United Nations
Educational, Scientific and
Cultural Organization

UNESCO INSTITUTE FOR STATISTICS

TECHNICAL COOPERATION GROUP

GLOBAL ALLIANCE TO MONITOR LEARNING

SUSTAINABLE DEVELOPMENT GOALS

A Measurement Strategy for SDG Thematic Indicators 4.7.4 and 4.7.5 Using International Large Scale Assessments in Education

March 2020

Prepared by

Andrés Sandoval-Hernández
University of Bath

Diego Carrasco
Pontificia Universidad Católica de Chile

Abstract

The aim of this document is to describe and implement a measurement strategy for the SDG Thematic Indicators 4.7.4 and 4.7.5 using International Large-Scale Assessments (ILSAs) in Education. Building on two reports previously published by the Global Alliance to Monitoring Learning (GAML) describing a proposal of a measurement strategy for these two indicators, we use items from Programme for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMSS) and the International Civic and Citizenship Education Survey (ICCS) to fit measurement models, generate scores, and propose a method to establish cut-off points for these indicators.

This document is divided into four main sections. In the first one, we describe the methods and tools we used for constructing both the scores to measure each indicator and the cut-off points to identify the individuals who reach the corresponding targets. The second and the third sections correspond to the implementation of the proposed methodological procedures for each of the thematic indicators covered by this document and for their subscales. As a way of summarizing the full set of scores, the fourth section includes a set of tables showing the average percentage of students who reach cut-off points set for any sub-scale for each indicator.

Acronyms and abbreviations

CIVED	IEA Civic Education
ESD	Education for Sustainable Development
ECV	Explained Common Variance
GAML	Global Alliance to Monitor Learning
GCED	Global Citizenship Education
GRM	Graded Response Model
ICCS	International Civic and Citizenship Education Study
IEA	International Association for the Evaluation of Educational Achievement
ILSA	International Large-Scale Assessments
IRT	Item response theory
OECD	Organisation for Economic Co-operation and Development
SDG	Sustainable Development Goals
TIMSS	Trends in International Mathematics and Science Study
WLSMV	Weighted least square mean and variance adjusted
1PL-GRM	One-parameter graded response model

Table of contents

Abstract ... 2
Acronyms and abbreviations ... 3
Introduction ... 8
A. Methods ... 13
A.1 Availability ... 13
A.2 Unidimensionality ... 13
A.2.1 Unidimensionality and interpretability of scores ... 15
A.3 Measurement model ... 16
A.4 Proficiency levels ... 18
A.4.1 Item-person maps ... 18
A.4.2 Inferences with subsequent applications of the instruments ... 20
A.4.3 Limitations of the provisional cut scores ... 21
B. SDG thematic indicator 4.7.5 ... 23
B.1 The selected items ... 23
B.2 SDG4 indicator 4.7.5 cognitive items ... 24
B.2.1 Availability ... 25
B.2.1 Unidimensionality ... 25
B.2.2 Proficiency classifications ... 26
B.3 SDG4 indicator 4.7.5 Non-cognitive items ... 34
B.3.1 Availability ... 34
B.3.2 Unidimensionality ... 36
B.3.3 Measurement Models ... 36
B.3.4 Proficiency classification ... 37
C. SDG thematic indicator 4.7.4 ... 43
C.1 Selected items ... 43
C.2 SDG indicator 4.7.4 cognitive items ... 45
C.3 SDG indicator 4.7.4 non-cognitive items ... 45
C.3.1 Availability ... 45
C.3.2 Unidimensionality ... 45
C.3.3 Measurement models ... 47
C.3.4 Proficiency classification ... 47
D. An overall indicator of standards met by students ... 77

Bibliography 80
Appendix I. MPLUS syntax for Gender Equality Items 88
Appendix II. Executive summary 90
Analytical strategy 90
Description of cut-off points 91
COGNITIVE (4.7.4) 91
NON-COGNITIVE (4.7.4) 91
COGNITIVE (4.7.5) 92
NON-COGNITIVE (4.7.5) 93
Summary tables 94

List of tables

Table 1. Global Content Framework for SDG indicators 4.7.1, 4.7.4 and 4.7.5 10
Table 2. Core conceptual learning dimensions for indicators 4.7.4 and 4.7.5 11
Table 3. Source of selected items to measure indicator 4.7.5 23
Table 4. Mapping of TIMSS 2015 scales into the indicator categories 24
Table 5. Selected test items to measure indicator 4.7.5 25
Table 6. TIMSS 2015 international benchmark variable codes 27
Table 7. Percentage of students meeting the indicator 4.7.4 based on the IRT scores Benchmark for Science in TIMSS 2015 (i.e. total score in science) 28
Table 8. Percentage of students meeting the indicator 4.7.5 based on the IRT scores of selected items of Science in TIMSS 2015 according to the mapping exercise 32
Table 9. Selected non-cognitive items to measure indicator 4.7.5 34
Table 10. Countries and Regions with available responses on enjoyment in learning and students' confidence in their scientific knowledge 35
Table 11. Alternative survey items to measure indicator 4.7.5 36
Table 12. Mapping of TIMSS 2015 motivation scales into the indicator categories 36
Table 13. Percentage of students meeting the indicator 4.7.5 Environmental Science (socio-emotional) 39
Table 14. Percentage of students meeting the indicator 4.7.5 Environmental Science (behavioural) 42
Table 15. Source of selected items to measure indicator 4.7.4 43
Table 16. Mapping of ICCS 2016 scales into the indicator categories 44

Table 17. Selected non-cognitive items to measure indicator 4.7.4 45
Table 18. Explained Common Variance and accounted Common Variance over the indicator 4.7.4 selected measures 46
Table 19. Percentage of students meeting the indicator 4.7.4 Global-local thinking (socio-emotional) 50
Table 20. Percentage of students meeting the indicator 4.7.4 Multicultural(ism) or intercultural(ism) (socio-emotional) 53
Table 21. Percentage of students meeting the indicator 4.7.4 Gender Equality (socio-emotional) 57
Table 22. Percentage of students meeting the indicator 4.7.4 Peace, Non-violence and Human Security (behavioural) 60
Table 23. Summary of fit indexes of the fitted latent class models 64
Table 24. Percentage of students meeting the indicator 4.7.4 Freedom (of expression, of speech, of press, of association/organisation) 66
Table 25. Percentage of students meeting the indicator 4.7.4 Social Justice (socio-emotional) 71
Table 26. Percentage of students meeting the indicator 4.7.4 Sustainable Development (socio-emotional and behavioural) 76
Table 27. Mean of students meeting any of the standards SDG 4.7.5 (Science scores with selected items, SLS, SCS) TIMSS 2015 78
Table 28. Mean of students meeting any of the standards SDG 4.7.4 ICCS 2016 79
Table 1A. Proportion of students reaching the targets of indicator 4.7.5 94
Table 2A. Table 1A. Proportion of students reaching the targets of indicator 4.7.4 95

List of figures

Figure 1. Gender Equality items in ICCS 2016 16
Figure 2. Latent variable model for Gender Equality items 17
Figure 3. item-person map for Gender Equality (ICCS 2016) 19
Figure 4. Scatter plot between Physics, Biology and Earth Science IRT scores 26
Figure 5. Item-person map for Science Scores (selected items) 30
Figure 6. Scatter between Science IRT scores from TIMSS 2015 33
Figure 7. Students Like Learning Science items in TIMSS 2015 for eighth-grade students 37
Figure 8. Item-person map for Students Like Learning Science 38
Figure 9. Students Confident in Science items in TIMSS 2015 for eighth-grade students 40
Figure 10. Item-person map for Students Confident in Science 41

Figure 11. Students' attitudes toward their country of residence in ICCS 2016 47
Figure 12. Item-person map for Students' attitudes toward their country of residence 49
Figure 13. Students' attitudes toward equal rights for all ethnic/racial groups in ICCS 2016 51
Figure 14. Item-person map for Students' attitudes toward equal rights for all ethnic/racial groups 52
Figure 15. Gender Equality items in ICCS 2016 54
Figure 16. Parallel analysis results over Gender Equality items in ICCS 2016 55
Figure 17. Item-person map for Gender Equality 56
Figure 18. Students' reports on personal experiences of bullying and abuse in ICCS 2016 58
Figure 19. Item-person map for Students' reports on personal experiences of bullying and abuse 59
Figure 20. Students' reports on students' opinions regarding what is good for democracy in ICCS 2016 61
Figure 21. Response patterns for What is good for democracy items from ICCS 2016 65
Figure 22. Students' perception of the importance of social movement related citizenship from ICCS 2016 67
Figure 23. Parallel analysis results over "Students' perception of the importance of social movement related citizenship" items in ICCS 2016 69
Figure 24. Item-person map for Students' perception of the importance of social movement related citizenship 70
Figure 25. Selected items for Sustainable Development from ICCS 2016 72
Figure 26. Parallel analysis results over the proposed Sustainable Development items from ICCS 2016 74
Figure 27. Item-person map for Sustainable Development items 75

Introduction

In September 2015, at the United Nations Sustainable Development Summit, Member States formally adopted the 2030 Agenda for Sustainable Development in New York. The Sustainable Development Goals (SDGs) are a call for action by all countries to promote prosperity while protecting the planet. They recognized that ending poverty must go hand-in-hand with strategies that build economic growth and address a range of social needs including education, health, social protection, and job opportunities while tackling climate change and environmental protection.

The Agenda contains 17 goals including a global education goal (SDG4). SDG4 establishes that by 2030 we have to "ensure inclusive and equitable quality education and promote lifelong learning opportunities for all" and has seven targets and three means of implementation. One of these targets, 4.7, refers to the knowledge and skills that are necessary for a sustainable future. Specifically, it states that by 2030, we have to "[...] ensure that all learners acquire the knowledge and skills needed to promote sustainable development, including, among others, through education for sustainable development and sustainable lifestyles, human rights, gender equality, promotion of a culture of peace and non-violence, global citizenship and appreciation of cultural diversity and of culture's contribution to sustainable development". SDG Indicator 4.7 has, in turn, one global and five specific indicators.

Global indicator

4.7.1 – Extent to which (i) global citizenship education and (ii) education for sustainable development, including gender equality and human rights, are mainstreamed at all levels in: (a) national education policies, (b) curricula, (c) teacher education and (d) student assessment

Thematic indicators

4.7.2 – Percentage of schools that provide life skills-based HIV and sexuality education

4.7.3 – Extent to which the framework on the World Programme on Human Rights Education is implemented nationally (as per the UNGA Resolution 59/113)

4.7.4 – Percentage of students by age group (or education level) showing an adequate understanding of issues relating to global citizenship and sustainability

4.7.5 – Percentage of 15-year-old students showing proficiency in knowledge of environmental science and geoscience

In this document, we focus on the last two thematic indicators (4.7.4 and 4.7.5), which refer to learning outcomes that are achieved as a result of the educational inputs described in the global indicator. The main objective of this document is to describe and implement a measurement strategy for these thematic indicators using data from International Large-Scale Assessments (ILSAs) in education. To do so, we build on two reports previously published by the Global Alliance to Monitor Learning (GAML) describing a proposal of a measurement strategy for these two

indicators[1] (See also Sandoval-Hernández et al., 2019). These two reports establish a global content framework for indicators 4.7.4 and 4.7.5 and carry out a mapping exercise to evaluate the extent to which the different concepts contained in the framework (i.e. categories and sub-categories) can be operationalised with the instruments and procedures of existing ILSAs.

The global content framework (see Table 1) is based on the extensive work already conducted by UNESCO to define and operationalise the Global Citizenship Education (GCED) and Education for Sustainable Development (ESD); it adopted the definitions and operationalization proposed in recent documents (e.g. Hoskins, 2016; IBE, 2016; Sandoval-Hernández & Miranda, 2018; UIS, 2017; UNESCO, 2012b, 2012a, 2013, 2014, 2015).

[1] Proposal for a Measurement Strategy for Thematic Indicator 4.7.4 using ILSAs. Available here: http://gaml.uis.unesco.org/wp-content/uploads/sites/2/2019/08/GAML6-WD-7-Measuring-4.7.4-using-International-Large-Scale-Assessments-in-Education.pdf
Proposal for a Measurement Strategy for Thematic Indicator 4.7.5 using ILSAs. Available here: http://gaml.uis.unesco.org/wp-content/uploads/sites/2/2019/05/GAML6-WD-8-Measuring-4.7.5-using-International-Large-Scale-Assessments-in-Education.pdf

Table 1. Global Content Framework for SDG indicators 4.7.1, 4.7.4 and 4.7.5

	Category	Sub-category
Global Citizenship Education (GCED)	Interconnectedness and Global Citizenship	Globalization
		Global/international citizen(ship), global culture/identity/community
		Global-local thinking, local-global, think global act local, glocal
		Multicultural(ism)/intercultural(ism)
		Migration, immigration, mobility, movement of people
		Global Competition/competitiveness/globally competitive/international competitiveness
		Global Inequalities/disparities
	Gender Equality	Gender equality / equallity / parity
		Empower(ment of) women/girls (female empowerment, encouraging female participation)
	Peace, Non-violence and Human Security	Peace, peace-building
		Awareness of forms of abuse/harassment/violence (school-based violence/bullying, household-based violence, gender-based violence, child abuse/harassment, sexual abuse/harassment)
	Human Rigts	Human rights, rights and responsibilities (children's rights, cultural rights, indigenous rights, women's rights, disability rights)
Education for Sustainable Development (ESD)		Freedom (of expression, of speech, of press, of association/organisation), civil liberties
		Social justice
		Democracy/democratic rule, democratic values/principles
	Health and Well-being	Physical health/activity/fitness
		Mental, emotional health, psychological health
		Healthy lifestyle (nutrition, diet, cleanliness, hygiene, sanitation, *clean water, being/staying healthy)
		Awareness of addictions (smoking, drugs, alcohol)
		Sexual and/or reproductive health
	Sustainable Development	Economic sustainability, sustainable growth, sustainable production/consumption, green economy
		Social sustainability, (social cohesion re sustainability)
		Environmental sustainability/environmentally sustainable
		Climate change (global warming, carbon emissions/footprint)
		Renewable energy, alternative energy (sources) (solar, tidal, wind, wave, geothermal, biomass…)
		Ecology, ecological sustainability (ecosystems, biodiversity, biosphere, ecology, loss of diversity)
		Waste management, recycling
	Environmental Science (geoscience)	Physical systems
		Living systems
		Earth and space systems

Source: Sandoval-Hernández, Isac & Miranda (2019)

Apart from the categories and sub-categories included in the global content framework, the mapping exercise also incorporated the three core dimensions proposed by UNESCO to measure learning outcomes in GCED (UNESCO, 2015): cognitive, socio-emotional and behavioural. These dimensions are interrelated and are presented below (see Table 2), each indicating the domain of learning they focus on for the two SDG indicators covered in this report.

Table 2. Core conceptual learning dimensions for indicators 4.7.4 and 4.7.5

	Target 4.7.4	Target 4.7.5
Cognitive	To acquire knowledge, understanding and critical thinking about global, regional, national and local issues and the interconnectedness and interdependency of different countries and populations.	To acquire knowledge, understanding and critical thinking necessary to encompassing the range of cognitive processes involved in learning environmental science concepts, and then applying these concepts and reasoning with them.
Socio-emotional	To have a sense of belonging to a common humanity, sharing values and responsibilities, empathy, solidarity and respect for differences and diversity.	To have intrinsic motivation to learn environmental science.
Behavioural	To act effectively and responsibly at local, national and global levels for a more peaceful and sustainable world.	To have self-confidence or self-concept in their ability to learn environmental science.

Source: Adapted from Sandoval-Hernández, Isac & Miranda (2019)

This mapping exercise identified International Association for the Evaluation of Educational Achievement (IEA) International Civic and Citizenship Education Study (ICCS) as the most valuable source of information for SGD indicator 4.7.4, and IEA's Trends in International Mathematics and Science Study (TIMSS) as the most informative for indicator 4.7.5, with some aspects covered by the Organisation for Economic Co-operation and Development (OECD) Programme for International Student Assessment (PISA). These studies were chosen due to their specific conceptual frameworks that showed the highest coverage of the topics relevant to these two indicators, as well as their potential to inform long-term monitoring. Two important observations included in these reports are that these ILSAs can provide high (but not total) coverage for indicators 4.7.4 and 4.7.5, but they can only be considered as proxy measures; and that the resulting measures cover only part of the intended population: ICCS and TIMSS are representative for eight-graders only, while PISA only offers representative information for 15-year-olds.

Thus, in this report, we fit a series of measurement models using items from ICCS, TIMSS and PISA to generate scores to measure each thematic indicator, thus a score for the cognitive domain of each thematic indicator, and a series of scores for each of the socio-emotional and behavioural domains of the sub-categories for each indicator.[2] In a second step, we propose a method to establish cut-off points to identify proficiency levels based on each respective score.

[2] Although it would have been more straightforward to produce one single score for each thematic indicator, this was not possible due to the lack of unidimensionality of the constructs. The second preferred option was to produce three scores for each thematic indicator, one for each learning domain: cognitive, socio-emotional and behavioural. This, however, was not possible either because of the same reason (i.e. lack of unidimensionality). So, in this report we produce one cognitive score for each thematic indicator and a series of scores for the socio-emotional and behavioural sub-categories within each thematic indicator.

Apart from this introduction, this document is divided into four sections. In the first one, we describe the methods and tools we used for constructing both the scores to measure each indicator and the cut-off points to identify the individuals who reach the corresponding targets. This includes the establishment of proficiency levels, the measurement models, the item-person maps, the test of unidimensionality, the availability of information and the limitations of the resulting scores. The second and the third sections correspond to the implementation of the proposed methodological procedures for each of the thematic indicators covered by this document and for their subscales. As a way of summary, the fourth section includes a set of tables showing the average percentage of students who reach the cut-off points set for any sub-scale for each indicator.

A. Methods

This section is structured according to the four main steps that we used to construct the scores and proficiency levels (cut-off points) for the SDG thematic indicators 4.7.4 and 4.7.5, These steps are: verifying the availability of observed responses to the items proposed by the mapping exercise described previously (Sandoval-Hernández et al., 2019), testing the unidimensionality of the intended constructs, fitting the corresponding measurement model, and estimating the proficiency levels for each score.

A.1 Availability

The use of item-person maps to establish cut-off scores requires that the depicted parameters come from a known population. For example, we can use data from a single country as a calibration sample. The generated realizations of θ_p would be then centred to this population latent mean. Likewise, the cumulative probabilities express in logit scores γ_{1k}–γ_{6k}, would be representative of this population. If the calibration sample is a representative sample, then we can produce an item-person map to make inferences to the represented population. It should be obvious then, that without observed data from a population, an item-person map cannot be used to make inferences to this population.

In practical terms, if for a certain country we do not have observed responses to the proposed items for each thematic indicator, is not possible to know how many people meet the standard.

A.2 Unidimensionality

Unidimensionality refers to the property of the random term θ_p to capture the common variance among a set of responses by a person *p*, while reaching local independence between the responses among persons. The main assumption of a response model is to treat a set of responses as repeated measures and explain these responses by a common source of variance of each respondent *p*. Thus, in essence, response models can be understood as special cases of analysis of variance, where the term θ_p is used to represent the propensity of people to respond in a certain direction (De Boeck & Wilson, 2004). This propensity is understood as abilities, attitudes, traits or other general constructs, conditional to the content of the items used to elicit the observed responses. Unidimensionality is a requirement, so a single propensity component is used to represent the pattern of responses to a set of items. If more than a single random term θ_p is included in the model, that is, when a multidimensional model is required, then the interpretation of the generated scores of this latter model have a different meaning than that of a unidimensional model (DeMars, 2013; Koch et al., 2018).

In this document, we used bifactor models to assess unidimensionality. More specifically, we used a Graded Response Model (GRM) (Samejima, 2016) with a probit link and the weighted least square mean and variance adjusted (WLSMV) estimator (Luo, 2018), to model responses of ordinal items. This option is computationally faster, and present negligible differences with full information

maximum likelihood methods (Forero & Maydeu-Olivares, 2009). Although, graded response models are different to partial credit models in terms of the expected model probabilities each model predicts (Rabe-Hesketh & Skrondal, 2012), the results of these two models present negligible differences regarding their results (see Baker et al., 2000 for an example). This is particularly true when these models are used to represent the cumulative probability of response for ordinal variables (idem). Moreover, if these models are specified with constrained $\lambda_1 - \lambda_n$ parameters to unity, thus, making the $\delta_1 - \delta_n$ parameters and $\theta_{..}$ terms the only informative entities of the model. This model is often referred to as the homogeneous case GRM (Samejima, 2016), or as the one-parameter graded response model (1PL-GRM) (Gochyyev, 2015). An equivalent model is the common slope GRM (Paek & Cole, 2020), which constrained $\lambda_1 - \lambda_n$ parameters to a single slope, while constraining the random variance of $\theta_{..}$ to unity. This model is a re-parametrization of the homogenous case and produces the same item thresholds ($\delta_1 - \delta_n$) and the same loglikelihood for the modelled responses.

In particular, we used bifactor models (Reise, 2012) to partition the variance of θ_p, between the general shared variance and the specific variance from each scale. In practice, if two scales were constructed as different scores, we would assess whether it is tenable to join these together in a single score. In practical terms, we used bifactor models to assess if these two collections of items, or more, shared enough variance. Using the index of Explained Common Variance (ECV) a collection of responses to a set of items can be considered essentially unidimensional if the common factor explains 85% of the variance (Toland et al., 2017). Simulation studies suggest that if 70% of the variance is accounted by a general factor, and 30% by the specific factors of the model, then reporting scores for the specific scales is more informative than a single score (Quinn, 2014).

To calculate the ECV index, we specified the common slope GRM (Paek & Cole, 2020), constraining to equality the $\lambda_1 - \lambda_n$ parameters of each factor, while fixing the variance of each $\theta_{..}$ to one. This model is just a re-parametrization of the homogenous case and produces the same item parameters ($\delta_1 - \delta_n$) and same loglikelihood. We use the following equation to produce this index (Reise et al., 2013):

$$ECV = \frac{\sum \lambda_g^2}{\sum \lambda_g^2 + \sum \lambda_{f_1}^2 + \cdots + \sum \lambda_{f_n}^2} \quad (2)$$

This is a measure of the strength of the general factor. This index is obtained as the ratio of the sum of the square of the factor loadings from the general factor, over the sum of the square of all factor loadings present in the model. The larger this index is, the more variance is explained by a common attribute than by a set of specific factors among responses. If this index lies between 1 and .9, essential unidimensionality is reached. For binary data, is recommended that if the ECV lies between .9 and .7 then more information needs to be used than the ECV alone to make a decision regarding creating a single score or different scores per factor. If ECV is .7 or less, then is advisable to generate different scores per factor (Quinn, 2014). In general, for Likert type items an ECV larger

or equal to .85 indicates enough unidimensionality to warrant a single factor model (Stucky & Edelen, 2015).

In this document, we used parallel analysis (Horn, 1965) as an additional procedure to assess unidimensionality. This procedure consists of comparing the number of extractable factors in an observed matrix of correlations, in contrast to the number of extractable factors from different simulated correlation matrices with similar characteristics of the observed correlation matrix. Specifically, we implemented the Timmerman & Lorenzo-Seva (2011) version, designed for polytomous responses. To implement this procedure, we select a random sample of 500 cases from each participating country and region, conditional to the survey weights each observation possess. With this random case selection, we can ensure all countries contribute equally to the parallel analysis This selection of cases for item analysis is a similar procedure used by the OECD in other large scale assessment studies (OECD, 2014).

A.2.1 Unidimensionality and interpretability of scores

The unidimensionality requirement means a single propensity term is sufficient to represent the pattern of response across a set of items. If this is not the case, then more dimensions are required to explain the observed responses. In this scenario, if we use a single model with only a random term θ_p, the specified model would fail to account for the shared variance across a set of responses. Consequently, the error between the expected responses and the observed responses would be larger, in comparison to a more complex model.

In the current study, SDG indicators 4.7.4 and 4.7.5 could potentially be represented by a single score. However, for this score to be interpretable, the unidimensionality requirement should be fulfilled. Otherwise, a single score would not be interpretable regarding the response pattern. Moreover, to develop a set of cut-off scores to establish standards that are interpretable over time (e.g. to subsequent applications of the instruments to other groups), we need a response model that allows such interpretations (Wilson & Draney, 2002).

There are different ways to assess the dimensionality of a set of responses. In the current report, we used a model-based approach when possible. In particular and as previously mentioned, we used bifactor models (Quinn, 2014; Rodriguez et al., 2016) to assess how much variance can be accounted by a single factor, in comparison to specific factors. Alternatively, scatter plots and correlations between measures were used when, by design, it was not possible to compute covariances between items (e.g., rotated booklet designs).

Apart from model estimates, substantive criteria were used to argue in favour or against the dimensionality and interpretability of scores that summarize responses to a set of items. Dimensionality analysis alone is not sufficient to ensure the interpretability of generated scores (Maul, 2017). As important as the common variance between responses is, interpretability of scores requires responses that are produced by a common construct or attribute (Wilson, 2005). In this document, this latter criteria was assessed based on the content of the proposed items, previous empirical research on the topic, and instrument development documentation from the original studies (M. O. Martin et al., 2016; Schulz, Carstens, et al., 2018). All the proposed items come from scales that measure defined intended constructs, that is, particular attributes that vary

over given populations (Cronbach & Meehl, 1955). Consequently, the information available for the proposed measures is integrated to generate scores and cut-off scores and to allow tenable interpretations.

A.3 Measurement model

To obtain person and items parameters, we propose to use a latent variable model approach. More specifically, we propose to use a partial credit model (Masters, 2016). This model allows us to obtain item and person parameters from items with two categories or more, or from a set of items with a different number of categories. Formally, this model can be described as follows (see Wu et al., 2016):

$$Pr(Y_{ip} = j|\theta_p) = \frac{\exp \sum_{k=0}^{j}(\theta_p - \delta_{ik})}{\sum_{h=0}^{m_i} exp \sum_{k=0}^{h}(\theta_p - \delta_{ik})} \quad (1)$$

In this model, the probability of answering an item (Y_{ip}), with a category of response 0, 1, 2, ..., m_i by a person p, depends on the propensity of the response of the person p (θ_p). For the first category of response, there is a constraint: $\sum_{k=0}^{0}(\theta_p - \delta_{ik}) = 1$. Thus, for the first category of response, the numerator in equation (1) is 1. The item parameters δ_{ik} needed are of one less the number of response categories for each item. Therefore, if all items are dichotomous, a single δ parameter is estimated per item. However, if all items have four categories of responses, then three δ parameters are estimated for each item.

The following is an example using the items proposed for the indicator category of Gender Equality in indicator 4.7.4:

Figure 1. Gender Equality items in ICCS 2016

Q24 There are different views about the roles of women and men in society.
How much do you agree or disagree with the following statements?

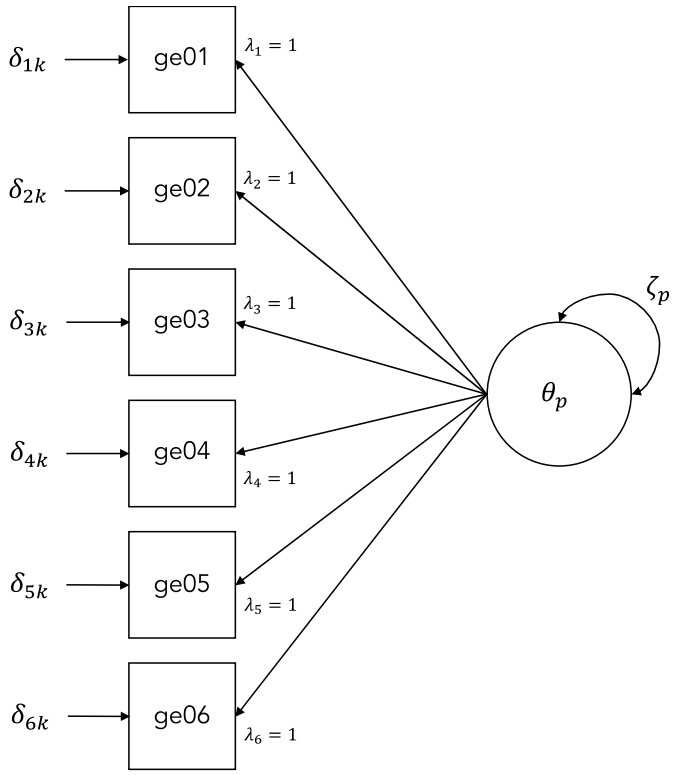
(Please tick only one box in each row.)

			Strongly agree	Agree	Disagree	Strongly disagree	
IS3G24A	a)	Men and women should have equal opportunities to take part in government.	☐ 1	☐ 2	☐ 3	☐ 4	ge01
IS3G24B	b)	Men and women should have the same rights in every way.	☐ 1	☐ 2	☐ 3	☐ 4	ge02
IS3G24C	c)	Women should stay out of politics.	☐ 1	☐ 2	☐ 3	☐ 4	ge04
IS3G24D	d)	When there are not many jobs available, men should have more right to a job than women.	☐ 1	☐ 2	☐ 3	☐ 4	ge05
IS3G24E	e)	Men and women should get equal pay when they are doing the same jobs.	☐ 1	☐ 2	☐ 3	☐ 4	ge03
IS3G24F	f)	Men are better qualified to be political leaders than women.	☐ 1	☐ 2	☐ 3	☐ 4	ge06

Students answer their level of agreement to these statements regarding women and men roles in society. With a partial credit model, we expect to represent the probability of response to each category. Each category of response for each item can be interpreted as an ordered response where a higher agreement expresses a higher endorsement of gender equality for items ge01, ge02 and ge03. Because ge04, ge05 and ge06 are reversed items, the response "Strongly Disagree" and "Disagree" express a higher endorsement of gender equality from respondents.

Using these items, we can represent the partial credit model as a latent variable model, with the following diagram:

Figure 2. Latent variable model for Gender Equality items



In this diagram (see Figure 2), the term θ_p represents the propensity of participants to provide a category of response of a higher value. To ensure this interpretation, all responses are recoded from 0 to 3, where higher values imply higher endorsement of gender equality for each item. The terms $\delta_{1k}-\delta_{6k}$ represents the step parameters in the partial credit model (Wu et al., 2016). These parameters represent where the two item characteristic curves intersect (Masters, 2016). That is, if we create a plot, where the probability of response is on the y-axis and the logit parameters are positioned on the x-axis, then the probability function of an item response is depicted as a curve. These curves would cross at the next category of responses; the $\delta_{1k}-\delta_{6k}$ marks these points on the logit scale. Using numerical methods, these parameters can be converted into cumulative probabilities, $\gamma_{1k}-\gamma_{6k}$, to build item-person maps (Wu et al., 2016). We use the term ζ_p to represent

the variance of θ_p, which is freely estimated in this model specification, and we leave θ_p, with a latent mean of zero. Parameters $\lambda_1 - \lambda_6$ are constrained to 1, to conform to a partial credit model.[3]

A.4 Proficiency levels

Proficiency levels refer to points on a scale used to classify participants between those who have a given level of capacity, and those who are less likely to have this same level of capacity (Zieky & Perie, 2006). These points on a scale or cut scores are similar to pass or fail these threshold on a test. In spite of being an uncommon practice, conceptually cut scores can be defined to establish levels to other type of attributes, different from academic outcomes such as mathematics, language or other common proficiency constructs. This is the case because levels of a theoretical attribute can be modelled for dichotomous and ordered responses (Diakow et al., 2013).

There are different ways in which these points on a scale can be defined. In general, these are referred to as different standard-setting procedures (Cizek et al., 2004). Popular methods used are the Bookmark method (Green et al., 2003), Angoff method (Ricker, 2006), and holistic methods (Torres Irribarra et al., 2015), among others (Zieky & Perie, 2006).

In this document, we followed an item-person map approach (Wyse, 2013). Unlike Bookmark and Angoff methods, the item-person map approach relies on judgments from experts to set the standards on scores and might be subject to revision once the results are obtained (Zieky & Perie, 2006). In this document, we propose standards with known results. That is, we build model-based construct maps (Torres Irribarra et al., 2015; Wilson, 2005) using responses from an ILSA with representative samples of students. Using the results of these construct maps, we proposed provisionary cut scores based on the criteria originally used in the corresponding ILSA.

In the following sections, we describe the measurement model we used to build item-person maps, we describe what item-person maps are and how we used these methodological tools to set the cut scores to identify the proposed proficiency levels. Additionally, we describe the characteristic of the measurement models used to produce these item-maps. Finally, we revise the conditions of responses availability and some limitations of the proposed cut scores.

A.4.1 Item-person maps

Item-person maps are a graphical display that orders items and respondents on a same scale. These are often called Wright Maps (Wilson & Draney, 2002), item-person maps (Desjardings & Bulut, 2018), or construct maps (Wyse, 2013). These figures order respondents and items on the same scale, aiding the interpretation of the location of responses. With these figures, it is easier to identify which items are more or less likely to be responding in a certain way. These plots can be created for responses on a test or questionnaire.

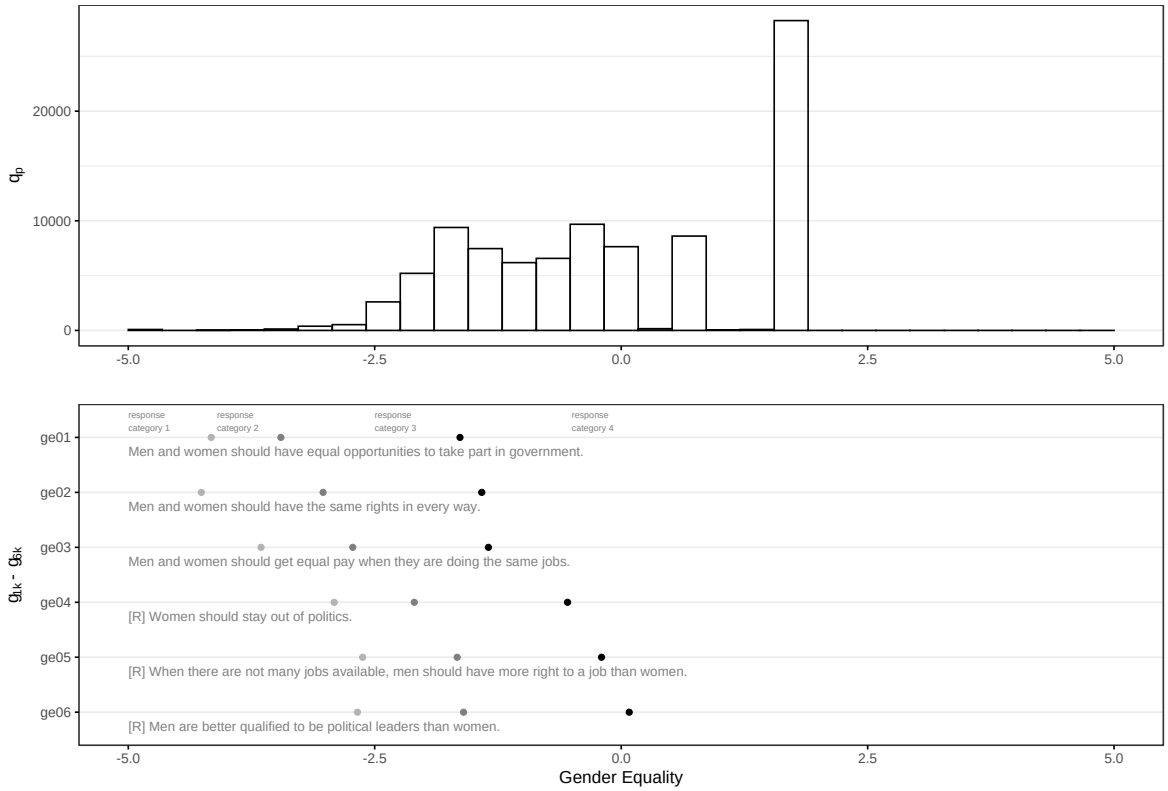
To build these figures, latent realizations of the respondents are generated, and item location parameters are extracted from the model. The two vectors are then plotted, one depicting a histogram or density of persons, while the second vector locates each response category given by

[3] See Appendix I for an example of how to fit this model with MPLUS.

the model. Then, persons' responses and items are located on the same scale. Traditionally, models from the Rasch family are used for these purposes. However, as longs as persons and item parameters are orthogonal in the measurement model, these item-person maps can be generated for special cases of the continuation ratio model (Kim, 2016), and the graded response model (Samejima, 2016). To keep a similar interpretation, the models should have one constraint: no covariance between items responses and the person locations.

The following is an example of a construct map, using the items proposed to measure the indicator 4.7.4 category of Gender Equality.

Figure 3. item-person map for Gender Equality (ICCS 2016)



Using a partial credit item response theory (IRT) model, person realizations are generated to create the first part of the figure, the histogram of θ_p. The second part is the *Thurstonian thresholds* for each response categories γ_{1k}–γ_{6k}. These estimates, γ_{1k}–γ_{6k}, depict the cumulative probabilities where each category of response reaches the median probability of response. That is, these locations demark when it is more likely that a participant has a 50% chance to answer a category of response or higher (Wu et al., 2016).

At the point 0, on the logit scale of the figure, we find the "most likely" response pattern. Items ge01-ge03 are Likert type items, where students rate their level of agreement to each affirmation presented in this figure. The response categories presented for each item were Strongly Disagree,

Disagree, Agree, and Strongly Agree. Items ge04-ge06 are reverse items. Thus, students who highly endorse gender equality would respond Disagree or Strongly Disagree to these given items.

Students with zero logit score have a 50% chance of *strongly* agreeing with items ge01, ge02 and ge03 and *strongly* disagreeing with items ge04, ge05, and ge06. Students with a logit score of about -2 have 50% chances of agreeing with items ge04, ge05, and ge06; which all express a low level of or no gender equality endorsement.

The requirements to build similar a figure includes the use of a unidimensional measurement model,[4] where persons and items parameters are orthogonal terms in the model. To use this approach and make inferences to a population, the modelled responses should come from a known population. With these two requirements, it is possible to infer which responses are most likely, less likely, and which ones occur the most often from a given population. This is possible because the item locations of a response category can be converted into the expected probability of a response.

A.4.2 Inferences with subsequent applications of the instruments

It is not possible to use a single country's item-person map to make inferences about another country. Item-person maps express how logit scores are tied to the expected pattern of responses from a sample. However, it is possible to use the parameters obtained from a new application of the instrument to evaluate the extent to which the parameters obtained from this new application are similar to the calibration sample. That is, we can locate the new application within the distribution generated from the model, using information from the calibrated sample. By using the parameters of the items, we can generate realizations of θ_p. If the two samples are very similar, then the latent means for both groups should be close to zero. If the new application presents a higher level of interest, then the latent mean of θ_p in the new one should be larger than zero.

In practical terms, if we used the full set of available responses from ICCS 2016 to obtain the parameters for Gender Equality, for example, then all other future applications of the items could be compared to the represented calibration sample.

In this exercise we assume that, under reasonable assumptions, the parameters resulting from two different applications are comparable, including the presence of invariant properties within a model. In this case, that a given set of parameters from a measurement model can be used for a second application of the same instrument to generate realizations of θ_p. So, if we assume that the instrument is applied to a sample of the same population, then generating additional realizations of θ_p using the parameters of the original application will result in the same interpretations of scores.

[4] It is also possible to build *Wrightmaps* with more measurement dimensions, including person and raters (see Engelhard & Wind, 2018), and for multidimensional models. However, these special cases are not applicable for the intended purpose of the current report.

A.4.3 Limitations of the provisional cut scores

In very simple terms, setting cut-off scores consists of choosing a point within a distribution that separates observations above a threshold from those below. As such, this threshold should correspond to a meaningful interpretation regarding the level of the attribute that is being tested. For example, in an ability test, this threshold should represent a pass or fail. Yet, reaching consensus on standards to assess where this threshold should be placed is a much more complicated matter. The critical requirement for a threshold within a distribution to become a standard is to establish a consensus among the users of these scores. When this consensus is reached, then one can expect this standard to be use in decision-making. However, if there is no consensus among users, then a cut-off score can hardly represent a standard to be used by users. The present document only discusses how to choose a cut-off score in a provisional manner; it does not propose a procedure to set a consensus among the users of the scores.

Several aspects of the proposed cut-off scores presented in this document can be subject of debate. First, the chosen cut-off score should be interpretable (Cizek et al., 2004). That is, we should be able to establish that participants above and below a threshold are different (or possess a different level of the attribute being tested). Moreover, we should be able to express what this difference means, and why it is relevant. In this document, we did not seek for agreement among potential users of the scores to establish the cut-off scores; instead, we used to item-person maps to establish the thresholds of the scales we generate. For this reason, their interpretation is limited to statistical criteria. So, the main limitation of the method proposed here relates to the lack of a content related interpretation of the cut-off scores[5].

If cut scores are intended to be used as a standards, it is desirable these could be used for future applications of the instruments (Wilson & Draney, 2002). If countries will use these cut-off scores as standards to monitor the achievement of SDG targets, then indicators should be comparable over time. In this report, we proposed the use of latent variable models to generate the scores. This option facilitates locating responses on the same scale to enable comparisons over time. That is, as long as there are enough anchor items (i.e. common items between applications), it is possible to produce comparable scales using data from different applications. However, it is important to consider that a low number of anchored items will result in a large linking error.

In summary, the proposed methods enable the interpretation of cut-off scores and allow to score responses on the same scales over time. However, producing cut-off scores that are interpretable in relation to the content of the items, cannot be achieved solely on the basis of the procedures described in this document. Producing content related cut-off scores would require users and/or experts to revise the content of the items included in each scale to agree on the standards that should be used for monitoring the SDG indicators. Nevertheless, if experts/users were going to discuss and agree on content-related standards for monitoring the SDG indicators, the resulting cut-off points could be located in the scales produced with the methodology proposed in this report.

---

[5] It is important to say that because the items belonging to the ILSAs' cognitive tests are not publicly available; it would have not been possible to establish cut-off scores with content-related interpretations (i.e. there is no information available to feed into such a discussion)

B. SDG thematic indicator 4.7.5

B.1 The selected items

The thematic indicator 4.7.5 refers to *Percentage of 15-year-old students showing proficiency in knowledge of environmental science and geoscience*. In general terms, this indicator taps into **Education for Sustainable Development (ESD)**. In previous documents (Sandoval-Hernández et al., 2019) this was described as:

> **Education for Sustainable Development (ESD)**: empowers learners to take informed decisions and responsible actions for environmental integrity, economic viability and a just society, for present and future generations, while respecting cultural diversity. It is about lifelong learning and is an integral part of quality education.

The operationalization of this indicator includes different items from the TIMSS 2015 TIMSS 2015.[6] The items selected to operationalize the cognitive domain of indicator 4.7.5 include items from the physics, biology and earth science tests.

Additionally, the socio-emotional and behavioural domains of this thematic indicator include items form the background questionnaires related to the motivation of students towards these disciplines and their self-efficacy on each of these subjects.

In total, the selection accounts for a total of 152 different items (see Table 3).

Table 3. Source of selected items to measure indicator 4.7.5

	Constructs	Number of items
a)	Physics	10
b)	Biology	34
c)	Earth Science	56
d)	Students Like Learning Physics	9
e)	Students Like Learning Biology	9
f)	Students Like Learning Earth Science	9
g)	Students Confident in Physics	8
h)	Students Confident in Biology	8
i)	Students Confident in Earth Science	8
	Total	**152**

Instruments “a” to “c” contain mostly dichotomous items, which are scaled in TIMSS 2015 using an IRT model (M. O. Martin et al., 2016). We labelled this collection of items as test items. The items contained in “d” to “i”, are Likert type items and were scaled using a partial credit model, producing

[6] See https://timssandpirls.bc.edu/

an IRT score for each item collection (M. O. Martin et al., 2016). We labelled this collection of items as questionnaire items.

Within the SDG framework, the selected items represent the indicator category Environmental Science and the sub-categories Physical Systems, Living systems, and Earth and Space Systems (see Table 4 and Sandoval-Hernández et al., 2019, p. 9).

Table 4. Mapping of TIMSS 2015 scales into the indicator categories

Category (indicator)	Test item collections	Questionnaire item collections
Environmental Science (geoscience)	Physics	Students Like Learning Physics
		Students Like Learning Biology
Environmental Science (geoscience)	Biology	Students Like Learning Earth Science
		Students Confident in Physics
Environmental Science (geoscience)	Earth Science	Students Confident in Biology
		Students Confident in Earth Science

Item collections from Physics, Biology and Earth Science are considered "cognitive" items in the "Proposal for a Measurement Strategy for Thematic Indicator 4.7.5 using International Large-Scale Assessments in Education" report (Sandoval-Hernández et al., 2019). While the rest of the items are classified as "non-cognitive" (i.e. socio-emotional and behavioural) in the same document. In the present document, these two groups of items will be treated differently. That is, we did not assume unidimensionality for cognitive and non-cognitive items at once. Despite being categorized in the same indicator category, these two groups of items are not aimed to produce a single interpretable score. The first collection of items assess proficiency to answer questions of Physics, Biology and Earth Science and consists of measures of maximal performance (Cronbach, 1984). In contrast, the second group of items are instruments designed to capture self-reports of students regarding their enjoyment and self-efficacy in Physics, Biology and Science, respectively. These latter constructs are different from the academic ability on each discipline (Yeager & Lee Duckworth, 2015).

B.2 SDG4 indicator 4.7.5 cognitive items

The content domains of the selected cognitive items are different. This includes Physics, Biology and Earth Science (Mullis & Martin, 2017). Although, all of these items refer to relevant knowledge to understand the environment. Before assessing the unidimensionality of the responses in the next section, we assess their availability.

Table 5. Selected test items to measure indicator 4.7.5

	Constructs	Number of items
a)	Physics	10
b)	Biology	34
c)	Earth Science	56

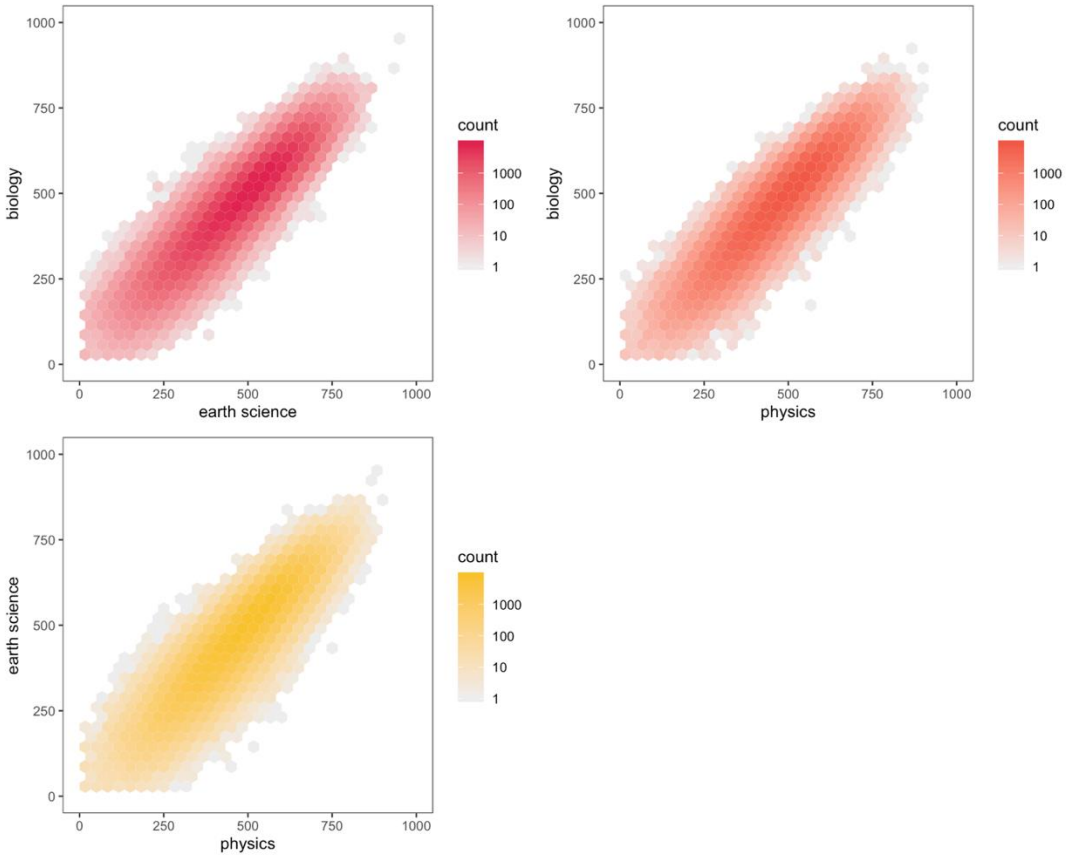
B.2.1 Availability

The selected items were answered, on average, by 13% of the participants in TIMSS 2015. This is the case because of TIMSS 2015 design, where block rotated items booklets are used (Rutkowski et al., 2010) To be precise, students answer one booklet out of 14 different booklets available (Foy, 2017).

B.2.1 Unidimensionality

Considering the block rotated design, there is a considerable amount of missing data by design and, therefore, it is not possible to estimate a covariance matrix for each pair of items. This is an obstacle to fit a bifactor model in a traditional way. However, there is a descriptive alternative to assess unidimensionality, namely, to evaluate if there is enough correlation between the selected items and the content domain IRT scores estimated by IEA as part of their normal scaling procedures. If there is a high correlation between the IRT scores generated with these collections of responses, then one can assume unidimensionality between the items. The overall correlation between theses scores varies between .90 and .91.

Figure 4. Scatter plot between Physics, Biology and Earth Science IRT scores



B.2.2 Proficiency classifications

To classify students as "showing proficiency in knowledge of environmental science and geoscience", we followed two strategies:

a) Classification of students above a threshold, using the IRT scores presented in TIMSS 2015 for Science (i.e. total score in science)
b) Classification of students above a threshold, using a unidimensional partial credit model for the selected items according to the mapping exercise in Sandoval-Hernandez, et al. (2019)

As mentioned before, the selected items for the cognitive component of indicator 4.7.5 are from the Physics, Biology and Earth Science tests and, therefore, are not public items. As such, the interpretability of item-person maps is limited under this condition. So, to identify a cut-off score that captures "proficiency in knowledge of environmental science and geoscience", we used the TIMSS 2015 anchored benchmark at 550 (Mullis et al., 2016).

At 550 points, students have 50% chances to, for example, "Explains why birds of prey cannot survive in an environment without plants". In more general terms, students at this level can:

> [...] apply and communicate their understanding of concepts from biology, chemistry, physics, and Earth science in everyday and abstract situations. Students apply knowledge

of cells and their functions and of the characteristics and life processes of organisms. They communicate their **understanding of ecosystems and the interaction of organisms with their environment** and apply some knowledge of human health related to nutrition and infectious disease. Students show some knowledge and understanding of the composition and properties of matter and chemical change. They apply basic knowledge of energy transformation and transfer and of light and sound in practical situations and demonstrate their understanding of simple electrical circuits and properties of magnets. Students apply their knowledge of forces and motion to everyday and abstract situations. They **apply knowledge of Earth's physical features, processes, cycles, and history, and show some understanding of Earth's resources, their use, and conservation** as well as some knowledge of the interaction between the Earth and the Moon. Students demonstrate some scientific inquiry skills, including selecting and justifying an appropriate experimental method. They combine and interpret information from various types of diagrams, graphs, and tables; select relevant information to analyse and draw conclusions; and provide short explanations conveying scientific knowledge (Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, 2016).

B.2.2.1 Proficiency Levels using the TIMSS 2015 IRT scores of Science (i.e. total score in science)

To classify students between those reaching the expected level of proficiency, we use the variable `BSSIBM01` available in TIMSS 2015 public use file. This is the first plausible value of the IRT science scores, recoded to classify students between the different international benchmarks. This variable presents five values, classifying students at different proficiency levels. The last two higher values were recoded as one while leaving the rest of the values as zero.

Table 6. TIMSS 2015 international benchmark variable codes

Value	Recode	Description
1	0	Student performed below the Low International Benchmark
2	0	Student performed at or above the Low International Benchmark, but below the Intermediate International Benchmark
3	0	Student performed at or above the Intermediate International Benchmark but below the High International Benchmark
4	1	Student performed at or above the High International Benchmark but below the Advanced International Benchmark
5	1	Student performed at or above the Advanced International Benchmark

Taylor Series Linearization is used to estimate the variance of the parameters, using pseudo strata, and primary sampling units' indicators. Proportions were estimated for all students reaching the High International Benchmark or above, for equally weighted countries, using senate weights scaled up to 1000 for each country.

Table 7. Percentage of students meeting the indicator 4.7.4 based on the IRT scores Benchmark for Science in TIMSS 2015 (i.e. total score in science)

Country or Region	Percentage	Lower limit	Upper limit
Morocco	0.03	0.03	0.04
Buenos Aires, Argentina	0.04	0.03	0.05
South Africa	0.05	0.03	0.07
Egypt	0.05	0.04	0.06
Botswana	0.05	0.05	0.06
Saudi Arabia	0.06	0.04	0.08
Lebanon	0.06	0.05	0.08
Jordan	0.10	0.08	0.11
Georgia	0.10	0.09	0.12
Kuwait	0.10	0.08	0.13
Chile	0.12	0.10	0.13
Thailand	0.12	0.10	0.16
Armenia	0.15	0.13	0.17
Iran, Islamic Rep. of	0.15	0.12	0.18
Oman	0.16	0.15	0.18
Abu Dhabi, UAE	0.21	0.17	0.24
Malaysia	0.21	0.19	0.23
Qatar	0.21	0.20	0.23
Bahrain	0.23	0.21	0.24
Italy	0.25	0.23	0.27
Norway	0.26	0.25	0.28
United Arab Emirates	0.27	0.25	0.28
Malta	0.27	0.26	0.29
Turkey	0.29	0.26	0.32
Australia	0.34	0.31	0.36
New Zealand	0.36	0.33	0.39
Lithuania	0.36	0.34	0.39
Israel	0.37	0.34	0.40
Ontario, Canada	0.38	0.35	0.41
Canada	0.39	0.36	0.41
Quebec, Canada	0.40	0.35	0.44
Sweden	0.41	0.37	0.44
Kazakhstan	0.42	0.38	0.46
Hungary	0.42	0.39	0.45
United States	0.42	0.39	0.45
Ireland	0.43	0.40	0.46
Dubai, UAE	0.44	0.41	0.46
England	0.45	0.41	0.50
Russian Federation	0.49	0.45	0.52
Hong Kong, SAR	0.52	0.48	0.56
Slovenia	0.52	0.50	0.54
Korea, Rep. of	0.54	0.52	0.56
Chinese Taipei	0.63	0.61	0.65
Japan	0.63	0.61	0.65
Singapore	0.74	0.71	0.77

We used a single plausible indicator for simplicity. Point estimates between plausible values vary maximum by .01; thus, if the five plausible values were used, the confidence interval would be expected to vary by a slightly larger error.

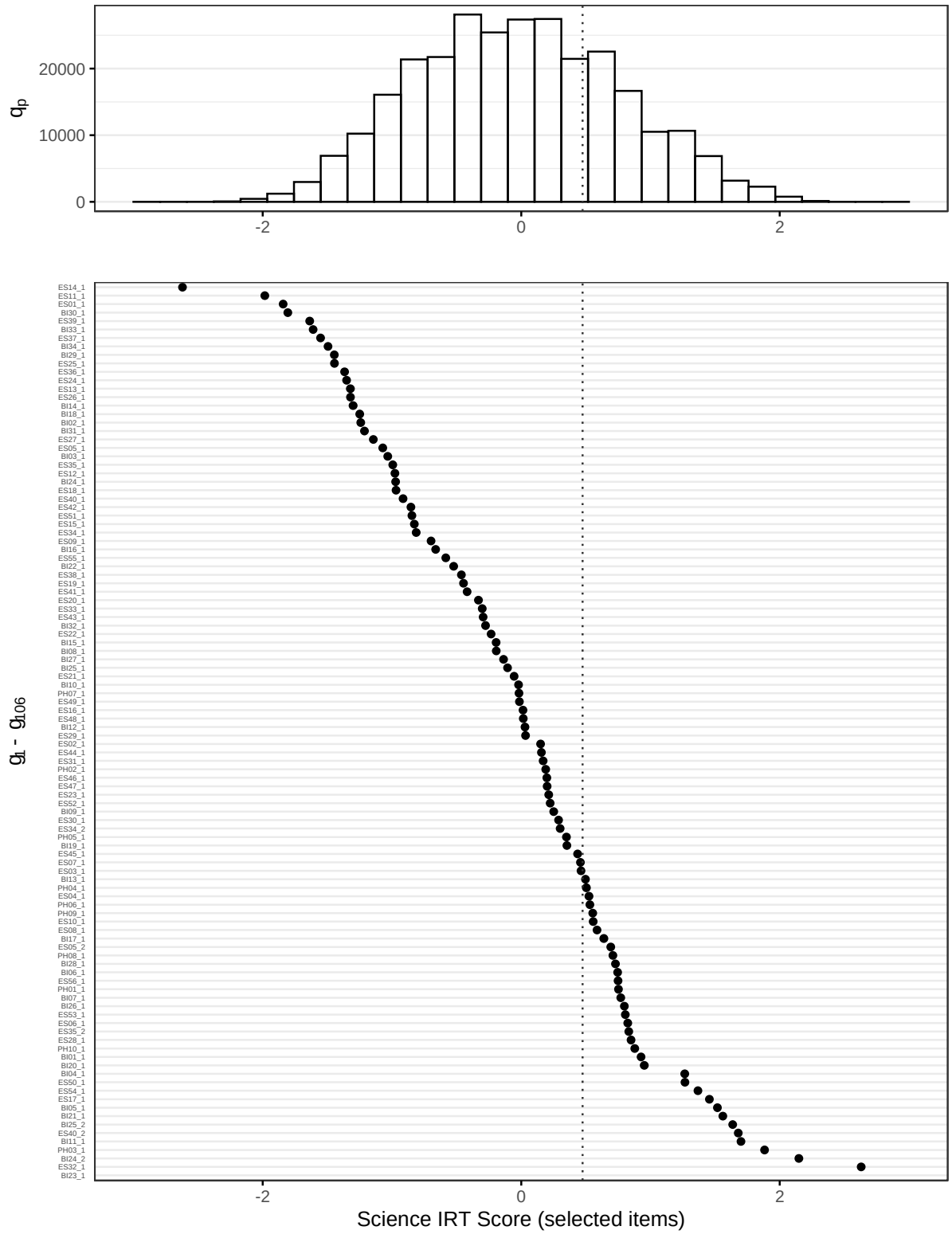
B.2.2.2 Proficiency Levels using the TIMSS 2015 selected items according to the mapping exercise

Given the rotated block design, the selected items to measure indicator 4.7.5 have observed responses, on average, from 13% of the nominal sample of TIMSS 2015. Assuming missing at random, the IRT model can generate values for θ_p and estimate $\gamma_1 - \gamma_{106}$ locations, conditional to θ_p. We fit a graded response model[7] (Samejima, 2016), with parameters λ constrained to 1. Thus, item parameters and person parameters are orthogonal. We use Taylor Series Linearization for variance estimation, using pseudo strata, and primary sampling unit indicators, for equally weighted countries using survey weights scaled up to 1000 for each participating country. Person realizations are generated as Expected a Posteriori values.

Because selected items cannot be inspected, a content judgment cannot be made for the selected items using the item-person map. Alternatively, we chose a similar distance from the latent mean as a cut-off score from the previous proficiency level. The High International Benchmark is located at half a standard deviation from the international scale (50 points). The fitted model presents a variance of .903 in the logit scale, thus having a standard deviation of .951 logits. Therefore, the expected location of half a standard deviation from the latent mean is at .475 logit scores. The chosen threshold is presented in the next figure.

[7] In the selected items, there are six items with partial credit scores. MPLUS v8.3 can't fit a partial credit score model for a mix of items with a mix of 2 and 3 categories. It is possible to replicate this same procedure with a different software. However, given the ability of MPLUS to fit latent variable models, while including the survey weights design, a graded response model was preferred.

Figure 5. Item-person map for Science Scores (selected items)



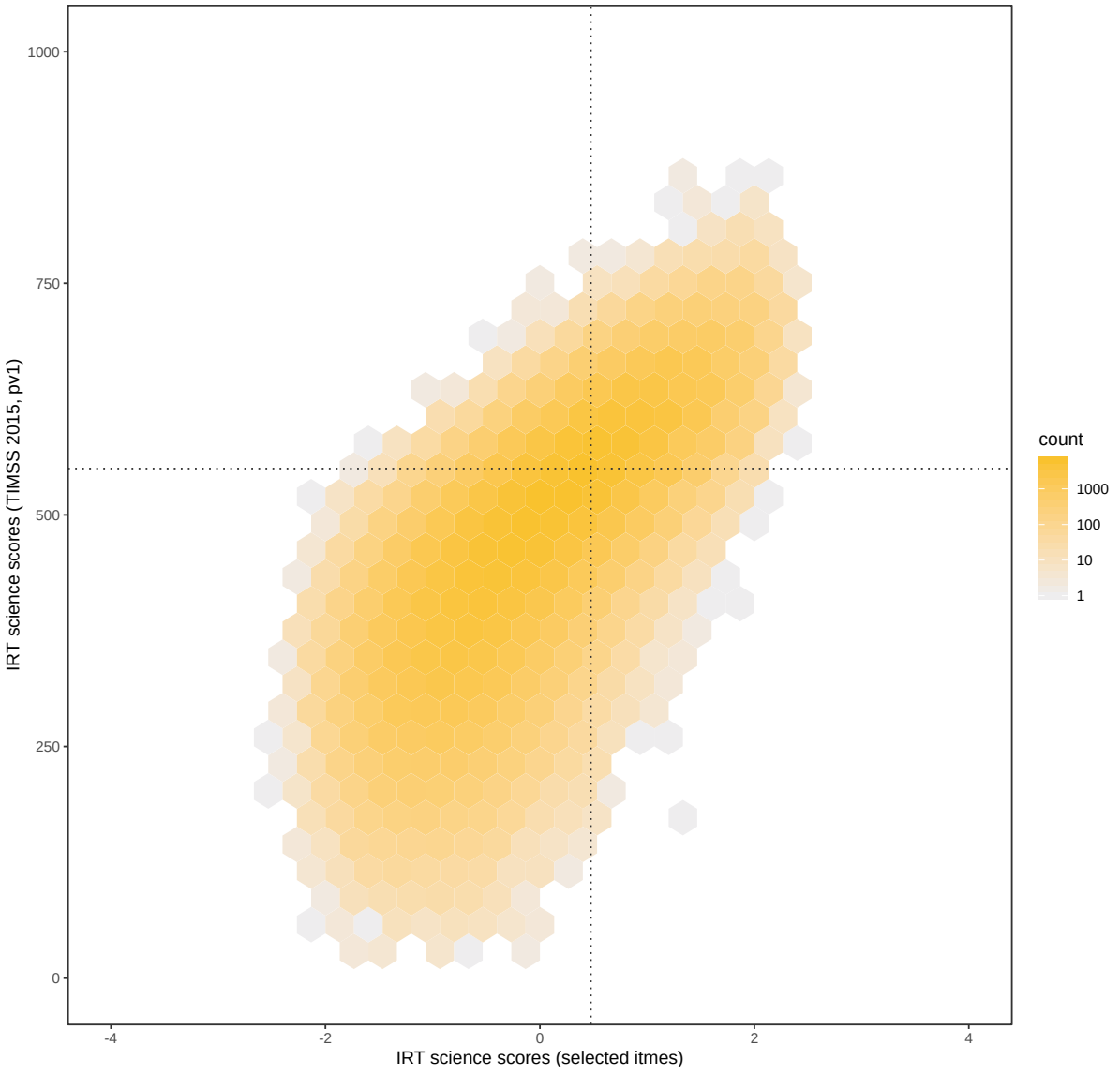
The proposed threshold distinguishes among all participant who presents similar levels of ability of the High International Benchmark for TIMSS 2015 on the selected items for Science. Estimates of what proportion of students are at the expected level are presented in the next table.

Table 8. Percentage of students meeting the indicator 4.7.5 based on the IRT scores of selected items of Science in TIMSS 2015 according to the mapping exercise

Country or Region	Percentage	Lower limit	Upper limit
South Africa	0.05	0.04	0.07
Egypt	0.06	0.05	0.07
Botswana	0.07	0.06	0.08
Saudi Arabia	0.07	0.06	0.09
Morocco	0.07	0.07	0.08
Lebanon	0.10	0.08	0.11
Jordan	0.11	0.10	0.13
Kuwait	0.12	0.10	0.15
Buenos Aires, Argentina	0.13	0.11	0.14
Georgia	0.13	0.12	0.14
Thailand	0.16	0.14	0.19
Oman	0.17	0.16	0.19
Chile	0.18	0.16	0.20
Iran, Islamic Rep. of	0.18	0.16	0.20
Abu Dhabi, UAE	0.19	0.17	0.21
Qatar	0.20	0.19	0.22
Malaysia	0.21	0.19	0.22
Bahrain	0.21	0.20	0.23
Malta	0.24	0.22	0.25
United Arab Emirates	0.24	0.23	0.25
Armenia	0.24	0.22	0.26
Turkey	0.25	0.22	0.27
Italy	0.31	0.29	0.32
Norway	0.33	0.31	0.34
Australia	0.34	0.32	0.36
Israel	0.34	0.32	0.37
Lithuania	0.35	0.33	0.37
New Zealand	0.36	0.34	0.39
Dubai, UAE	0.36	0.34	0.39
Kazakhstan	0.37	0.33	0.40
Ontario, Canada	0.37	0.35	0.40
Hungary	0.38	0.36	0.41
Ireland	0.39	0.37	0.41
Canada	0.39	0.37	0.41
England	0.39	0.36	0.43
United States	0.40	0.38	0.43
Sweden	0.41	0.38	0.43
Quebec, Canada	0.42	0.39	0.46
Hong Kong, SAR	0.45	0.41	0.48
Russian Federation	0.45	0.42	0.48
Korea, Rep. of	0.45	0.43	0.47
Japan	0.49	0.47	0.51
Slovenia	0.50	0.48	0.52
Chinese Taipei	0.55	0.53	0.57
Singapore	0.59	0.56	0.61

The correlation between the scores produced with the two approaches (i.e. scores from the selected items and the total TIMSS Science scores) is very high. It is estimated at 1, and if we regress the IRT scores of Science on the IRT scores of Science (selected items), we observed a beta coefficient of .75 in its standardized scale. We can therefore conclude that using the scores estimated with the selected items according to the mapping exercise constitutes a reliable option. The next figure depicts the relationship between these two scores (see Figure 6).

Figure 6. Scatter between Science IRT scores from TIMSS 2015



B.3 SDG4 indicator 4.7.5 Non-cognitive items

The selected items for this domain comprise a total of 51 items. These different items were used to generate IRT scores representing six different constructs (Michael O Martin & Foy, 2016). Three of these constructs (d, g and f) represent the students' enjoyment of learning Physics, Biology, and Earth Science. Similarly, the other three constructs (g, h and i) represent students' confidence in their knowledge of Physics, Biology, and Earth Science (see Table 9).

Table 9. Selected non-cognitive items to measure indicator 4.7.5

	Constructs	Number of items
d)	Students Like Learning Physics	9
e)	Students Like Learning Biology	9
f)	Students Like Learning Earth Science	9
g)	Students Confident in Physics	8
h)	Students Confident in Biology	8
i)	Students Confident in Earth Science	8
	Total	**51**

B.3.1 Availability

Out of the total sample of TIMSS 2015, on average, 17% cases present responses to all these items. These items were answered by some countries only including Georgia, Hungary, Kazakhstan, Lebanon, Lithuania, Malta, Morocco, Russian Federation, Slovenia and Sweden (Michael O Martin & Foy, 2016). In contrast, the scales of enjoyment of learning science and students' confidence in their science knowledge, in general, present a larger coverage across countries (see Table 10). Out of the 46 countries and regions that participated in TIMSS 2015, 76% of these have responses on liking the learning of science and self- report measures of students' confidence in their science knowledge.

Table 10. Countries and Regions with available responses on enjoyment in learning and students' confidence in their scientific knowledge

Country or Region	Science	Physics	Biology	Earth Science
Australia	yes	no	no	no
Bahrain	yes	no	no	no
Armenia	no	yes	yes	yes
Botswana	yes	no	no	no
Canada	yes	no	no	no
Chile	yes	no	no	no
Chinese Taipei	yes	no	no	no
Georgia	no	yes	yes	yes
Hong Kong, SAR	yes	no	no	no
Hungary	no	yes	yes	yes
Iran, Islamic Rep. of	yes	no	no	no
Ireland	yes	no	no	no
Israel	yes	no	no	no
Italy	yes	no	no	no
Japan	yes	no	no	no
Kazakhstan	no	yes	yes	yes
Jordan	yes	no	no	no
Korea, Rep. of	yes	no	no	no
Kuwait	yes	no	no	no
Lebanon	no	yes	yes	no
Lithuania	no	yes	yes	yes
Malaysia	yes	no	no	no
Malta	no	yes	yes	yes
Morocco	no	yes	yes	yes
Oman	yes	no	no	no
New Zealand	yes	no	no	no
Norway	yes	no	no	no
Qatar	yes	no	no	no
Russian Federation	no	yes	yes	yes
Saudi Arabia	yes	no	no	no
Singapore	yes	no	no	no
Slovenia	no	yes	yes	yes
South Africa	yes	no	no	no
Sweden	no	yes	yes	no
Thailand	yes	no	no	no
United Arab Emirates	yes	no	no	no
Turkey	yes	no	no	no
Egypt	yes	no	no	no
United States	yes	no	no	no
England	yes	no	no	no
Norway (8th grade)	yes	no	no	no
Dubai, UAE	yes	no	no	no
Abu Dhabi, UAE	yes	no	no	no
Ontario, Canada	yes	no	no	no
Quebec, Canada	yes	no	no	no
Buenos Aires, Argentina	yes	no	no	no

Considering the presented scenario, we alternatively propose the "Students Like Learning Science" and "Students Confident in Science" as complementary indicators of the SDG indicator 4.7.5.

Table 11. Alternative survey items to measure indicator 4.7.5

	Constructs	Number of items
j)	Students Like Learning Science	9
k)	Students Confident in Science	8
	Total	**17**

B.3.2 Unidimensionality

We fitted a common slope GRM (Paek & Cole, 2020) with a probit link, using the WLSMV estimator. We relied on Taylor Series Linearization to get corrected standard errors including clusters and pseudo strata indicators (Stapleton, 2013). Survey total weights were scaled to 1000 for each country and region, so each representative sample contributes equally to all estimations. We specified a bifactor model, where all item responses are conditioned by a general common factor, with two additional factors to account for the responses to the "Students Like Learning Science" items, and the responses to the "Students Confident in Science" items.

The general factor accounts for 69% of the variance (ECV = .69 CI95% [.68, .69]). Thus, it is not advisable to represent responses to all these items into a single score due to a loss of information. We fit a two-factor model, using the same model parametrization. These two factors have a correlation of .72 (SE=.03, p <.001). Although these are two highly correlated factors, the results of these analyses provide evidence to consider these two collections of items as different constructs.

In other words, the enjoyment of learning science is a different construct than that of student's self-evaluation regarding their scientific knowledge. The first expresses if students have a positive inclination towards the school subject of science (Osborne et al., 2003) or the extent to which students like learning science. In contrast, student's science self-efficacy consists of how students assess themselves regarding their competence in science (Wigfield et al., 2015). These are beliefs held by students regarding their capabilities or expectations of personal mastery (Bandura, 1977).

B.3.3 Measurement Models

Considering the availability of responses of students among different participating countries and regions, and the dimensionality between the proposed measures, we assess the enjoyment levels and students' self-efficacy levels as two separate constructs.

Table 12. Mapping of TIMSS 2015 motivation scales into the indicator categories

Category (indicator)	Item collections
Environmental Science (socio-emotional)	Students Like Learning Science
Environmental Science (behavioural)	Students Confident in Science

These two measures are presented separately in the following section.

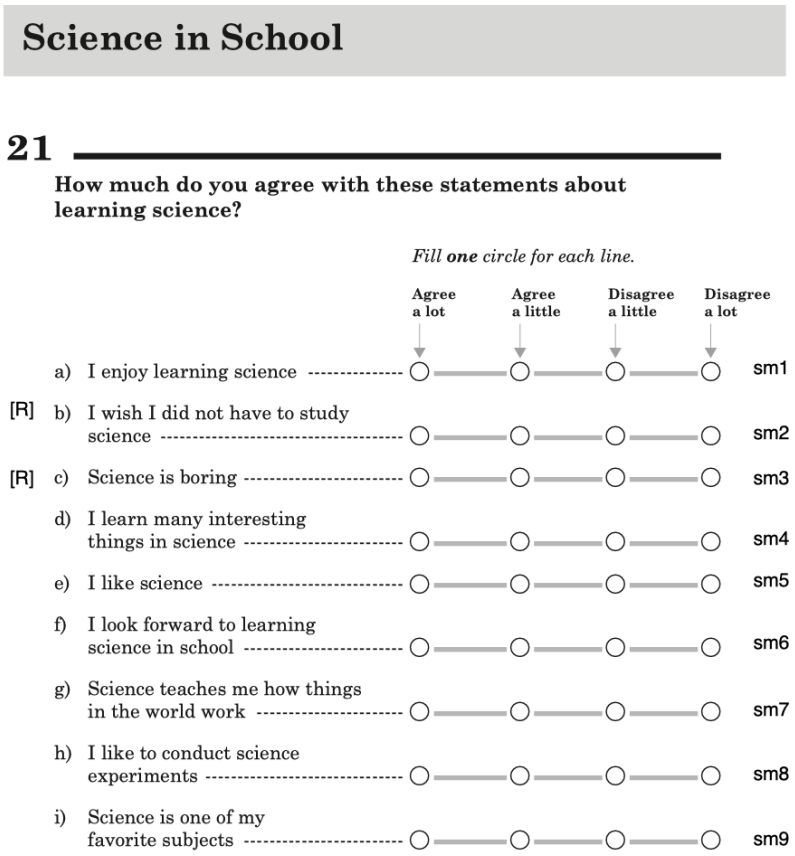
B.3.4 Proficiency classification

The proficiency classification exercise, or establishment of cut-off points, is organised according to the non-cognitive conceptual learning dimensions established for SDG indicators 4.7.4 and 4.7.5: socio-emotional and behavioural (see Table 2).

B.3.4.1 Proficiency classifications of Environmental Science (socio-emotional)

The items measuring Environmental Science (socio-emotional) are presented in the next figure.

Figure 7. Students Like Learning Science items in TIMSS 2015 for eighth-grade students



Science in School

21

How much do you agree with these statements about learning science?

*Fill **one** circle for each line.*

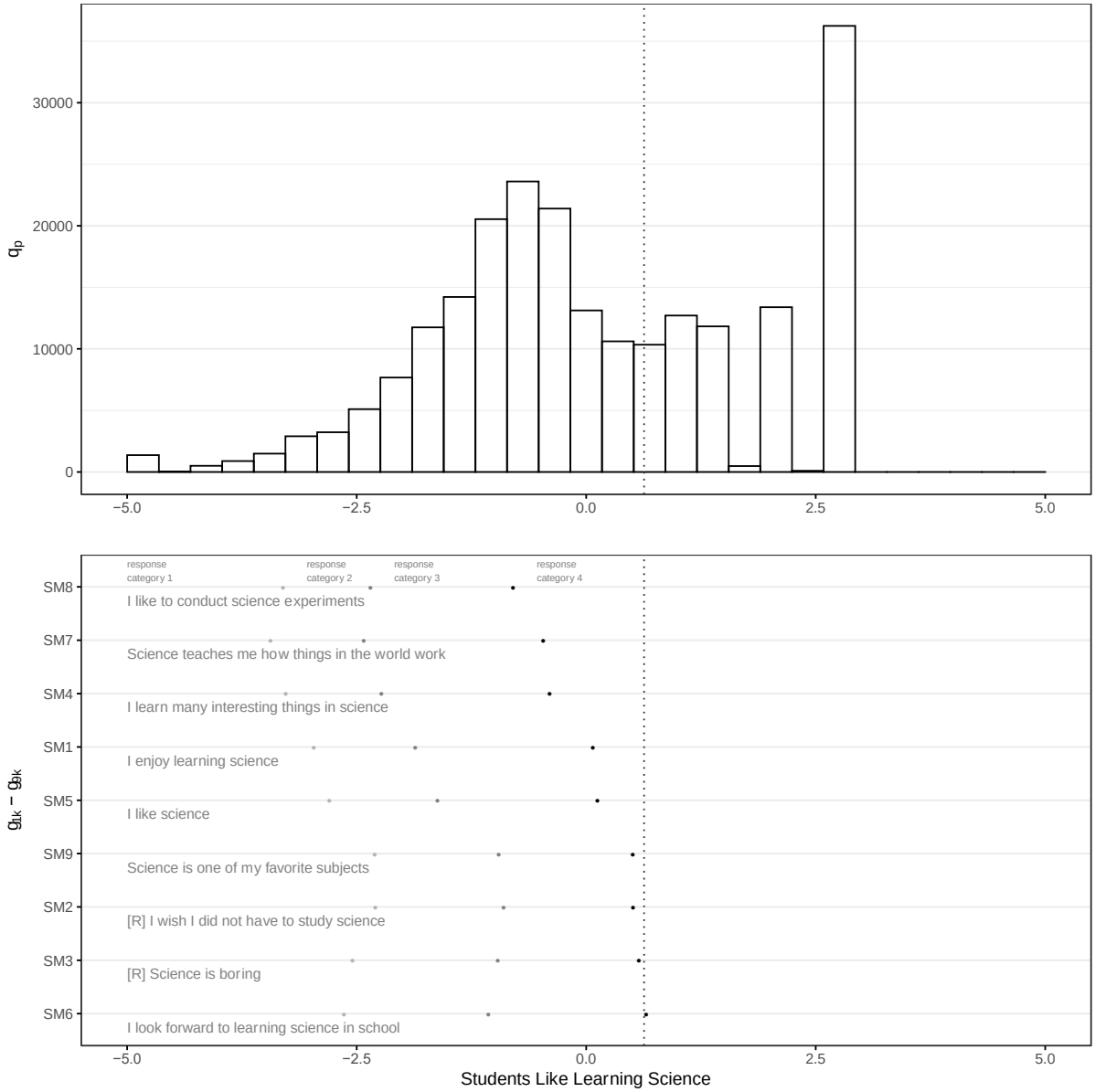
		Agree a lot	Agree a little	Disagree a little	Disagree a lot	
	a) I enjoy learning science	○	○	○	○	sm1
[R]	b) I wish I did not have to study science	○	○	○	○	sm2
[R]	c) Science is boring	○	○	○	○	sm3
	d) I learn many interesting things in science	○	○	○	○	sm4
	e) I like science	○	○	○	○	sm5
	f) I look forward to learning science in school	○	○	○	○	sm6
	g) Science teaches me how things in the world work	○	○	○	○	sm7
	h) I like to conduct science experiments	○	○	○	○	sm8
	i) Science is one of my favorite subjects	○	○	○	○	sm9

Note: [R] = are reverse score items. sm1-sm9 = variable names assigned to the responses to these items.

A partial credit model for equally weighted countries, using senate weights scaled up to 1000 for each country fitted. Taylor Series Linearization is used to estimate the variance of parameters, using pseudo strata, and primary sampling unit indicators. Person realizations are generated as Expected a Posteriori, and delta parameters are converted into Thurstonian thresholds.

The proposed threshold is presented in the following item-person map and is located at the highest category of response, after item sm6 (“I look forward to learning science in school”).

Figure 8. Item-person map for Students Like Learning Science



The proposed threshold distinguishes among all participant who presents 50% chances to highly express science learning enjoyment and those students who are less likely enjoy learning science. The majority of the students meeting the proposed standard agree a lot to expressions such as “I like to conduct science experiments”, “I learn many interesting things in science” and “I like Science”. Students meeting the proposed standard have equal chances to express that they agree a little or agree a lot to expressions such as “Science is of one my favorite subjects” and “I look forward to learning science in school”. Complementary, the students meeting the proposed standard, express disagreement to expressions such as “Science is boring” and “I wish I did not have to study science”. In the next table, we estimate the proportion of students meeting the standard in TIMSS 2015.

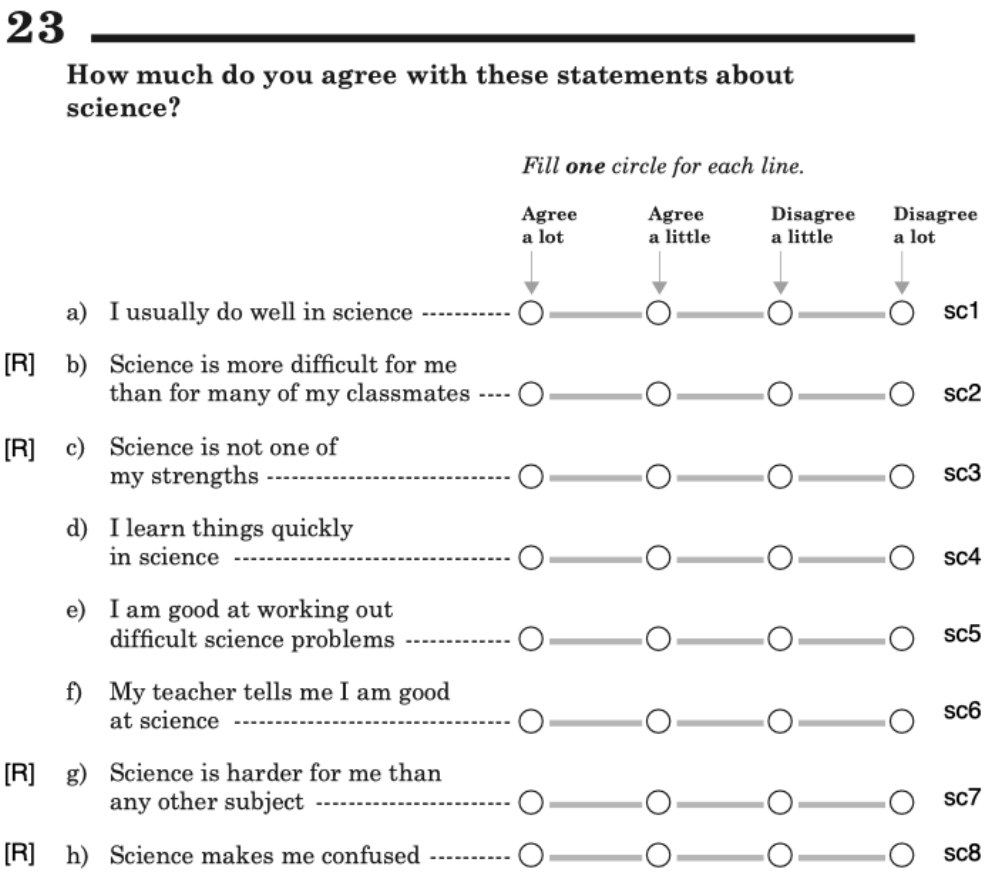
Table 13. Percentage of students meeting the indicator 4.7.5 Environmental Science (socio-emotional)

Country or Region	Percentage	Lower limit	Upper limit
Korea, Rep. of	0.09	0.08	0.10
Japan	0.13	0.12	0.15
Chinese Taipei	0.16	0.15	0.17
Buenos Aires, Argentina	0.18	0.17	0.20
Norway	0.24	0.22	0.26
Australia	0.24	0.22	0.26
Italy	0.24	0.22	0.26
Israel	0.25	0.23	0.27
Chile	0.25	0.23	0.28
Quebec, Canada	0.25	0.22	0.29
Hong Kong, SAR	0.26	0.24	0.28
New Zealand	0.27	0.25	0.29
England	0.28	0.26	0.30
Ireland	0.28	0.26	0.31
Canada	0.29	0.27	0.30
Ontario, Canada	0.30	0.28	0.32
Norway (8th grade)	0.31	0.29	0.33
Thailand	0.31	0.29	0.34
United States	0.32	0.31	0.34
Abu Dhabi, UAE	0.33	0.29	0.37
Singapore	0.34	0.32	0.35
Qatar	0.34	0.32	0.37
Saudi Arabia	0.37	0.33	0.40
Bahrain	0.37	0.35	0.39
United Arab Emirates	0.37	0.35	0.39
South Africa	0.41	0.39	0.43
Iran, Islamic Rep. of	0.43	0.41	0.46
Kuwait	0.43	0.41	0.46
Dubai, UAE	0.44	0.42	0.46
Egypt	0.44	0.42	0.47
Oman	0.45	0.43	0.48
Malaysia	0.46	0.43	0.48
Turkey	0.46	0.44	0.48
Jordan	0.49	0.47	0.51
Botswana	0.51	0.49	0.53

B.3.4.1 Proficiency classifications of Environmental Science (behavioural)

The items measuring Environmental Science (behavioural) are presented in the next figure.

Figure 9. Students Confident in Science items in TIMSS 2015 for eighth-grade students



23

How much do you agree with these statements about science?

*Fill **one** circle for each line.*

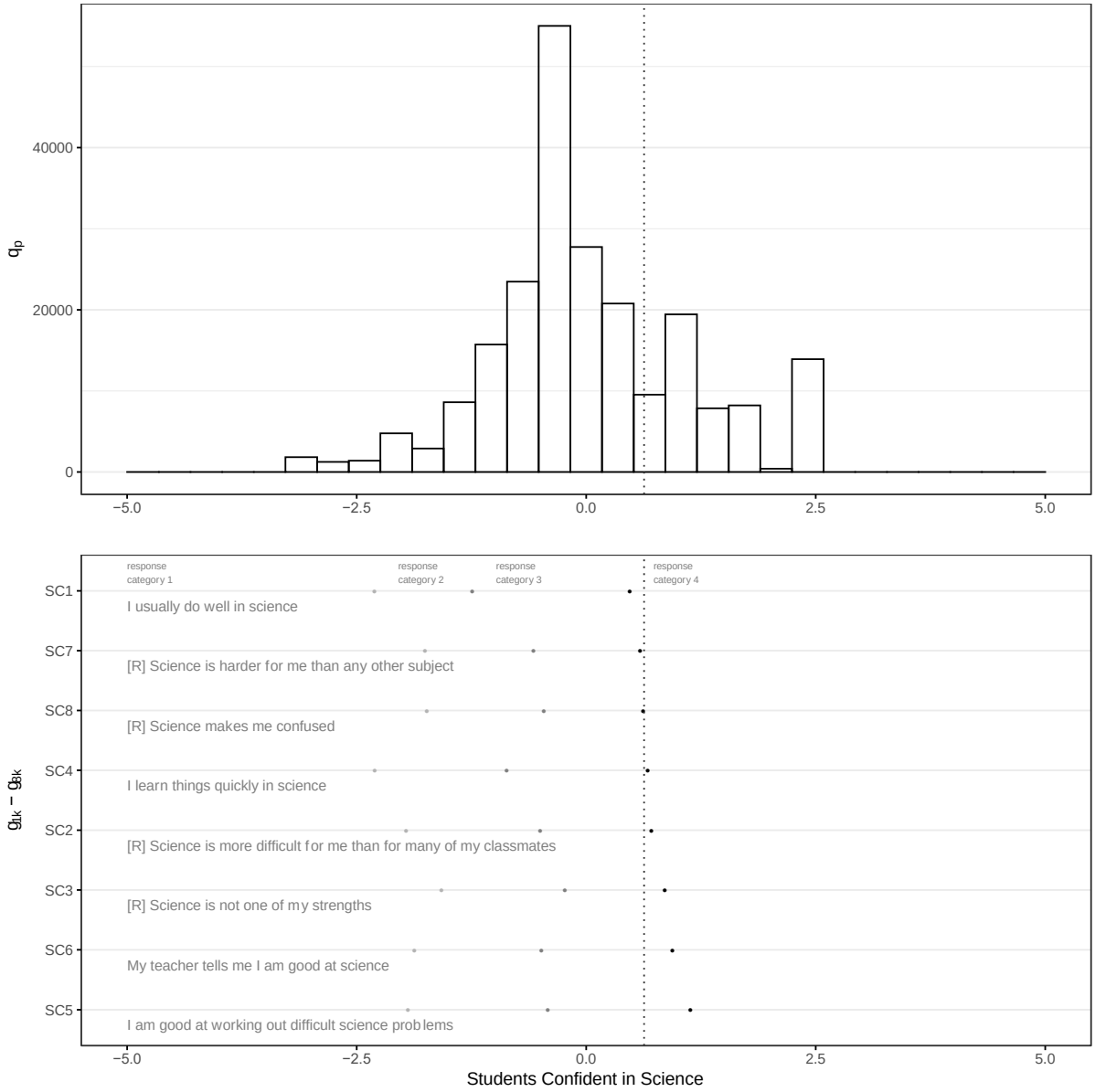
		Agree a lot	Agree a little	Disagree a little	Disagree a lot	
	a) I usually do well in science	○	○	○	○	sc1
[R]	b) Science is more difficult for me than for many of my classmates	○	○	○	○	sc2
[R]	c) Science is not one of my strengths	○	○	○	○	sc3
	d) I learn things quickly in science	○	○	○	○	sc4
	e) I am good at working out difficult science problems	○	○	○	○	sc5
	f) My teacher tells me I am good at science	○	○	○	○	sc6
[R]	g) Science is harder for me than any other subject	○	○	○	○	sc7
[R]	h) Science makes me confused	○	○	○	○	sc8

Note: [R] = are reverse score items. sc1-sc8 = variable names assigned to the responses to these items.

A partial credit model for equally weighted countries, using senate weights scaled up to 1000 for each country was fitted. Taylor Series Linearization is used to estimate the variance of parameters, using pseudo strata, and primary sampling unit indicators. Person realizations are generated as Expected a Posteriori, and delta parameters are converted into Thurstonian thresholds.

The proposed threshold is presented in the following item-person map, and is located at the highest category of response, after item sc8 ("[R] Science makes me confused").

Figure 10. Item-person map for Students Confident in Science



The proposed threshold distinguishes among all who express high confidence in their competence in science. In particular, students meeting the proposed standards have 50% chances to highly disagree with the statement “Science makes me confused”, and express agreement to statements such as “I learn things quickly in science”, “I usually do well in science”, and “I’m good to work out difficult science problems”. In the next table, we estimate the proportion of students meeting the standard in TIMSS 2015.

Table 14. Percentage of students meeting the indicator 4.7.5 Environmental Science (behavioural)

Country or Region	Percentage	Lower limit	Upper limit
Japan	0.07	0.06	0.08
Malaysia	0.07	0.06	0.08
Korea, Rep. of	0.09	0.08	0.10
Thailand	0.09	0.08	0.10
Chinese Taipei	0.11	0.10	0.12
Hong Kong, SAR	0.16	0.14	0.17
Botswana	0.18	0.17	0.20
New Zealand	0.19	0.17	0.20
Chile	0.19	0.18	0.21
Singapore	0.20	0.19	0.22
Australia	0.21	0.19	0.22
Buenos Aires, Argentina	0.22	0.20	0.24
England	0.25	0.23	0.27
South Africa	0.25	0.24	0.27
Quebec, Canada	0.29	0.26	0.31
Canada	0.29	0.27	0.30
Abu Dhabi, UAE	0.29	0.26	0.33
Ontario, Canada	0.29	0.27	0.31
Ireland	0.30	0.28	0.33
Qatar	0.31	0.28	0.33
Italy	0.31	0.28	0.33
Egypt	0.31	0.28	0.34
Saudi Arabia	0.31	0.28	0.34
United Arab Emirates	0.32	0.31	0.33
Bahrain	0.32	0.31	0.34
Jordan	0.34	0.31	0.36
Norway	0.34	0.32	0.37
United States	0.35	0.33	0.37
Iran, Islamic Rep. of	0.36	0.34	0.38
Oman	0.36	0.35	0.38
Israel	0.37	0.34	0.39
Turkey	0.37	0.35	0.39
Dubai, UAE	0.38	0.36	0.40
Kuwait	0.39	0.36	0.42
Norway (8th grade)	0.39	0.37	0.42

C. SDG thematic indicator 4.7.4

C.1 Selected items

The SDG 4 Thematic Indicator 4.7.4 refers to Global Citizenship Education (GCED). In previous documents, this was described as:

> **Global Citizenship Education (GCED):** nurtures respect for all, building a sense of belonging to a common humanity and helping learners become responsible and active global citizens. GCED aims to empower learners to assume active roles to face and resolve global challenges and to become proactive contributors to a more peaceful, tolerant, and inclusive and secure world.

The operationalization to these indicators includes different items from the IEA ICCS[8] 2016. The items selected to operationalize the cognitive domain of SDG indicator 4.7.4 include items from the four content domains of ICCS: civic society and systems, civic principles, civic participation and civic identities.

Additionally, the socio-emotional and behavioural domains of this thematic indicator include items form the background questionnaires related to the categories and sub-categories included in indicator 4.7.4.

The selection of items accounts for a total of ~46 items. These different items were originally developed to generate scores representing different constructs (see Table 15).

Table 15. Source of selected items to measure indicator 4.7.4

	Constructs	Number of items
a)	Civic society and systems (content domain 1)	X
b)	Civic principles (content domain 2)	X
c)	Civic participation (content domain 3)	X
d)	Civic identities (content domain 4)	X
e)	Students' attitudes toward their country of residence	5
f)	Students' attitudes toward equal rights for all ethnic/racial groups	5
g)	Students' attitudes toward gender rights	6
h)	Students' reports on personal experiences of bullying and abuse	6
i)	Students' perception of the importance of social movement related citizenship	4
j)	What is good for democracy	9
k)	Threats to the world future	11
	Total	46

[8] See: https://iccs.iea.nl/home.html

Instruments “a” to “d” contain mostly dichotomous items, which are scaled in ICCS 2016 using an IRT model (Schulz, Carstens, et al., 2018). We labelled this collection of items as test items. Constructs “e” to “k” contain a set of Likert-type items, which are already scaled as unidimensional latent traits in the ICCS 2016 public data file, using a partial credit model (Schulz, Carstens, et al., 2018). In contrast, items included in sections labelled here as “What is good for democracy” and “Threats to the world” were not scaled into an IRT score. We labelled this collection of items as questionnaire items.

These different items are expected to represent other categories under the SDG framework (see Table 16 and Sandoval-Hernández et al., 2019, pp. 13–16).

Table 16. Mapping of ICCS 2016 scales into the indicator categories

Category (indicator)	Test item collections	Questionnaire item collections
Interconnectedness and Global Citizenship	Students' attitudes toward their country of residence	Civic society and systems Civic identities
	Students' attitudes toward equal rights for all ethnic/racial groups	
Gender Equality	Students' attitudes toward gender rights	Civic principles
Peace, Non-violence and Human Security	Students' reports on personal experiences of bullying and abuse	Civic participation
Human Rights	What is good for democracy	Civic society and systems Civic principles
	Students' perception of the importance of social movement related citizenship	
Sustainable Development	Threats to the world future	Civic society and systems Civic principles

Item collections from Civic Society and Systems, Civic Principles, Civic Participation and Civic Identities (content domains) are considered “cognitive” items in the “Proposal for a Measurement Strategy for Thematic Indicator 4.7.5 using International Large-Scale Assessments in Education” report (Sandoval-Hernández et al., 2019). While the rest of the items are classified as “non-cognitive” (i.e. socio-emotional and behavioural) in the same document. In the present document, these two groups of items will be treated differently. That is, we did not assume unidimensionality for cognitive and non-cognitive items at once. Despite being categorized in the same indicator categories, these two groups of items are not aimed to produce a single interpretable score. The first collection of items assess proficiency to answer questions related to the ICCS content domains and consists of measures of maximal performance (Cronbach, 1984). In contrast, the second group of items are instruments designed to capture self-reports of students regarding their attitudes and behaviours. These latter constructs are different from the academic ability on each discipline (Yeager & Lee Duckworth, 2015).

C.2 SDG indicator 4.7.4 cognitive items

This section is pending until we receive the classification of the test items from the IEA

C.3 SDG indicator 4.7.4 non-cognitive items

The selection of items accounts for a total of 46 items. These different items were used to generate scales representing seven different constructs. The first 5 constructs have already been scaled by the ICCS team (Köhler et al., 2018), while the last two were not scaled in the ICSS 2016 database (see Table 17).

Table 17. Selected non-cognitive items to measure indicator 4.7.4

	Constructs	Number of items
e)	Students' attitudes toward their country of residence	5
f)	Students' attitudes toward equal rights for all ethnic/racial groups	5
g)	Students' attitudes toward gender rights	6
h)	Students' reports on personal experiences of bullying and abuse	6
i)	Students' perception of the importance of social movement related citizenship	4
j)	What is good for democracy	9
k)	Threats to the world's future	11
	Total	46

C.3.1 Availability

The proposed measures are available for all countries and regions who participated in ICCS 2016. These include Bulgaria, Chile, Chinese Taipei, Colombia, Croatia, Denmark, Dominican Republic, Estonia, Finland, Hong Kong SAR, Italy, Republic of Korea, Latvia, Lithuania, Malta, Mexico, Netherlands, Norway, Peru, Russian Federation, Slovenia, Sweden, Belgium (Flemish) and North Rhine-Westphalia, for a total of 24 countries and regions.

C.3.2 Unidimensionality

To assess the dimensionality of the proposed measures, we followed a two-fold strategy. We first assessed the ECV across all items, by specifying a general factor while including specific factors for each of the proposed scales. Thus, we fitted a bifactor model including all measures. Then, we assessed the ECV for each indicator category, in particular for the indicator categories of "Interconnectedness and Global Citizenship" and "Human Rights", these are the only SDG categories that were mapped into more than one ICCS scale and, therefore, the ones that could

potentially have a single score. In particular, we used a common slope GRM model (Paek & Cole, 2020), and calculate the ECV following equation (1) from this document (Reise et al., 2013). Complementary, we included a measure of common variance between the proposed items for each original scale (Brown, 2006). For the case of the common slope GRM model, this index is obtained as the square of the common slope estimate. This index expresses the average common variance on each item accounted by the specified factor.

All estimates were obtained considering the study survey sampling design using a Taylor Series Linearization. Both stratification and clusters indicators were declared for these purposes (Stapleton, 2013). Survey weights were re-scaled up to 1000 for each country (Gonzalez, 2012), so all countries contribute equally to the estimations. All estimations were carried out with MPLUS 8.3 (Muthén & Muthén, 2017), using the WLSMV estimator.

Table 18. Explained Common Variance and accounted Common Variance over the indicator 4.7.4 selected measures

Category (indicator)	Item collections	ECV	CV
Total SDG	All scales	.20	.20
Interconnectedness and Global Citizenship	Students' attitudes toward their country of residence	.23	.74
Interconnectedness and Global Citizenship	Students' attitudes toward equal rights for all ethnic/racial groups	.23	.70
Gender Equality[1]	Students' attitudes toward gender rights	---	.62
Peace, Non-violence and Human Security[1]	Students' reports on personal experiences of bullying and abuse	---	.56
Human Rights	What is good for democracy	.18	.20
Human Rights	Students' perception of the importance of social movement related citizenship	.18	.57
Sustainable Development[1]	Threats to the world future	---	.45

Note: All SDG categories flagged with [1] were operationalized by a single scale from ICCS 2016. Therefore, these categories cannot be assessed with a bifactor model that accounts for known sources of variance attributed to other known scales.

Considering the resulting ECV values from the indicator categories, it is not advisable to summarize these original scales into a single score. The common variance between the included measures in each of these categories is too low to be represented by a single score without a substantive loss of information (Quinn, 2014; Stucky & Edelen, 2015).

C.3.3 Measurement models

Assuming unidimensionality across all the proposed items is not advisable. That is, establishing a single standard with a single score, which represents the different proposed attributes that can be monitored over time in an interpretable manner is not feasible. Therefore, we estimated scores for each proposed scale individually (one scale per indicator sub-category) and proposed a provisional threshold of proficiency for each of these attributes. As a consequence, instead of producing a single standard for all the indicator 4.7.4 measures, we suggest assessing the feasibility of developing standards for each of the proposed measures.

This assessment of feasibility follows the same steps applied earlier in this document. We assess the dimensionality of the presented measures, we fit a partial credit model and produce the corresponding item-person maps, and then we classify students in reference to the provisory threshold of proficiency. Thus, we produced a section for each of the selected scales from ICCS 2016.

C.3.4 Proficiency classification

Because indicator 4.7.4 has more than one category (see Table 1), the proficiency classification exercise is organised according to each of these categories, with the scores grouped into socio-emotional and behavioural dimensions (see Table 2).

C.3.4.1 Proficiency classification of Global-local thinking (socio-emotional)

One of the proposed measures for indicator 4.7.4 for the category of "Interconnectedness and Global Citizenship" and sub-category "Global-local thinking" is the original scale of "Students' attitudes toward their country of residence" present in ICCS 2016. This scale is composed of the following items:

Figure 11. Students' attitudes toward their country of residence in ICCS 2016

Q27 How much do you agree or disagree with the following statements about <country of test>?

(Please tick only one box in each row.)

			Strongly Agree	Agree	Disagree	Strongly disagree	
IS3G27A	a)	The <flag of country of test> is important to me.	☐1	☐2	☐3	☐4	ca01
IS3G27B	b)	I have great respect for <country of test>.	☐1	☐2	☐3	☐4	ca02
IS3G27C	c)	In <country of test> we should be proud of what we have achieved.	☐1	☐2	☐3	☐4	ca03
IS3G27D	d)	I am proud to live in <country of test>.	☐1	☐2	☐3	☐4	ca04
IS3G27E	e)	Generally speaking, <country of test> is a better country to live in than most other countries.	☐1	☐2	☐3	☐4	ca05

Note: Variables names in the left side of each of the items are the original names present in public data files from ICCS 2016. In the right-hand side, we include the names ca01-ca05 to refer to the recoded responses

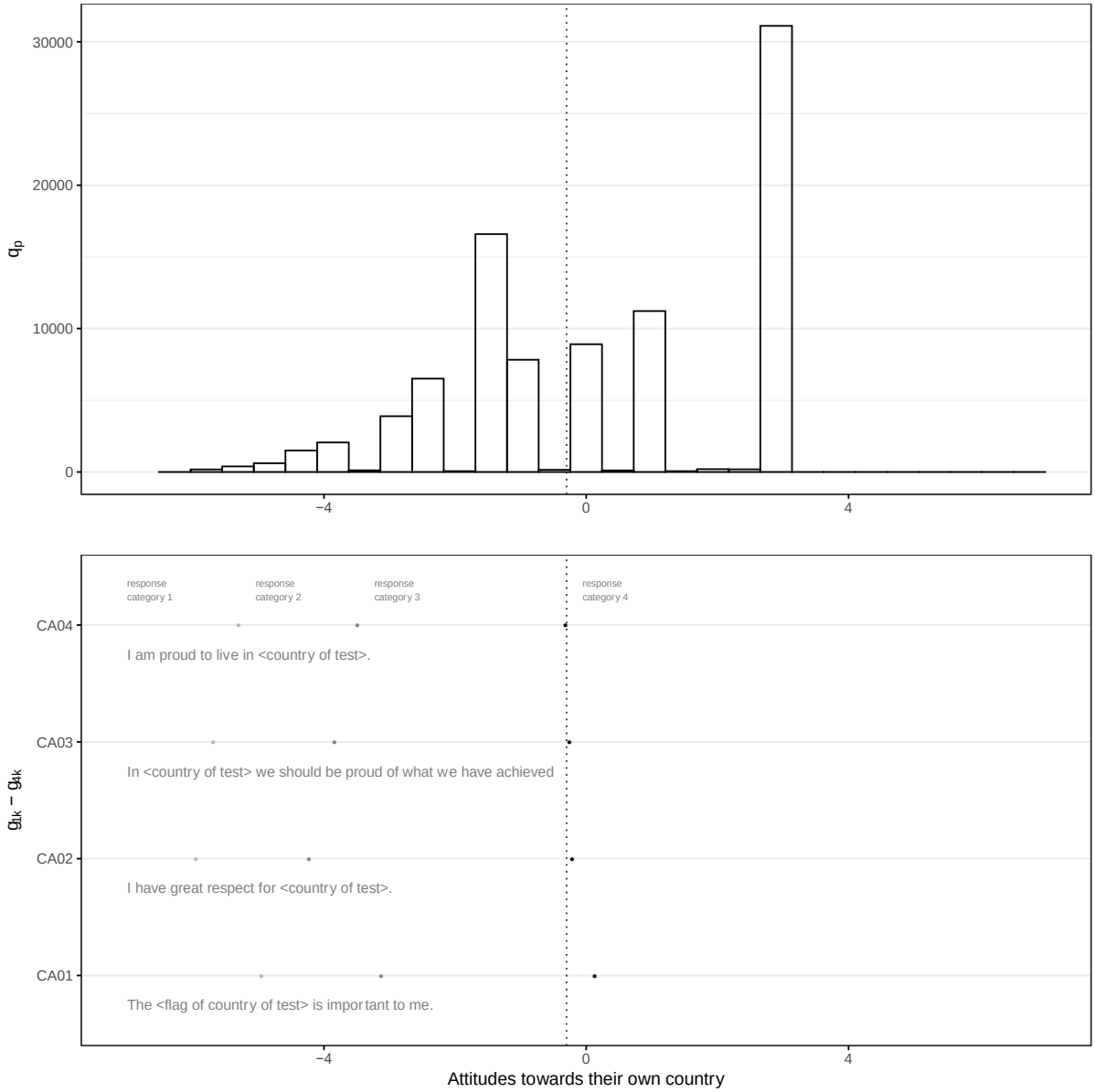
analysed in the present document. These responses were recoded so higher value expresses a higher presence of the self-reported attribute.

All the presented items assess patriotism (Janmaat & Mons, 2011), with a series of items generally expressing a positive attitude from the respondent towards their country of residence. However, item ca05 deviates from this meaning. This latter item includes a comparative component, which can elicit a response closer to the construct of nationalism instead of patriotism. This latter constructs is characterized by including derogatory components towards other countries (Mummendey et al., 2001), and is a predictor of prejudice towards immigrants when essentialist beliefs of nationality are endorsed (Pehrson, Brown, et al., 2009; Pehrson, Vignoles, et al., 2009). In the present document, this item is removed from the scale, to ensure a high quality measure of patriotism.

For the remaining items, a partial credit model is fitted. We scale survey weights up to 1000, so each country and region contributes equally to the estimates. We use Taylor Series Linearization to estimate the variance of the parameters, using pseudo strata, and primary sampling unit indicators. Person realizations are generated as Expected a Posteriori, and delta parameters are converted into Thurstonian thresholds.

The proposed threshold is presented in the following item-person map, and is located at the highest category of response after item ca04 'I am proud to live in <country of test>', which is often presented as a representative item for patriotism (Janmaat & Mons, 2011; Mummendey et al., 2001).

Figure 12. Item-person map for Students' attitudes toward their country of residence



The proposed threshold distinguishes among all participants who express highly positive attitudes towards their country of residence. These are students who feel proud of their country of residence, and express respect for their own country of residence. In terms of the response model, these are students who have 50% chances to respond “Strongly agree”, in contrast to other response categories such as “I am proud to live in <country of test>.”, “In <country of test> we should be proud of what we have achieved”, and to “I have great respect for <country of test>.”

In the next table, we estimate the proportion of students meeting the proposed standard.

Table 19. Percentage of students meeting the indicator 4.7.4 Global-local thinking (socio-emotional)

Country or Region	Percentage	Lower limit	Upper limit
Hong Kong SAR	0.22	0.20	0.23
North Rhine-Westphalia	0.29	0.26	0.32
Netherlands	0.30	0.28	0.33
Sweden	0.33	0.30	0.35
Belgium (Flemish)	0.35	0.33	0.37
Denmark	0.38	0.36	0.40
Italy	0.45	0.43	0.47
Slovenia	0.48	0.45	0.50
Estonia	0.49	0.46	0.52
Latvia	0.52	0.49	0.55
Chinese Taipei	0.52	0.50	0.54
Korea, Republic of	0.53	0.50	0.55
Finland	0.53	0.51	0.55
Lithuania	0.54	0.52	0.57
Malta	0.57	0.56	0.59
Norway	0.61	0.59	0.62
Russian Federation	0.63	0.61	0.66
Chile	0.64	0.62	0.66
Mexico	0.66	0.64	0.69
Croatia	0.68	0.65	0.71
Bulgaria	0.71	0.68	0.73
Colombia	0.76	0.74	0.78
Peru	0.79	0.77	0.80
Dominican Republic	0.87	0.86	0.89

C.3.4.2 Proficiency classification of Multicultural(ism)/intercultural(ism) (socio-emotional)

For the indicator category "Interconnectedness and Global Citizenship" and sub-category "Multicultural(ism)/intercultural(ism)", the next proposed measure is "Students' attitudes toward equal rights for all ethnic/racial groups". The present scale was created using the following items:

Figure 13. Students' attitudes toward equal rights for all ethnic/racial groups in ICCS 2016

Q25 There are different views on the rights and responsibilities of different <ethnic/racial groups> in society.

How much do you agree or disagree with the following statements?

(Please tick only one box in each row.)

			Strongly agree	Agree	Disagree	Strongly disagree	
IS3G25A	a)	All <ethnic/racial groups> should have an equal chance to get a good education in <country of test>..	☐1	☐2	☐3	☐4	et01
IS3G25B	b)	All <ethnic/racial groups> should have an equal chance to get good jobs in <country of test>.....	☐1	☐2	☐3	☐4	et02
IS3G25C	c)	Schools should teach students to respect <members of all ethnic/racial groups>.....	☐1	☐2	☐3	☐4	et03
IS3G25D	d)	<Members of all ethnic/racial groups> should be encouraged to run in elections for political office.	☐1	☐2	☐3	☐4	et04
IS3G25E	e)	<Members of all ethnic/racial groups> should have the same rights and responsibilities.	☐1	☐2	☐3	☐4	et05

Note: Variables names in the left side of each of the items are the original names present in public data files from ICCS 2016. In the right-hand side, we include the names et01-et05 to refer to the recoded responses analysed in the present document. These responses were recoded so higher value expresses a higher presence of the self-reported attribute.

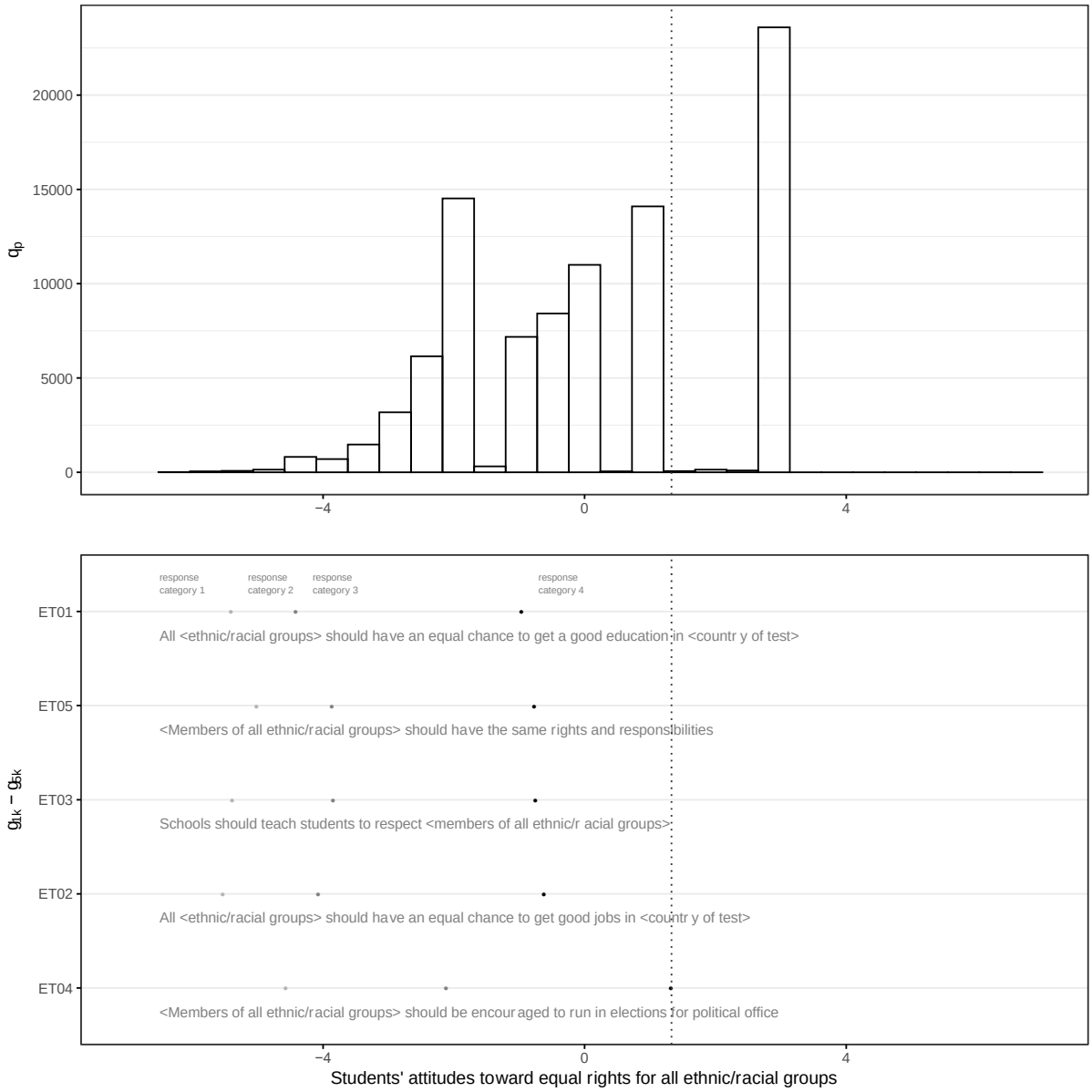
These different items measure students endorsement of equal rights and opportunities for all ethnic and racial groups (Schulz, Carstens, et al., 2018). The endorsement of equal rights to specific groups is also referred to as measures of tolerance to other groups (see Zalk & Kerr, 2014). Studies of measurement invariance of the responses to these items favour comparability between countries. Using data from ICCS 2016, the scalar model specified with a confirmatory factor analysis between countries, fits the model adequately within the acceptable range (Schulz, Carstens, et al., 2018, p. 168). Similar studies carried out using data from ICCS 2009 reach similar conclusions using items et01, et02, et03 and et05 (Miranda & Castillo, 2018).

Similarly to previous sections, we fit a partial credit model to compare the distributions of persons to the response pattern over all the proposed items. To this end, we scale survey weights up to 1000, so each country and region contributes equally to the estimates. We use Taylor Series Linearization to estimate the variance of the parameters, using pseudo strata, and primary sampling unit indicators. We generate person realizations as Expected a Posteriori, and we transform delta parameters into Thurstonian thresholds to express cumulative probabilities of the response of items. With the results of this response model, we build an item-person map.

Interpreting the presented items as a measure of tolerance to other ethnic and racial groups, we proposed the highest threshold from the present item-person map. The proposed standard is

located after the item et04 "<Members of all ethnic/racial groups>, should be encouraged to run in elections for political office".

Figure 14. Item-person map for Students' attitudes toward equal rights for all ethnic/racial groups



The proposed standard distinguishes among all students who express the highest social tolerance to other ethnic/racial groups, in contrast to the rest of the participants. Students meeting the standard present 50% chances to respond "Strongly agree", in contrast to other response categories such as "<Members of all ethnic/racial groups> should be encouraged to run in elections for political office". Considering the location on the scale of the other items, students

meeting the standard believe that all ethnic/racial groups should have equal access to education, have the same rights and responsibilities, and have equal access to the labour market.

In the following table, we present the population estimates of students meeting the proposed standard in each participating country and region.

Table 20. Percentage of students meeting the indicator 4.7.4 Multicultural(ism) or intercultural(ism) (socio-emotional)

Country or Region	Percentage	Lower limit	Upper limit
Latvia	0.09	0.08	0.10
Bulgaria	0.12	0.11	0.14
Belgium (Flemish)	0.13	0.11	0.15
Netherlands	0.13	0.11	0.15
Italy	0.15	0.14	0.16
Slovenia	0.16	0.15	0.18
Croatia	0.17	0.15	0.19
Malta	0.18	0.17	0.19
Denmark	0.20	0.19	0.22
Lithuania	0.21	0.19	0.23
Peru	0.21	0.20	0.23
Estonia	0.21	0.19	0.24
Dominican Republic	0.22	0.20	0.24
Colombia	0.22	0.21	0.24
Russian Federation	0.24	0.22	0.26
North Rhine-Westphalia	0.25	0.22	0.28
Finland	0.26	0.24	0.28
Mexico	0.27	0.26	0.29
Norway	0.38	0.36	0.40
Hong Kong SAR	0.39	0.37	0.41
Korea, Republic of	0.41	0.38	0.44
Chile	0.44	0.42	0.47
Chinese Taipei	0.45	0.43	0.47
Sweden	0.50	0.48	0.52

C.3.4.3 Proficiency classification of Gender Equality (socio-emotional)

The proposed items to measure the indicator category of Gender equality are the items present in ICCS for the scale of “Students' attitudes toward gender rights”. These items are presented in the following figure.

Figure 15. Gender Equality items in ICCS 2016

Q24 There are different views about the roles of women and men in society.
How much do you agree or disagree with the following statements?

(Please tick only one box in each row.)

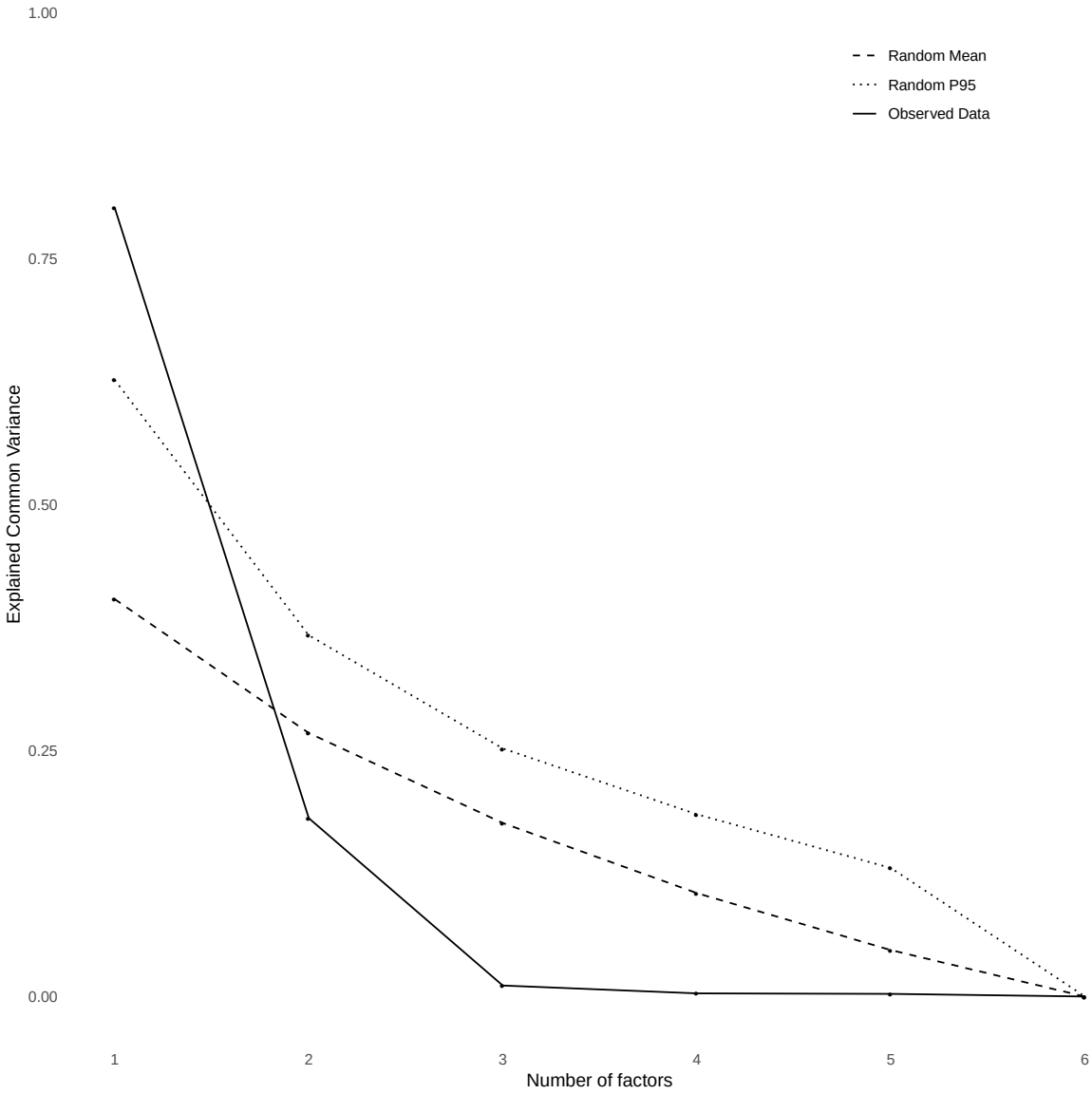
			Strongly agree	Agree	Disagree	Strongly disagree	
IS3G24A	a)	Men and women should have equal opportunities to take part in government.	☐1	☐2	☐3	☐4	ge01
IS3G24B	b)	Men and women should have the same rights in every way.	☐1	☐2	☐3	☐4	ge02
IS3G24C	c)	Women should stay out of politics.	☐1	☐2	☐3	☐4	ge04
IS3G24D	d)	When there are not many jobs available, men should have more right to a job than women.	☐1	☐2	☐3	☐4	ge05
IS3G24E	e)	Men and women should get equal pay when they are doing the same jobs.	☐1	☐2	☐3	☐4	ge03
IS3G24F	f)	Men are better qualified to be political leaders than women.	☐1	☐2	☐3	☐4	ge06
IS3G24G	g)	Women's first priority should be raising children.	☐1	☐2	☐3	☐4	

Note: Variables names in the left side of each of the items are the original names present in public data files from ICCS 2016. In the right-hand side, we include the names ge01-ge06 to referred to the recoded responses analysed in the present document. These responses were recoded so higher value expresses a higher presence of the self-reported attribute. As such, items ge04, ge05 and ge06 are reverse code items, where higher values indicate a higher endorsement of gender equality.

Responses to items ge01, ge02 and ge03 represent students support for gender rights equality (Sandoval-Hernández et al., 2018). This selection of items presents scalar invariance, allowing between-country comparison of latent means (Miranda & Castillo, 2018). Response to items ge04, ge05, and ge06, resemble hostile sexism items (Brandt, 2011; Napier et al., 2010). In essence, these are prescriptive stereotypes regarding women gender roles (Rudman & Phelan, 2007). Thus, considering the content of the items is plausible this scale contains more than one factor.

We fitted a bifactor model to compare the explained variance by the common factor, in comparison to the specific facets of sexism and gender equality support. Similar to previous sections, we specified a common slope graded response model for these purposes. The results of this exercise show that 63% of the variance is explained by the common factor, while 18% is explained by the responses to the gender equality support, and 19% is explained by the responses over the sexism items. We complement the bifactor model results with a parallel analysis. This latter procedure assesses how many latent factors are required to explain a matrix of correlations, in comparison to a set of simulated correlation matrices with a similar structure. These simulated correlation matrices serve the purposes of bringing a baseline regarding how many latent factors are expected by chance over random data. The results of this procedure suggest that the responses to these items are explained mainly by a general factor.

Figure 16. Parallel analysis results over Gender Equality items in ICCS 2016

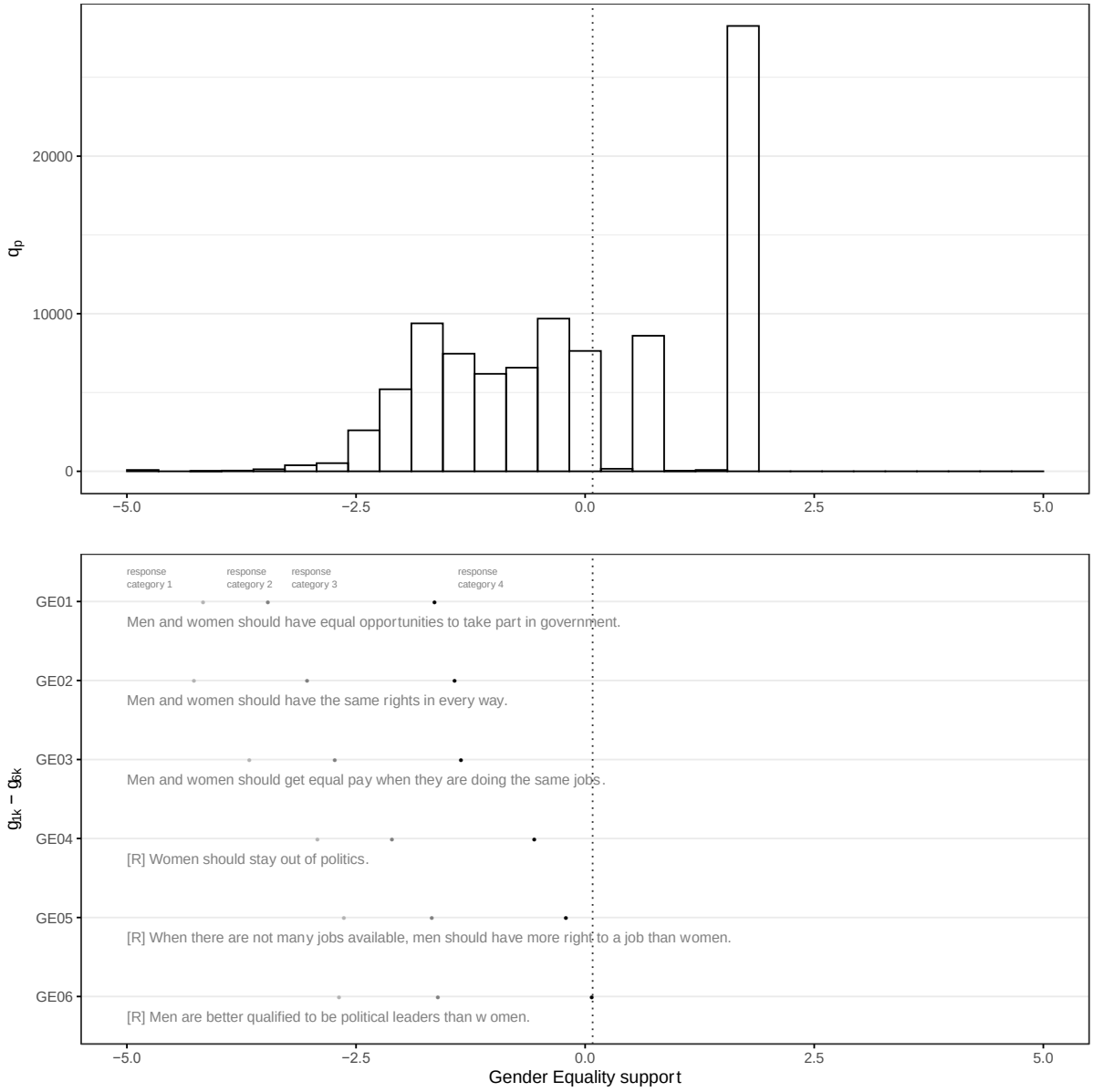


Unidimensionality assessment of the proposed items is not univocal. The bifactor model, on one hand, favours the generation of different scores per facet. One for the gender rights equality responses, and a second score for the responses over the sexism items. Nevertheless, the parallel analysis suggests specifying a single factor to account for the common variance among the proposed items. Separating the original scale into different components can imply a drawback in reliability (Cortina, 1993). Indeed, this is the case: the Expected a Posteriori reliability of the gender equality support items alone, reaches a value of .61, whereas the Expected a Posteriori reliability of the sexism items reaches a value of .76. Taking these results together, in the present report, we will assume enough dimensionality between the proposed items and used a single latent trait model. That is a single latent factor that accounts for the common variance across the responses of the proposed items.

A partial credit model for equally weighted countries, using senate weights scaled up to 1000 for each country, is used to model these responses. Taylor Series Linearization is used to estimate the variance of the parameters, using pseudo strata, and primary sampling unit indicators. Person realizations are generated as Expected a Posteriori, and delta parameters are converted into Thurstonian thresholds.

The proposed threshold is presented in the following item-person map and is located at the highest category of response, after item ge06.

Figure 17. Item-person map for Gender Equality



The proposed threshold distinguishes among all participants who present 50% chances to highly endorse gender equality and those students who are less likely to highly to endorse gender equality.

Table 21. Percentage of students meeting the indicator 4.7.4 Gender Equality (socio-emotional)

Country or Region	Percentage	Lower limit	Upper limit
Dominican Republic	0.16	0.14	0.18
Russian Federation	0.16	0.14	0.18
Mexico	0.17	0.16	0.19
Latvia	0.25	0.23	0.27
Bulgaria	0.26	0.24	0.28
Peru	0.36	0.34	0.39
Lithuania	0.37	0.34	0.39
Colombia	0.41	0.38	0.44
Hong Kong SAR	0.45	0.42	0.48
Estonia	0.47	0.44	0.51
Chile	0.52	0.50	0.54
Netherlands	0.53	0.50	0.56
Korea, Republic of	0.55	0.52	0.57
Slovenia	0.56	0.54	0.59
Malta	0.57	0.55	0.59
Croatia	0.58	0.55	0.60
Italy	0.59	0.56	0.61
Belgium (Flemish)	0.62	0.59	0.65
Finland	0.63	0.61	0.66
North Rhine-Westphalia	0.67	0.64	0.70
Chinese Taipei	0.69	0.67	0.71
Denmark	0.71	0.69	0.73
Norway	0.72	0.71	0.74
Sweden	0.74	0.71	0.76

The present measure may need some revision. That is, to evaluate if it is better to use all the items and assume unidimensionality, or to separate the selected items into different scores. The current option may have a cost regarding cross-country comparability. The ICCS 2016 Technical reports present results that favour the suspicion that the responses to these items together may not be comparable between countries (Schulz, Carstens, et al., 2018, p. 168) and some differential item functioning might be expected. The unidimensionality assessment results from the bifactor model and the parallel analysis do not converge to the same conclusion. Thus, the presented standard might need another iteration to produce more studies regarding these measures and provide further evidence to reach a more satisfactory conclusion on this regard.

C.3.4.3 Proficiency classification of Peace, Non-violence and Human Security (behavioural)

The indicator category "Peace, Non-violence and Human Security" includes measures of students bullying victimization from the ICCS 2016 scale "Students' reports on personal experiences of bullying and abuse". This scale is generated using responses from the following items:

Figure 18. Students' reports on personal experiences of bullying and abuse in ICCS 2016

Q20 During the last three months, how often did you experience the following situations at your school?

(Please tick only one box in each row.)

			Not at all	Once	2 to 4 times	5 times or more	
IS3G20A	a)	A student called you by an offensive nickname.	☐1	☐2	☐3	☐4	ab01
IS3G20B	b)	A student said things about you to make others laugh.	☐1	☐2	☐3	☐4	ab02
IS3G20C	c)	A student threatened to hurt you.	☐1	☐2	☐3	☐4	ab03
IS3G20D	d)	You were physically attacked by another student.	☐1	☐2	☐3	☐4	ab04
IS3G20E	e)	A student broke something belonging to you on purpose.	☐1	☐2	☐3	☐4	ab05
IS3G20F	f)	A student posted offensive pictures or text about you on the Internet.	☐1	☐2	☐3	☐4	ab06

Note: Variables names in the left side of each of the items are the original names present in public data files from ICCS 2016. In the right-hand side, we include the names ab01-ab06 to referred rename variables generated for this report. These responses are coded as higher values expressing a higher frequency of bullying experiences.

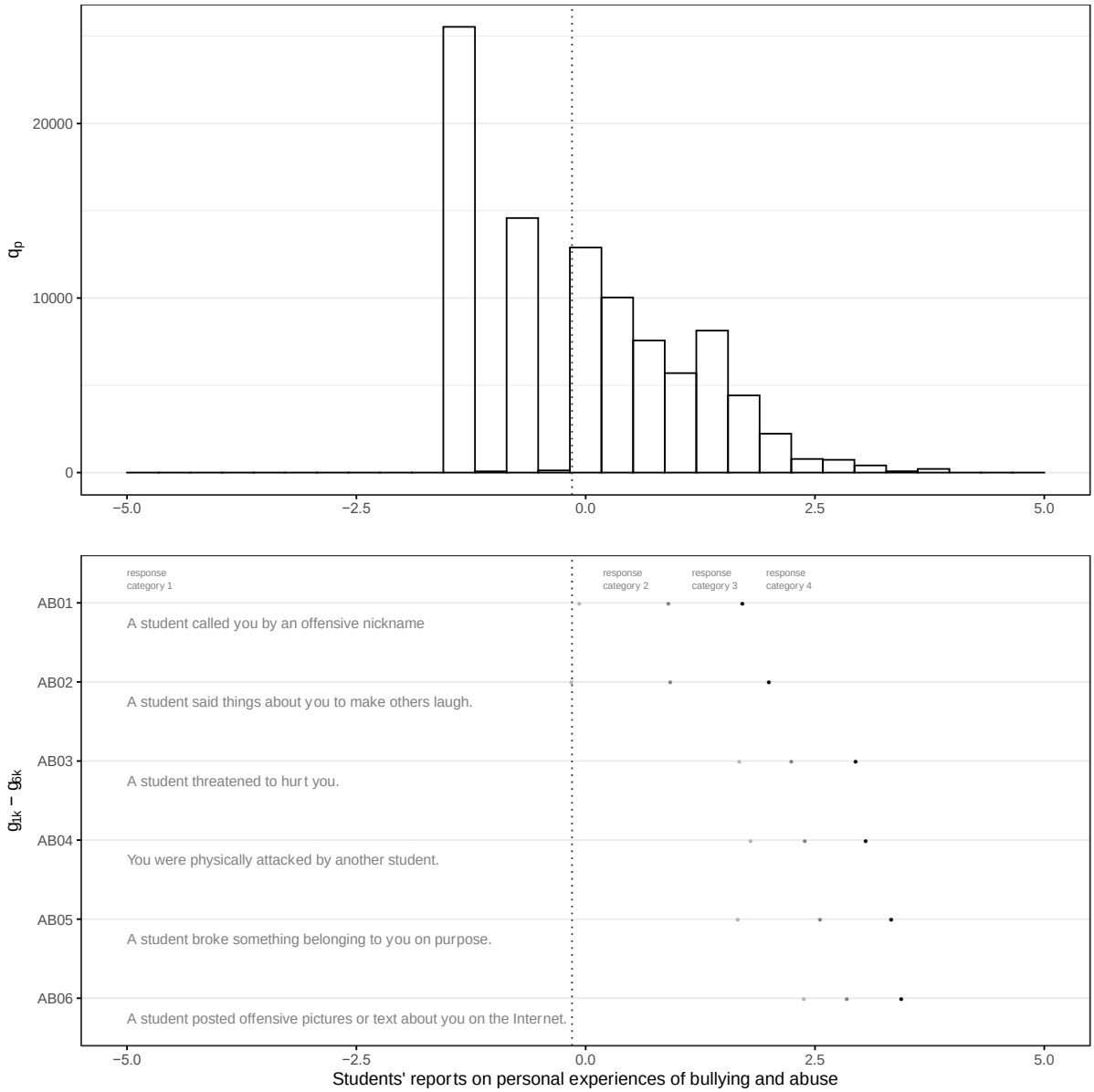
Students' experience of bullying presents adverse effects on students' wellbeing. Anxiety, depression and suicides are related to students' experience of bullying (Espelage et al., 2013; Hertz et al., 2013). Bullying victimization affects students' academic achievement in a negative way; meta-analytic estimates have found a size effect of = -.10 (Nakamoto & Schwartz, 2010). Casual inferences studies, matching students from similar characteristics, found differences between non-bullied and bullied students from 9 to 13 points in TIMSS and PIRLS 2006 among Italian students (Ponzo, 2013). Moreover, experiences of bullying have also been linked to lifelong consequences such as violence, convictions, drug user and low job status (Farrington & Ttofi, 2011; Ttofi et al., 2012). In summary, the experience of bullying at schools is a detrimental factor for youth development.

Invariance studies over the proposed items (Schulz, Carstens, et al., 2018) suggest the collected responses present a high degree of comparability between the participating countries of ICCS 2016.

We fit a partial credit model over the pooled sample. We scale survey weights up to 1000, so participant countries and regions contributed equally to estimations. To get correct standard errors, we use Taylor Series Linearization, specifying pseudo strata and primary sampling unit indicators. With the fitted model, person realizations are generated as Expected a Posteriori, and

delta parameters are converted into Thurstonian thresholds. The results are presented in the next item-person map.

Figure 19. Item-person map for Students' reports on personal experiences of bullying and abuse



In contrast to previous measures, where higher scores implied a more desired level of the attribute, the scores for students' bullying experiences are in the response direction of the items. That is, a higher score expresses a higher frequency of different bullying events. For the present measure, we proposed a threshold close to the lowest category of response. Students meeting the proposed standard have 50% chances of reporting not experiencing nickname calling and events of ridicule. All students below the proposed threshold are students with low rates of bullying experiences. As such, these are students more likely to attend a safer school environment of this regard.

In the next table, we report the population estimates of the students meeting the proposed standard.

Table 22. Percentage of students meeting the indicator 4.7.4 Peace, Non-violence and Human Security (behavioural)

Country or Region	Percentage	lower limit	upper limit
Croatia	0.33	0.31	0.35
Malta	0.36	0.34	0.37
Mexico	0.36	0.34	0.38
Hong Kong SAR	0.36	0.34	0.38
Peru	0.36	0.34	0.38
Dominican Republic	0.37	0.35	0.39
Colombia	0.37	0.35	0.40
Lithuania	0.38	0.36	0.40
Slovenia	0.39	0.37	0.41
Estonia	0.41	0.38	0.43
Norway	0.42	0.40	0.44
Bulgaria	0.42	0.39	0.45
Latvia	0.44	0.42	0.46
Chile	0.44	0.43	0.46
Belgium (Flemish)	0.45	0.42	0.47
Russian Federation	0.45	0.43	0.47
Denmark	0.47	0.45	0.49
Sweden	0.48	0.44	0.51
Italy	0.49	0.47	0.52
North Rhine-Westphalia	0.49	0.46	0.53
Finland	0.53	0.51	0.55
Netherlands	0.55	0.52	0.58
Korea, Republic of	0.59	0.56	0.62
Chinese Taipei	0.59	0.57	0.61

C.3.4.5 Proficiency classification of Freedom (of expression, of speech, of press, of association/organisation) (socio-emotional)

For the indicator category "Human Rights" and sub-category "Freedom (of expression, of speech, of press, of association/organisation)", we proposed to consider the responses to the items present in "What is good for democracy" section from ICCS 2016 (Schulz, Carstens, et al., 2018). These are presented in the following figure:

Figure 20. Students' reports on students' opinions regarding what is good for democracy in ICCS 2016

Q22 Below is a list of things that may happen in a democratic country. Some of them may be good for and strengthen democracy, some may be bad for and weaken democracy, while others are neither good nor bad for democracy.

Which of the following situations do you think would be good, neither good nor bad, or bad for democracy?

(Please tick only one box in each row.)

			Good for democracy	Neither good nor bad for democracy	Bad for democracy	
IS3G22A	a)	Political leaders give government jobs to their family members.	☐ 1	☐ 2	☐ 3	td06
IS3G22B	b)	One company or the government owns all newspapers in a country.	☐ 1	☐ 2	☐ 3	td07
IS3G22C	c)	People are allowed to publicly criticize the government.	☐ 1	☐ 2	☐ 3	td01
IS3G22D	d)	All adult citizens have the right to elect their political leaders.	☐ 1	☐ 2	☐ 3	td02
IS3G22E	e)	People are able to protest if they think a law is unfair.	☐ 1	☐ 2	☐ 3	td03
IS3G22F	f)	The police have the right to hold people suspected of threatening national security in jail without trial. ...	☐ 1	☐ 2	☐ 3	td08
IS3G22G	g)	Differences in income between poor and rich people are small.	☐ 1	☐ 2	☐ 3	td04
IS3G22H	h)	The government influences decisions by courts of justice.	☐ 1	☐ 2	☐ 3	td09
IS3G22I	i)	All <ethnic/racial> groups in the country have the same rights.	☐ 1	☐ 2	☐ 3	td05

Note: Variables names in the left side of each of the items are the original names present in public data files from ICCS 2016. In the right-hand side, we include the names td01-td09 to referred to the recoded responses analysed in the present document. These responses were recoded so higher value expresses what is good for democracy. Items td06-td09 are reverse coded items, thus, for these items, higher values express what is bad for democracy.

The proposed items measure what a democratic system should look like (Schulz, Ainley, et al., 2018) to measure students' conceptions of democracy (Judith Torney-Purta et al., 2006), and what is the meaning of democracy for students (Quaranta, 2019).

This collection of items have been present in the IEA Civic Education (CIVED) Study (J. Torney-Purta et al., 2001), and in the ICCS, with different variations (Schulz, Carstens, et al., 2018; Schulz et al., 2011). These items present less research in comparison to other items and scales present in CIVED and ICCS studies (Knowles et al., 2018). We think this is the case because these responses present low common variance in a single trait model (ECV = .20), and throughout CIVED and ICCS studies IRT scores were not generated for these items (Schulz, Carstens, et al., 2018; Schulz et al., 2011; Judith Torney-Purta et al., 2006). Thus, most of the previous research regarding these items exist using composite scores (Judith Torney-Purta et al., 2006) and descriptive results per item (Schulz, Ainley, et al., 2018; J. Torney-Purta et al., 2001; Torney-purta & Amadeo, 2004). However, two

exemptions exist. One is the work from Husfeldt & Nikolova (2003), and more recently the work of Quaranta (2019). Husfeldt & Nikolova (2003). In the next section, we describe the approaches taken by these authors to provide a sensible alternative regarding how to produce a standard for democracy conceptions measurement.

C.3.4.5.1 Previous modelling approaches

Husfeldt & Nikolova (2003) used data from CIVED 1999 and proposed three latent factors to modelled responses to a larger battery of items where most of the proposed items were included. These factors were "rights and opportunities", "limited government" and "threats to democracy". In the first factor, their work included items alluding to free speech, electing political leaders and protest against unjust laws. The second factor included items referring to free press, separation of church and state, and business having no restrictions. Finally, the third factor, for example, included items denoting nepotism, media control, coercion of justice by the government, among other indicators.

Quaranta (2019) followed a different approach. The author used a person-centered analysis to uncover interpretable patterns of responses between students. The author used a latent class analysis (Vermunt & Magidson, 2002) to reduce the observed responses to twelve items presented in ICCS 2009 of similar content to those items presented in ICCS 2016. In its research, the author found five different latent groups that distinguish students' responses regarding these items. These latent groups were named limited, free speech, minimalist, complex and uncritical. The limited group consists of students with low rates of 'strongly agree' responses to all items. The free speech group was characterized with a high rate of strongly agree only for the item referring to free speech ("Everyone should always have the right to express their opinions freely"). The minimalist group are students who strongly agree to items of free speech, that political rights should be respected for all people, and people should elect their political leaders and protest should never be violent. The complex group highly agree to items from the previous group, while also including a strong agreement to items referring no news media concentration, and people are able to criticize the government, protest against unfair laws and agree that differences in income between the rich and the poor should be small. Finally, the uncritical group are students who strongly agree to all the items, including positive and negative attributes for democratic systems. Out of these five latent groups, it seems the "complex" class seems to be the one closer to the intended interpretation of indicator4.7.4 and sub-category of Democracy/democratic rule, democratic values/principles.

C.3.4.5.2 Item response theory modelling

In the present exercise, we explore the results from a unidimensional model including all items, and separate models for two different factors, a similar approach to the one using by Husfeldt and Nikolova (2003). The generated scores using a partial credit model presented low reliability (Expected a Posteriori reliability = .57). This means respondents are too similar within this model, given the measurement error of the generated scores. We generated an IRT score including responses from items td01-td05, thus resembling factor 1 from Husfeldt and Nikolova (2003). Its resulting Expected a Posteriori reliability was also considerably low (Expected a Posteriori

reliability = .52) to provide trustworthy scores to generate standards. We proceed similarly with items td06-td09, resembling factor 3 ("threats to democracy") and we observed similar results regarding the reliability (Expected a Posteriori reliability = .56). In summary, single latent trait models for all these items and separate factors models distinguish with difficulty students' responses in a reliable manner. In conclusion, these model approaches are discarded to represent democracy conceptions of students in a reliable manner.

C.3.4.5.3 Latent class analysis modelling

In the present report, we follow the approach of Quaranta (2019) and we fit a series of latent class analysis over the proposed items, including 1 to 10 latent classes. In particular, we specified a structurally homogenous model (Kankaraš & Vermunt, 2015). In practical terms, this model's specifications searches for the same number of latent classes across countries, while keeping the types of expected response patterns across countries constant. Other models, such as the partially homogenous model specification (Kankaraš et al., 2011; Kankaraš & Vermunt, 2015), are less interpretable models because they allow the pattern of responses to be different between countries while fixing only the amount of latent classes. Therefore, this latter model allows differential item functioning for all items in all countries (Masyn, 2017). In practical terms, the structurally homogenous model specification allows the same interpretation of the pattern of responses across countries for each latent class. This property cannot be fulfilled with the partially homogenous model because it conforms to a country-specific model where all latent class can be different response patterns.

To estimate these models, we use Latent Gold 5.1 software (Vermunt & Magidson, 2013), which includes scaled survey weights (up to 1000), so that each country contributes equally to the estimates (Gonzalez, 2012). For variance estimation, we use Taylor Series Linearization specifying primary sampling unit, and pseudo strata indicators (Asparouhov & Muthén, 2010; Stapleton, 2013). Before fitting the different latent class models, we recode the responses of each proposed item as dummy variables. Items td01-td05, where the response 1 equals "Good for democracy" and the rest of the response categories were assigned a value of zero. Complementary, items td06-td09, were reverse coded, so a value of 1 was assigned to responses of "Bad for Democracy", while the rest of the response categories were coded as zero. We recoded responses in this manner to avoid cells sparseness.

In the next table, the fit indexes of the ten fitted models are displayed.

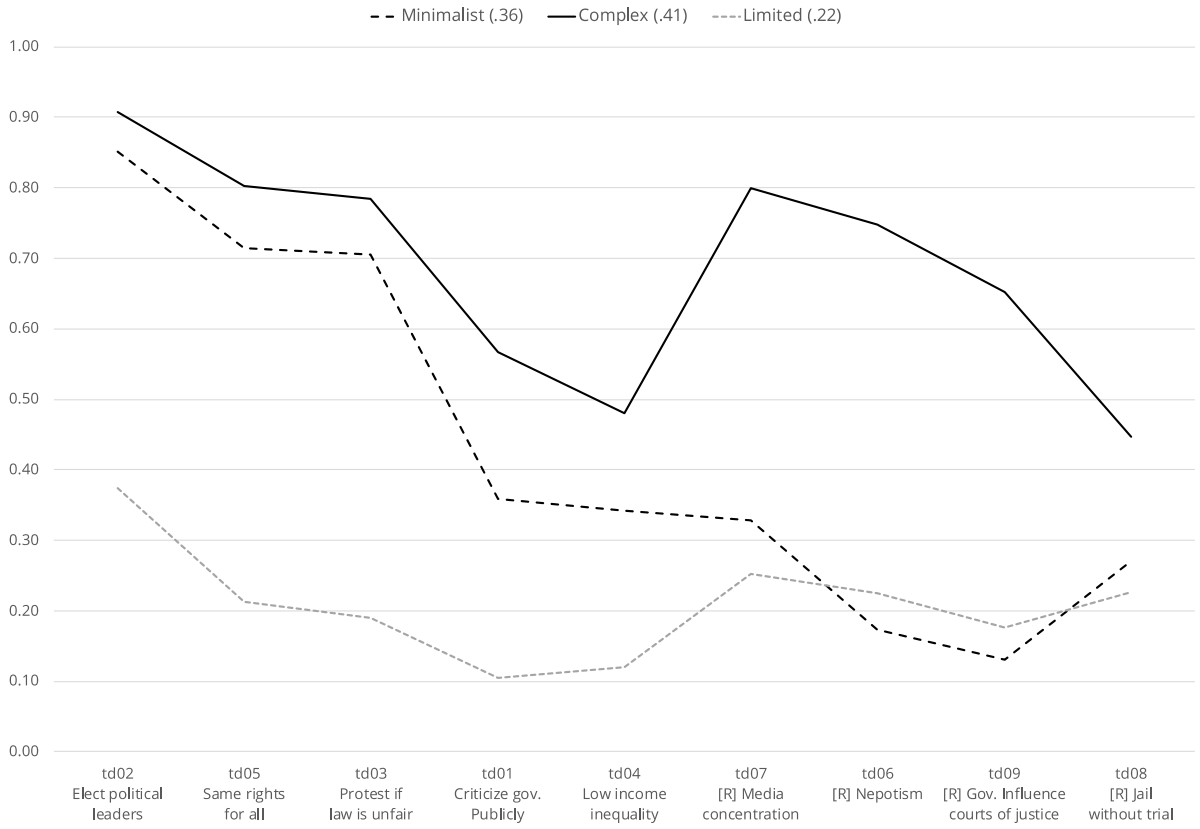
Table 23. Summary of fit indexes of the fitted latent class models

Classes	LL	BIC	Number of parameters	L^2	df	p-value	Classification error
1	-135137.91	270366.59	9	39679.30	23991	0.00	0.00
2	-127485.83	255395.25	42	24375.13	23958	0.03	0.12
3	**-125644.53**	**252045.51**	**75**	**20692.55**	**23925**	**1.00**	**0.18**
4	-124525.23	250139.74	108	18453.95	23892	1.00	0.23
5	-123904.30	249230.69	141	17212.07	23859	1.00	0.26
6	-123416.67	248588.28	174	16236.83	23826	1.00	0.29
7	-122943.33	247974.42	207	15290.14	23793	1.00	0.29
8	-122621.86	247664.31	240	14647.20	23760	1.00	0.31
9	-122300.49	247354.41	273	14004.47	23727	1.00	0.32
10	-122074.11	247234.48	306	13551.71	23694	1.00	0.33

Note: selected latent class model is highlighted in bold. LL = loglikelihood, BIC = Bayesian information criterion, L^2 = Likelihood ratio chi-square, df = degrees of freedom, p-value of the Likelihood ratio chi-square test. Classification error =

To decide which is the most appropriate number of latent classes, we assess the models in terms of their fit to the observed data. A three latent class model fits the data well, presenting a good absolute fit to the observed data (L2= 20692.55, df = 23925, p = 1.00). These fit index results mean that the observed data can be generated by a fitted model of three latent classes (Masyn, 2013). This model presents a classification error of .18, which is the lowest classification error among all the fitted models with a satisfactory fit to the observed data (models with 3-10 latent classes). In the following figure, we present the response profile of the three latent classes model.

Figure 21. Response patterns for What is good for democracy items from ICCS 2016



Because we are using a structurally homogenous across countries, the response pattern or response profile is the same across countries. What is different is the number of cases on each of these presented latent groups. To assigned names to the fitted latent classes, we used Quaranta (2019) latent group names. The minimalist group highly endorse the election of political leaders, the equal access to rights, and protest to unfair laws. However, it is a less critical group, with less than 40% of endorsement for criticizing the government, and lower rates to threats for democracy, such as media concentration, nepotism, the influence of courts of justice by the government and jailing people without trial. This group represent 36% of the students. The limited class, present low rates across all proposed items, thus failing to identify good and bad situations in democracy. In contrast, the students in the complex latent category identify as good for democracy electing political leaders, access to equal rights, and protesting if a law is unfair. Simultaneously, this group also identifies as bad for democracy news media concentration, nepotism in the government, and the influence of government over the justice system.

We proposed to use the response pattern of the latent complex group as the standard for the indicator 4.7.4 and sub-category of Democratic principles. These are students who are more likely to identify situations that are deemed good for democracy, while at the same time, they are more likely to identify situations that are bad for democracy. In the next table, we include the expected percentages of these latent group outcomes at the population level.

Table 24. Percentage of students meeting the indicator 4.7.4 Freedom (of expression, of speech, of press, of association/organisation)

Country or Region	Complex	Minimalist	Limited
Dominican Republic	0.03	0.81	0.16
Peru	0.09	0.75	0.16
Colombia	0.11	0.74	0.16
Mexico	0.11	0.63	0.26
Malta	0.25	0.46	0.28
Norway	0.30	0.51	0.19
Chile	0.34	0.41	0.25
Belgium (Flemish)	0.39	0.48	0.13
Latvia	0.40	0.10	0.50
Russian Federation	0.41	0.32	0.28
Lithuania	0.42	0.34	0.24
Bulgaria	0.42	0.34	0.23
Korea, Republic of	0.47	0.36	0.17
Italy	0.47	0.38	0.15
Sweden	0.51	0.35	0.14
Hong Kong SAR	0.51	0.25	0.24
Estonia	0.52	0.22	0.26
Netherlands	0.54	0.30	0.16
Slovenia	0.54	0.30	0.16
Denmark	0.56	0.24	0.20
Croatia	0.60	0.26	0.14
Finland	0.61	0.17	0.22
North Rhine-Westphalia	0.61	0.22	0.17
Chinese Taipei	0.77	0.06	0.18

C.3.4.6 Proficiency classification of Social Justice (socio-emotional)

The items to measure one of the components of the indicator category for “Human Rights”, sub-category “Social Justice” come from the ICCS 2016 scale “Students' perception of the importance of social movement related citizenship”. We present these items in the following figure:

Figure 22. Students' perception of the importance of social movement related citizenship from ICCS 2016

Q23 How important are the following behaviors for being a good adult citizen?

(Please tick only one box in each row.)

			Very important	Quite important	Not very important	Not important at all	
IS3G23A	a)	Voting in every national election	☐1	☐2	☐3	☐4	
IS3G23B	b)	Joining a political party	☐1	☐2	☐3	☐4	
IS3G23C	c)	Learning about the country's history	☐1	☐2	☐3	☐4	
IS3G23D	d)	Following political issues in the newspaper, on the radio, on TV or on the Internet	☐1	☐2	☐3	☐4	
IS3G23E	e)	Showing respect for government representatives	☐1	☐2	☐3	☐4	
IS3G23F	f)	Engaging in political discussions	☐1	☐2	☐3	☐4	
IS3G23G	g)	Participating in peaceful protests against laws believed to be unjust	☐1	☐2	☐3	☐4	cn01
IS3G23H	h)	Participating in activities to benefit people in the <local community>	☐1	☐2	☐3	☐4	cn02
IS3G23I	i)	Taking part in activities promoting human rights	☐1	☐2	☐3	☐4	cn03
IS3G23J	j)	Taking part in activities to protect the environment	☐1	☐2	☐3	☐4	cn04
IS3G23K	k)	Working hard	☐1	☐2	☐3	☐4	
IS3G23L	l)	Always obeying the law	☐1	☐2	☐3	☐4	
IS3G23M	m)	Ensuring the economic welfare of their families	☐1	☐2	☐3	☐4	
IS3G23N	n)	Making personal efforts to protect natural resources (e.g. through saving water or recycling waste)	☐1	☐2	☐3	☐4	
IS3G23O	o)	Respecting the rights of others to have their own opinions	☐1	☐2	☐3	☐4	
IS3G23P	p)	Supporting people who are worse off than you	☐1	☐2	☐3	☐4	
IS3G23Q	q)	Engaging in activities to help people in less developed countries	☐1	☐2	☐3	☐4	

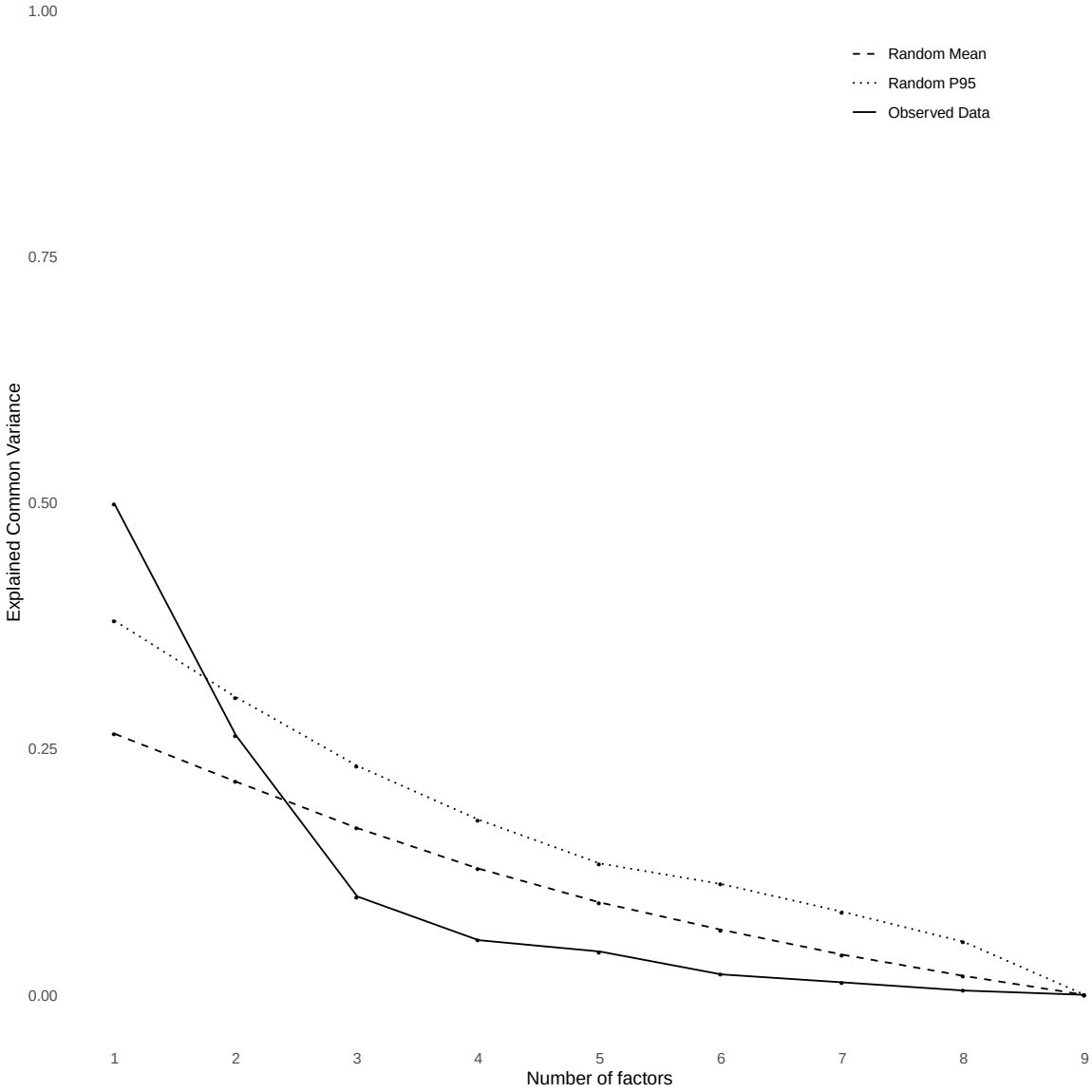
Note: Variables names in the left side of each of the items are the original names present in public data files from ICCS 2016. In the right-hand side, we include the names cn01-cn04 to refer to the recoded responses analysed in the present document. These responses were recoded so higher value expresses a higher presence of the self-reported attribute.

These different items represent the endorsement of different citizenship norms of participation in social movements. In particular, these are injunctive norms, because these items express what people ought to do, instead of what people tend to do (Cialdini & Goldstein, 2004). The content of the items resembles political participation norms with overlapping targets. Using Miranda, Castillo & Sandoval-Hernandez (2017) political participation taxonomy, we can argue that the content of the proposed items are directed to the civil society and to influence the government. “Participating in activities to benefit people in the <local community>” (cn02) is an exemplary item for civic engagement directed to civil society. In contrast, "Taking part in activities to protect the environment" (cn04) can be thought of as a civil society and government-directed action. This is

the case under the assumption that protecting the environment when it is under threat may require some changes in the law. Thus, it is challenging to participate in activities to safeguard the environment without any intention to influence the law. Finally, "Taking parts in activities promoting human rights" (cn03) and "Participating in peaceful protests against laws believed to be unjust" (cn01) can be classified as directed to influence governments because human rights guarantors are governments who adhere to the "Universal Declaration of Human Rights", and protest against unjust laws as it appeals to government bodies' decisions. All in all, these are injunctive citizenship norms regarding political involvement.

Current invariance report from these measures present in the ICCS 2016 report (Schulz, Carstens, et al., 2018) indicates a certain lack of measurement invariance. However, these results were obtained in a larger model where more items and factors were included. Parallel analysis for polytomous items (Timmerman & Lorenzo-Seva, 2011) favours a single latent trait model (see next figure).

Figure 23. Parallel analysis results over "Students' perception of the importance of social movement related citizenship" items in ICCS 2016

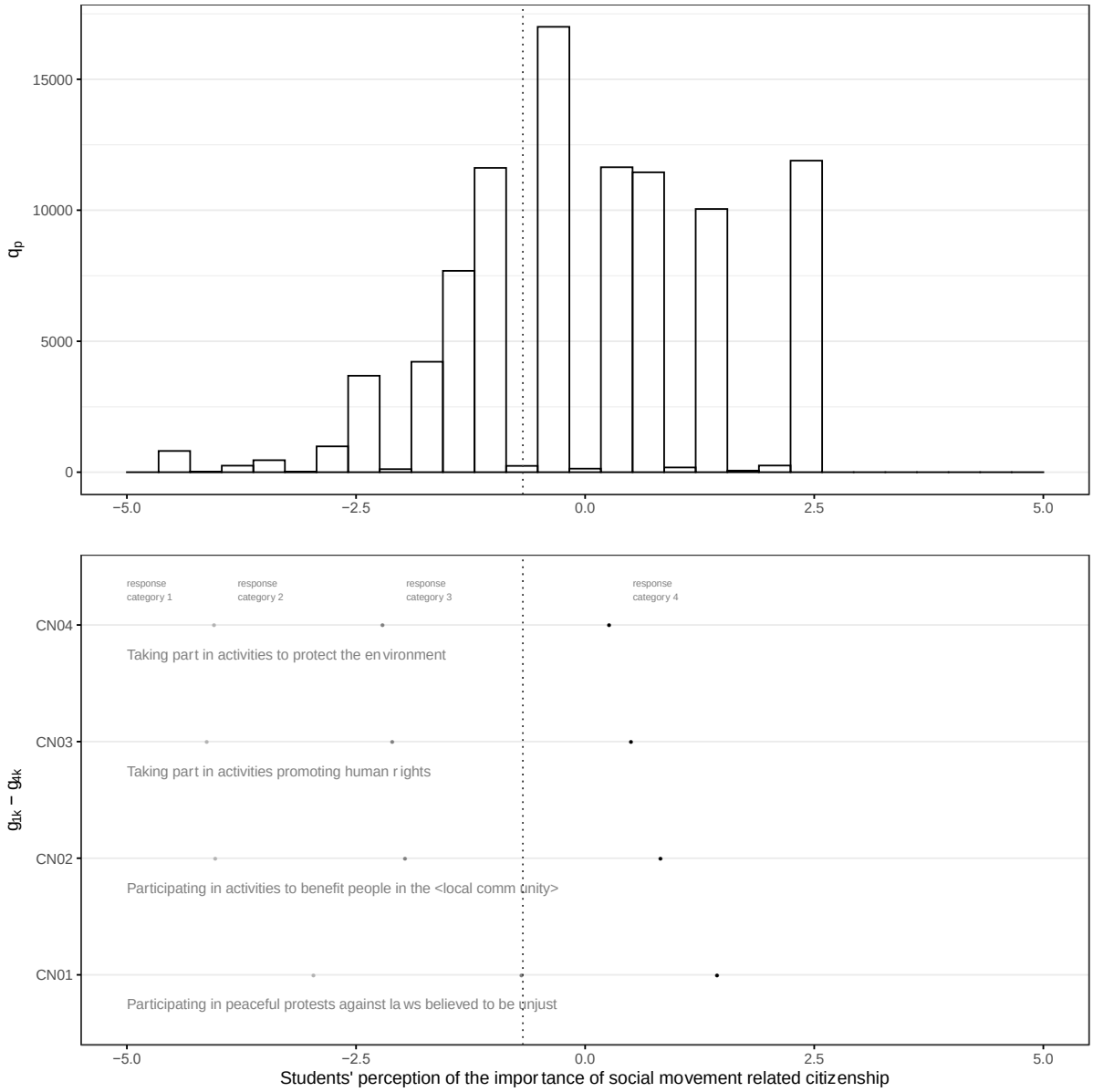


We fitted a common slope graded response model as a multigroup for all participating countries and regions and found that it fits indexes close to satisfactory results (RMSEA = .071 CI95% [.069, .072], CFI = .97, TLI = .99, SRMR = .042). As such, the proposed measures seem to reach enough invariance to compare countries.

We fit a partial credit model for equally weighted countries, using senate weights scaled up to 1000 for each country. We use Taylor Series Linearization to estimate the variance of the parameters, using pseudo strata, and primary sampling unit as indicators. Person realizations are generated as Expected a Posteriori, and delta parameters are converted into Thurstonian thresholds. With the results from this model, we produced an item-person map for this scale. As a standard, we

proposed the location parameter $\gamma_{1.2}$, of item cn01 "Participating in peaceful protests against laws believed to be unjust".

Figure 24. Item-person map for Students' perception of the importance of social movement related citizenship



This standard is similar to the classification of citizenship norms adherence from Hooghe and colleagues (Hooghe et al., 2016; Hooghe & Oser, 2015) for the latent groups of "all-around" and "engaged". These are students who highly endorsed the importance of participating in their local community, to protect the environment, promote human rights, and to protest against unjust laws. In the present measurement model, the standard distinguished between students who have 50% chances to agree that "Participating in peaceful protests against laws believed to be unjust" is quite important over previous categories of responses.

Citizenship norms are a special case of social norms and are expected to influence behaviour. Different social norms predict the likelihood to obey the law, vote and participate in protests (Gerber & Rogers, 2009; Köbis et al., 2015; Rees & Bamberg, 2014; Wenzel, 2005). The present standard is citizenship norms; these are expected to predict involvement to benefit the local community, to protect the environment, promote human rights, and to protest against unjust laws. In the next table, we provide the population estimates for students meeting this standard.

Table 25. Percentage of students meeting the indicator 4.7.4 Social Justice (socio-emotional)

Country or Region	Percentage	Lower limit	Upper limit
Denmark	0.39	0.37	0.41
Netherlands	0.40	0.38	0.42
Finland	0.53	0.52	0.55
Latvia	0.57	0.55	0.59
Estonia	0.57	0.55	0.60
North Rhine-Westphalia	0.58	0.54	0.61
Belgium (Flemish)	0.58	0.55	0.61
Lithuania	0.62	0.60	0.64
Slovenia	0.63	0.61	0.65
Malta	0.65	0.63	0.67
Sweden	0.65	0.63	0.68
Russian Federation	0.65	0.63	0.68
Hong Kong SAR	0.66	0.64	0.68
Norway	0.66	0.65	0.68
Chile	0.71	0.70	0.73
Chinese Taipei	0.74	0.72	0.76
Croatia	0.78	0.76	0.80
Italy	0.79	0.77	0.80
Korea, Republic of	0.79	0.77	0.81
Bulgaria	0.80	0.77	0.82
Mexico	0.81	0.79	0.82
Peru	0.81	0.79	0.82
Colombia	0.84	0.83	0.86
Dominican Republic	0.87	0.85	0.88

C.3.4.7 Proficiency classification of Sustainable Development (socio-emotional and behavioural)

In the indicator category of Sustainable Development, the proposed measures are 11 items from the ICCS 2016 section "Threats to the world future" (10 items), and a single item from students' future participation. These items are presented in the following figures.

Figure 25. Selected items for Sustainable Development from ICCS 2016

Q28 To what extent do you think the following issues are a threat to the world's future?

(Please tick only one box in each row.)

			To a large extent	To a moderate extent	To a small extent	Not at all	
IS3G28A	a)	Pollution	☐1	☐2	☐3	☐4	ft01
IS3G28B	b)	Energy shortages	☐1	☐2	☐3	☐4	ft02
IS3G28C	c)	Global financial crises	☐1	☐2	☐3	☐4	ft03
IS3G28D	d)	Crime	☐1	☐2	☐3	☐4	ft04
IS3G28E	e)	Water shortages	☐1	☐2	☐3	☐4	ft05
IS3G28F	f)	Violent conflict	☐1	☐2	☐3	☐4	ft06
IS3G28G	g)	Poverty	☐1	☐2	☐3	☐4	ft07
IS3G28H	h)	Food shortages	☐1	☐2	☐3	☐4	ft08
IS3G28I	i)	Climate change	☐1	☐2	☐3	☐4	ft09
IS3G28J	j)	Unemployment	☐1	☐2	☐3	☐4	ft10
IS3G28K	k)	Overpopulation	☐1	☐2	☐3	☐4	
IS3G28L	l)	Infectious diseases (e.g. <bird flu>, <AIDS>)	☐1	☐2	☐3	☐4	
IS3G28M	m)	Terrorism	☐1	☐2	☐3	☐4	

Q31 Listed below are different ways adults can take an active part in society. When you are an adult, what do you think you will do?

(Please tick only one box in each row.)

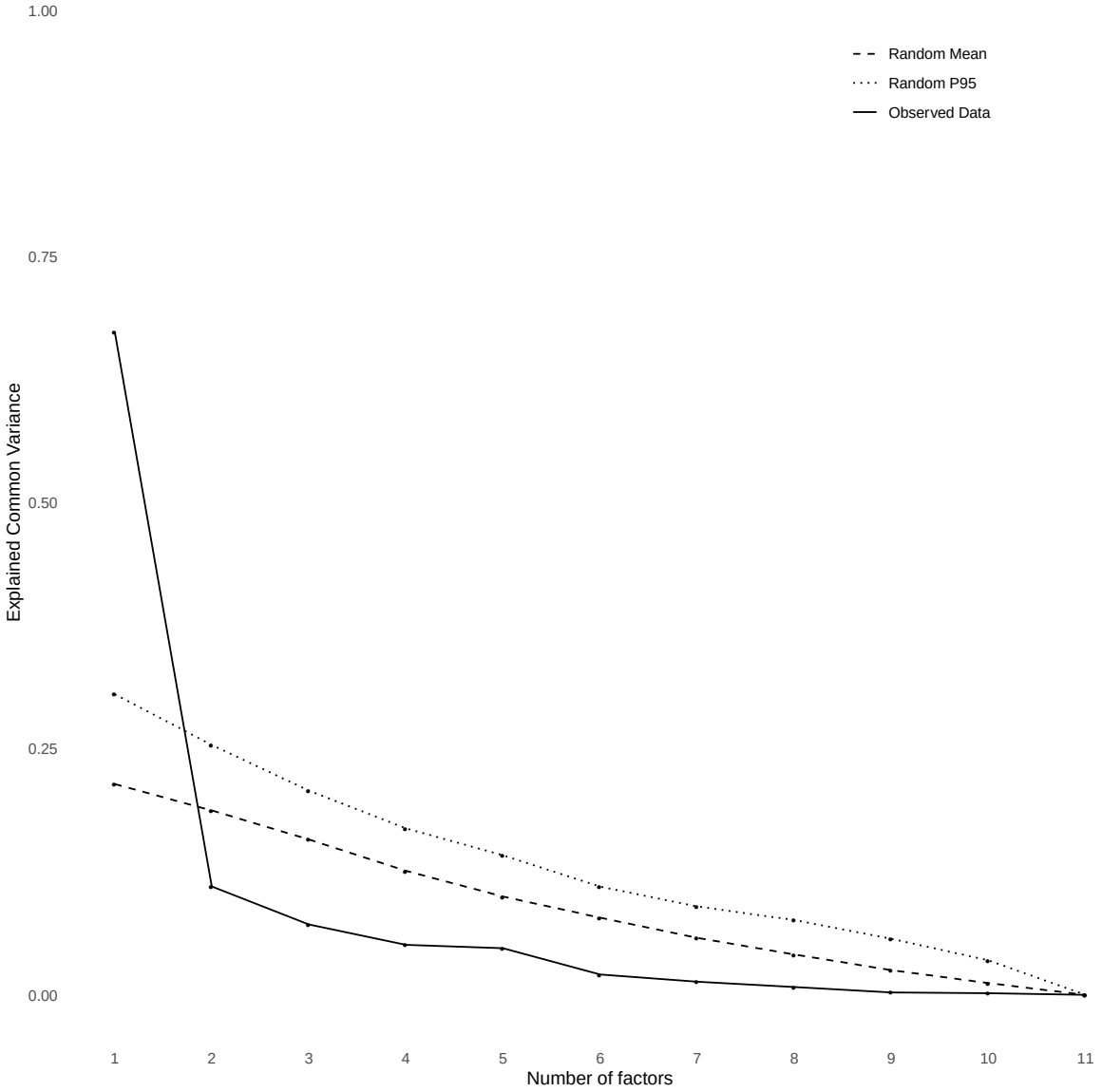
			I would certainly do this	I would probably do this	I would probably not do this	I would certainly not do this	
IS3G31A	a)	Vote in <local elections>	☐1	☐2	☐3	☐4	
IS3G31B	b)	Vote in <national elections>	☐1	☐2	☐3	☐4	
IS3G31C	c)	Get information about candidates before voting in an election	☐1	☐2	☐3	☐4	
IS3G31D	d)	Help a candidate or party during an election campaign	☐1	☐2	☐3	☐4	
IS3G31E	e)	Join a political party	☐1	☐2	☐3	☐4	
IS3G31F	f)	Join a trade union	☐1	☐2	☐3	☐4	
IS3G31G	g)	Stand as a candidate in <local elections>	☐1	☐2	☐3	☐4	
IS3G31H	h)	Join an organization for a political or social cause	☐1	☐2	☐3	☐4	
IS3G31I	i)	Volunteer time to help other people in the <local community>	☐1	☐2	☐3	☐4	
IS3G31J	j)	Make personal efforts to help the environment (e.g. through saving water)	☐1	☐2	☐3	☐4	ft11
IS3G31K	k)	Vote in <state, province elections>	☐1	☐2	☐3	☐4	
IS3G31L	l)	Vote in European elections	☐1	☐2	☐3	☐4	

Note: Variables names in the left side of each of the items are the original names present in public data files from ICCS 2016. In the right-hand side, we include the names ft01-ft11 to refer to the recoded responses

analysed in the present document. These responses were recoded so higher value expresses a higher presence of the intended attribute.

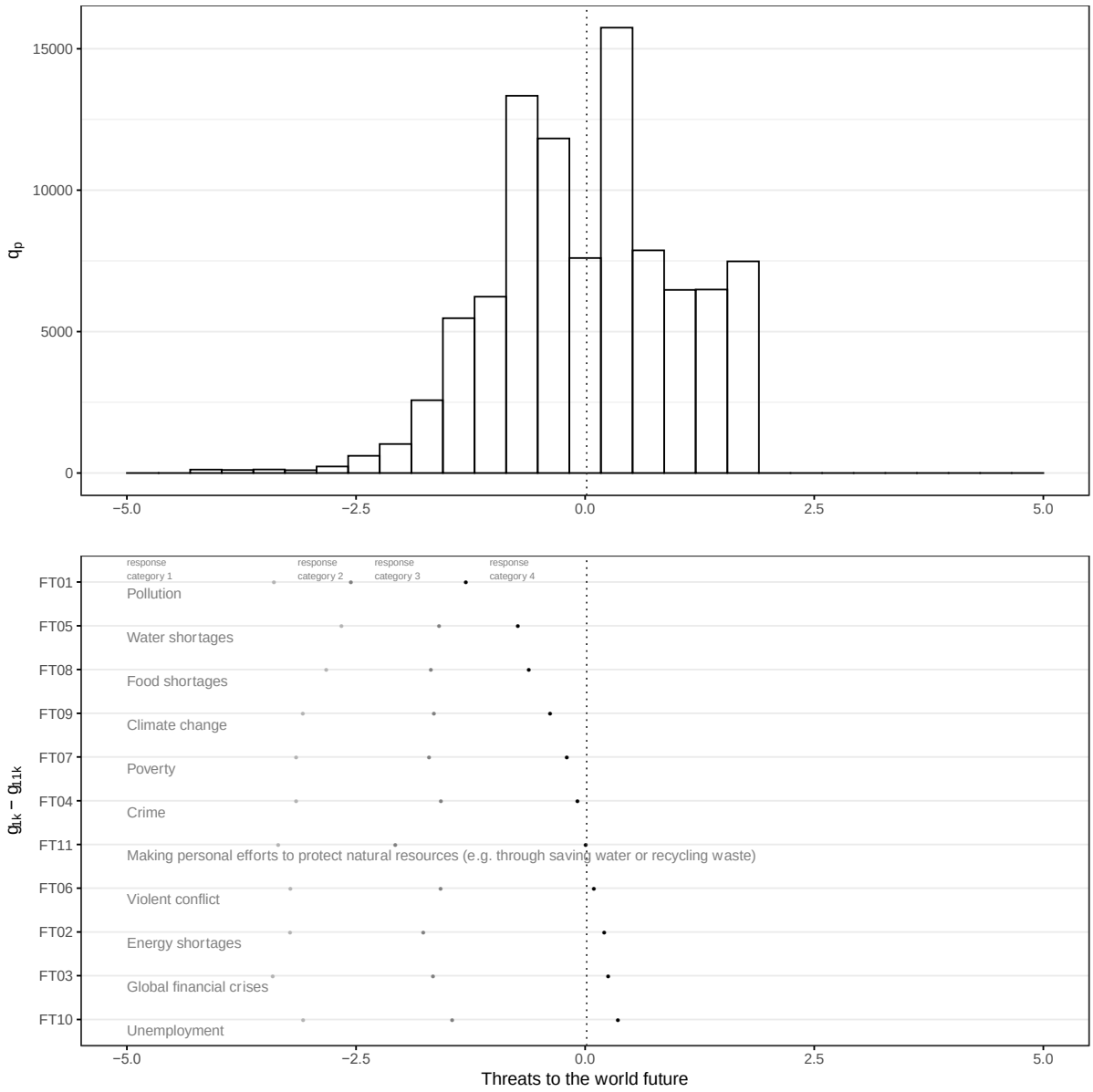
Unlike previously proposed scales, these collections of responses do not conform to a generated scale within the ICCS 2016 study (Schulz, Carstens, et al., 2018). As a consequence, there is less research about the response patterns observed for these items. So, before fitting a latent variable model, we assess the unidimensionality of the proposed items. To this end, we select a random sample of 500 cases per participating country and region, a similar procedure followed by the OECD (OECD, 2014). However, in this report, we select this random sample conditional to their survey weights. With this collection of cases, we build a calibration sample. Using this randomly selected case we produced a parallel analysis for polytomous items (Timmerman & Lorenzo-Seva, 2011). The results of this procedure favour the presence of a main factor.

Figure 26. Parallel analysis results over the proposed Sustainable Development items from ICCS 2016



We fit a partial credit model for equally weighted countries, using survey weights scaled up to 1000 observations. We use Taylor Series Linearization for variance estimation and correct standard errors, specifying pseudo strata and primary sampling unit indicators. Latent variable realizations were generated as Expected a Posteriori, and delta parameters were turned into Thurstonian thresholds. The results of this model are presented in the following figure.

Figure 27. Item-person map for Sustainable Development items



To set a standard, we rely on item ft11. Students meeting the standard present 50% chances of answering that they would definitely make personal efforts to protect natural resources. We interpreted this pattern of response as expressing a positive inclination to sustainable development. All students meeting the standard express high awareness regarding different threats to the world's future, including Pollution, Water Shortages, Food Shortages, Climate Change, Poverty and Crime. Students meeting the standard are likely to consider violent conflicts, energy shortages, global financial crises and unemployment as threats of the world's future at least, to a moderated extent. In the following table, we provide population estimates of students meeting the proposed standards.

Table 26. Percentage of students meeting the indicator 4.7.4 Sustainable Development (socio-emotional and behavioural)

Country or Region	Percentage	Lower limit	Upper limit
Netherlands	0.21	0.19	0.23
North Rhine-Westphalia	0.27	0.25	0.29
Denmark	0.29	0.28	0.31
Sweden	0.31	0.29	0.33
Norway	0.31	0.30	0.33
Finland	0.32	0.30	0.34
Estonia	0.36	0.34	0.38
Belgium (Flemish)	0.39	0.36	0.41
Peru	0.44	0.42	0.46
Russian Federation	0.44	0.42	0.46
Malta	0.44	0.43	0.46
Latvia	0.46	0.43	0.48
Dominican Republic	0.48	0.46	0.50
Korea, Republic of	0.49	0.47	0.52
Chinese Taipei	0.50	0.48	0.52
Bulgaria	0.52	0.49	0.55
Croatia	0.53	0.51	0.55
Slovenia	0.55	0.53	0.57
Italy	0.55	0.53	0.57
Hong Kong SAR	0.56	0.53	0.58
Lithuania	0.61	0.58	0.63
Mexico	0.63	0.61	0.64
Chile	0.71	0.70	0.73
Colombia	0.73	0.71	0.75

D. An overall indicator of standards met by students

In this section, we estimate the proportion of students that meet any of the standards stipulated for indicators 4.7.5 and 4.7.4, for each country and region for which data is available. To this end, we use a mean score that summarizes all the standards that a student has met. This mean score varies from 0 to 1, where the maximum is achievable by a student if and only if this student has met all the standards where he or she was classified. Zero is assigned if a student has not met any of the proposed standards. Likewise, if a student satisfies two out of three, then it receives a .66 (2/3). This calculation is expressed in the next equation:

$$\overline{D}_i = \frac{\sum_i^{n_D} D_i}{n_D} \quad (3)$$

In this equation, D_i represents a binary variable that determines if student *i* met a standard. This variable is equal to 1 if student *i* meets the standard, and a value of zero if he or she does not. n_D represents the number of standards. Because D_i is a binary variable, the mean score can be interpreted as the proportion of standards student *i* meets.

For the case of the indicator 4.7.5, where only three standards are proposed, the possible values per students are .00 (none), .33 (one out of three), .66 (two out of three), 1 (all). In the case of the indicator 4.7.4, more values are possible because more (seven) standards were proposed.

In the next tables, we include the overall mean of this mean score per country and region. The overall mean score for indicator 4.7.5 is presented in Table 27 and the overall mean score for indicator 4.7.4, in Table 28.

The interpretation of the proposed overall indicator needs some caution. First, because this is a central tendency measure at the country and region level, it is not informative of the spreading of the standards being met at the observation level. That is, a mean of .33 at the country and region level could be the result of different possible distributions, some of these could be more homogenous than others. For example, distributions where most of the cases only meet one standard would be more homogenous. In the same manner, a country and region could display a mean of .33 by having a third of the students meeting all the standards, and two thirds not fulfilling any of the proposed standards. Second, this overall indicator, should not be interpreted as a unidimensional variable. Many of the original scores used to produce these standards are not correlated to each other (e.g. Isac et al., 2014). As such, extremes values are easier to interpret than middle values. Values which are closer to 1 mean more students are likely to meet the proposed standards; and conversely, values closer to zero mean a smaller number of students meet the proposed standards. Whereas values closer to the middle of the distribution could imply a mix of students meeting some of the proposed standards, or alternatively, that some students meet all the standards and others do not.

Similar to previous sections, we produced the reported overall mean score $\overline{D}_i$ as the mean per country and region using Taylor Series Linearization to obtain corrected standard errors, including clusters and pseudo strata indicators (Stapleton, 2013).

Table 27. Mean of students meeting any of the standards SDG 4.7.5 (Science scores with selected items, SLS, SCS) TIMSS 2015

Country or Region	Percentage	Lower limit	Upper limit
Buenos Aires, Argentina	0.18	0.17	0.19
Thailand	0.19	0.17	0.20
Korea, Rep. of	0.21	0.20	0.22
Chile	0.21	0.19	0.23
Japan	0.23	0.22	0.24
South Africa	0.24	0.22	0.25
Malaysia	0.25	0.23	0.26
Saudi Arabia	0.25	0.23	0.27
Botswana	0.25	0.24	0.27
Australia	0.26	0.25	0.28
Abu Dhabi, UAE	0.27	0.24	0.30
Egypt	0.27	0.25	0.29
Chinese Taipei	0.27	0.26	0.28
New Zealand	0.27	0.26	0.29
Italy	0.28	0.27	0.30
Qatar	0.29	0.27	0.30
Hong Kong, SAR	0.29	0.27	0.31
Bahrain	0.30	0.29	0.32
England	0.31	0.29	0.33
United Arab Emirates	0.31	0.30	0.32
Jordan	0.31	0.30	0.33
Kuwait	0.32	0.30	0.34
Norway	0.32	0.30	0.33
Israel	0.32	0.30	0.34
Canada	0.32	0.31	0.33
Quebec, Canada	0.32	0.30	0.35
Ontario, Canada	0.32	0.31	0.34
Iran, Islamic Rep. of	0.32	0.31	0.34
Ireland	0.33	0.31	0.35
Norway (8th grade)	0.33	0.32	0.34
Oman	0.33	0.32	0.35
Turkey	0.36	0.34	0.38
United States	0.36	0.34	0.38
Singapore	0.38	0.36	0.39
Dubai, UAE	0.40	0.38	0.41

Table 28. Mean of students meeting any of the standards SDG 4.7.4 ICCS 2016

Country or Region	Percentage	Lower limit	Upper limit
Netherlands	0.38	0.37	0.39
Latvia	0.39	0.38	0.40
Belgium (Flemish)	0.41	0.40	0.43
Dominican Republic	0.42	0.41	0.43
Russian Federation	0.42	0.41	0.44
Denmark	0.43	0.42	0.44
Mexico	0.43	0.42	0.44
Malta	0.43	0.42	0.44
Peru	0.43	0.42	0.45
Estonia	0.43	0.42	0.45
Hong Kong SAR	0.45	0.43	0.46
Lithuania	0.45	0.44	0.46
North Rhine-Westphalia	0.45	0.44	0.46
Bulgaria	0.46	0.45	0.48
Slovenia	0.47	0.46	0.49
Norway	0.49	0.48	0.49
Finland	0.49	0.48	0.50
Colombia	0.49	0.48	0.50
Italy	0.50	0.49	0.51
Sweden	0.50	0.49	0.51
Croatia	0.52	0.51	0.54
Chile	0.55	0.53	0.56
Korea, Republic of	0.55	0.53	0.56
Chinese Taipei	0.61	0.60	0.62

Bibliography

Asparouhov, T., & Muthén, B. (2010). Resampling Methods in Mplus for Complex Survey Data. In *Mplus Technical Report*. Muthén & Muthén.

Baker, J. G., Rounds, J. B., & Zevon, M. A. (2000). A Comparison of Graded Response and Rasch Partial Credit Models with Subjective Well-Being. *Journal of Educational and Behavioral Statistics*, *25*(3), 253–270. https://doi.org/10.3102/10769986025003253

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191–215. https://doi.org/10.1037/0033-295X.84.2.191

Brandt, M. J. (2011). Sexism and Gender Inequality Across 57 Societies. *Psychological Science*, *22*(11), 1413–1418. https://doi.org/10.1177/0956797611420445

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. Guilford Press.

Cialdini, R. B., & Goldstein, N. J. (2004). Social Influence: Compliance and Conformity. *Annual Review of Psychology*, *55*(1), 591–621. https://doi.org/10.1146/annurev.psych.55.090902.142015

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting Performance Standards: Contemporary Methods. *Educational Measurement: Issues and Practice*, *23*(4), 31. https://doi.org/10.1111/j.1745-3992.2004.tb00166.x

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98–104. https://doi.org/10.1037/0021-9010.78.1.98

Cronbach, L. J. (1984). *Essentials of psychological testing* (4th ed.). Harper & Row.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

De Boeck, P., & Wilson, M. (2004). *Explanatory Item Response Models* (P. De Boeck & M. Wilson (Eds.)). Springer New York. https://doi.org/10.1007/978-1-4757-3990-9

DeMars, C. E. (2013). A Tutorial on Interpreting Bifactor Model Scores. *International Journal of Testing*, *13*(4), 354–378. https://doi.org/10.1080/15305058.2013.799067

Desjardings, C. D., & Bulut, O. (2018). *Handbook of Educational Measurement and Psychometrics Using R*. CRC Press, Taylor & Francis Group.

Diakow, R., Irribarra, D. T., & Wilson, M. (2013). Some Comments on Representing Construct Levels in Psychometric Models. In *New Developments in Quantitative Psychology* (Vol. 66, pp. 319–334). https://doi.org/10.1007/978-1-4614-9348-8_20

Engelhard, G. J., & Wind, S. A. (2018). *Invariant Measurement with raters and rating scales*. Routledge.

Espelage, D. L., Hong, J. S., Rao, M. A., & Low, S. (2013). Associations Between Peer Victimization and Academic Performance. *Theory Into Practice*, *52*(4), 233–240. https://doi.org/10.1080/00405841.2013.829724

Farrington, D. P., & Ttofi, M. M. (2011). Bullying as a predictor of offending, violence and later life outcomes. *Criminal Behaviour and Mental Health*, *21*(2), 90–98. https://doi.org/10.1002/cbm.801

Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: limited versus full information methods. *Psychological Methods*, *14*(3), 275–299. https://doi.org/10.1037/a0015825

Foy, P. (2017). *User Guide for the International Database Timss 2015*.

Gerber, A. S., & Rogers, T. (2009). Descriptive social norms and motivation to vote: Everybody's voting and so should you. *Journal of Politics*, *71*(1), 178–191. https://doi.org/10.1017/S0022381608090117

Gochyyev, P. (2015). *Essays in psychometrics and behavioral statistics*. University of California, Berkeley.

Gonzalez, E. J. (2012). Rescaling sampling weights and selecting mini-samples from large-scale assessment databases. *IERI Monograph Series Issues and Methodologies in Large-Scale Assessments*, *5*, 115–134.

Green, D. R., Trimble, C. S., & Lewis, D. M. (2003). Interpreting the Results of Three Different Standard-Setting Procedures. *Educational Measurement: Issues and Practice*, *22*(1), 22–32. https://doi.org/10.1111/j.1745-3992.2003.tb00113.x

Hertz, M. F., Donato, I., & Wright, J. (2013). Bullying and Suicide: A Public Health Approach. *Journal of Adolescent Health*, *53*(1), S1–S3. https://doi.org/10.1016/j.jadohealth.2013.05.002

Hooghe, M., & Oser, J. (2015). The rise of engaged citizenship: The evolution of citizenship norms among adolescents in 21 countries between 1999 and 2009. *International Journal of Comparative Sociology*, *56*(1), 29–52. https://doi.org/10.1177/0020715215578488

Hooghe, M., Oser, J., & Marien, S. (2016). A comparative analysis of 'good citizenship': A latent class analysis of adolescents' citizenship norms in 38 countries. *International Political Science Review*, *37*(1), 115–129. https://doi.org/10.1177/0192512114541562

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185. https://doi.org/10.1007/BF02289447

Hoskins, B. (2016). *Towards the development of an international module for assessing learning in Global Citizenship Education (GCE) and Education for Sustainable Development (ESD): A critical review of current measurement strategies*. UNESCO. https://www.gcedclearinghouse.org/sites/default/files/resources/245620e.pdf

Husfeldt, V., & Nikolova, R. (2003). Students' Concepts of Democracy. *European Educational Research Journal*, *2*(3), 396–409. https://doi.org/10.2304/eerj.2003.2.3.6

IBE. (2016). *Global Monitoring of Target 4.7: Themes in National Curriculum Frameworks: Vol. IBE/2016/WP/CD/06*. International Bureau of Education - UNESCO.

Isac, M. M., Maslowski, R., Creemers, B., & van der Werf, G. (2014). The contribution of schooling to secondary-school students' citizenship outcomes across countries. *School Effectiveness & School Improvement*, *25*(January 2015), 29–63. https://doi.org/10.1080/09243453.2012.751035

Janmaat, J. G., & Mons, N. (2011). Promoting Ethnic Tolerance and Patriotism: The Role of Education System Characteristics. *Comparative Education Review*, *55*(1), 056–081. https://doi.org/10.1086/657105

Kankaraš, M., Moors, G., & Vermunt, J. K. (2011). Testing for Measurement Invariance with Latent Class Analysis. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), *Crosscultural analysis Methods and applications* (pp. 359–384). Psychology Press. https://doi.org/10.4324/9781315537078

Kankaraš, M., & Vermunt, J. K. (2015). Simultaneous Latent-Class Analysis Across Groups. *Encyclopedia of Quality of Life and Well-Being Research*, *1974*, 5969–5974. https://doi.org/10.1007/978-94-007-0753-5_2711

Kim, S. (2016). *Continuation Ratio Model in Item Response Theory and Selection of Models for Polytomous Items* (L. A. van der Ark, M. Wiberg, S. A. Culpepper, J. A. Douglas, & W.-C. Wang (Eds.); Vol. 196, pp. 1–13). Springer International Publishing. https://doi.org/10.1007/978-3-319-38759-8_1

Knowles, R. T., Torney-Purta, J., & Barber, C. (2018). Enhancing citizenship learning with international comparative research: Analyses of IEA civic education datasets. *Citizenship Teaching and Learning*, *13*(1), 7–30. https://doi.org/10.1386/ctl.13.1.7

Köbis, N. C., Van Prooijen, J. W., Righetti, F., & Van Lange, P. A. M. (2015). "Who doesn't?" - The impact of descriptive norms on corruption. *PLoS ONE*, *10*(6), 1–14. https://doi.org/10.1371/journal.pone.0131830

Koch, T., Holtmann, J., Bohn, J., & Eid, M. (2018). Explaining general and specific factors in longitudinal, multimethod, and bifactor models: Some caveats and recommendations. *Psychological Methods*, *23*(3), 505–523. https://doi.org/10.1037/met0000146

Luo, Y. (2018). A Short Note on Estimating the Testlet Model With Different Estimators in Mplus. *Educational and Psychological Measurement*, *78*(3), 517–529. https://doi.org/10.1177/0013164417717314

Martin, M. O., Mullis, I. V. S., & Hooper, M. (2016). Methods and procedures in TIMSS 2015. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Chestnut Hill, MA: Boston College, TIMSS & PIRLS International Study Center* (Vol. 21). TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

Martin, Michael O, & Foy, P. (2016). Creating and Interpreting the TIMSS 2015 Context Questionnaire Scales Students ' Sense of School Belonging Scale ,. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and Procedures in TIMSS 2015* (pp. 15.1-15.312). TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

Masters, G. N. (2016). Partial Credit Model. In W. J. van der Linden (Ed.), *Handbook of Item Response Theory. Volume One. Models* (pp. 109–126). CRC Press.

Masyn, K. E. (2013). Latent Class analysis and finite mixture modeling. In T. D. Little (Ed.), *The Oxford Handbook of Quantitative Methods* (1st ed., Vol. 2, pp. 551–611). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199934898.013.0025

Masyn, K. E. (2017). Measurement Invariance and Differential Item Functioning in Latent Class Analysis With Stepwise Multiple Indicator Multiple Cause Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(2), 180–197. https://doi.org/10.1080/10705511.2016.1254049

Maul, A. (2017). Rethinking Traditional Methods of Survey Validation. *Measurement*, *15*(2), 51–69. https://doi.org/10.1080/15366367.2017.1348108

Miranda, D., & Castillo, J. C. (2018). *Measurement Model and Invariance Testing of Scales Measuring Egalitarian Values in ICCS 2009* (A. Sandoval-Hernández, M. M. Isac, & D. Miranda (Eds.); Vol. 4, pp. 19–31). Springer International Publishing. https://doi.org/10.1007/978-3-319-78692-6_3

Miranda, D., Castillo, J. C., & Sandoval-Hernández, A. (2017). Young Citizens Participation: Empirical Testing of a Conceptual Model. *Youth & Society*, *October*, 0044118X1774102. https://doi.org/10.1177/0044118X17741024

Mullis, I. V. S., Cotter, K., Centurino, V. A. S., Fishbein, Bethany, G., & Liu, J. (2016). Using Scale Anchoring to Interpret the TIMSS 2015 Achievement Scales. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in TIMSS 2015* (pp. 1–47). TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

Mullis, I. V. S., & Martin, M. O. (Eds.). (2017). *TIMSS 2019 Assessment Frameworks*. TIMSS & PIRLS International Study Center and IEA.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016). Exhibit 2.8: Descriptions of the TIMSS 2015 International Benchmarks of Mathematics Achievement 625. In M. Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper (Ed.), *TIMSS 2015 International Results in Mathematics*.

Mummendey, A., Klink, A., & Brown, R. (2001). Nationalism and patriotism: National identification and out-group rejection. *British Journal of Social Psychology*, *40*(2), 159–172. https://doi.org/10.1348/014466601164740

Muthén, L. K., & Muthén, B. (2017). *Mplus User's Guide* (Eight). Muthén & Muthén.

Nakamoto, J., & Schwartz, D. (2010). Is peer victimization associated with academic achievement? A meta-analytic review. *Social Development*, *19*(2), 221–242. https://doi.org/10.1111/j.1467-9507.2009.00539.x

Napier, J. L., Thorisdottir, H., & Jost, J. T. (2010). The joy of sexism? A multinational investigation of hostile and benevolent justifications for gender inequality and their relations to subjective well-being. *Sex Roles*, *62*(7–8), 405–419. https://doi.org/10.1007/s11199-009-9712-7

OECD. (2014). *PISA 2012 Technical Report*. OECD Publishing.

Osborne, J., Simon, S., & Collins, S. (2003). Attitudes towards science: A review of the literature and its implications. *International Journal of Science Education*, *25*(9), 1049–1079. https://doi.org/10.1080/0950069032000032199

Paek, I., & Cole, K. (2020). *Using R for Item Response Theory Model Applications*. Routledge.

Pehrson, S., Brown, R., & Zagefka, H. (2009). When does national identification lead to the rejection of immigrants? Cross-sectional and longitudinal evidence for the role of essentialist in-group definitions. *British Journal of Social Psychology*, *48*(1), 61–76. https://doi.org/10.1348/014466608X288827

Pehrson, S., Vignoles, V. L., & Brown, R. (2009). National identification and anti-immigrant prejudice: Individual and contextual effects of national definitions. *Social Psychology Quarterly*, *72*(1), 24–38. https://doi.org/10.1177/019027250907200104

Ponzo, M. (2013). Does bullying reduce educational achievement? An evaluation using matching estimators. *Journal of Policy Modeling*, *35*(6), 1057–1078. https://doi.org/10.1016/j.jpolmod.2013.06.002

Quaranta, M. (2019). What makes up democracy? Meanings of democracy and their correlates among adolescents in 38 countries. *Acta Politica*, *0123456789*. https://doi.org/10.1057/s41269-019-00129-4

Quinn, H. O. C. (2014). *Bifactor Models, Explained Common Variance (ECV), and the Usefulness of Scores From Unidimensional Item Response Theory Analyses*. The University of North Carolina at Chapel Hill, Chapel Hill, NC.

Rabe-Hesketh, S., & Skrondal, A. (2012). *Multilevel and Longitudinal Modeling Using Stata, Volumes I and II, Third Edition* (3rd ed.). Stata Press.

Rees, J. H., & Bamberg, S. (2014). Climate protection needs societal change: Determinants of intention to participate in collective climate action. *European Journal of Social Psychology*, *44*(5), 466–473. https://doi.org/10.1002/ejsp.2032

Reise, S. P. (2012). The Rediscovery of Bifactor Measurement Models. *Multivariate Behavioral Research*, *47*(5), 667–696. https://doi.org/10.1080/00273171.2012.715555

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and Structural Coefficient Bias in Structural Equation Modeling A Bifactor Perspective.

Educational and Psychological Measurement, *73*(1), 5–26. https://doi.org/10.1177/0013164412449831

Ricker, K. L. (2006). Setting cut-scores: A critical review of the Angoff and modified Angoff methods. *Alberta Journal of Educational Research*, *52*(1), 53–64.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, *21*(2). https://doi.org/10.1037/met0000045

Rudman, L. A., & Phelan, J. E. (2007). Sex Differences, Sexism, and Sex: The Social Psychology of Gender from Past to Present. *Advances in Group Processes*, *24*, 19–45. https://doi.org/10.1016/S0882-6145(07)24002-0

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International Large-Scale Assessment Data: Issues in Secondary Analysis and Reporting. *Educational Researcher*, *39*(2), 142–151. https://doi.org/10.3102/0013189X10363170

Samejima, F. (2016). Graded Response Models. In W. J. van der Linden (Ed.), *Handbook of Item Response Theory. Volume One. Models* (pp. 95–107). CRC Press. https://doi.org/10.1201/9781315374512-16

Sandoval-Hernández, A., Isac, M. M., & Miranda, D. (2018). *Teaching Tolerance in a Globalized World* (A. Sandoval-Hernández, M. M. Isac, & D. Miranda (Eds.); Vol. 4). Springer International Publishing. https://doi.org/10.1007/978-3-319-78692-6

Sandoval-Hernández, A., Isac, M. M., & Miranda, D. (2019). *Measurement Strategy for SDG Global Indicator 4.7.1 and Thematic Indicators 4.7.4 and 4.7.5 using International LargeScale Assessments in Education*. UNESCO Institute for Statistics. http://gaml.uis.unesco.org/wp-content/uploads/sites/2/2019/08/GAML6-REF-9-measurement-strategy-for-4.7.1-4.7.4-4.7.5.pdf

Sandoval-Hernández, A., & Miranda, D. (2018). *Exploring ICCS 2016 to measure progress toward target 4.7*. UNESCO.

Schulz, W., Ainley, J., & Fraillon, J. (2011). *ICCS 2009 Technical Report* (W. Schulz, J. Ainley, & J. Fraillon (Eds.)). International Association for the Evaluation of Educational Achievement (IEA).

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*. Springer International Publishing.

Schulz, W., Carstens, R., Losito, B., & Fraillon, J. (2018). *ICCS 2016 Technical Report* (W. Schulz, R. Carstens, B. Losito, & J. Fraillon (Eds.)). International Association for the Evaluation of Educational Achievement (IEA).

Stapleton, L. M. (2013). Incorporating Sampling Weights into Single- and Multilevel Analyses. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of International Large scale*

Assessment: background, technical issues, and methods of data analysis (pp. 363–388). Chapman and Hall/CRC.

Stucky, B. D., & Edelen, M. O. (2015). Using hierarchical IRT models to create unidimensional measures from multidimensional data. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of Item Response Theory modeling* (pp. 183–206). Routledge. https://doi.org/10.4324/9781315736013.ch9

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods*, *16*(2), 209–220. https://doi.org/10.1037/a0023353

Toland, M. D., Sulis, I., Giambona, F., Porcu, M., & Campbell, J. M. (2017). Introduction to bifactor polytomous item response theory analysis. *Journal of School Psychology*, *60*. https://doi.org/10.1016/j.jsp.2016.11.001

Torney-purta, J., & Amadeo, J. (2004). *Strengthening Democracy in the Americas through Civic Education: An Empirical Analysis Highlighting the Views of Students and Teachers*. Organization of American States.

Torney-Purta, J., Lehmann, R., Oswald, H., & Schulz, W. (2001). *Citizenship and education in twenty-eight countries: Civic knowledge and engagement at age fourteen.* Eburon Publishers, Delft, Netherlands.

Torney-Purta, Judith, Barber, C., & Wilkenfeld, B. (2006). Differences in the civic knowledge and attitudes of adolescents in the United States by immigrant status and hispanic background. *Prospects*, *36*(3), 343–354. https://doi.org/10.1007/s11125-006-0015-2

Torres Irribarra, D., Diakow, R., Freund, R., & Wilson, M. (2015). Modeling for Directly Setting Theory-Based Performance Levels. *Psychological Test and Assessment Modeling*, *57*(3), 396.

Ttofi, M. M., Farrington, D. P., & Lösel, F. (2012). School bullying as a predictor of violence later in life: A systematic review and meta-analysis of prospective longitudinal studies. *Aggression and Violent Behavior*, *17*(5), 405–418. https://doi.org/10.1016/j.avb.2012.05.002

UIS. (2017). *Measurement strategy for SDG Target 4.7: Vol. GAML4/17*. UNESCO Institute for Statistics.

UNESCO. (2012a). *Education for Sustainable Development Sourcebook*. UNESCO.

UNESCO. (2012b). *Exploring Sustainable Development: a Multiple-Perspective Approach*. UNESCO.

UNESCO. (2013). *Global Citizenship Education: An Emerging Perspective, Outcome document of the Technical Consultation on Global Citizenship Education*. UNESCO.

UNESCO. (2014). *Global Citizenship Education, Preparing Learners for the 21st Century*. UNESCO.

UNESCO. (2015). *Global Citizenship Education, Topics and Learning Objectives*. UNESCO.

Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. In J. Hagenaars & A. McCutcheon (Eds.), *Applied latent class analysis* (pp. 89–106). Cambridge University Press.

Vermunt, J. K., & Magidson, J. (2013). *LG-Syntax User's Guide: Manual for Latent GOLD 5.0 Syntax Module*. Statistical Innovations Inc.

Wenzel, M. (2005). Misperceptions of social norms about tax compliance: From theory to intervention. *Journal of Economic Psychology*, *26*(6), 862–883. https://doi.org/10.1016/j.joep.2005.02.002

Wigfield, A., Eccles, J. S., Fredricks, J. A., Simpkins, S., Roeser, R. W., & Schiefele, U. (2015). Development of Achievement Motivation and Engagement. In *Handbook of Child Psychology and Developmental Science* (pp. 1–44). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118963418.childpsy316

Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach*. Lawrence Erlbaum Associates, Publishers.

Wilson, M., & Draney, K. (2002). A Technique for Setting Standards and Maintaining Them over Time. *Measurement and Multivariate Analysis*, 325–332. https://doi.org/10.1007/978-4-431-65955-6_35

Wu, M., Tam, H. P., & Jen, T.-H. (2016). Partial Credit Model. In *Educational Measurement for Applied Researchers* (pp. 159–185). Springer Singapore. https://doi.org/10.1007/978-981-10-3302-5_9

Wyse, A. E. (2013). Construct Maps as a Foundation for Standard Setting. *Measurement*, *11*(4), 139–170. https://doi.org/10.1080/15366367.2013.850287

Yeager, D. S., & Lee Duckworth, A. (2015). Measurement Matters: Assessing Personal Qualities Other Than Cognitive Ability for Educational Purposes. *Educational Researcher*, *44*(4), 237–251. https://doi.org/10.3102/0013189X15584327.Measurement

Zalk, M. H. W. van, & Kerr, M. (2014). Developmental Trajectories of Prejudice and Tolerance Toward Immigrants from Early to Late Adolescence. *Journal of Youth and Adolescence*, *43*(10), 1658–1671. https://doi.org/10.1007/s10964-014-0164-1

Zieky, M., & Perie, M. (2006). *A primer on setting cut scores on tests of educational achievement excerpts from passing scores : A manual for setting standards of performance on educational and occupational tests*.

Appendix I. MPLUS syntax for Gender Equality Items

Code	Comments
TITLE: pcm geneql; DATA: FILE = sgd_474.dat; VARIABLE: NAMES = ge01 !item 1 ge02 !item 2 ge03 !item 3 ge04 !item 4 ge05 !item 5 ge06 !item 6 id_i !id case id_k !id country id_s !id strata id_r !id pseudo cluster id_j !id school ws ! senate weight scaled up to 1000 ;	Section to give a title for the model, specify the data file and the variable names in the data file. All text after a "!" is interpreted as a comment within MPLUS syntax. As such, these have no effect on model specification. In the current code, these are included as notes to remind the analyst of the content of each data vector.
MISSING=.; CATEGORICAL = ge01 (gpcm) ge02 (gpcm) ge03 (gpcm) ge04 (gpcm) ge05 (gpcm) ge06 (gpcm) ; USEVARIABLES = ge01 ge02 ge03 ge04 ge05 ge06 ; ! id variable IDVARIABLE = id_i;	In this section, variables are declared. In the categorical section, items are declared as a categorical and the term "(gpcm)" is used. This latter specification allows MPLUS to fit a partial credit model by using an adjacent logit between the response categories to the items.
!survey method taylor WEIGHT = ws; STRATIFICATION = id_s; CLUSTER = id_r; ANALYSIS: TYPE = COMPLEX; ESTIMATOR = MLR; STSEED = 382;	Section to describe the variance method for estimation. In this example, Taylor Series Linearization is specified.

MODEL: !loadings theta_p by ge01-ge06@1; !variance theta_p; !latent mean [theta_p@0]; !delta [ge01$1]; [ge01$2]; [ge01$3]; [ge02$1]; [ge02$2]; [ge02$3]; [ge03$1]; [ge03$2]; [ge03$3]; [ge04$1]; [ge04$2]; [ge04$3]; [ge05$1]; [ge05$2]; [ge05$3]; [ge06$1]; [ge06$2]; [ge06$3];	In the model section, the latent variable *model* is specified, based in results from Figure 2 presented in the current document: 1. Loadings are constrained to 1 2. Variance of theta_p is freely estimated 3. Latent mean is centred Delta parameters are declared in the model, yet these are assumed by the model and it is not necessary to declare these. These lines are redundant for the code to run the partial credit model. These are declared in this code for clarity, so there is no ambiguity regarding which parameters are fixed and which are freely estimated.
OUTPUT: CINTERVAL RESIDUAL ; !item characteristic curves PLOT: TYPE = PLOT3; !saves realizations of theta_p SAVEDATA: SAVE = FSCORES; FILE = sgd_474_geneql_eap.dat;	The output is requesting the estimates, the confidence interval of the estimates and the residuals of the model. The PLOT: statement, is requesting item characteristic curves and other IRT plots. Finally, the SAVEDATA command is saving the latent realizations of theta_p, used to generate the item-person maps. These are Expected a Posteriori values of the random term theta_p.

Appendix II. Executive summary

In this report, we use data from ICCS, TIMSS and PISA to estimate the proportion of students who reach the targets set by SDG Thematic Indicators 4.7.4 and 4.7.5 for each country and region with available data. In what follows, we briefly describe our analytical strategy, the description of the content and types of cognitive processing skills and strategies demonstrated by students at the cut-off points estimated for each target, and present summary tables with the proportion of students who reach each of the specified targets in each country or region.

Analytical strategy

The analytical strategy includes five main steps: verify the availability of observed responses to the items proposed by the mapping exercise described above (Sandoval-Hernández et al., 2019), test the unidimensionality of the intended constructs, fit the corresponding measurement models to obtain scores for each target, estimate the cut-off points to identify the students who reach each of the targets evaluated.

To obtain the scores, we use a latent variable model approach. More specifically, we use a partial credit model (Masters, 2016).[9] Formally, this model can be described as follows (see Wu et al., 2016):

$$Pr(Y_{ip} = j|\theta_p) = \frac{\exp\sum_{k=0}^{j}(\theta_p - \delta_{ik})}{\sum_{h=0}^{m_i} exp\sum_{k=0}^{h}(\theta_p - \delta_{ik})} \quad (1)$$

The proportion of students reaching the targets within each country or region is then calculated as a simple proportion.

$$P = \frac{X}{n}$$

We also estimate the proportion of students who meet any of the standards stipulated by indicators 4.7.5 and 4.7.4, for each country and region for which data is available. To this end, we use a mean score that summarizes all the standards that a student has met. This mean score varies from 0 to 1, where the maximum is achievable by a student if and only if this student has met all the standards where he or she was classified. Zero is assigned if a student has not met any of the proposed standards. Likewise, if a student satisfies two out of three, then he or she is attributed a score of .66 (2/3). This calculation is expressed in the next equation:

$$\overline{D}_i = \frac{\sum_i^{n_D} D_i}{n_D} \quad (3)$$

[9] The exception is indicator 4.7.4 sub-category 'Freedom', for which we used a series of latent class analysis. See the main report for details.

Description of cut-off points

4.7.4 – Percentage of students by age group (or education level) showing an adequate understanding of issues relating to global citizenship and sustainability.

COGNITIVE (4.7.4)

This section is pending until we receive the classification of the test items from the IEA

NON-COGNITIVE (4.7.4)

Interconnectedness and Global Citizenship

This category is measured through two sub-categories: 'Global-local thinking' and 'Multicultural(ism)/intercultural(ism)'.

Global-local thinking

At the threshold, students have more than 50% chances to express positives attitudes towards their country of residence. Most of the students at or above the cut-off score agree a lot to expressions such as "I am proud to live in <country of test>.", "In <country of test> we should be proud of what we have achieved", or "I have great respect for <country of test>."

Multicultural(ism)/intercultural(ism)

At the threshold, students have more than 50% chances to express positives attitudes towards ethnic/racial minorities. Most of the students at or above the cut-off score agree a lot to expressions such as "<Members of all ethnic/racial groups> should be encouraged to run in elections for political office", "<Members of all ethnic/racial groups> should have equal access to education", or "<Members of all ethnic/racial groups> should have equal chances to get a good job in <country of test>."

Gender Equality

At the threshold, students have more than 50% chances to strongly endorse gender equality. Most of the students at or above the cut-off score agree a lot to expressions such as "Men and women should have equal opportunities to take part in government" or "Men and women should get equal pay when they are doing the same jobs". Complementary, most of the students at or above the cut-off score express strong disagreement to expressions such as "Women should stay out of politics" or "Men are better qualified to be political leaders than women".

Peace, Non-violence and Human Security

At the threshold, students have more than 50% chances of reporting not experiencing bullying. Most of the students at or above the cut-off score report not having experienced at all

situations such as "being called by an offensive nickname", "being threatened to be hurt", or "other students posting offensive pictures or texts about them".

Human Rights

This category is measured through two sub-categories: 'Freedom (of expression, of speech, of press, of association/organisation)' and 'Social Justice'.

Freedom (of expression, of speech, of press, of association/organisation)

At the threshold, students have more than 50% chances of identifying situations that are deemed good for democracy, as well as those situations that are deemed bad for democracy. Most of the students at or above the cut-off score consider that situations like "People are allowed to publicly criticise the government" or "All adult citizens have the right to elect their political leaders" are good for democracy. Complementary, most of the students at or above the cut-off score consider that situations like "Political leaders give government jobs to their family members" or "One company or the government owns all newspapers in the country" are bad for democracy.

Social Justice

At the threshold, students have more than 50% chances to highly endorse the importance of social participation in social movements. Most of the students at or above the cut-off score consider that behaviours such as "Participating in protests against laws believed to be unjust", "Participating in activities to benefit people in the local community" or " Taking part in activities to protect the environment" are very important for being a good citizen.

Sustainable Development

At the threshold, students have more than 50% chances of identifying threats to the world's future and reporting that they would definitely make personal efforts to avoid them. Most of the students at or above the cut-off score consider that, to a large extent, issues like "Pollution", "global financial crisis", "Violent conflicts" or "climate change" are a threat to the world's future; and that they would certainly make personal efforts to help the environment.

4.7.5 – Percentage of 15-year-old students showing proficiency in knowledge of environmental science and geoscience

COGNITIVE (4.7.5)

At the threshold, students apply and communicate their understanding of concepts from environmental science and geoscience in everyday and abstract situations. They communicate their understanding of ecosystems and the interaction of organisms with their environment and apply some knowledge of human health related to nutrition and infectious disease. Students show some knowledge and understanding of the composition and properties of matter and chemical change. They apply knowledge of Earth's physical features, processes, cycles, and history, and show some understanding of Earth's resources, their use, and conservation as well as some knowledge of the interaction between the Earth and the Moon.

NON-COGNITIVE (4.7.5)

Enjoy environmental science and geoscience

At the threshold, students have more than 50% chances to express high enjoyment of learning environmental science and geoscience. Most of the students at or above the cut-off score agree a lot to expressions such as "I like to conduct science experiments", "I learn many interesting things in science" or "I like Science". Complementary, most of the students at or above the cut-off score express disagreement to expressions such as "Science is boring" or "I wish I did not have to study science".

Confidence in environmental science and geoscience

At the threshold, students have more than 50% chances to report high confidence in learning environmental science and geoscience. Most of the students at or above the cut-off score highly disagree with the statement "Science makes me confused", and express agreement to statements such as "I learn things quickly in science", "I usually do well in science", or "I'm good to work out difficult science problems".

Summary tables

Table 29A. Proportion of students reaching the targets of indicator 4.7.5

Country	Cognitive	Non-Cognitive		Global %
		Enjoyment	Confidence	
Abu Dhabi, UAE	0.19	0.33	0.29	0.27
Armenia	0.24			
Australia	0.34	0.24	0.21	0.26
Bahrain	0.21	0.37	0.32	0.3
Botswana	0.07	0.51	0.18	0.25
Buenos Aires, Argentina	0.13	0.18	0.22	0.18
Canada	0.39	0.29	0.29	0.32
Chile	0.18	0.25	0.19	0.21
Chinese Taipei	0.55	0.16	0.11	0.27
Dubai, UAE	0.36	0.44	0.38	0.4
Egypt	0.06	0.44	0.31	0.27
England	0.39	0.28	0.25	0.31
Georgia	0.13			
Hong Kong, SAR	0.45	0.26	0.16	0.29
Hungary	0.38			
Iran, Islamic Rep. of	0.18	0.43	0.36	0.32
Ireland	0.39	0.28	0.3	0.33
Israel	0.34	0.25	0.37	0.32
Italy	0.31	0.24	0.31	0.28
Japan	0.49	0.13	0.07	0.23
Jordan	0.11	0.49	0.34	0.31
Kazakhstan	0.37			
Korea, Rep. of	0.45	0.09	0.09	0.21
Kuwait	0.12	0.43	0.39	0.32
Lebanon	0.1			
Lithuania	0.35			
Malaysia	0.21	0.46	0.07	0.25
Malta	0.24			
Morocco	0.07			
New Zealand	0.36	0.27	0.19	0.27
Norway	0.33	0.24	0.34	0.32
Oman	0.17	0.45	0.36	0.33
Ontario, Canada	0.37	0.3	0.29	0.32
Qatar	0.2	0.34	0.31	0.29
Quebec, Canada	0.42	0.25	0.29	0.32
Russian Federation	0.45			
Saudi Arabia	0.07	0.37	0.31	0.25
Singapore	0.59	0.34	0.2	0.38
Slovenia	0.5			
South Africa	0.05	0.41	0.25	0.24
Sweden	0.41			
Thailand	0.16	0.31	0.09	0.19
Turkey	0.25	0.46	0.37	0.36
United Arab Emirates	0.24	0.37	0.32	0.31
United States	0.4	0.32	0.35	0.36

Table 30A. Table 1A. Proportion of students reaching the targets of indicator 4.7.4

Country	Cognitive	Non-Cognitive							Global %
		Global-local	Multiculturalism	Gender equality	Peace	Freedom	Social justice	Sustainable dev.	
Belgium (Flemish)		0.35	0.13	0.62	0.45	0.39	0.58	0.39	0.41
Bulgaria		0.71	0.12	0.26	0.42	0.42	0.8	0.52	0.46
Chile		0.64	0.44	0.52	0.44	0.34	0.71	0.71	0.55
Chinese Taipei		0.52	0.45	0.69	0.59	0.77	0.74	0.5	0.61
Colombia		0.76	0.22	0.41	0.37	0.11	0.84	0.73	0.49
Croatia		0.68	0.17	0.58	0.33	0.6	0.78	0.53	0.52
Denmark		0.38	0.2	0.71	0.47	0.56	0.39	0.29	0.43
Dominican Republic		0.87	0.22	0.16	0.37	0.03	0.87	0.48	0.42
Estonia		0.49	0.21	0.47	0.41	0.52	0.57	0.36	0.43
Finland		0.53	0.26	0.63	0.53	0.61	0.53	0.32	0.49
Hong Kong SAR		0.22	0.39	0.45	0.36	0.51	0.66	0.56	0.45
Italy		0.45	0.15	0.59	0.49	0.47	0.79	0.55	0.5
Korea, Republic of		0.53	0.41	0.55	0.59	0.47	0.79	0.49	0.55
Latvia		0.52	0.09	0.25	0.44	0.4	0.57	0.46	0.39
Lithuania		0.54	0.21	0.37	0.38	0.42	0.62	0.61	0.45
Malta		0.57	0.18	0.57	0.36	0.25	0.65	0.44	0.43
Mexico		0.66	0.27	0.17	0.36	0.11	0.81	0.63	0.43
Netherlands		0.3	0.13	0.53	0.55	0.54	0.4	0.21	0.38
North Rhine-Westphalia		0.29	0.25	0.67	0.49	0.61	0.58	0.27	0.45
Norway		0.61	0.38	0.72	0.42	0.3	0.66	0.31	0.49
Peru		0.79	0.21	0.36	0.36	0.09	0.81	0.44	0.43
Russian Federation		0.63	0.24	0.16	0.45	0.41	0.65	0.44	0.42
Slovenia		0.48	0.16	0.56	0.39	0.54	0.63	0.55	0.47
Sweden		0.33	0.5	0.74	0.48	0.51	0.65	0.31	0.5