Large-scale Assessments in Education

RESEARCH

Open Access

Indicators of (in)tolerance toward immigrants among European youth: an assessment of measurement invariance in ICCS 2016

Maria Magdalena Isac[1*], Laura Palmerio[2] and M. P. C. (Greetje) van der Werf[1]

*Correspondence: m.m.isac@rug.nl
[1] University of Groningen, Grote Rozenstraat 3, 9712 TG Groningen, The Netherlands
Full list of author information is available at the end of the article

Abstract

Background: Promoting tolerance is an important goal of European education policies focused on education for democratic citizenship and human rights. In this article, we argue that cross-cultural comparability must be empirically assessed and ensured for the measurement of highly relevant indicators that serve to monitor inter-European and international differences in young people's tolerant attitudes toward immigrants.

Methods: Using the framework and data provided by the recent International Civic and Citizenship Education Study (ICCS 2016), we examine the extent to which average comparisons of cross-national differences in young people's tolerant attitudes toward immigrants are empirically justified. Multiple-group confirmatory factor analysis (MGCFA) is applied to estimate the measurement model of the concept and test its measurement invariance across fourteen European countries.

Results: In line with prior research, our findings show that cross-cultural comparability can be achieved with some modifications. Results of measurement invariance analysis point to the achievement of full scalar invariance with the implication that average scores can be validly compared across the European educational systems under investigation. These findings are largely corroborated by robustness analyses.

Conclusions: We conclude by providing information on further scale refinement and improvement. Limitations and implications for further research are outlined and discussed.

Keywords: Tolerance, Attitudes toward immigrants, Measurement invariance, International Civic and Citizenship Education Study (ICCS), International large-scale assessments

Introduction

Tolerance, like freedom and equality, is a fundamental feature of a mature citizenship in democratic societies (Almond and Verba 1963; Sherrod and Lauckhardt 2009). Tolerance of diversity is expected to promote democratic interaction and equitable participation in multicultural societies while intolerant attitudes may lead to racism and violence and pose threats to the stability of democratic institutions (Berry 2011; Berry and Sam 2014; Van Zalk et al. 2013). In a European context challenged by unpreceded

Springer Open

migration, monitoring and promoting tolerance of immigrants is an essential part of the policies focused on education for democratic citizenship and human rights (Council of Europe 2017; European Commission/EACEA/Eurydice 2017; European Council 2015). Therefore, comparative studies gauging the extent of cross-national differences in young people's tolerant attitudes toward immigrants are highly needed. Cross-national studies attempting to chart such attitudes among adult populations are relatively frequent (Ceobanu and Escandell 2010). European indicators of attitudes toward immigrants among youth are, in contrast, much rarer (Elchardus et al. 2013). Yet, a fair amount of research indicates that tolerance emerges in and is most malleable from early adolescence onwards, which points to the strategic importance of monitoring such attitudes among youth (Allport 1954; Côté and Erickson 2009; Elchardus et al. 2013; Van Zalk et al. 2013).

The Civic Education Study (CIVED 1999) and the International Civic and Citizenship Education Studies (ICCS, 2009 and 2016) (Schulz et al. 2010, 2018; Torney-Purta et al. 2001) conducted by the International Association for the Evaluation of Educational Achievement (IEA) are among the few exceptions in this respect. These studies investigate the ways in which young people (Grade 8 students, approximately 14 years of age) are prepared to undertake their roles as citizens in a range of countries, and they also inquire into young people's beliefs about equal rights and opportunities for different groups in society based on gender, ethnic/racial status and immigration background (Schulz 2016).

Over the years, indicators based on these civic and citizenship education studies served to monitor inter-European and international differences in young people's tolerant attitudes toward immigrants. In the current European context, such indicators often provide valuable tools to different stakeholders in their efforts to monitor, to contextualize and to explain differences and polarization in such attitudes in several countries. Nevertheless, a large body of research warns about the risks of directly comparing scores on constructs of interest across educational systems especially in the context of international large-scale assessments (ILSAs) such as ICCS (He and Van de Vijver 2013; Rutkowski and Rutkowski 2017). Meaningful comparisons of means across countries require that the construct is understood and operationalized in a similar way in each context. Yet, measurement instruments can be sensitive to cultural, linguistic, and geographic differences. For this reason, secondary users of data collected in such studies are urged to test the assumption of cross-cultural comparability or measurement invariance (most commonly investigated by means of multiple-group confirmatory factor analysis, MGCFA) before proceeding to cross-national comparisons. This issue is even more relevant when attitudinal measures (such as tolerant attitudes toward immigrants) collected by background questionnaires are the object of investigation. To a large extent, attitudinal measures can be culturally and context specific, and measurement invariance is often not achieved when a large number of countries are considered (He and Van de Vijver 2013). Moreover, such constructs are often measured using Likert scales, which require analytic considerations of the ordinal character of the data. In addition and in contrast with current operational scaling procedures in ICCS, recent research points to the need of reconsidering the guidelines for the evaluation of measurement invariance testing in the context of ILSAs involving many groups and often categorical indicators

(Rutkowski and Svetina 2017). Consequently, more should be done to better account for these developments when investigating the comparability of scales derived from ILSA data.

Against this background, the present study examined the extent to which average comparisons of cross-national differences in young people's tolerant attitudes toward immigrants in the context of the ICCS 2016 study are justified. We applied multiple-group confirmatory factor analysis (MGCFA) (Jöreskog 1971; Steenkamp and Baumgartner 1998) to assess whether comparisons of average scale scores across fourteen European countries participating in ICCS 2016 can be made with confidence. In response to recent developments in the field, the current study considered the ordinal character of the data and followed the most recent guidelines for model fit evaluation (Rutkowski and Svetina 2017). Moreover, we aimed to further add to current research by providing information regarding the robustness of our findings. To this end, we tested whether the comparisons were defensible also within four sub-groups of country clusters that show similarities in terms of linguistic, and geographic and cultural characteristics. In addition, we illustrated cross-national differences in young people's tolerant attitudes both in terms of overall levels and degree of polarization.

In the following section we provide a brief conceptual overview of tolerance toward immigrants and its measurement in the context of the IEA citizenship education studies. Next, we briefly describe the application of measurement invariance in the context of ILSAs and review previous research into the measurement invariance of the attitudes toward immigrants construct applied to data from the IEA civic and citizenship education studies. We then describe our samples and instruments, illustrate our analytic strategy and report our findings. In the concluding section we address implications and limitations of the research.

Theoretical background and previous research

Tolerance toward immigrants. Conceptualization and measurement in the current study

In broad terms, tolerance is described as respect, acceptance and appreciation of diversity (UNESCO 1995), while tolerance toward immigrants is generally defined as positive feelings toward immigrants as well as an understanding and endorsement of equality between immigrants and non-immigrants (Côté and Erickson 2009; Van Zalk et al. 2013). Overall, tolerance is a controversial and complex concept (Forst 2003; Green et al. 2006; Mutz 2001; Van Driel et al. 2016). Although studying tolerance is a multidisciplinary endeavor, insightful leads of particular relevance for the current work are provided by political socialization research. More specifically, scholars of the field (Gibson 2006, 2013; Weldon 2006) make the important distinction between political and social tolerance. Political tolerance concerns the granting of democratic and political rights to different groups in society while social tolerance refers more to the evaluation of the direct contact with people from out-groups (e.g. inter-ethnic friendships). The two forms are rather distinct in the sense that political tolerance involves a higher level of abstract understanding. While people may be socially intolerant (e.g. not willing to create family ties with immigrants) or even xenophobic (e.g. irrationally fearing immigrants), they may still be able to understand and extend political and civil rights to immigrants such as the right to education or the right to participate in the political life.

One of the most common approaches to the measurement of tolerance is the fixed-group approach (for an overview see Gibson 2013). In this approach, the measures intend to capture the degree to which respondents will support the extension of political and civil rights to different groups in society. The groups to be tolerated are predefined by the researcher (e.g. immigrants) and construct indicators are developed to capture whether certain rights and liberties should be tolerated with respect to the reference group. A similar strategy is implemented in the context of the IEA citizenship education studies. Tolerant attitudes toward immigrants are conceptualized in a larger framework of respecting civic principles such as equity, freedom and the rule of law. The construct reflects young people's beliefs about equal political and cultural rights and opportunities for (three) different groups in society based on immigration background, ethnic/racial status and gender (Schulz et al. 2016b). Three scales are used to measure this three-dimensional construct: (a) student attitudes toward equal rights for immigrants, (b) student attitudes toward equal rights for all ethnic/racial groups, and (c) student attitudes toward gender equality. In ICCS 2016, student attitudes towards equal rights for immigrants are captured by items focused on civil and political liberties such as equal rights to education, rights to linguistic and cultural diversity and the right to vote. Similar sets of indicators, tailored to rights and opportunities relevant for each of the groups, are used to capture tolerant attitudes toward ethnic/racial groups (e.g. equal opportunities to labor market participation) and toward gender equality (e.g. equal opportunities to political participation). Different levels of agreement with these items are measured by means of a 4-point Likert scale.

Measurement invariance

In the context of comparative research (concerning more than one group), meaningful comparisons of mean scores require that the items used to operationalize a scale's underlying construct capture the same latent trait across groups or are measurement invariant (Millsap 2011). Measurement invariance holds when a questionnaire measures a construct in the same way regardless of country membership and fails when different sets of people from different countries respond to the items in a dissimilar manner.

In the context of ILSAs, such as ICCS, the intention is often to measure a construct (e.g. young people's tolerant attitudes toward immigrants) across countries; however, a specific construct (or the items underlying it) may very well have a different meaning for the different groups involved or be measurement non-invariant. Causes for measurement non-invariance may relate to the fact that participants do not consider some of the items to be indicative of the construct due to linguistic differences (e.g. inconsistencies in translation that may change the meaning of the items), other cultural differences or country-specific response styles (e.g. social desirability) (He and Van de Vijver 2013; Putnick and Bornstein 2016). For example, societal features such as cross-national differences in the implementation of immigrant integration policies may shape the way in which young people in different countries conceptualize and understand particular aspects of tolerance toward immigrants in that context. Young people from some countries may be wrongly labelled as "less tolerant" only because certain indicators are less relevant to their contextual operationalization of tolerance. Therefore, an essential feature of comparative studies in this area is the establishment of measurement invariance of constructs measured across participants from different countries.

Multiple-group confirmatory factor analysis (MGCFA) is most commonly used for measurement invariance testing of attitudinal measures (Brown 2014; Putnick and Bornstein 2016). MGCFA assumes equality of model parameters in all groups (full measurement invariance) and allows the evaluation of three hierarchical levels of measurement invariance trough the comparison of different models with increasing constraints: (a) configural invariance, (b) metric invariance, and (c) scalar invariance.

The *configural invariance* model tests if the instrument measures the same latent factors and if the set of items associated with each factor is similar across countries. If the configural level of invariance is not achieved, it may be that a different pattern of item loadings is identified in some of the groups (e.g. in one culture one item may load on a different factor or cross-loads on several factors). Meeting the assumption of configural invariance justifies the subsequent tests of metric and scalar invariance but does not guarantee any valid cross-group comparisons.

The *metric invariance* model tests whether the factors have the same meaning and the same measurement unit in all groups. It assumes that each item contributes to the latent factor (has equal item loadings) to a similar degree in all groups. If the metric level of invariance is not achieved, it is likely that some item loadings are not equivalent in some groups (e.g. one item may be strongly related to a factor in some groups but not in other). Reaching this level of measurement invariance justifies the subsequent tests of scalar invariance and also allows for comparisons of latent constructs across groups (e.g. exploring associations of these concepts and other theoretical constructs across countries). However, it does not justify country mean comparisons.

At the *scalar invariance*, in addition to equal item loadings, the item thresholds (the levels of the categorical items; intercepts in the continuous approach) are assumed to be equal in all countries. If scalar invariance is not demonstrated, at least one item threshold (intercept in the continuous approach) may differ across groups. Reaching the level of scalar measurement invariance allows for valid cross-country comparisons of factor scores (scale means).

When measurement invariance tests fail to support the three different assumptions, several options can be considered. One may assume that the construct is non-invariant and refrain from group comparisons, redefine the construct (e.g. omitting some of the items and retesting the models), seek measurement invariance within smaller, more homogeneous, number of groups (e.g. excluding countries and/or focusing on similar groups), and seek only partial measurement invariance and investigate the potential sources of non-invariance (e.g. by relaxing some of the model parameters and retesting the model) (Byrne and Van de Vijver 2010; Kim et al. 2017; Marsh et al. 2017; Putnick and Bornstein 2016).

Full measurement invariance is evaluated by assessing how well the hypothesized models fit the observed data and by testing whether different constraints significantly affect model fit (Brown 2014; Millsap 2011; Putnick and Bornstein 2016). For this purposes, the simultaneous consideration of several overall and comparative fit statistics are recommended (Brown 2014). Nevertheless, we note that model evaluation can be a cumbersome endeavor in practice due to the fact that many fit statistics and their associated cutoffs tend to vary depending on many aspects of the model (e.g. sample size, number of factors, number of groups, continuous versus categorical indicators). Typical overall fit measures are the Chi square test, the root mean square error of approximation (RMSEA), the comparative fit index (CFI), and the Tucker-Lewis index (TLI). These are complemented by measures of comparative fit such as the Chi square difference test, and change in RMSEA and CFI, ΔRMSEA and ΔCFI respectively. The most commonly accepted cutoff criteria (see Brown 2014 for a review) suggest that RMSEA values close to or below 0.060 and CFI and TLI values close to or above 0.950 indicate good model fit. Yet, values of RMSEA in the range of 0.080 and 0.010 and CFI and TLI values in the range of 0.900 and 0.950 can also be considered to indicate acceptable model fit (see also Bentler 1990; Browne and Cudeck 1993). The Chi square statistic is found to be oversensitive to sample size and is less useful when sample sizes are large (Brown 2014). In turn, measures of comparative fit such as ΔRMSEA and ΔCFI were suggested to be indicative of non-invariance when changes in RMSEA (ΔRMSEA) are less than or equal to 0.015 and changes in CFI (ΔCFI) are equal to or greater than −0.010 (Chen 2007; Cheung and Rensvold 2002; French and Finch 2006). However, most of these recommendations apply to MGCFA measurement invariance testing involving smaller number of groups. When the number of groups is larger (such as in ILSAs involving more than 10 countries), different criteria seem to apply. For example, Rutkowski and Svetina (2014) showed that when the number of groups is larger than 10 and data is assumed to be continuous, more liberal criteria such as RMSEA ≤ 0.010 and CFI/TLI ≥ 0.950 can be applied. Moreover, they showed that ΔCFI ≤ 0.020 and ΔRMSEA ≤ 0.030 can be used for evaluating metric invariance while traditional criteria (ΔRMSEA ≤ 0.015; ΔCFI ≥ −0.010) still apply for scalar invariance tests. These cutoffs are overall consistent with the criteria used in current IEA operational procedures for scaling where observed variables are assumed to follow a normal distribution (e.g. Schulz 2016; Schulz et al. 2011). Nevertheless, recent research (Rutkowski and Svetina 2017) shows that the state-of-the-art is changing and that these guidelines must be reconsidered when the character of the data is ordered categorical (e.g. 4-point Likert scales), which is the case with most of the items underlying attitudinal scales in ILSAs. Based on simulation studies applied to ILSA data, Rutkowski and Svetina (2017) recommended a cutoff of 0.055 for the RMSEA, changes in the RMSEA (ΔRMSEA) of 0.050 for metric invariance tests and of 0.010 for scalar invariance test and a ΔCFI threshold of −0.004 for both metric and scalar invariance tests.

Previous research on measurement invariance applied to tolerant attitudes in IEA civic and citizenship education studies

Information regarding the measurement invariance of the scales used to measure student attitudes toward equal rights for immigrants, equal rights for all ethnic/racial groups, and gender equality comes from research conducted in the framework of the IEA civic and citizenship education studies and a few secondary analyses of these data.

The IEA ICCS operational scaling procedures are preceded by a careful a-priory consideration of potential sources of bias (e.g. involvement of national research coordinators representing each participating country in instrument refinement, rigorous translation and piloting procedures) and include investigations of measurement invariance (Schulz 2009, 2016). Although these investigations focus more strongly on detecting item bias by means of differential item functioning (DIF) in an item response theory (IRT) framework, some analyses are carried out in a CFA framework. Findings emerging from research into the measurement invariance of ICCS 2009 and 2016 questionnaire data point to a certain lack of measurement invariance especially at scalar levels of invariance (Schulz 2009, 2016). Against this background, researchers making use of IEA ICCS data are urged to engage in analytic efforts to understand the degree to which questionnaire constructs are impacted by cultural differences in measurement and to determine when such differences become problematic (Schulz 2016).

In terms of secondary data analyses, two previous studies specifically dealt with the evaluation of comparability of attitudes toward immigrants within and across the two previous waves of the IEA citizenship education studies, CIVED 1999 and ICCS 2009 (Miranda and Castillo 2018; Munck et al. 2017). They provided evidence that (at least a number of) items are comparable across countries. Moreover, this work signaled potential sources of non-invariance and provided useful hints for further evaluations. For example, to reach measurement invariance, these studies either had to reconsider the concept (e.g. excluding some ill-fitting items due small item loadings) or operate in a partial or approximate invariance framework (e.g. applying the alignment method to identify an optimal partial measurement model). Using both CIVED 1999 and ICCS 2009 data, Munck et al. (2017) pointed out the difficulty to consistently measure cultural aspects of tolerance toward immigrants (e.g. endorsing rights to linguistic and cultural diversity), which could be due to cross-national differences in endorsing generic versus cultural rights. Using data of the 38 countries participating in ICCS 2009, Miranda and Castillo (2018) demonstrated in turn the importance of taking into account the dependency between the three aspects of the multidimensional construct of tolerant attitudes toward equal rights that includes student attitudes toward equal rights for immigrants, but also student attitudes toward equal rights for all ethnic/racial groups, and student attitudes toward gender equality. More specifically, they find that the three different attitudinal measures are highly interdependent showing strong correlations (well above 0.600, on average) among each other with a particularly strong association (0.800, on average) between attitudes toward equal rights for immigrants and attitudes toward equal rights for all ethnic/racial groups.

The current study

In the current study, we aimed to add to current research along several lines. Our main purpose was to examine the extent to which average comparisons of cross-national differences in young people's tolerant attitudes toward immigrants are justified also in the context of the latest wave of the IEA civic and citizenship education study, ICCS 2016. To this end, we responded to the current state-of-the-art by taking into account the ordered categorical character of the data and followed the most recent guidelines for model fit evaluation for measurement invariance testing in a MGCFA framework in the context of ILSAs (Rutkowski and Svetina 2017). Moreover, we considered insights from previous research (Miranda and Castillo 2018) and specifically tested the appropriateness of estimating a multidimensional construct of tolerant attitudes toward equal rights composed of three interrelated factors: student attitudes toward equal rights for immigrants, student attitudes toward equal rights for all ethnic/racial groups, and student attitudes toward gender equality. Furthermore, acknowledging that cross-cultural comparability among such heterogeneous contexts is often a difficult task guided by model fit evaluation criteria that are being often reconsidered (in light of specific characteristics of models), we carried out investigations into the robustness of our findings by means of sub-group analysis (see "Selection of country clusters" section for details). In doing so, we aimed to provide secondary-users of ICCS 2016 data with sufficient information regarding the cross-cultural comparability of tolerant attitudes toward immigrants in the context of this survey and potentially provide relevant information for future scale development in forthcoming studies.

Method

Data

Sample

The International Civic and Citizenship Education Study (ICCS) 2016 (Schulz et al. 2016b, 2018) conducted in 24 countries by the International Association for the Evaluation of Educational Achievement (IEA) was the principal data source for all the analyses reported here. Nevertheless, in this research, we used only data from the 14 European countries that participated in the European Module of the ICCS 2016 (Losito et al. 2018; Schulz et al. 2016a) study where students completed questionnaires inquiring into their attitudes toward equal rights for immigrants, student attitudes toward equal rights for all ethnic/racial groups, and student attitudes toward gender equality.

In each country, the surveyed students are representative samples of the population of grade 8 students. More specifically, the study followed a two-stage cluster sampling strategy. In a first stage probability proportional to size (PPS) procedures were used to select schools within each country. In the second stage, within each sampled school, an intact class from the target grade was selected at random, with all the students in this class participating in the study. In total, 51,040 students clustered in 14 countries were included in this research. Table 2 indicates the distribution of students per country.

Selection of country clusters

One strategy to test generalizability of findings obtained in a full measurement invariance approach consists of examining if the findings obtained from the full sample are consistent with findings estimated across smaller, more culturally homogeneous groupings. Therefore, we used different sources of information to identify clusters of countries that show similarities in terms of language, and geographic location. We also consider information regarding the democratic tradition, immigration patterns, integration policies, and attitudes towards immigration. For these purposes we were mainly guided by the classification of immigrant destination countries introduced by the Organisation for Economic Co-operation and Development (OECD) and the European Union (EU) (OECD and European Union 2015). Moreover, whenever possible, we updated and complemented this information using several sources such as the Human Development Index (HDI) (Jāhāna 2016), the Migrant Integration Policy Index (MIPEX) (Migration Policy Group 2015), the Democracy Index (Economist Intelligence Unit 2017), the European Social Survey (ESS) (Heath et al. 2016), and the Special Eurobarometer 469 (European Commission 2018).

The following clusters of countries were identified:

a. Nordic Countries: Denmark, Finland, Norway, Sweden.

   Other than sharing geographical and linguistic similarities, these countries share a long and stable tradition with democracy, high levels of development and high levels of egalitarianism. They are also characterized by significant recent and humanitarian migration. Most of the immigrants are non-native speakers and humanitarian immigrants struggle to integrate. Overall (with slightly lower scores for Denmark), integration policies are strong and long-standing providing access to citizenship, education and training and equal opportunities. The levels of support for immigration among adults are the highest in Europe.
b. Western European Countries: Belgium (Flemish), The Netherlands.
   The two countries share high levels of linguistic and geographic proximity. They have a strong democratic tradition and high levels of development. In the European context, they are long-standing immigration destinations that received the inflows of immigrants or "guest workers" in the wake of World War II and afterwards (family reunion). Most of the immigrants and their families are low-educated and face integration issues such as lower labor market participation and higher relative poverty rates. Integration policies are slightly favorable with relatively strong anti-discrimination laws and support for education, but they are rather restrictive in terms of access to long-term residence and family reunion. The levels of support for immigration among adults are higher in the Netherlands than in Belgium and slightly lower relative to the Nordic countries.
c. Central and Eastern European Countries: Bulgaria, Estonia, Latvia, Lithuania, Croatia, Slovenia.
   This cluster of countries is by far the most heterogeneous. Nevertheless, they share a recent and less stable democratic tradition and relatively lower levels of economic development as compared with the Nordic and Western European clusters. Moreover, in all these countries the immigrant population was shaped by border changes (in the late twentieth century) and/or by national minorities. Some countries (Bulgaria) experience recent major humanitarian migration. Overall, most of the immigrants show outcomes (e.g. education, labour market) similar to those of the native-born. Integration policies are, on average, the least favorable in Europe. With some exemptions (Slovenia), overall levels of support for immigration are lower than in most of the other European countries.
d. Southern European Countries: Italy, Malta.
   Italy and Malta show similarities in terms of geographic proximity, level of development and some linguistic overlap. Until recently, they were characterized as being new destination countries with many immigrants arriving at the beginning of the twentyfirst century. Nevertheless, they (and particularly Italy) currently experience massive intakes of humanitarian migration. Most immigrants tend to be less educated and show lower integration outcomes, especially in Italy. Integration policies are evaluated to be halfway favorable in Italy but among the best among Europe's major countries of immigration. Malta's integration policies are rated as being slightly

Table 1 Measures of attitudes toward equal opportunities

Item code	Item text
Response categories: 1 = Strongly disagree; 2 = Disagree; 3 = Agree; 4 = Strongly agree	
Domain 1: Attitudes toward equal rights for all ethnic/racial groups	
IS3G25A[a]	All <ethnic/racial groups> should have an equal chance to get a good education in <country of test>
IS3G25B[a]	All <ethnic/racial groups> should have an equal chance to get good jobs in <country of test>
IS3G25C[a]	Schools should teach students to respect <members of all ethnic/racial groups>
IS3G25D[a,c]	<Members of all ethnic/racial groups> should be encouraged to run in elections for political office
IS3G25E[a]	<Members of all ethnic/racial groups> should have the same rights and responsibilities
Domain 2: Attitudes toward gender equality	
IS3G24A[a,b]	Men and women should have equal opportunities to take part in government
IS3G24B[a,b]	Men and women should have the same rights in every way
IS3G24C	Women should stay out of politics
IS3G24D	When there are not many jobs available, men should have more right to a job than women
IS3G24E[a,b]	Men and women should get equal pay when they are doing the same jobs
IS3G24F	Men are better qualified to be political leaders than women
IS3G24G[c]	Men and women should have equal opportunities to take part in government
Domain 3: Attitudes toward equal rights for immigrants	
ES3G04A[a,c]	<Immigrants> should have the opportunity to continue speaking their own language
ES3G04B[a]	<Immigrant> children should have the same opportunities for education that other children in the country have
ES3G04C[a]	<Immigrants> who live in a country for several years should have the opportunity to vote in elections
ES3G04D[a]	<Immigrants> should have the opportunity to continue their own customs and lifestyle
ES3G04E[a]	<Immigrants> should have the same rights that everyone else in the country has

[a] Item reversed coded

[b] Item excluded from the analysis due to extremity scoring

[c] Item excluded from the analysis due to low item loadings

unfavorable. In both contexts, the most needed developments seem to be in the area of equality and anti-discrimination policies. Immigration tends to be perceived as problematic in these countries.

Variables

The variables used as indicators for the three dimensions of "attitudes toward equal rights" are described in Table 1. Each construct is captured by a set of items measured on 4-point Likert scales ranging from *strongly disagree* to *strongly agree*. As indicated in Table 1, some items were reverse coded to ensure that high scores on each item reflect positive attitudes toward the three groups. Moreover, preliminary descriptive analyses led to the exclusion of some items. In a first step, descriptive analysis showed that student responses to items IS3G24A (Men and women should have equal opportunities to take part in government), IS3G24B (Men and women should have the same rights in every way) and IS3G24E (Men and women should get equal pay when they are doing the same jobs) showed high rates (exceeding 70%) of agreement ("strongly agree"). These items were therefore excluded from subsequent analysis due to extremity scoring. In a second step, preliminary country specific confirmatory factor analysis pointed out three items with moderate factor loadings (well below or at the threshold of 0.600; MacCallum et al. 1999, 2001) in a majority of countries. These items were: IS3G25D (Members of all ethnic/racial groups should be encouraged to run in elections for political office)—showing factor loadings below the threshold in seven countries, IS3G24G (Men and women should have equal opportunities to take part in government)—showing factor loadings below the threshold in ten countries, and ES3G04A (Immigrants should have the opportunity to continue speaking their own language)—showing factor loadings below the threshold in six countries. These items were also excluded from further analyses.

Analytical strategy

To establish if average scores on attitudes toward immigrants are comparable across the contexts, measurement invariance was investigated in a factor analytical framework. We considered attitudes toward immigrants to be one aspect of a three-dimensional construct of attitudes toward equal rights. Data preparation was done with the IEA IDB analyzer (IEA 2017) and IBM SPSS Statistics 23.00 (IBM Corp. 2015). All CFA and MGCFA analyses were performed in Mplus 7.4 (Muthén and Muthén 2017). To handle missing data, we used the full information maximum likelihood (FIML) method implemented in Mplus 7.4. This method uses all available information for any variable. Only cases with missing data on all variables are not included in the analysis. The number of cases with missing data on all variables for this research was 297. Moreover, we took into account the multilevel character of the data (students nested within schools within countries) by implementing the TYPE = COMPLEX option of Mplus 7.4 that adjusts model goodness-of-fit statistics and standard errors of the parameter estimates for the dependency in the data (see also Brown 2014).

An initial step in assessing the measurement invariance of the instrument involved country-specific analysis. This entailed specifying a first-order correlated three-factor model of attitudes toward equal rights with the 11 indicator variables loading on the three dimensions: (a) attitudes toward equal rights for all ethnic/racial groups (4 items), (b) attitudes toward gender equality (3 items) and c) attitudes toward equal rights for immigrants (4 items). Confirmatory factor analysis (CFA) was used to test the factor structure of this model in each of the 14 countries. The ordered categorical character of the data (4 point Likert scale) was taken into account by using an extension of the CFA model that estimates polychoric correlations and asymptotic covariance matrices to reflect the relations between response variables with a weighted least square mean variance (WLSMV) estimator. The fit of this model in each country was compared to a first-order uncorrelated three-factor model. To evaluate model fit we used the following overall goodness of fit measures: the root mean square error of approximation (RMSEA), the comparative fit index (CFI), and the Tucker-Lewis index (TLI). Common guidelines for model fit evaluation were applied: RMSEA ≤ 0.060 CFI ≥ 0.950; TLI ≥ 0.950 (Brown 2014; Wang and Wang 2012).

To investigate the measurement invariance of the construct, we applied multiple-group confirmatory factor analysis (MGCFA) taking into account the ordered categorical character of the data. The three-factor model was estimated simultaneously for the 14 countries. Data was weighted so that countries contributed equally to the analysis. In this framework, the assessment of measurement invariance involved the comparison of the three nested competing models, i.e. configural, metric and scalar models. In order to evaluate model fit, we considered both overall fit measures (i.e. CFI, TLI, RMSEA) and relative fit measures such as changes in CFI (ΔCFI) and RMSEA (ΔRMSEA). We followed the guidelines indicated by Rutkowski and Svetina (2017) that are more applicable in the context of ILSA and MGCFA with categorical indicators: RMSEA ≤ 0.055; CFI ≥ 0.950; TLI ≥ 0.950; ΔRMSEA ≤ 0.010; ΔCFI ≥ − 0.004. To further asses model-fit, model-based item reliability (item loadings which capture the strength of the association between the indicators and the underlying latent variable) and construct/scale reliability (assessing the reliability of a construct underlying a set of observed indicators) were estimated following Wang and Wang (2012) based on the MGCFA scalar model. To provide further evidence on the robustness of our results, we conducted measurement invariance tests by means of MGCFA within the four sub-groups of country clusters that show contextual similarities. An approach similar to the one applied to the analysis of all countries was followed. In addition to the model fit criteria relevant for larger number of groups (Rutkowski and Svetina 2017), when applicable, we also observed guidelines relevant in the case of comparing two groups: RMSEA ≤ 0.060; CFI ≥ 0.950; TLI ≥ 0.950; ΔRMSEA ≤ 0.015; ΔCFI ≥ − 0.010 (see Brown 2014). Factor scores obtained with this approach were compared with the ones estimated across all countries.

In a last step, to ensure greater interpretation and comparability with the estimations reported in the IEA ICCS 2016 documentation, model-based factor scores were saved and rescaled to a T-scale with a mean of 50 and a standard deviation of 10. These scores were used to estimate and illustrate average comparisons of cross-country differences in attitudes toward immigrants in the fourteen European countries. These comparisons were estimated and tested using the IEA IDB analyzer (IEA 2017) and focused on two aspects: (a) overall (mean) differences and (b) disparities in terms of the distance between the 5th and the 95th percentile.

Results

Results of measurement invariance testing

Country-specific models

Table 2 presents the model data fit for the 14 country-specific CFA models estimated on the full samples. We tested the fit of the first-order correlated three-factor model (Table 2, M2) and compared it with a first-order uncorrelated three-factor model (Table 2, M1). We can observe that the first-order correlated three-factor model (Table 2, M2) showed an adequate fit in most samples. Fit indices largely fell within acceptable ranges with RMSEA values below 0.060 and CFI and TLI well above 0.950. The findings from these separate CFAs indicate that the same number of (three) correlated factors

Table 2 Results of confirmatory factor analysis

Country	N	M1: First-order uncorrelated three-factor model			M2: First-order correlated three-factor model		
		RMSEA	TLI	CFI	RMSEA	TLI	CFI
Bulgaria	2958	0.130	0.824	0.780	0.054	0.972	0.962
Croatia	3893	0.173	0.788	0.735	0.036	0.991	0.988
Denmark	6125	0.189	0.755	0.694	0.029	0.994	0.993
Estonia	2854	0.178	0.785	0.731	0.051	0.984	0.978
Finland	3162	0.264	0.709	0.636	0.030	0.997	0.995
Italy	3446	0.224	0.662	0.577	0.053	0.982	0.976
Latvia	3208	0.119	0.870	0.837	0.044	0.983	0.978
Lithuania	3624	0.179	0.796	0.745	0.044	0.989	0.985
Malta	3749	0.114	0.786	0.745	0.058	0.969	0.958
The Netherlands	2800	0.186	0.757	0.696	0.043	0.988	0.984
Norway	6235	0.247	0.842	0.803	0.056	0.992	0.990
Slovenia	2842	0.217	0.736	0.670	0.046	0.989	0.985
Sweden	3218	0.233	0.806	0.757	0.035	0.996	0.994
Belgium (Flemish)	2926	0.158	0.793	0.741	0.044	0.985	0.980

Country-specific models

N, sample size; RMSEA, root mean square error of approximation; CFI, comparative fit index; TLI, Tucker-Lewis index

Table 3 Results of multiple-group analysis, overall model

Model	Full sample	RMSEA	CFI	TLI
M1	Configural	0.045	0.990	0.986
M2	Metric	0.041	0.990	0.989
M3	Scalar	0.043	0.985	0.987
	Nested models comparisons	ΔRMSEA	ΔCFI	
	Metric vs configural	0.004	0.000	
	Scalar vs metric	− 0.002	0.005	

RMSEA, root mean square error of approximation; CFI, comparative fit index; TLI, Tucker-Lewis index; ΔRMSEA, change in RMSEA; ΔCFI, change in CFI

with similar patterns of item loadings can be identified in all countries. In contrast, the first-order uncorrelated three-factor model (Table 2, M1) showed unacceptable model fit in all countries. We therefore accepted the first-order correlated three-factor model (Table 2, M2) and retained it for the subsequent analyses.

Results of multiple-group analysis across all countries

Table 3 provides a summary of the findings for the competing measurement invariance models: configural, metric and scalar. The results at the configural, metric and scalar levels of invariance largely comply with the model fit evaluation criteria both in terms of overall fit indices (e.g. RMSEA ≤ 0.055; CFI ≥ 0.950; TLI ≥ 0.950) as well as some of the relative fit indices (e.g. ΔRMSEA ≤ 0.010. An exception is one of the relative fit indices (ΔCFI) that exceeds slightly the threshold of ΔCFI ≥ − 0.004 taking a value of − 0.005 when comparing the fit of the scalar model to the one of the metric model. Nevertheless, all other indices are well within acceptable boundaries (RMSEA = 0.043; CFI = 0.985; TLI = 0.987; ΔRMSEA = − 0.002). Following Brown (2014), we considered model fit to be expressed by the highest level of consistency among all other fit indices, and we accepted the scalar model. Moreover, we considered other aspects of model evaluation (see Brown 2014) by examining the strength of the associations between the items and the latent variables (item loadings) and model-based scale reliability based on the scalar model.

Model-based scale reliabilities and item loadings are illustrated by Fig. 1. Scale reliabilities for three scales were above 0.800 in all countries ranging from 0.809 to 0.960. These measures capture the proportion of scale variance not attributable to measurement error. In the current case, the estimates suggest high reliability of the three latent variables underlying the three sets of observed indicators (items). Moreover, the scales showed to be particularly reliable in most of the Nordic countries (i.e. Finland, Norway and Sweden) with the reliability measure well above 0.900 for the scales capturing attitudes toward equal rights for all ethnic/racial groups and attitudes toward gender equality and above 0.880 for the scale capturing attitudes toward equal rights for immigrants.

Item loadings, were well above the 0.500 for all scales and countries ranging from 0.577 to 0.966. This finding indicates sufficient indicator reliability. For the attitudes toward equal rights for all ethnic/racial groups and attitudes toward gender equality scales, item reliability was high (above 0.780, on average) and rather consistent across countries. In contrast, for the attitudes toward equal rights for immigrants scale, findings were more heterogeneous across countries and some items were clearly stronger measures than other. More specifically, the strongest indicator of the scale in the majority of countries was item ES3G04B (<Immigrant> children should have the same opportunities for education that other children in the country have) with item loadings exceeding 0.800 while the weakest indicators was item ES3G04D (<Immigrants> should have the opportunity

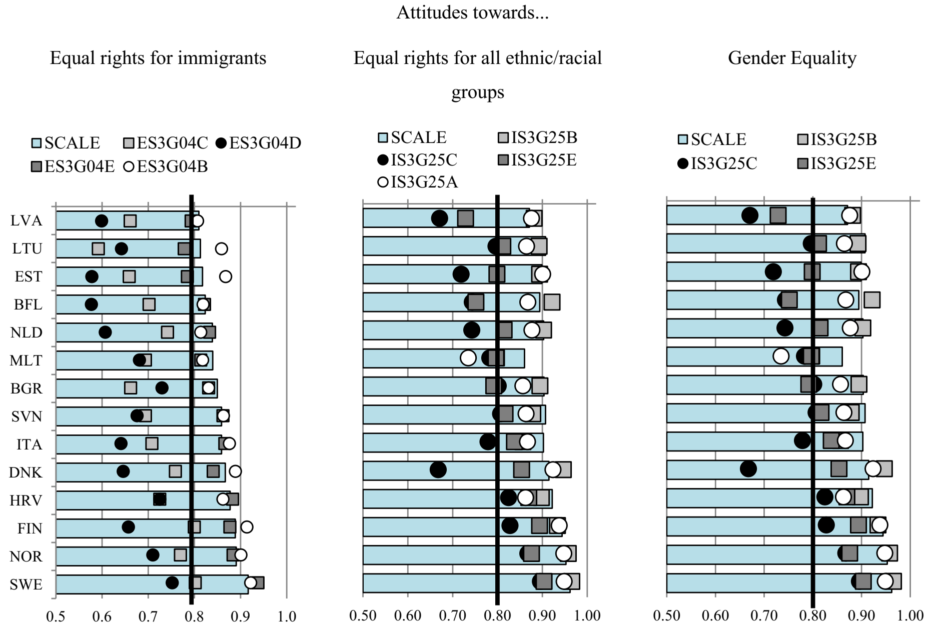

Fig. 1 Scale reliability and item loadings

to continue their own customs and lifestyle) with item loadings ranging from 0.577 to 0.720.

In addition, supportive to the suitability of the hypothesized three-dimensional model, results pointed out that the three factors are related with strong associations in most countries, with a correlation of 0.600, on average. This was especially true for the associations between attitudes toward equal rights for immigrants and attitudes toward equal rights for all ethnic/racial groups. More specifically, across the 14 countries, associations between attitudes toward equal rights for immigrants and attitudes toward equal rights for all ethnic/racial groups ranged from 0.465 to 0.748 with an average of 0.628. Associations exceeding the value of 0.700 were registered for Denmark, Finland, Italy, Norway and Sweden while an association below 0.500 was recorded for Latvia.

Robustness analysis—country clusters-specific models

Table 4 provides a summary of the findings for the competing measurement invariance models: configural, metric and scalar for each country cluster. For the "Nordic" and "Western European" country clusters the results at the configural, metric and scalar levels of invariance, all of the overall and relative fit indices showed values well within established criteria (e.g. RMSEA ≤ 0.055; CFI ≥ 0.950; TLI ≥ 0.950; ΔRMSEA ≤ 0.010; ΔCFI ≥ − 0.004) confirming a very good model fit for the scalar model. For the "Central and Eastern European" cluster, the majority of indices were well within acceptable boundaries. Only ΔCFI exceeded slightly the threshold of ΔCFI ≥ − 0.004 taking a value of 0.007 when comparing the fit of the scalar model to the one of the metric model. Nevertheless, we considered once more the highest level of consistency among all other fit indices as an indication of good model fit and accepted the scalar model. For the "Southern European" cluster, the comparison included only two countries (Italy and Malta).

Table 4 Results of multiple-group analysis

Model	RMSEA	CFI	TLI	ΔRMSEA	ΔCFI
Nordic Countries					
Configural	0.041	0.994	0.992	–	–
Metric	0.037	0.994	0.993	0.004	0.000
Scalar	0.039	0.992	0.993	− 0.002	0.002
Western European Countries					
Configural	0.043	0.987	0.982	–	–
Metric	0.040	0.988	0.985	0.003	− 0.001
Scalar	0.038	0.986	0.986	0.002	0.002
Central and Eastern European Countries					
Configural	0.045	0.986	0.981	–	–
Metric	0.041	0.986	0.984	0.004	0.000
Scalar	0.045	0.979	0.981	− 0.004	0.007
Southern European Countries					
Configural	0.056	0.976	0.968	–	–
Metric	0.053	0.977	0.972	0.003	− 0.001
Scalar	0.050	0.974	0.974	0.003	0.003

Country clusters-specific models

RMSEA, root mean square error of approximation; CFI, comparative fit index; TLI, Tucker-Lewis index; ΔRMSEA, change in RMSEA; ΔCFI, change in CFI

The value of RMSEA for the configural model marginally exceeded the strict threshold of RMSEA ≤ 0.055, taking a value of 0.056, but it still indicated a reasonably good model fit (RMSEA ≤ 0.060, Brown 2014). All other fit indices were well within acceptable boundaries. We therefore accepted the scalar model.

Although, taken together, all findings from the country–cluster specific analysis confirmed that measurement invariance can be achieved at the scalar level, we note that a very good model fit was particularly registered for the "Nordic" and "Western European" country clusters. To provide further validation of findings, we compared the association between the factor scores (based on the scalar models) obtained from the measurement invariance analysis conducted across all countries with the ones extracted from the country-cluster specific analysis. For all countries, we found a perfect correlation ($r = 1.000$) among the two solutions.

Cross-country comparisons

Having established that latent mean scores can be compared, we proceeded to illustrate the differences in attitudes toward immigrants in the fourteen European countries.

For further interpretation of the distribution of attitudes across groups Fig. 2, shows the national and European average as well as the threshold scores for the 5th percentile (the 5% of students with the most negative attitudes) and the 95th percentile (the 5% of students with the most positive attitudes). The percentile ranges are specific to each education system's distribution of scores; the education systems are ordered by national

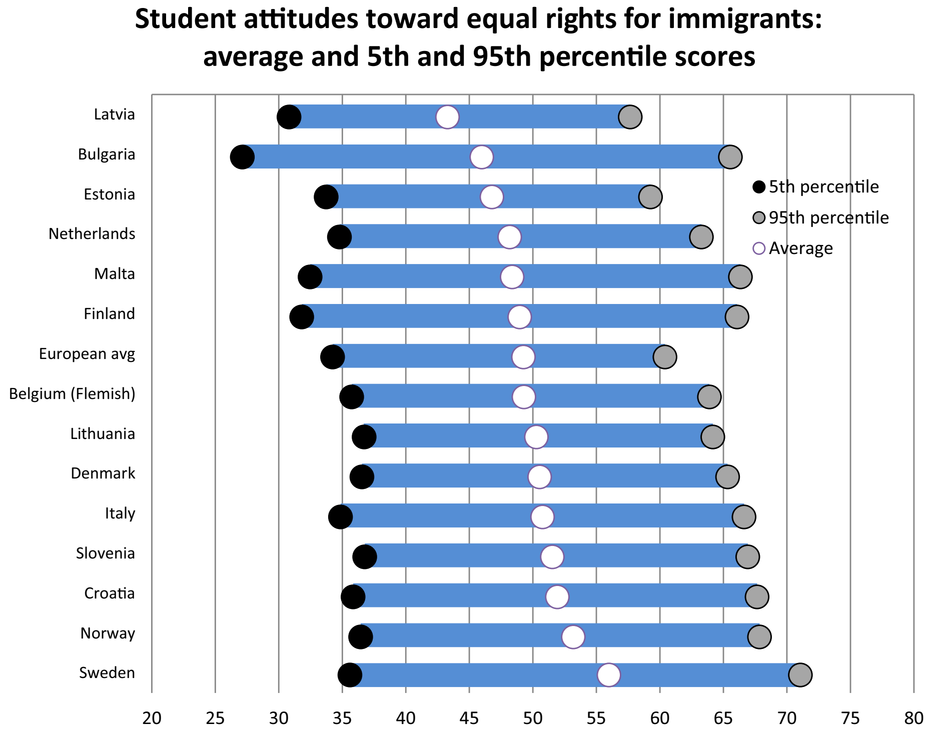

Fig. 2 Cross-country comparisons: average and 5th and 95th percentile scores

average from smallest to largest. The European average is the average of the national averages of the participating European countries, with each country weighted equally.

Across all countries, student attitudes toward immigrants are rather positive with average scores close to the mean of 50. Moreover, on average, the attitudes toward immigrants of the students in Sweden, Norway, Croatia, Slovenia, Italy, Denmark, Lithuania, and Belgium (Flemish), are significantly more positive than the European average of 49.26. In these countries the variability of the scores is moderate, the attitudes of the students tend to be relatively concentrated around the mean, and the score gap between the 95th and the 5th percentile is relatively small (around 30 points). Among these countries, Sweden has both the highest mean score (55.99) and the largest variability (between 35.45 and 71.06).

Concerning the countries with average scores significantly lower than the European average, the situation is more varied. In some countries (Latvia, Estonia, Netherlands) the attitudes are relatively concentrated around the mean and the distance between the 5th and 95th percentiles is less than 30 points. In other countries (Bulgaria, Malta, Finland) the variability is much higher as depicted by the distance between the 5th and 95th percentiles, which ranges between 34 and 38 points.

Discussion and conclusion

In the current study we argued that cross-cultural comparability must be ensured for the measurement of highly relevant indicators that serve to monitor inter-European and international differences in young people's tolerant attitudes toward immigrants. To this end, we examined the extent to which average comparisons of cross-national differences in young people's tolerant attitudes toward immigrants in the context of the recent ICCS 2016 study are justified. We tested for measurement invariance and applied multiple-group confirmatory factor analysis (MGCFA) (Jöreskog 1971; Steenkamp and Baumgartner 1998) to assess whether comparisons of average scale scores across fourteen European countries participating in ICCS 2016 can be made with confidence. Moreover, in response to recent developments in the field, we aimed to contribute to further research by considering new methodological challenges (e.g. the ordered categorical character of the data, recent guidelines for model fit evaluation), incorporating previous research (e.g. by considering the multidimensionality of the construct of tolerant attitudes toward equal rights) and investigating the robustness of our findings by means of sub-group analysis. Our results pointed out that cross-cultural comparability can be achieved for (most items of) this scale with ICCS 2016 data among the 14 investigated countries. More specifically, results of measurement invariance tests using MGCFA pointed to the achievement of full scalar invariance with the implication that average scores based on three interrelated scales (attitudes toward equal rights for immigrants, attitudes toward equal rights for all ethnic/racial groups, and attitudes toward gender equality) can be validly compared across the educational systems under investigation. These findings were largely corroborated by information of item and scale reliability and the results emerging from our robustness analysis. Notably, in addition to providing validation to the results obtained across all countries, the country cluster-specific analyses indicated that cross-country comparisons are defensible also among more homogeneous groupings of countries and that these comparisons are particularly strong (as indicated by model fit evaluation parameters as well as item and scale reliability, in some cases) in the "Nordic" and "Western" European clusters of countries. Such findings may be due to higher proximity between these countries in terms of, for example, linguistic similarity, democratic tradition, and/or experience with immigration and integration policies. In this respect, future research involving a higher number of countries may permit exploring the validity of such explanatory mechanisms.

Moreover, the analysis also revealed information relevant for further scale refinement. First, we confirmed that tolerant attitudes toward immigrants is one aspect of a larger three-dimensional concept encompassing three (correlated) factors: attitudes toward equal rights for immigrants, attitudes toward equal rights for all ethnic/racial groups and attitudes toward gender equality. Given the long tradition of this conceptualization in the IEA civic and citizenship education studies (most indicators remained almost identical across CIVED 1999, ICCS 2009 and ICCS 2016), this finding corroborates existing assumptions and previous research (Miranda and Castillo 2018). Second, in line with extant findings (Munck et al. 2017), we showed that items capturing cultural aspects of tolerance toward immigrants (e.g. endorsing rights to linguistic and cultural diversity) are either unreliable in most countries or show substantial variability in terms of factor loadings. This finding stands in contrast with the performance of items tapping into more generic and less debated rights such as the right to education. One may only speculate that such items are more susceptible to influences defined by public opinion or the characteristics of the application of immigrant integration policies in some countries but this issue certainly needs further investigation especially in developmental phases of future studies.

In addition, our illustration of the differences in attitudes toward immigrants in the fourteen European countries showed that young people have, on average, largely tolerant attitudes in all European countries. Moreover, these results provided evidence that both low (e.g. Bulgaria) and high (e.g. Sweden) average scores on tolerance toward immigrants could be coupled with high degrees of polarization (measured as the gap between the 95th and 5th percentile) of these attitudes within the countries. These findings point to the need of considering multiple ways of describing these indicators, especially when improvement efforts are targeting the entire distribution of the young population in a country (limited here to the ICCS 2016 sample). For further research, this information could be supplemented with other indicators of attitudinal polarization shaped, for example, by socioeconomic background or gender.

Lastly, the current study encountered a number of limitations which may provide additional avenues for further research in the field. First, we acknowledge that our substantive contribution to the conceptualization and measurement of tolerance toward immigrants is limited as it had to be embedded within the boundaries set by the ICCS 2016 study. Although we have sufficient confidence that this measure is valid and reliable over the years, we acknowledge that conceptualizations of tolerance toward immigrants could be broader (e.g. including additional measures of social tolerance) and that measurement strategies could be improved (e.g. by applying a "least liked" approach, see Gibson 2013). We therefore welcome further reflections on these issues both outside and within the ILSA community. Second, this article illustrated that investigating measurement invariance in ILSAs such as ICCS is a vibrant emerging field of research. As such, guidelines for the evaluation of measurement invariance testing are rather imprecise (e.g. when considering various characteristics of the model such as number of factors, number of groups, sample size, number of items per factors, continuous versus categorical data, etc.) and are constantly being reconsidered. In this study, we followed guidelines for model fit evaluation that were largely based on several studies investigating smaller groups and mostly continuous measures and one study that documented cutoff criteria applied to large groups of countries and unidimensional categorical data. We were unable to identify studies that document model fit evaluation criteria that are closely applicable to our specific models (e.g. three-dimensional construct, different number of items underlying several concepts, large sample sizes, large number of groups and categorical indicators). We therefore acknowledge that future research may challenge our findings. We consider further simulation studies (e.g. considering several conditions) and well documented measurement invariance studies (e.g. reporting potential sources of non-invariance) applied to ILSA data to be fruitful avenues for advancing further knowledge. Third, in contrast to previous waves of the IEA civic and citizenship education studies (i.e. CIVED 1999 and ICCS 2019), in ICCS 2016, the measure tapping into student's attitudes toward immigrants was administered only within the European module of the survey. For this reason, measurement invariance investigations applied to the larger set of countries (located also in Latin America or Asia) included in the ICCS 2016 main survey could not be performed. Therefore, we must acknowledge that our results are only applicable to the fourteen European countries surveyed in ICCS 2016. While different, the European countries share a common core of cultural and contextual characteristics. The non-European ICCS 2016 countries may instead show stronger contextual differences that could challenge comparability. Future ICCS studies could potentially provide opportunities to extend this analysis to a larger, more heterogeneous, group of countries. To conclude, for the time being, we are confident to have provided different stakeholders with sufficiently reliable and relevant information regarding the cross-cultural comparability of inter-European differences in young people's tolerant attitudes toward immigrants in the context of ICCS 2016 and we highly encourage further research along the lines outlined above.

Authors' contributions

MMI conducted the literature review, carried out the analyses, and prepared the manuscript. GvdW and LP revised and provided substantial input for all aspects of the manuscript. All authors read and approved the final manuscript.

Author details

[1] University of Groningen, Grote Rozenstraat 3, 9712 TG Groningen, The Netherlands. [2] INVALSI, Via Ippolito Nievo, 35, 00153 Rome, Italy.

Acknowledgements

MMI would like to thank Dr. Andres Sandoval-Hernandez for his useful feedback at the initial stages of drafting the manuscript.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The datasets generated and/or analysed during the current study are available in the IEA Data Repository, https://www.iea.nl/data.

Funding
Not applicable.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 30 November 2018 Accepted: 16 March 2019
Published online: 28 March 2019

References

Allport, G. W. (1954). *The nature of prejudice*. New York, NY: Addison.

Almond, G. A., & Verba, S. (1963). *The civic culture: Political attitudes and democracy in five nations*. Princeton, New York: Princeton University Press.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238.

Berry, J. W. (2011). Integration and multiculturalism: Ways towards social solidarity. *Papers on Social Representations, 20,* 1–21.

Berry, J. W., & Sam, D. L. (2014). *Multicultural societies* (pp. 97–117). Handbook of Multicultural Identity: Basic and Applied Perspectives.

Brown, T. A. (2014). *Confirmatory factor analysis for applied research. Methodology in the social sciences*. London: Guilford.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. *Sage Focus Editions, 154,* 136.

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*(2), 107–132.

Ceobanu, A. M., & Escandell, X. (2010). Comparative analyses of public attitudes toward immigrants and immigration using multinational survey data: A review of theories and research. *Annual Review of Sociology, 36*(1), 309–328. https://doi.org/10.1146/annurev.soc.012809.102651.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modelling, 9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5.

Côté, R. R., & Erickson, B. H. (2009). Untangling the Roots of Tolerance. *American Behavioral Scientist, 52*(12), 1664–1689. https://doi.org/10.1177/0002764209331532.

Council of Europe. (2017). *Learning to Live Together. Council of Europe Report on the state of citizenship and human rights education in Europe*. Retrieved from https://rm.coe.int/the-state-of-citizenship-in-europe-e-publication/168072b3cd.

Economist Intelligence Unit. (2017). *Democracy Index 2016: Revenge of the "Deplorables"*. The Economist Intelligence Unit. Retrieved from https://www.eiu.com/public/topical_report.aspx?campaignid=DemocracyIndex2016.

Elchardus, M., Franck, E., Groof, S. D., & Kavadias, D. (2013). The acceptance of the multicultural society among young people. A comparative analysis of the effect of market-driven versus publicly regulated educational systems. *European Sociological Review, 29*(4), 767–779. https://doi.org/10.1093/esr/jcs056.

European Commission. (2018). *Special Eurobarometer 469: Integration of immigrants in the European Union*. Retrieved from http://data.europa.eu/euodp/en/data/dataset/S2169_88_2_469_ENG.

European Commission/EACEA/Eurydice. (2017). *Citizenship Education at School in Europe—2017*. Bruxelles.

European Council. (2015). Declaration on Promoting citizenship and the common values of freedom, tolerance and non-discrimination through education. Retrieved from http://ec.europa.eu/dgs/education_culture/repository/education/news/2015/documents/citizenship-education-declaration_en.pdf.

Forst, R. (2003). Toleration, justice and reason. In C. McKinnon & D. Castiglione (Eds.), *The culture of toleration in diverse societies* (pp. 71–85). Manchester, UK: Manchester University Press. http://www.manchesteruniversitypress.co.uk/9780719080623/.

French, B. F., & Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modelling, 13*(3), 378–402. https://doi.org/10.1207/s15328007sem1303_3.

Gibson, J. L. (2006). Enigmas of intolerance: Fifty years after Stouffer's communism, conformity, and civil liberties. *Perspectives on Politics*, *4*(1), 21–34. Retrieved from http://www.journals.cambridge.org/abstract_S153759270606004X.

Gibson, J. L. (2013). Measuring political tolerance and general support for pro-civil liberties policiesnotes, evidence, and cautions. *Public Opinion Quarterly, 77*(S1), 45–68. https://doi.org/10.1093/poq/nfs073.

Green, A., Preston, J., & Janmaat, J. (2006). *Education, equality and social cohesion: A comparative analysis*. Berlin: Springer.

He, J., & Van de Vijver, F. J. R. (2013). Methodological issues in cross-cultural studies in educational psychology. In G. A. D. Liem & A. B. I. Bernardo (Eds.), *Advancing cross-cultural perspectives on educational psychology: A festschrift for Dennis McInerney* (pp. 39–56). NC: Information Age Publishing Charlotte.

Heath, A., Schmidt, P., Butt, S., Dorer, B., Fitzgerald, R., Prestage, Y., et al. (2016). Attitudes towards Immigration and their Antecedents: Topline results from round 7 of the European Social Survey. *European Social Survey Topline Results, 7,* 1–16. https://doi.org/10.5167/uzh-93716.

IBM Corp. (2015). *IBM SPSS statistics for windows, Version 23.0*. Armonk: IBM Corp.

IEA. (2017). *IDB analyzer*. Amsterdam and Hamburg: International Association for the Evaluation of Educational Achievement. Retrieved from http://www.iea.nl/data.html.

Jāhāna, S. (2016). *Human development report 2016: Human development for everyone*. New York: United Nations Publications.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*(4), 409–426.

Kim, E. S., Cao, C., Wang, Y., & Nguyen, D. T. (2017). Measurement invariance testing with many groups: A comparison of five approaches. *Structural Equation Modeling, 24*(4), 524–544.

Losito, B., Agrusti, G., Damiani, V., & Schulz, W. (2018). *Young People's Perceptions of Europe in a time of change: IEA international civic and citizenship education study 2016 European Report*. Berlin: Springer.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*(1), 84.

MacCallum, R. C., Widaman, K. F., Preacher, K. J., & Hong, S. (2001). Sample size in factor analysis: The role of model error. *Multivariate Behavioral Research, 36*(4), 611–637.

Marsh, H. W., Guo, J., Parker, P. D., Nagengast, B., Asparouhov, T., Muthén, B., et al. (2017). What to do when scalar invariance fails: The extended alignment method for multi-group factor analysis comparison of latent means across many groups. *Psychological Methods, 1,* 1. https://doi.org/10.1037/met0000113.

Migration Policy Group. (2015). Migrant integration policy index 2015. Retrieved from http://www.mipex.eu/.

Millsap, R. E. (2011). *Statistical Approaches to Measurement Invariance*. Routledge. Retrieved from https://www.taylorfrancis.com/books/9780203821961.

Miranda, D., & Castillo, J. C. (2018). Measurement model and invariance testing of scales measuring egalitarian values in ICCS 2009. *Teaching Tolerance in a Globalized World, 2018,* 19–31.

Munck, I., Barber, C., & Torney-Purta, J. (2017). Measurement invariance in comparing attitudes toward immigrants among youth across Europe in 1999 and 2009. *Sociological Methods & Research*. https://doi.org/10.1177/0049124117729691.

Muthén, L. K., & Muthén, B. O. (2017). *Mplus User's Guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.

Mutz, D. C. (2001). Tolerance. In N. J. Smelser & B. Baltes (Eds.), *International Encyclopedia of the Social and Behavioral Sciences* (pp. 15766–15771).

OECD & European Union. (2015). *Indicators of immigrant integration 2015 Settling in*. Paris: OECD Publishing Paris.

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review, 41,* 71–90.

Rutkowski, L., & Rutkowski, D. (2017). Improving the comparability and local usefulness of international assessments: A look back and a way forward. *Scandinavian Journal of Educational Research*. https://doi.org/10.1080/00313831.2016.1261044.

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement, 74*(1), 31–57.

Rutkowski, L., & Svetina, D. (2017). Measurement invariance in international surveys: Categorical indicators and fit measure performance. *Applied Measurement in Education, 30*(1), 39–51. https://doi.org/10.1080/08957347.2016.1243540.

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

Schulz, W. (2016). *Reviewing Measurement Invariance of Questionnaire Constructs in Cross - National Research: Examples from ICCS 2016*. Retrieved from https://iccs.acer.org/files/aera16_proceeding_1064466.pdf.

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 38 countries. ICCS Intn Report*. Amsterdam, The Netherlands, The Netherlands: International Association for the Evaluation of Educational Achievement (IEA). Retrieved from http://www.iea.nl/fileadmin/user_upload/Publications/Electronic_versions/ICCS_2009_International_Report.pdf.

Schulz, W., Ainley, J., & Fraillon, J. (2011). *ICCS 2009 technical report*. International Association for the Evaluation of Educational Achievement. Herengracht 487, Amsterdam, 1017 BT, The Netherlands, 2010.

Schulz, W., Ainley, J., Fraillon, J., Losito, B., & Agrusti, G. (2016a). *Assesment framework: IEA international civic and citizenship education study 2016*. Paris: IEA.

Schulz, W., Ainley, J., Fraillon, J., Losito, B., & Agrusti, G. (2016b). *IEA international civic and citizenship education study 2016 assessment framework*. Cham: Springer Open. https://doi.org/10.1007/978-3-319-39357-5.

Schulz, W., Ainley, J., Fraillon, J., Losito, B., Agrusti, G., & Friedman, T. (2018). *Becoming citizens in a changing world: IEA international civic and citizenship education study 2016 international report*. Berlin: Springer.

Sherrod, L. R., & Lauckhardt, J. (2009). The development of citizenship. In L. S. R. M. Lerner (Ed.), *Handbook of adolescent psychology* (pp. 372–407). Hoboken, NJ: Wiley.

Steenkamp, J. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25*(1), 78–107. https://doi.org/10.1086/209528.

Torney-Purta, J., Lehmann, R., Oswald, H., & Schulz, W. (2001). *Citizenship and Education in Twenty-eight Countries*. Delft, Netherlands: International Association for the Evaluation of Educational Achievement (IEA). Retrieved from http://pub.iea.nl/fileadmin/user_upload/Publications/Electronic_versions/CIVED_Phase2_Age_Fourteen.pdf.

UNESCO. (1995). Declaration of principles on tolerance. *Culture of Peace Programme*.

Van Driel, B., Darmody, M., & Kerzil, J. (2016). Education policies and practices to foster tolerance, respect for diversity and civic responsibility in children and young people in the EU. *NESET II Report*, 2012–2015.

Van Zalk, M. H. W., Kerr, M., Van Zalk, N., & Stattin, H. (2013). Xenophobia and tolerance toward immigrants in adolescence: Cross-influence processes within friendships. *Journal of Abnormal Child Psychology, 41*(4), 627–639.

Wang, J., & Wang, X. (2012). *Structural equation modeling: Applications using Mplus*. Chichester: Wiley. https://doi.org/10.1002/9781118356258.

Weldon, S. A. (2006). The institutional context of tolerance for ethnic minorities: A comparative, multilevel analysis of Western Europe. *American Journal of Political Science, 50*(2), 331–349.