B. M., & Stewart, S. M. (2006). The MACS approach to testing for multigroup invariance of a second-order factor structure: A walk through the process. Structural Equation Modeling, 13, 287−321.

Byrne, B. M., & van de Vijver, F. J. R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. International Journal of Testing, 10, 107–132.

Byrne, B. M., & Watkins, D. (2003). The issue of measurement invariance revisited. Journal of Cross-Cultural Psychology, 34, 155−175.

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.

Byrne, B.M. (2003). Testing for Equivalent Self-Concept Measurement across Culture. In: H.W. Marsh, R.G. Craven, & D.M. McInerney (Eds.). International advances in self-research: speaking to the future, 291-314. Greenwich: Information Age Publishing.

Chan, D. (1998). Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. Journal of Applied Psychology, 83, 234−246

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. Journal of Personality and Social Psychology.

Chen, F. F., Sousa, K. H., & West, S. G. (2005). Testing measurement invariance of second-order factor models. Structural Equation Modeling, 12, 471−492.

Chen, F.F., (2007). Sensitivity of goodness of fit indices to lack of measurement invariance. Structural Equation Modeling, 14, 464-504.

Davidov, E. (2008). A cross-country and cross-time comparison of the human values measurements with the second round of the European social survey. Survey Research Methods, 2(1), 33–46.

Davidov, E., Dulmer, H., Schluter, E., Schmidt, P., & Meuleman, B. (2012). Using a multilevel structural equation modeling approach to explain cross-cultural measurement noninvariance. Journal of Cross-Cultural Psychology, 43(4), 558–575

Davidov, E., Meuleman, B., Billiet, J., & Schmidt, P. (2008). Values and support for immigration: A cross-country comparison. European Sociological Review, 24(5), 583-599.

Davidov, E., Meuleman, B, Cieciuch, J., Schmidt, P., &. Billiet, J. (2014). Measurement Equivalene in Cross-National Research. Annual Review of Sociology, 40, 55-75.

Davidov, E., Schmidt, P., & Schwartz, S. H. (2008). Bringing values back in the adequacy of the European Social Survey to measure values in 20 countries. Public opinion quarterly, 72(3), 420-445.

de Jong, M. G., Steenkamp, J.-B. E. M., & Fox, J.-P. (2007). Relaxing measurement invariance in cross national consumer research using a hierarchical IRT model. Journal of Consumer Research, 34(August), 260–278

Dejaeghere, Y., & Quintelier, E. (2008). Does European citizenship increase tolerance in young people? European Union Politics, 9(3), 329–336.

Foshay, A.W., Thorndike, R.L., Hotyat, F., Pigeon, D.A. & Walker, D.A. (1962). Educational Achievements of Thirteen-Year-Olds in Twelve Countries. Hamburg: Unesco Institute for Education

French, B. F., & Finch, W. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. Structural Equation Modeling, 13, 378-402

Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory analysis framework. Medical Care, 44, S78−S94.

Hanushek, E. & Woessmann, L. (2009) Do Better Schools Lead to More Growth? Cognitive Skills, Economic Outcomes and Causation. NBER Working Paper 14633. Cambridge, MA: National Bureau of Economics Research.

Harkness, J., Pennell, B., & Schoua-Glusberg, A. (2004). Survey questionnaire translationand assessment. In J. Presser, M. Rothgeb, J. Couper, E. Lessler, E. Martin, & E. Singer(Eds.), Questionnaire development evaluation and testing methods (pp. 453–473) Hoboken, NJ: Wiley.

Heyneman, S. P., & Lee, B. (2012). The Impact of International Studies of Academic Achievement on Policy and Research. In Handbook of International Large Scale Assessment: Background, Technical Issues, and Methods of Data Analysis (pp. 37-74). Chapman and Hall Publishers London.

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

Hoskins, B., Jesinghaus, J., Mascherini, M., Munda, G., Nardo, M., Saisana, M., Van Nijlen, D., & Villalba, E. (2006). Measuring active citizenship in Europe (CRELL Research Paper No. 4). Ispra, Italy: Joint Research Centre/CRELL.

Hoskins, B., Villalba, E., Van Nijlen, D., & Barber, C. (2008). Measuring civic competence in Europe: A composite indicator based on the IEA Civic Education Study 1999 for 14 year olds in school (CRELL Research Paper No. EUR 23210). Ispra, Italy: European Commission.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling, 6(1), 1–55.

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

Isac, M. M., Maslowski, R., Creemers, B., & van der Werf, G. (2014). The contribution of schooling to secondary-school students' citizenship outcomes across countries. School Effectiveness and School Improvement, 25(1), 29-63.

Isac, M. M., Maslowski, R., & van der Werf, G. (2012). Native Students Attitudes towards Equal Rights for Immigrants. A Study in 18 European Countries. JSSE-Journal of Social Science Education, 11(1).

Isac, M. M., Maslowski, R., & van der Werf, G. (2011). Effective civic education: an educational effectiveness model for explaining students' civic knowledge. School Effectiveness and School Improvement, 22(3), 313-333.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. Psychometrika, 34, 183-202.

Jöreskog, K., & Sörbom, D. (1996). LISREL 8 user's guide. Chicago: Scientific Software

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

Kamens, D. (2013) Globalization and the Emergence of an Audit Culture: PISA and the search for 'best practices' and magic bullets, in H.D. Meyer & A. Benavot (Eds) PISA, Power, and Policy: the emergence of global educational governance. Oxford: Symposium Books.

Kamens, D.H. & McNeely, C. (2010) Globalization and the Growth of International Educational Testing and National Assessment, Comparative Education Review, 54(1), 5-27

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

Kankaraš, M., & Moors, G. (2010). Researching measurement equivalence in cross-cultural studies. Psihologija 43.2 : 121-136.

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

Kankaraš, M., Moors, G.,& Vermunt, J.K., (2010). Testing for measurement invariance with latent class analysis. Cross-cultural analysis: Methods and applications, 359-384.

Kaplan, D. (2000). Structural equation modeling: Foundations and extensions. Thousand Oaks, CA: Sage

Kaplan, D., & George, R (1995) A study of the power associated with testing factor mean differences under violations of factorial invariance Structural Equation Modeling: A Multidisciplinary Journal 2 (2), 101-118

Kerr, D., & Losito, B. (forthcoming). Policy tool for education for democratic citizenship and human rights (EDC/HRE): Key strategic support for decision makers. Strasbourg, France: Council of Europe

Kerr, M., Stattin, H., & Burk, W.J. (2010). A reinterpretation of parental monitoring in longitudinal perspective. Journal of Research on Adolescence 20.1, 39-64.

Kline, R. B. (2011). Principles and practice of structural equation modeling (3rd ed., pp. 1–427). New York: The Guilfor.

Kirsch, I., Lennon, M., von Davier, M., Gonzalez, E., & Yamamoto, K. (2013). On the growing importance of international large-scale assessments. In The Role of International Large-Scale Assessments: Perspectives from Technology, Economy, and Educational Research (pp. 1-11). Springer Netherlands.

Lehmann, R. (2011) The impact of IEA on educational policy making in Germany. In: IEA 1958-2008:50 Years of Experiences and Memories, eds. C. Papanastasiou T. Plomp, and E. Papanastasiou, 411-430. Nicolsia: Cultural Center of the Kykkos Monestery.

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. Multivariate Behavioral Research, 32, 53−76

Little, T. D., Card, N.A., Slegers, D.W., & Ledford, E.C. (2007). Representing contextual effects in multiple-group MACS models. In T.D. Little, J.A. Bovaird, & N.A. Card (Eds.), Modeling contextual effects in longitudinal studies 121-147. Mahwah, NJ: Erlbaum.

Lockheed, M. E.,& Wagemaker, H., (2013) International Large-Scale Assessments:thermometers, whips or useful policy tools? Research in Comparative and International Education Volume 8 Number 3 2013, 296-306

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

MacCallum, R.C., Browne, M.W., & Sugawara, H., M. (1996), "Power Analysis and Determination of Sample Size for Covariance Structure Modeling," Psychological Methods, 1 (2), 130-49.

Marsh, H. W., & Grayson, D. (1994). Longitudinal stability of means and individual differences: A unified approach. Structural Equation Modeling: Concepts, issues, and applications, 177-198. Thousand Oaks, CA: Sage.

Marsh, H. W., Muthén, B.O., Asparouhov, T., Ludtke, O, Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA. Application to students' evaluations of university teaching. Structural Equation Modeling, 16, 439-476.

Marsh, H. W., Nagengast, B., Morin, A. J. S., Nahengast, B. M., & Morin A. J. S. (2012). Measurement invariance of big-five factors over the life span: ESEM tests of gender, age, plasticity, maturity, and La Dolce Vita effects. Development Psychology

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

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. Medical Care, 44(11 Suppl 3), 69–77.

Meuleman, B. (2012). When are intercept differences substantively relevant in measurement invariance testing? In Methods, Theories, and Empirical Applications in the Social Sciences: Festschrift for Peter Schmidt, ed.S. Salzborn, E. Davidov, J. Reinecke, 97-104. Heidelberg, Ger.: Springer VS.

Milfont, T.L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. International Journal of psychological research 3.1, 111-130.

Millsap, R. E.,& Hartog, S. B. (1988). Alpha, beta, and gamma change in evaluation research: A structural equation approach. Journal of Applied Psyc hology, 73, 574-584.

Millsap, R.E. (2011). Statistical Approaches to Measurement Invariance. New York: Routledge.

Mohler, P. P., Smith, T. W., & Harkness, J. A. (1998). Respondent's ratings of expressions from response scales: A two-country, two-language investigation on equivalence and translation. In J. A. Harkness (Ed.), ZUMA-Nachrichten spezial No.3: Cross-cultural survey equivalence (pp. 159–184). Mannheim: ZUMA.

Muthén, B., & Asparouhov, T. (2013). BSEM measurement invariance analysis. Mplus Web Notes, 1–48.

Muthén, L. K., & Muthén, B. O. (2012). Mplus User's Guide (7th ed., pp. 1–856). Los Angeles: Muthén&Muthén.

Noah, H.J. & Eckstein, M.A. (1969) Toward a Science of Comparative Education. New York: Macmillan.

Penninx, R. (2005). Integration of migrants: Economic, social, cultural and political dimensions. In M. Macura, A. L. MacDonald, & W. Haug (Eds.), The new demographic regime: Population challenges and policy responses (pp. 137–152). New York/Geneva, Switzerland: United Nations.

Penninx, R., Berger, M., & Kraal, K. (Eds.). (2006). The dynamics of intergenerational migration and settlement in Europe: A state of the art. Amsterdam, The Netherlands: AUP (IMISCOE Joint StudiesSeries)

Phillips, D. & Ochs, K. (2003) Processes of Policy Borrowing in Education: some explanatory and analytical devices, Comparative Education, 49(4), 451-461

Ployhart, R. E., & Oswald, F. L. (2004). Applications of mean and covariance structure analysis: Integrating correlational and experimental approaches. Organizational Research Methods, 7, 27−65.

Postlethwaite, T. N. (1999). International studies of educational achievement: Methodological issues. Vol. 6. Comparative Education Research Centre, University of Hong Kong.

Postlethwaite, T.N. & Ross, K.N. (1992) Effective Schools in Reading: implications for educational planners. The Hague: International Association for the Evaluation of Educational Achievement.

Postlethwaite, T.N. & Wiley, D.E. (1992) The IEA Study of Science II: science achievement in twenty-three countries. New York: Pergamon Press

Robitaille, D.F. & Garden, R.A. (1989) The IEA Study of Mathematics II: contexts and outcomes of school mathematics. New York: Pergamon Press.

Rutkowski, L., von Davier, M., & Rutkowski, D. (Eds.). (2013). Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis. CRC Press.

Rutkowski, L., Gonzalez, E., Joncas, M. & von Davier, M. (2010 A brief Introduction to Modern International Large-Scale Assessment Educational Researcher, 39(2), 142-151.

Sadler, M. (1900) How Far Can We Learn Anything of Practical Value from the Study of Foreign Systems of Education? Reprinted in 1964, Comparative Education Review, 7, 307-31

Schaubroeck, J., & Green, S.G. (1989). Confirmatory factor analytic procedures for assessing change during organizational entry. Journal of Applied Psychology, 74(6), 892-900.

Schmitt, N. (1982). The use of analysis of covariance structures to assess beta and gamma change. Multivariate Behavioral Research, 17, 343-358.

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. Human Resource Management Review, 18(4), 210-222.

Schulz, W. (2003). Validating questionnaire constructs in international studies: Two examples from PISA 2000. Paper presented at the annual meeting of the American Educational Research Association (AERA), Chicago, April 21–25, 2003

Schulz, W. (2009). Questionnaire construct validation in the International Civic and Citizenship Education Study. IERI Monograph Series Volume 2, 113–135.

Schulz, W., Ainley, J., & Fraillon, J. (Eds.). (2010). ICCS 2009 Technical Report. Amsterdam, The Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

Schulz, W., Ainley, J., & Fraillon, J. (Eds.). (2010). ICCS 2009 Technical Report. Amsterdam, The Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

Schulz, W., Ainley, J., Fraillon, J., Kerr, D. & Losito, B. (2011) Civic Knowledge, Attitudes, and Engagement Among Lower-secondary School Students in 38 Countries. Amsterdam: IEA.

Schulz, W., Ainley, J., Fraillon, J., Kerr, D., & Losito, B. (2010). ICCS 2009 international report: Civic knowledge, attitudes and engagement among lower secondary school students in thirty-eight countries. Amsterdam, The Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

Schulz, W., Ainley, J., Fraillon, J., Kerr, D., & Losito, B. (2010b). ICCS 2009 international report:Civic knowledge, attitudes and engagement among lower secondary school students in thirty-eight countries .Amsterdam, The Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

Schulz, W., Fraillon, J., Ainley, J., Losito, B., & Kerr, D. (2008). International Civic and Citizenship Education Study: Assessment framework. Amsterdam, The Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

Selig, J.P., Card, N.A., & Little, T.D. (2008). Latent variable structural equation modelling in cross-cultural research: Multigroup and multilevel approaches. In F.J.R. vand de Vijver, D.A., van Hemert, & Y.H. Poortinga (Eds.) Multilevel analysis of individuals and cultures, 93-119. Mahwah, NJ: Erlbaum.

SPSS (Version 16.0) [Computer software and manual]. (2007). Chicago: SPSS, Inc.

Stanat, P., & Lüdtke,O. (2013). International large-scale assessment studies of student achievement. International guide to student achievement, 481-483.

Steenkamp, J. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. Journal of Consumer Research, 25, 78−90.

Steiger, J.H., & Lind, J.C. (1980). Statistically based tests for the number of common factors. Paper presented at the Psychometric Society annual meeting, Iowa City, IA.

Torney-Purta, J. & Amadeo, J., (2013) International Large-Scale Assessments: challenges in reporting and potentials for secondary analysis Research in Comparative and International Education Volume 8 Number 3 2013, 248-258

Torney-Purta, J., Lehmann, R., Oswald, H., & Schulz, W. (2001). Citizenship and education in twenty-eight countries. Amsterdam, The Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

Torney-Purta, J., Wilkenfeld, B., & Barber, C. (2008). How adolescents in twenty-seven countries understand, support, and practice human rights. Journal of Social Issues, 64, 857–880.

Tucker, L. R., & Lewis, C. (1973). The reliability coefficient for maximum likelihood factor analysis. Psychometrika, 38, 1-10.

Van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthén, B. (2013). Facing off with Scylla and Charybdis: a comparison of scalar, partial, and the novel possibility of approximate measurement invariance. Frontiers in Psychology, 4, 770.

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

Van de Vijver, F. J. R., & Tanzer, N. K. (1997). Bias and equivalence in cross-culturalassessment: An overview. European Review of Applied Psychology, 47, 263–279.

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

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. Organizational Research Methods, 5, 139−158.

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, 4−69.

Vandenberg, R. J., & Self, R. M. (1993). Assessing newcomers' changing commitment to the organization during the first 6 months of work. Journal of Applied Psychology, 78, 557−568.

Wagemaker, H. (2013) International Large-Scale Assessments (ILSAs) and the challenge of consequential validity, in M. Chatterji (Ed.) Validity and Test Use: an international dialogue on educational assessments, accountability, and equity. Bingley: Emerald.

Wagemaker, H. (2013). International Large-Scale Assessments: From Research to Policy. Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis, 11.

Weijters, B., Schillewaert, N., & Geuens, M. (2008). Assessing response styles across models of data collection. Journal of Academy of Marketing Science, 36(3), 409-422.

Welkenhuysen-Gybels, J. (2003). The detection of differential item functioning in Likert score items. PhD Thesis, Kathol. Univ. Leuven, Belgium.

Welkenhuysen-Gybels, J., Billiet, J., & Cambre, B. (2003). Adjustment for acquiescence in the assessment of the construct equivalence of Likert-type score items. Journal of Cross-Cultural Psychology, 34, 702−722.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), The science of prevention (pp. 281−324). Washington, DC: American Psychological Association.

Weziak-Bialowolska, D., & Isac, M.M. (2013). Cross-National Equivalence of Students' Perceptions of Good Citizenship in ICCS. Conference Proceedings 5th IEA International Research Conference - Singapore

Weziak-Bialowolska, D. (2014). Differences in Gender Norms Between Countries: Are They Valid? The Issue of Measurement Invariance. European Journal of Population 31.1, 51-76.

Williams, J. H, & Engel, L. (2013) Testing to Rank, Testing to Learn, Testing to Improve: an introduction and overview Research in Comparative and International Education Volume 8 Number 3 2013, 214-235

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. Practical Assessment Research & Evaluation, 12(3), 1–22