Edukacja 2014, 4(129), 5–20
ISSN 0239-6858

School autonomy – a cross-national perspective. Can we compare the opinion of school principals?

Dorota Węziak-Białowolska, Maria Magdalena Isac

European Commission – Joint Research Centre, Econometrics and Applied Statistics Unit*

Perception of school autonomy was measured by the International Civic and Citizenship Education Study (ICCS) 2009, allowing potential cross-national comparison. The possibility of a common, general scale for all countries participating in the study was investigated. Using multi-group confirmatory and exploratory factor analysis, measurement invariance was tested for countries, such that meaningful comparisons for the concept could be made. The results show that the concept is not necessarily comparable between all countries involved in the ICCS but that secondary data analysis is generally feasible depending on the research questions posed and the methodology applied. The scientific and practical implications of this reach are discussed.

Keywords: school autonomy, measurement invariance, factor analysis, ICCS.

The increasing availability and complexity of the datasets from international large-scale educational assessment studies allow comparative investigation of numerous cross-national research questions. However, such studies are only valuable with awareness of their limitations and the proviso that appropriate methodology can be applied; that is, the establishment of measurement invariance in questionnaire data allows comparison of constructs between countries.

In this study, the equivalence of perception of school autonomy between principals in different countries was investigated. The choice of study concept was justified for several reasons. It has recently received much attention in international comparative studies and the issue is of particular relevance to contemporary education policy and practice. As described later, this concept is also relatively new in terms of its theoretical grounding, maturity of the research field and susceptibility to contextual influences, factors that may hinder comparability at country level.

The school autonomy concept relates to the degree of decision-making authority that school management (principals, teachers and possibly the school council) exerts over school operations, including the hiring and firing of personnel, decisions about the curriculum, assessment of teachers and teaching (Arcia, Macdonald, Patrinos and Porta, 2011; Barrera, Tazeen and Patrinos, 2009; Di Gropello, 2004; 2006). It reflects the relative independence of an institution in its operation and is a measure related to decentralisation of decision-making

* Address: Via E. Fermi 2749, TP 361, Ispra (VA), I-21027, Italy. E-mail: dorota.bialowolska@jrc.ec.europa.eu

to schools, which is official policy in many countries. It is, therefore, natural and commonplace to incorporate the concept of school autonomy into comparison between countries. As over the past two decades many countries have decentralised decision--making to schools (Maslowski, Scheerens and Luyten, 2007), monitoring progress in these countries in this respect is of particular interest (e.g., Arcia et al., 2011; European Commission, 2007; OECD, 2012). In this regard, benchmarking countries according to level of school autonomy implies comparison of country averages for such measures. Further, to assess the impact of policy initiatives focussing on school autonomy, the research tradition in the field has been to link them with measures of student achievement and other education policy indicators such as accountability. Particular advances have been made in this field by secondary analysis of data derived from international comparative studies such as OECD-PISA and IEA-TIMMS (Fuchs and Woessmann, 2007; Hanushek, Link and Woessmann, 2011; Maslowski et al., 2007). Because of its overall relevance to student learning, school autonomy has also been included in international large-scale assessments such as the International Civic and Citizenship Education Study (ICCS) 2009.

However, researchers note that school autonomy is a rather complex concept to measure, as it is contingent on national legal frameworks and their implementation. In addition, not only are the available quantitative measures, largely provided by the Organisation for Economic Cooperation and Development (OECD), "rather rudimentary measures of autonomy in the various domains" (Maslowski et al., 2007). They also offer dichotomous measures of different types of decision-making (e.g., either having or not having some basic discretion on curriculum planning or financial resource allocation) and as the literature suggests, understanding of the concept may vary according to country, which may cloud the justification for comparison of country averages.

Therefore, the purpose of this paper is to assess cross-country equivalence of the school autonomy scale (*scauton*) as operationalized in the ICCS 2009 study with respect to mean level of school autonomy. Guided by the analyses of Schulz and Friedman (2011), the *scauton* scale is initially considered as one-dimensional and comparability between countries is explored for the latent country mean of the *scauton* scores. As results related to measurement invariance testing raise the issue of multidimensionality, this issue is also addressed by investigation of dimensionality for the *scauton* scale.

In the following sections, we present the conceptualisation of school autonomy and its operationalization, leading to an account of the available data, methodology for the measurement invariance assessment and analysis strategy. We then report our results, formulate the main conclusions and finally discuss the scientific and practical implications.

School autonomy in the ICCS

The International Civic and Citizenship Education Study

The International Civic and Citizenship Education Study (ICCS) is currently the most comprehensive international source for information on civic and citizenship education and student civic outcomes. Its aim was to investigate "the ways in which young people are prepared to undertake their roles as citizens in democracy" in a range of countries (Schulz, Ainley, Fraillon, Kerr and Losito, 2010). It built on experience from two previous international civic education studies conducted by the International Association for the Evaluation of Educational Achievement (IEA) in 1971 (Torney, Oppenheim and Farnen, 1975) and in 1999 (Torney-Purta, Lehmann, Oswald and Schulz, 2001). Guided by broad theoretical models

(Schulz, Ainley, Fraillon, Kerr and Losito, 2010; Torney-Purta et al., 2001), the ICCS conceptual framework covered a wide range of concepts (Schulz, Ainley, Fraillon, Kerr and Losito, 2008). Not only did they comprise student civic competences (outputs), contextual factors characterising individual students and their learning experiences in the family, school and wider community but also information about the classroom climate, school resources and school governance.

From the outset, the ICCS team paid scrupulous attention to high methodological quality (e.g., employing a rigorous sampling strategy, consulting international expert groups for instrument development, conducting pilot and field trial studies) (Schulz et al., 2010; Schulz, 2009). Nevertheless, although the comparative validity of some constructs was assessed during the field trial stage of the ICCS (Schulz and Friedman, 2011), to the knowledge of the authors, the available ICCS documentation does not provide sufficient information on comparability of national averages for the measurement of school autonomy. The implication is that the school autonomy scale requires validation.

Conceptualisation of school autonomy

The conceptualisation of school autonomy is not necessarily straightforward, as the means by which autonomy is devolved to schools and the objectives of this policy vary greatly depending on the national legal framework and context in which the policy is implemented, making country comparisons difficult (Di Gropello, 2006). Nevertheless, working with data mainly provided by the Programme for International Student Assessment (PISA) of the OECD, which includes a set of items on school autonomy, a few explorative taxonomies have been proposed for identification and segmentation of tasks that could be devolved to management at a school level. For instance, Winkler and Gershberg (2000), based on the PISA 2000 and a review of educational decentralisation with a particular focus on the Latin American countries, identified four main aspects of school autonomy: organisation of instruction, personnel management, planning and structures and resources. Adopting a quantitative perspective, Maslowski et al. (2007) investigated the international PISA 2000 data and using principal component analysis, arrived at a similar segmentation for four domains of school autonomy: the curriculum, personnel management, student policies and financial resources. In the 2009 edition of PISA, the OECD (2010) proposed a more aggregated measure of autonomy, identified by only two dimensions: school resources and school organisation and assessment.

Table 1 summarises the types of decisions considered in each study, their corresponding association into domains and the dimensions considered in the ICCS 2009 study.

Operationalization of the school autonomy concept in the ICCS study

An instrument to measure school autonomy was included in the ICCS 2009 study. It was developed in connection with other measures of the educational proceses such as school climate and teacher, parent or student participation with the aim of characterising the context for implementation of civic and citizenship education at school. The topic of particular interest was school autonomy in terms of curriculum development and delivery, since schools with a potentially high level of autonomy in this domain can exercise wider discretion regarding the implementation of civic and citizenship education (European Commission, 2007). Nevertheless, the instrument measuring school autonomy from the perspectives of broader school effectiveness and school improvement (see Reezigt and Creemers, 2005 in Schulz et al., 2010), includes 12 items which describe most types of decision, as summarised in Table 1. Table 2 presents the 12 items used by the ICCS.

Table 1
Domains of the school autonomy concept

Type of decision	Present in the ICCS	Winkler and Gershberg (2000) – Theoretical perspective based on the PISA 2000 grouping	Maslowski, Scheerens and Luyten (2007) – Theoretical and data-driven perspective based on the PISA 2000	PISA 2009
Allocating personnel budget, allocating non-personnel budget	yes	Resources	Financial resources	Allocation of resources
Allocating resources for training	yes	Resources	Financial resources	Allocation of resources
Appointing teachers	yes	Personnel management	Personnel management	Allocation of resources
Dismissing teachers	yes	Personnel management	Personnel management	Allocation of resources
Assigning teacher responsibilities		Personnel management		
Choosing textbooks	yes	Organisation of instruction	Curriculum	Instruction and assessment
Establishing or closing a school		Planning and structures		
Determining course content	yes	Organisation of instruction	Curriculum	Instruction and assessment
Determining training provision		Personnel management		
Determining teaching methods		Organisation of instruction		
Determining the content of in-service training for teachers	yes			
Developing a school improvement plan		Resources		
Establishing student assessment policies	yes	Planning and structures	Student policies	Instruction and assessment
Establishing student disciplinary policies			Student policies	Instruction and assessment
Establishing teacher starting salaries and salary rises	yes	Personnel management	Personnel management	Allocation of resources
Formulating the school budget	yes		Financial resources	Allocation of resources
Selection of programmes offered	yes	Planning and structures	Curriculum	Instruction and assessment
Determining instruction time		Organisation of instruction		
Student admission policies	yes	Organisation of instruction	Student policies	Instruction and assessment
Teacher appraisal	yes			

The items describing school autonomy were included in the ICCS school questionnaire (see Schulz, Ainley and Fraillon, 2011) to which school principals were asked to respond. It should be noted that unlike other measures for school autonomy (e.g., developed by the OECD) this measure aims to capture more information. In the PISA studies, school principals were asked to specify whether decisions in several areas were a school's responsibility and to identify which actors in schools had main responsibility in these areas, resulting in several dichotomous items. In the ICCS, each item is measured on a 4-point Likert scale ranging from "full" autonomy to "none". Moreover, using these 12 items, the ICCS experts constructed a scale of "principals' perceptions of school autonomy" (*scauton*) (see Schulz and Friedman, 2011). This scale is available in the ICCS data set. It is important to note that the *scauton* scale was conceptualised and operationalized as one-dimensional. The measurement quality of the scale was estimated by means of confirmatory factor analysis (with only one latent factor reflected by all 12 items) on the pooled dataset (Schulz and Friedman, 2011). This implied that the 12 items which measured the school autonomy phenomenon could be grouped together into one single scale, i.e., without any subscales. Yet, proof for this assumption for each country was not provided and subsequent analysis in this study examines the issue.

Data

All data selected for the analysis were drawn from the 2009 IEA-ICCS study, which was conducted in 38 European, Asian, South and Central American and Oceanic countries.

In the ICCS study, data measuring *school autonomy/scauton* came from school principals. The sample size varied from nine schools in Liechtenstein to 214 in Mexico.

Preliminary analysis based on the descriptive statistics led to the exclusion of countries with small sample sizes. These were Lichtenstein, Luxembourg, Netherlands, Cyprus, Malta and Hong Kong, for which the sample size was below 100.

Testing for measurement invariance

In order to compare country scores on a scale, it is necessary to establish the cross-country comparability of the scale. Scale comparability between countries, also defined as measurement invariance (MI), infers

Table 2
School autonomy (SCAUTON) — item codes and wording

How much autonomy does this school have in relation to the following issues?	IC2G04A Curriculum planning IC2G04B Curriculum delivery IC2G04C Choice and use of textbooks IC2G04D Appointing teachers IC2G04E Dismissing teachers IC2G04F Establishing student assessment policies IC2G04G Determining the content of in-service professional development programmes for teachers IC2G04H Teacher appraisal IC2G04I Budget allocations within the school IC2G04J Extracurricular activities IC2G04K Student admittance policies IC2G04L Establishing teachers' salaries	1. Full autonomy 2. Quite a lot of autonomy 3. Little autonomy 4. No autonomy

that scale scores from different countries measure the same construct with the same measurement unit (Byrne, Shavelson and Muthen, 1989; Meredith, 1993). Meredith (1993) distinguished four levels of measurement invariance, which may be tested under factor analytical framework from both confirmatory (Davidov, Meuleman, Cieciuch, Schmidt and Billiet, 2014) and exploratory (Marsh, Nagengast and Morin, 2012) perspectives:

- configural invariance – the same factor model is specified across compared groups (Davidov et al., 2014; Horn and McArdle, 1992);
- weak invariance (called also metric factorial invariance (Davidov, 2008; De Jong, Steenkamp and Fox, 2007; Meredith and Teresi, 2006) – requires invariant factor loadings across groups;
- strong invariance (also called strong factorial invariance (Meredith and Teresi, 2006) or scalar invariance (Davidov, 2008; De Jong et al., 2007) – requires factor intercepts to be identical for all groups;
- strict invariance – in addition to equal factor loadings and factor intercepts, requires the manifest variable residuals to be equal for all groups[1].

Looking at its typology, the process to establish measurement invariance is clearly hierarchical. This usually starts by separately establishing a well-fitting baseline model for each group (in this study – country) and then proceeds to test subsequent types of invariance (Byrne, 2008; Cieciuch Davidov, Vecchione, Beierlein and Schwartz, 2014). However, an approach based on first establishing the strong MI and then testing less restricted models, although less frequently applied, is also accepted.

Establishing configural invariance ensures that common factors are associated with the same items for all groups but it is not sufficient for meaningful statistical comparisons. Establishing weak measurement invariance endorses that the common factors have the same meanings for groups and the same measurement unit. Therefore, comparison of the relationships between factor scores/scale scores and other observable variables between groups is validated. For example, only after establishing the existence of a weak MI can the statement "High level of school autonomy is positively linked with students' achievement, but this relationship is stronger in Scandinavian countries than in Southern European countries." be supported. Establishing strong measurement invariance permits meaningful comparison of the latent factor group means, as the factors have both the same measurement unit and the same reference point. This implies that, in addition to the same one-unit difference in the question and factor scores for all analysed groups, the same responses (e.g., "I agree", "full autonomy") to a given question by respondents from different groups are calibrated to the same factor scores. Therefore, to conduct valid inter-group comparisons of scale scores (e.g., country rankings based on mean school autonomy scale scores), a strong measurement invariance is required. For example, only after establishing a strong MI can the statement, "The level of school autonomy in the European countries is higher than in the Southern European countries", be supported. Strict MI means that, in addition to what is stated above, the reliabilities of the scales that are indirectly reflected by error variances are comparable between groups. However, it should be noted that there is no consensus on whether a strict MI is necessary to perform valid inter-group comparison of the latent mean scale scores. Lubke and Dolon (2003), Meredith (1993) and Wu (2007) state a strict MI requirement, whereas Byrne and van de Vijver (2010), Davidov, Meuleman, Billiet and Schmidt (2008) and Davidov

[1] In this case, groups should be understood as countries.

(2008) discuss that meaningful information can only be obtained by assuming a strong MI. Byrne and van de Vijver (2010) also claim that there is widespread consensus that testing for strict equivalence in international settings is not only of the least importance but is also somewhat unreasonable and, citing Selig, Card and Little (2008), not recommended. Therefore, taking into consideration the aim of the paper (i.e., determination of whether it is valid to compare mean levels of school autonomy between countries using the *scauton* scale) in this study it was decided to follow the guidelines suggested by Byrne and van de Vijver (2010), Davidov et al. (2008) and Davidov (2008); concentrating on the strong MI.

All types of invariance can be verified either fully or partially (Baumgartner and Steenkamp, 1992; Byrne et al., 1989; Byrne, 2008; De Jong et al., 2007; Gregorich, 2006; Millsap and Kwok, 2004). In the full version of measurement invariance, equality constraints apply to all manifest variables, whereas in the partial version some can be relaxed. This means that only the subset of items meeting the weak, strong, or strict factorial invariance criteria are used to estimate group differences, which, under the conditions for partial invariance, provide substantive and defensible information. Partial measurement invariance is commonly used particularly in the area of cross-cultural research (Byrne and van de Vijver, 2010; Rutkowski and Svetina, 2014). It must be noted, however, that in large-scale cross-cultural studies, due to the many countries subject to assessment, not only do measurement scales often not demonstrate adequate measurement equivalence properties but also determination of which model parameters to relax is too cumbersome. This is owing to the many possible violations of invariance and many possible modifications (Byrne and van de Vijver, 2010; Rutkowski and Svetina, 2014).

When measurement invariance is not satisfied, subgroups of countries have to be found that are measurement-invariant (Welkenhuysen-Gybels, Billiet and Cambré, 2003). This approach is particularly valid in cultural equivalence studies, where (a) the construct of interest may be structurally and psychometrically inappropriate, or (b) the clusters of countries may exhibit both intra-cluster homogeneity and inter-cluster heterogeneity (Byrne and van de Vijver, 2010). However, when the steps to establish a configural model or configural invariance tests are not satisfied, multi-group exploratory structural equation modeling (MG-ESEM) with multi-group exploratory factor analysis (MG-EFA)[2] included, may be applied. This is an approach that integrates exploratory factor analysis with confirmatory factor analysis (Asparouhov and Muthén, 2009; Marsh et al., 2009; Marsh et al., 2010). It allows the CFA assumption for independent cluster models, in which it is permissible for each item to load on only one single factor, to be relaxed. This differs from typical CFA in that all factor loadings are estimated, subject to the constraints necessary for identification (Marsh et al., 2012).

Methodology

In testing the one-dimensional *scauton* scale for strong measurement invariance properties, the intention was to use a multi-group factor analytical framework. It should be noted that the aim here was to find configural invariance first (see Byrne and van de Vijver, 2010; Davidov et al., 2014), because to continue to check higher levels of equivalence this had to be established.

[2] Here the expression the MG-EFA is used, which is more often referred to as the MG-ESEM (Asparouhov and Muthén, 2009), following Asparauhov and Muthen (2014) and Muthen and Muthen (2012). It was decided to distinguish MG-EFA from the more general set of methods, i.e., MG-ESEM, to clarify the procedure adopted.

All EFA, CFA, MG-EFA and MG-CFA analyses were run using Mplus 6.1 and descriptive statistics were obtained using IBM SPSS Statistics, Version 20. Since item values were classed as categorical data, the robust weighted least squares estimator was used for the estimation procedure. This is the default estimator for analysis of categorical indicators in Mplus (Muthén and Muthén, 2012). Sampling weights were included in the analyses.

Among the broad range of goodness-of-fit indices reported were the root-mean-square error of approximation (RMSEA), the 90% confidence interval for the RMSEA, the Tucker-Lewis index (TLI) and the comparative-fit index (CFI), as implemented in Mplus. For RMSEA, values below 0.08 indicate that model performance is satisfactory (Browne and Cudeck, 1992) and very good below 0.05 (Hu and Bentler, 1999). It is also desirable for the upper boundary of the 90% confidence interval to be below 0.08 (Hu and Bentler, 1999). As for the CFI and TLI, the model is satisfactory if these figures are over 0.95. Values over 0.90 are also considered acceptable (Hu and Bentler, 1999; Marsh et al., 2012; Marsh, 2004). However, along with others (Hu and Bentler, 1999; Kline, 2011; Marsh et al., 2004; 2012), cut-off values are only treated as rough guidelines. This applies especially to ESEM and MG-EFA, for which a suitable evaluation of this type is not available.

In addition, following the suggestions of Chen (2007) and Nagengast and Marsh (2014), the change in the CFI and the RMSEA should be analysed. The lack of weak MI or strong MI was indicated by ΔCFI > 0.01, ΔTLI > 0.01, ΔRMSEA > 0.015 with priority given to the CFI.

Although the school autonomy concept is not new, the 12-item scale used in the ICCS to measure it is recent in large-scale educational assessments. The conceptual model for this concept is decidedly one-dimensional (see Schulz and Friedman, 2011). Therefore, the analysis began with confirmation of configural invariance for a one-factor model, then proceeding to higher levels of measurement invariance. However, analysis of the frequency distribution for each item per country showed that in 31 out of 38 countries at least one item occurred without incidence of all response categories represented. This had serious implications for verification of the measurement invariance in these countries using MG-CFA. This analysis, when performed on categorical data, requires incidence of all response categories for each country and for each item. Without this condition, satisfied estimation of all model parameters was impossible, item thresholds[3] here in particular. Accordingly, two different approaches were adopted to achieve this. In the first approach, A, data dimensionality was manipulated. In the second, B, measurement scales of items were manipulated.

In approach A, the analysis was carried out on the original items measured on a 4-point scale, in reverse order to ensure orientation — the higher, the better. This meant however, that due to the limitation described above, only seven of 38 countries participating in the ICCS could be investigated in this way: Chile, the Dominican Republic, Guatemala, Indonesia, Korea, Mexico and Sweden. In spite of diverse geography and legal systems, they still formed a valid subset for which the measurement invariance properties of the school autonomy scale could be tested.

In approach B, different recoding procedures were attempted, reconsidering the full set of the ICCS countries:

- a dichotomous scale consisting of "no autonomy" and "a little autonomy" + "quite a lot of autonomy" + "full autonomy", resulting in 32 countries to be analysed (approach B1);

[3] Threshold is understood as the transition point between adjacent answer scale categories.

- a dichotomous scale consisting of "no autonomy" + "a little autonomy" and "quite a lot of autonomy" + "full autonomy", resulting in 23 countries to be analysed (approach B2);
- a dichotomous scale consisting of "no autonomy" + "a little autonomy" + "quite a lot of autonomy" and "full autonomy", resulting in 9 countries to be analysed (approach B3);
- a 3-point scale consisting of "no autonomy" + "a little autonomy" and "quite a lot of autonomy" and "full autonomy", resulting in 9 countries to be analysed (approach B4).

Where no configural invariance was found, two different strategies were adopted depending on the approach:

- In approach A an exploratory factor analysis was separately carried out for each country in order to identify the configural model. In the absence of consistency with either the number of factors or the pattern of factor loadings, MG-EFA was run in a following step. The purpose was to test the multi-dimensionality of the *scauton* scale because, based on the conceptualisation of the school autonomy concept presented in the previous section, the *scauton* scale may be multi-dimensional, which may explain the poor fit of the one-dimensional model run using multi-group framework.
- In approach B the analysis was completed. The dimensionality of the recoded data remained intentionally unexplored. This would have constituted a further violation (apart from the changes in the measurement scale of the items) in the operationalization of the model of school autonomy (see Schulz and Friedman, 2011).

Results

Approach A

The one-factor MG-CFA model for school autonomy fitted poorly under conditions of configural invariance (see first row of Table 3). Therefore, in a next step, using the EFA, the structure of the data was explored for each country separately. One-, two- and three--factor solutions were analysed. In no countries did the one-factor model fit the data (see analysis by country in Table 3). Moreover, the two-factor solution only fitted the data satisfactorily for Mexico. For the remaining six countries, the solution was three-dimensional but comparison of the pattern of factor loadings did not reveal any similarities and led to the conclusion that the groupings of items were not similar in different countries. Therefore, it was concluded that school autonomy did not have the same conceptual meaning in each country and so, under the EFA framework, it would be difficult to establish a valid configural model for all eight countries.

To identify a structure for the concept in order to allow comparison between countries, MG-EFA was used. As the aim was to establish methodological justification for comparison of countries with respect to the level of school autonomy using the *scauton* scale, strong measurement invariance conditions were the focus. Nevertheless, metric and configural measurement invariance were also tested by applying one-, two- and three-dimensional solutions.

Although the fit of the three-factor model estimated under the full measurement invariance conditions was not entirely satisfactory by normal standards (see the RMSEA and 90% CI in Table 3), for reasons of parsimony, it was preferable to a four-factor model. Moreover, following inspection of the loading pattern in the scalar measurement model, it was decided to exclude items C and J, because they loaded the "Personnel management" dimension more, rather than "Curriculum", as had been anticipated. However, this decision led to a slight decrease in the RMSEA and an improvement in the CFI and TLI. As the fit indices for the ESEM models (with MG-EFA included) are still not well investigated (Marsh 2009; Marsh et al. 2010; 2012), it was assumed that the model adequately fitted the data. The results are presented in Table 3.

Then, the three-factor model estimated under conditions of metric measurement invariance was also not fully satisfactory due to the slightly excessive RMSEA value.

Table 3
Approach A: fit statistics

Countries	RMSEA	90% CI	CFI	TLI
One-factor model with configural invariance – MG-CFA				
8 countries	0.175	0.168 – 0.181	0.897	0.896
Analysis by country:				
One-factor model – EFA				
Chile	0.181	0.163–0.198	0.966	0.958
Dominican Republic	0.171	0.097–0.138	0.968	0.961
Guatemala	0.150	0.130–0.170	0.850	0.816
Indonesia	0.147	0.127–0.167	0.769	0.718
Korea	0.108	0.088–0.129	0.946	0.934
Mexico	0.081	0.063–0.099	0.978	0.973
Sweden	0.152	0.133–0.171	0.901	0.879
Two-factor model – EFA				
Chile	0.073	0.050–0.097	0.995	0.993
Dominican Republic	0.079	0.053–0.105	0.988	0.982
Guatemala	0.094	0.069–0.118	0.953	0.928
Indonesia	0.092	0.066–0.117	0.928	0.890
Korea	0.088	0.063–0.112	0.972	0.957
Mexico	0.048	0.020–0.071	0.994	0.991
Sweden	0.107	0.084–0.130	0.961	0.940
Three-factor model – EFA				
Chile	0.062	0.031–0.089	0.998	0.995
Dominican Republic	0.062	0.024–0.094	0.995	0.989
Guatemala	0.074	0.042–0.104	0.978	0.955
Indonesia	0.042	0.000–0.079	0.988	0.977
Korea	0.059	0.019–0.090	0.990	0.981
Mexico	0.026	0.000–0.059	0.999	0.997
Sweden	0.066	0.032–0.096	0.989	0.977
Two-factor model with full strong measurement invariance – MG-EFA				
8 countries	0.095	0.089–0.102	0.963	0.969
Three-factor model with full strong measurement invariance – MG-EFA				
8 countries	0.078	0.071–0.085	0.976	0.979
8 countries (without C and J items)	0.090	0.082–0.099	0.979	0.982
Three-factor model with full metric measurement invariance – MG-EFA				
8 countries (without C and J items)	0.096	0.086–0.105	0.985	0.980
Three-factor model with full configural measurement invariance – MG-EFA				
8 countries (without C and J items)	-	-	-	-

Nevertheless, the observed difference in the CFA was 0.006, below the recommended 0.01. Differences in the RMSEA and the TLI were negative, which implied improvement as opposed to the expected deterioration in model fit. Unfortunately, the configural model failed to converge, which may imply a problem with either the specification or the structure of the data. (Rutkowski and Rutkowski, 2013).

Although the classification of items between factors is not disjunctive, it creates an idea of item grouping. The first factor, labelled "Allocation of resources", is loaded by all items relating to decisions on teaching staff (D, E, G, H and L) and budget allocation within the school (I). The second factor, "Curriculum", is loaded heavily by items concerning curriculum planning and delivery (A, B). The third factor, "Student assessment policies", is loaded by two items for decisions on student assessment (F and K).

These results implied that for comparison of the mean level of school autonomy using the *scauton* scale from the ICCS for the eight countries investigated, the recommended method (of those tested in this study) was MG-EFA. Mean values for each of all the three dimensions could then be estimated using MG-EFA. Only then could the investigation accommodate (a) the three-dimensionality of the school autonomy concept, (b) cross-loading between items and (c) correlation between dimensions of the concept. Countries' average values for all school autonomy dimensions taken from the MG-EFA could then be applied, for example, in two-level regression analysis at country level.

Table 4
Parameter estimates (non-standardised) for the MG-EFA three-factor model with full strong measurement invariance for seven countries (Chile, the Dominican Republic, Guatemala, Indonesia, Korea, Mexico and Sweden)

Dimension	Item	Factor loadings		
		F1	F2	F3
Curriculum	A. Curriculum planning	-0.105	2.722*	-0.008
	B. Curriculum delivery	0.012	2.799*	0.038
	C. Choice and use of textbooks	-	-	-
	J. Extracurricular activities	-	-	-
Allocation of resources	D. Appointing teachers	6.349*	0.118	-0.014
	E. Dismissing teachers	5.246*	-0.172	0.123
	G. Determining the content of in-service professional development programmes for teachers	1.043*	0.379	0.186
	H. Teacher appraisal	0.895*	0.252	0.331
	L. Establishing teachers' salaries.	7.065*	0.001	-0.143
	I. Budget allocations within the school	1.261*	0.166	0.030
Student assessment policies	F. Establishing student assessment policies	0.264	0.316	0.548*
	K. Student admittance policies	0.006	0.007	1.388*

* The highest factor loading per variable.

Approach B

The results shown in Table 5 indicated that, regardless of the recoding scheme (B1, B2, B3 or B4), the one-factor model was not characterised by configural measurement invariance. For all scenarios, the fit statistics were outside the acceptable range. This meant that although different recoding strategies could conceivably be used to represent the structure of the data better, this would not ensure equivalence for the meaning of the one-dimensional school autonomy construct between countries.

Discussion

This article was a response to the call from experts involved with the ICCS study for research on the assessment of measurement invariance of the ICCS data (see Schulz, 2009). More specifically, it investigated whether valid comparison of country specific mean levels of school autonomy, as often used for comparative secondary data analysis, was possible.

Confirming the assumption, results described a complex picture indicating that the concept was not necessarily comparable between all countries involved in the ICCS and that its potential use for secondary data analysis depends on the research questions posed and the method applied.

The first obstacle encountered owed to the revelation by analysis of the distribution of responses that for at least one item on the school autonomy scale, not all response categories were used in 31 out of 38 countries. From a technical point of view, this seriously affected verification of measurement invariance in these countries using the MG-CFA framework but an attempt to tackle the problem was made using alternative approaches. More importantly, there was the serious implication that either anchoring of the answer scale was not well matched to the phenomenon measured or that some dimensions of school autonomy in some countries were entirely governed by law. This latter case would further imply that opinions were of little significance, since decisions left for school principals would then have little place there.

Nevertheless, considering all reservations relating to the limited number of countries in the study and the problematic distribution of responses to school autonomy related questions, the equivalence of the school autonomy concept was investigated using the multi-group CFA model on countries meeting the requirements demanded by the method. The results did not support the most basic type of configural invariance, showing that the concept of school autonomy did not have a uniform accepted understanding in the investigated countries. Moreover, investigations using country-specific exploratory factor analysis demonstrated that a common specification for a model to apply to each country was difficult to apply. Furthermore, exploration of the data under the MG-EFA framework revealed that the concept appeared to be three-dimensional

Table 5
Approach B: fit statistics

Approach and number of countries	RMSEA	90%CI	CFI	TLI
One-factor model with configural invariance – MG-CFA				
B1 (32 countries)	0.106	0.102–0.110	0.886	0.861
B2 (23 countries)	0.098	0.093–0.102	0.895	0.871
B3 (9 countries)	0.130	0.123–0.137	0.838	0.802
B4 (9 countries)	0.131	0.124–0.137	0.934	0.919

for the seven countries examined (Curriculum, Allocation of Resources and Student Assessment Policies), although loading of items tended to be onto more than one dimension and these dimensions were correlated. This was the only model to demonstrate strong measurement invariance.

Based on these results it follows that associations with other variables can be tested (e.g., relationship with student achievement) and average values of the three dimensions of the school autonomy concept compared between seven countries. Additionally however, the results indicated that even a strategy, which involved differently recoding data as applied to the full set of the ICCS countries, was insufficient to guarantee the observability of measurement invariance.

There are many potential explanations for such results. The most often quoted are either of a technical nature (e.g., differential interpretation of scale anchors, differential response style, differential familiarity with item scale format and translation errors) or related to cultural and institutional bias (e.g., differential extent to which respondents from a particular country have inculcated its social values and norms; Byrne and van de Vijver, 2010; Rutkowski and Svetina, 2014). In this respect, problematic country-specific distributions for responses related to scale anchoring, as mentioned in section 4, may be explained by the former group of reasons, whereas inter-country comparability of the dimensionality of the school autonomy concept – by the latter.

Conclusion

To conclude, although the study provides extensive information about using MI for the assessment of the concept of school autonomy, it is not without its limitations. This should point to new directions for future research. The biggest limitation encountered here was due to the restrictions imposed by the MG-CFA for categorical data (i.e., all response possibilities for each item needed to be recorded in the data from each country), our analysis yielded only limited coverage of countries (seven, rather different countries). As illustrated, we approached this problem by repeating MI testing on recoded data for all countries, but without obtaining better results for the one-dimensional school autonomy scale. Another further limitation owed to the failure of the configural invariance model to achieve convergence when run using the exploratory framework.

Considering our results and the limitations encountered, further research should investigate the problems with the aim of both developing better instruments and the testing of new methods, such as, MG-CFA with alignment, as implemented recently in the Mplus software[4] for the same type of data (Asparouhov and Muthén, 2014; Van de Schoot, Kluytmans, Tummers, Lugtig, Hox and Muthén, 2013; Węziak-Białowolska, 2014). Nevertheless, based on the data used here, future single-country analyses complemented by qualitative research could expose reasons for non-invariance and item cross-loading.

Literature

Arcia, G., Macdonald, K., Patrinos, H. and Porta, E. (2011). *School autonomy and accountability. System assesment and benchmarking for education results (SABER)*. Washington, D. C.: The World Bank.

Asparouhov, T. and Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(3), 397–438. doi:10.1080/10705510903008204

Asparouhov, T. and Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation*

[4] It should be noted that this method, despite being promising for approximation of measurement invariance, at present has considerable limitations. Namely, the means by which it is implemented does not allow introduction of either weights or imputations into computation. It is applicable only, therefore, to either continuous or dichotomous data.

Modeling: A Multidisciplinary Journal, *21*(4), 495–508. doi:10.1080/10705511.2014.919210

Barrera, F., Tazeen, F. and Patrinos, H. (2009). *Decentralized Decision-making in schools. the theory and evidence on school-based management*. Washington, D.C.: The World Bank.

Baumgartner, H. and Steenkamp, J.-B. E. M. (1992). The role of optimum stimulation level in exploratory consumer behavior. *The Journal of Consumer Research*, *19*(3), 434–448.

Browne, M. W. and Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods and Research*, *21*(2), 230–258.

Byrne, B. M. (2008). Testing for multigroup equivalence of a measuring instrument: a walk through the process. *Psicothema*, *20*(4), 872–882. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18940097

Byrne, B. M., Shavelson, R. J. and Muthen, 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. and 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. doi:10.1080/15305051003637306

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, *14*(3), 464–504.

Cieciuch, J., Davidov, E., Vecchione, M., Beierlein, C. and Schwartz, S. H. (2014). The cross-national invariance properties of a new scale to measure 19 basic human values: a test across eight countries. *Journal of Cross-Cultural Psychology*, *45*(5), 764–776. doi:10.1177/0022022114527348

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., Meuleman, B., Billiet, J. and Schmidt, P. (2008). Values and support for immigration: a cross-country comparison. *European Sociological Review*, *24*(1), 583–599. doi:10.1093/esr/jcn020

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P. and Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology*, *40*(1), 55–75. doi:10.1146/annurev-soc-071913-043137

De Jong, M. G., Steenkamp, J.-B. E. M. and Fox, J.-P. (2007). Relaxing measurement invariance in cross--national consumer research using a hierarchical IRT model. *Journal of Consumer Research*, *34*(2), 260–278.

Di Gropello, E. (2004). Education Decentralization and accountability relationships in Latin America. *World Bank Policy Research Working Paper*, *3453*. Washington, D. C.: The World Bank.

Di Gropello, E. (2006). A comparative analysis of school-based management in Central America. *World Bank Working Paper*, *72*. Washington, D. C.: The World Bank.

European Commission (2007). *Eurydice – School autonomy in Europe: Policies and measures*. Brussels: European Commission.

Fuchs, T. and Woessmann, L. (2007). What accounts for international differences in student performance? A re-examination using PISA data. *Empirical Economics*, *32*(2–3), 433–462.

Gregorich, S. E. (2006). Do Self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care*, *44*(11), 78–94.

Hanushek, E. A., Link, S. and Woessmann, L. (2011). Does school autonomy make sense everywhere? Panel estimates from Pisa. *National Bureau of Economic Research Working Paper*, *17591*. Retrieved from http://www.nber.org/papers/w17591

Horn, J. L. and McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research: An International Journal Devoted to the Scientific Study of the Aging Process*, *18*(3), 117–144.

Hu, L. and 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.

Kline, R. B. (2011). *Principles and Practice of Structural Equation Modeling* (3rd ed., pp. 1–427). New York–London: The Guilfor Press.

Lubke, G. H. and Dolon, C. V. (2003). Can unequal residual variances across groups mask differences in residual means in the common factor model? *Structural Equation Modeling: A Multidisciplinary Journal*, *10*(2), 175–192.

Marsh, H. W. (2004). In Search of golden rules: comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, *11*(3), 320–341.

Marsh, H. W. (2009). Exploratory structural equation modeling, Integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling*, *16*(3), 439–476.

Marsh, H. W., Hau, K. and Wen, Z. (2004). In search of golden rules: comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, *11*(3), 320–341.

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S. and Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: application to students' evaluations of university teaching. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(3), 439–476. doi:10.1080/10705510903008220

Marsh, H. W., Nagengast, B. and 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*, *49*(6), 1194–1218. doi:10.1037/a0026913

Marsh, W. H., Ludtke, O., Muthen, B., Asparouhov, T., Morin, A. J. S., W., H. and Trautwein, U. (2010). A new look at the Big-Five factor structure through exploratory structural equation modeling. *Psychological Assessment*, *22*(3), 471–491.

Maslowski, R., Scheerens, J. and Luyten, H. (2007). The effect of school autonomy and school internal decentralization on students' reading literacy. School effectiveness and school improvement. *International Journal of Research, Policy and Practice*, *18*(3), 303–334.

Meredith, W. (1993). MI, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543.

Meredith, W. and Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care*, *44*(11 Suppl 3), 69–77. doi:10.1097/01.mlr.0000245438.73837.89

Millsap, R. E. and Kwok, O.-M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods*, *9*(1), 93–115. doi:10.1037/1082-989X.9.1.93

Muthén, L. K. and Muthén, B. O. (2012). *Mplus user's guide* (7th ed., pp. 1–856). Los Angeles, CA: Muthen and Muthen.

Nagengast, B. and Marsh, H. W. (2014). Motivation and engagement in science aroung the globe: testing measurement invariance with multigroup structural equation models across 57 countries using PISA 2006. In L. Rutkowski, M. von Davier and D. Rutkowski (eds.), *Handbook of International large-scale assessment: background, technical issues and mathods of data analysis* (pp. 317–244). Taylor and Francis Group.

OECD (2010). *PISA 2009 results: what makes a school successful? resources, policies and practices (vol. IV)*. Retrieved from http://dx.doi.org/10.1787/9789264091559-en

OECD. (2012). *Education at a glance*. Paris: OECD Publishing. Retrieved from: http://www.keepeek.com/Digital-Asset-Management/oecd/education/education-at-a-glance-2012_eag-2012-en#page1

Rutkowski, D. and Rutkowski, L. (2013). Measuring socioeconomic background in PISA: one size might not fit all. *Research in Comparative and International Education*, *8*(3), 259–278.

Rutkowski, L. and 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. doi:10.1177/0013164413498257

Schulz, W. (2009). Questionnaire construct validation in the International Civic and Citizenship Education Study. *IERI Monograph Series Issues and Methodologies in Large-Scale Assessments*, *2*, 113–136.

Schulz, W., Ainley, J. and Fraillon, J. (eds.). (2011). *ICCS 2009 Technical Report*. Amsterdam: IEA, ACER, NFER, Universita degli Studi Roma Tre. Retrieved from http://www.iea.nl/fileadmin/user_upload/Publications/Electronic_versions/ICCS_2009_Technical_Report.pdf

Schulz, W., Ainley, J., Fraillon, J., Kerr, D. and Losito, B. (2008). *International Civic and Citizenship Education Study: assessment framework*. Amsterdam: IEA.

Schulz, W., Ainley, J., Fraillon, J., Kerr, D. and Losito, B. (2010). *ICCS 2009 international report: civic knowledge, attitudes and engagement among lower-secondary school students in 38 countries*. Amsterdam: IEA.

Schulz, W. and Friedman, T. (2011). Scaling procedure for ICCS questionnaire items. In W. Schulz, J. Ainley and J. Fraillon (eds.), *ICCS Technical Report* (pp. 157–259). IEA, ACER, nfer, Universita degli Studi Roma Tre. http://www.iea.nl/fileadmin/user_upload/Publications/Electronic_versions/ICCS_2009_Technical_Report.pdf

Selig, J. P., Card, N. A. and Little, T. D. (2008). Latent variable structural equation modeling in cross-cultural research: multigroup and multilevel approaches. In F. J. R. van de Vijver, D. A. van Hemert, and Y. H. Poortinga (eds.), *Multilevel analysis of individuals and cultures* (pp. 93–119). New York, NY: Taylor & Francis Group/Lawrence Erlbaum Associates.

Torney, J. V, Oppenheim, A. N. and Farnen, R. F. (1975). *Civic education in ten countries: an empirical study*. Stockholm: Almqvist and Wiksell.

Torney-Purta, J., Lehmann, R., Oswald, H. and Schulz, W. (2001). *Citizenship and education in twenty-eight countries: civic knowledge and engagement at age fourteen*. Amsterdam: IEA.

Van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J. and 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*(Ml), 770. doi:10.3389/fpsyg.2013.00770

Welkenhuysen-Gybels, J., Billiet, J. and Cambré, B. (2003). Adjustment for acquiescence in the assessment of the construct equivalence of Likert-type score items. *Journal of Cross-Cultural Psychology*, *34*(6), 702–722. doi:10.1177/0022022103257070

Węziak-Białowolska, D. (2014). Differences in gender norms between countries – are they valid? The issue of measurement invariance. *European Journal of Population*. doi:10.1007/s10680-014-9329-6

Winkler, D. and Gershberg, A. (2000). Education decentralization in Latin America: the effects on the quality of schooling. *LCSHD Paper Series*, *59*.

Wu, C. (2007). An empirical study on the transformation of Likert-scale data to numerical scores. *Applied Mathematical Sciences*, *58*(1), 2851–2862.

Appendix

Table 1A
Groups of countries

Group	Countries
European countries (25)	Austria, Belgium (Flemish), Bulgaria, Cyprus, Czech Republic, Denmark, England, Estonia, Finland, Greece, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Norway, Poland, Russian Federation, Slovakia, Slovenia, Spain, Sweden, Switzerland
South and Central American countries (6)	Chile, Colombia, Dominican Republic, Guatemala, Mexico, Paraguay
Asian countries (5)	Chinese Taipei, Hong Kong, Indonesia, the Republic of Korea, Thailand
	New Zealand