Annex A6. Are PISA mathematics scores comparable across countries and languages?

The validity and reliability of PISA scores, and their comparability across countries and languages are the key concerns that guide the development of assessment instruments and selection of the statistical model for scaling students' responses. The procedures used by PISA to meet these goals include qualitative reviews conducted by national experts on the final main study items and statistical analyses of model fit in the context of multi-group item-response-theory models, which indicate the measurement equivalence of each item across groups defined by country and language.

Countries' preferred items

National mathematics experts conducted qualitative reviews of the full set of items included in the PISA 2022 assessment at different stages of their development. The ratings and comments submitted by national experts determined the revision of items and coding guides for the main study, and guided the final selection of the item pool. In many cases, these changes mitigated cultural concerns and improved test fairness.

At the end of 2021, the PISA consortium asked national experts to confirm or revise their original ratings with respect to the final instruments. Sixty-eight national centres submitted ratings of the relevance of PISA 2022 mathematics items for measuring students' "preparedness for life" – a key aspect of the validity of PISA (response options were: "not relevant", "somewhat relevant", "highly relevant"). National experts also indicated whether the specific competences addressed by each item were within the scope of official curricula ("not in curriculum", "in some curricula", "standard curriculum material"). While PISA does not intend to measure only what students learn as part of the school curriculum, ratings of curriculum coverage for PISA items provide contextual indicators to understand countries' strengths and weaknesses in the assessment.

On average across countries/economies, 81% of items were rated as "highly relevant for students' preparedness for life" (the highest possible rating); only 2% received a low rating on this dimension (rating equal to 1, i.e. "not relevant").

On the other hand, national experts indicated high overlap between national curricula and the PISA mathematics item set. On average, 86% of items were rated as "standard curriculum material", and only 3% of items were identified as "not in curriculum". National experts from five countries – Kazakhstan, Norway, Peru, the Philippines, and Thailand – indicated that all items used in PISA could be considered standard curriculum material in their country.

Table I.A6.1 provides a summary of the ratings received from national centres about the PISA 2022 set of reading items.

Table I.A6.1. How national experts rated PISA mathematics items

Percentage of test items, by rating

		In curriculum?			Relevant to "preparedness for life"?		
		Not in curriculum	In some curricula	Standard curriculum material	Not at all relevant	Somewhat relevant	Highly relevant
		(%)	(%)	(%)	(%)	(%)	(%)
OECD	**Austria**	6.4	9.2	84.5	8.0	23.9	68.1
	Belgium	0.4	4.2	95.5	4.2	6.4	89.4
	Canada	5.3	18.6	76.1	4.2	20.5	75.4
	Chile	0.0	1.9	98.1	15.5	15.5	68.9
	Colombia	4.2	24.6	71.2	1.1	13.6	85.2
	Costa Rica	14.0	16.3	69.7	16.3	35.6	48.1
	Czech Republic	1.5	5.3	93.2	0.4	11.4	88.3
	Denmark	1.5	6.8	91.6	0.0	12.2	87.8
	Estonia	0.5	0.0	99.5	0.5	2.7	96.7
	Finland	0.0	9.1	90.9	0.0	6.4	93.6
	Germany	14.6	35.4	50.0	15.7	38.6	45.7
	Greece	2.3	8.7	89.0	0.0	0.4	99.6
	Hungary	0.4	6.4	93.2	12.1	19.3	68.6
	Iceland	0.0	1.9	98.1	0.0	6.1	93.9
	Ireland	0.0	3.0	97.0	1.1	9.1	89.8
	Israel	6.5	4.2	89.4	4.2	22.8	73.0
	Italy	5.3	0.8	93.9	2.7	8.0	89.4
	Korea	0.0	3.2	96.8	3.8	61.3	34.9
	Lithuania	0.0	1.5	98.5	1.1	10.2	88.6
	Mexico	7.2	6.4	86.5	1.6	11.2	87.3
	New Zealand	4.9	4.9	90.2	2.3	37.9	59.8
	Norway	0.0	0.0	100.0	0.4	0.0	99.6
	Poland	12.1	14.8	73.1	0.0	3.8	96.2
	Portugal	0.8	7.2	92.0	0.0	7.2	92.8
	Slovak Republic	2.3	3.4	94.3	3.0	3.4	93.6
	Slovenia	1.1	17.8	81.1	0.0	25.4	74.6
	Spain	1.1	13.3	85.6	3.0	25.4	71.6
	Sweden	0.8	23.1	76.1	0.0	33.3	66.7
	Switzerland	3.0	54.2	42.8	0.0	37.1	62.9
	Türkiye	1.1	0.4	98.5	0.4	1.1	98.5
	United Kingdom (Excl. Scotland)	4.2	12.5	83.3	3.0	22.0	75.0
	United States	3.3	14.3	82.4	3.3	25.8	70.9
Partners	Albania	3.4	12.5	84.1	3.0	4.5	92.4
	Argentina	3.0	17.8	79.2	0.4	19.0	80.6
	Brazil	0.4	0.0	99.6	0.4	7.2	92.4
	Brunei Darussalam	2.0	2.0	96.1	9.8	52.9	37.3
	Bulgaria	4.5	34.8	60.6	2.3	39.4	58.3
	Croatia	0.0	0.4	99.6	0.0	3.0	97.0
	Cyprus	3.8	11.0	85.2	0.0	8.7	91.3
	Dominican Republic	0.0	50.0	50.0	0.4	37.9	61.7
	El Salvador	20.2	0.0	79.8	0.8	9.9	89.3
	Georgia	2.3	4.5	93.2	0.0	15.9	84.1
	Hong Kong (China)	0.0	1.9	98.1	1.9	3.0	95.1
	Indonesia	2.7	5.7	91.7	1.9	10.2	87.9
	Jamaica	5.7	0.8	93.4	0.4	11.5	88.1
	Jordan	0.4	6.4	93.2	0.0	6.1	93.9
	Kazakhstan	0.0	0.0	100.0	0.0	0.0	100.0
	Macao (China)	1.5	62.9	35.6	0.0	42.0	58.0
	Malaysia	3.0	17.5	79.5	0.0	9.1	90.9
	Moldova	1.1	1.9	97.0	0.0	5.7	94.3
	Mongolia	1.5	3.4	95.1	3.0	36.0	61.0
	Montenegro	3.4	6.4	90.2	0.4	6.8	92.8
	Morocco	7.5	33.2	59.3	0.8	30.6	68.7
	Palestinian Authority	4.9	34.1	61.0	2.3	14.8	83.0
	Panama	0.0	1.5	98.5	0.0	10.2	89.8
	Peru	0.0	0.0	100.0	0.0	5.7	94.3
	Philippines	0.0	0.0	100.0	0.0	1.5	98.5
	Qatar	4.9	3.8	91.3	0.8	3.4	95.8
	Romania	10.6	6.8	82.6	1.1	12.1	86.7
	Saudi Arabia	0.0	6.8	93.2	0.0	94.3	5.7
	Serbia	7.6	0.4	92.0	1.1	14.0	84.8
	Singapore	11.8	16.7	71.5	11.8	38.0	50.2
	Chinese Taipei	25.0	14.8	60.2	0.0	5.3	94.7
	Thailand	0.0	0.0	100.0	0.0	3.8	96.2
	Ukraine	0.4	15.5	84.1	0.4	2.7	97.0
	United Arab Emirates	0.4	12.1	87.5	8.7	22.3	68.9
	Uruguay	1.9	14.0	84.1	0.4	9.1	90.5
	Uzbekistan	0.4	6.8	92.8	0.8	2.7	96.6
	Overall average	3.4	10.6	86.0	2.3	16.5	81.2

Note: Percentages may not add to 100% due to rounding. Percentages are reported as a proportion of all test items that received a rating. For countries that delivered the test on paper, only ratings for trend items were considered. Countries and economies that are not included in this table did not submit ratings on the final set of items.

National item deletions, item misfit, and item-by-country interactions

PISA reporting scales in mathematics, reading and science are linked across countries, survey cycles and delivery modes (paper and computer) through common items whose parameters are constrained to the same values and which can therefore serve as "anchors" on the reporting scale. A large number of anchor items support the validity of cross-country comparisons and trend comparisons.

The unidimensional multi-group item-response-theory (IRT) models used in PISA, with groups defined by language within countries and by cycle also result in model-fit indices for each item-group combination. These indices can indicate tensions between model constraints and response data, a situation known as "misfit" or "differential item functioning" (DIF).

In cases where the international parameters for a given item did not fit well for a particular country or language group, or for a subset of countries or language groups, PISA allowed for a "partial invariance" solution in which the equality constraints on the item parameters were released and group-specific item parameters were estimated. This approach was favoured over dropping the group-specific item responses for these items from the analysis in order to retain the information from these responses. While items with DIF treated in this way no longer contribute to the international set of comparable responses, they help reduce measurement uncertainty for the specific country-by-language group.

In rare instances where the partial invariance model was not sufficient to resolve the tension between students' responses and the IRT model, the group-specific response data for that particular item were dropped.

An overview of the number of international/common (invariant) item parameters and group-specific item parameters in mathematics for PISA 2022 is given in Figure I.A6.1 and Figure I.A6.2; the corresponding figures for other domains can be found in the PISA 2022 Technical Report (OECD, Forthcoming[1]). Each set of stacked bars in these figures represents a country or economy; countries and economies with multiple language groups have one bar for each country-by-language group.

The bars represent the items used in the country. A colour code indicates whether international item parameters were used in scaling ("invariant items"), or whether, due to misfit when using international parameters, national item parameters were used. For items where international equality constraints were released, a distinction is made between two groups:

- group-specific new items: items that received unique parameters for the particular group defined by country/language and year (in many cases, equality constraints across a subset of misfit groups defined by country/language and year, e.g. across all language groups in a country, could be implemented)
- group-specific trend items: items for which the "non-invariant" item parameters used in 2022 could be constrained to the same values used in 2018 for the particular country/language group (these items contribute to measurement invariance over time but not across groups).

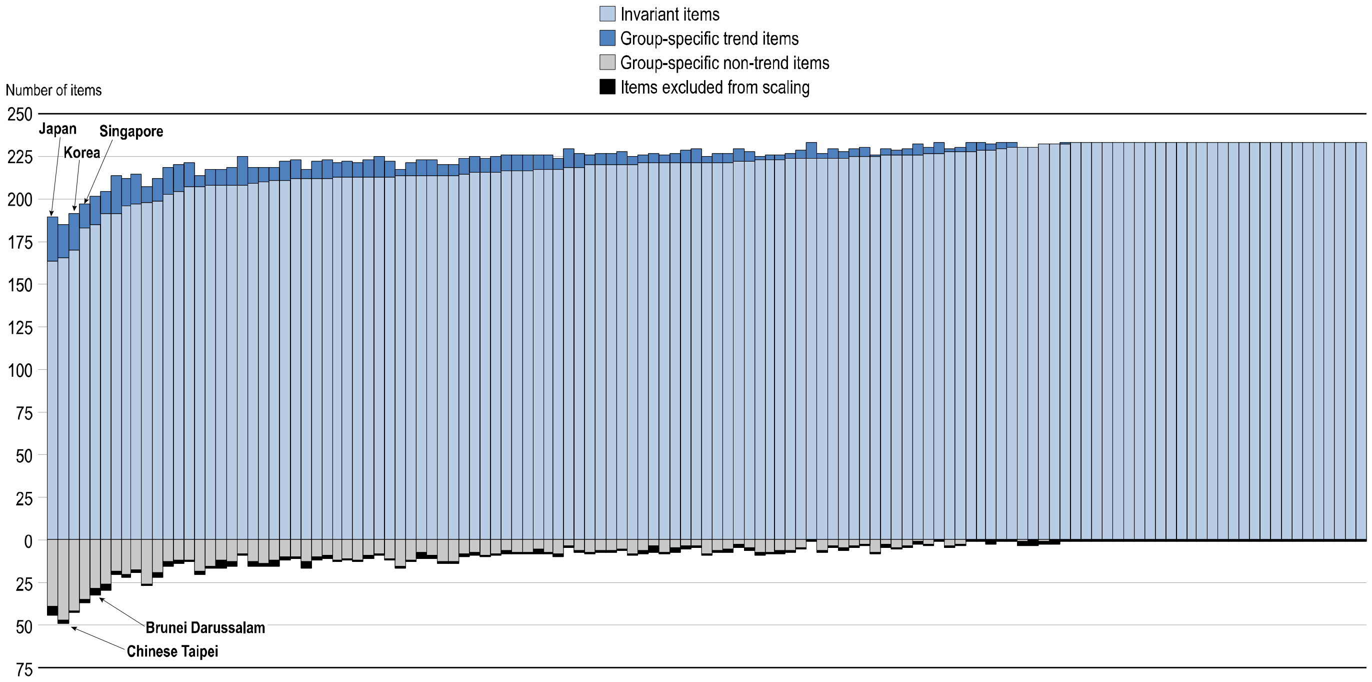
For any pair of countries/economies, the larger the number and share of common ("invariant") item parameters, the more comparable the PISA scores. As the figures show, comparisons between most countries' results are supported by strong links involving many items (in 115 of 125 country-by-language group, over 85% of the items use international, invariant item parameters).

Across every domain, international/common (invariant) item parameters dominate and only a small proportion of the item parameters are group-specific. The *PISA 2022 Technical Report* (OECD, Forthcoming[1]) includes an overview of the number of deviations per item across all country-by-language groups.

The country/language group with the largest amount of misfit across items is Viet Nam in reading (this was not the case in mathematics and science). In reading, almost 40% of items (34 of 87) were assigned unique parameters in Viet Nam. As a result, a strong linkage to the international PISA scale could not be established.

Figure I.A6.1. Invariance of items in the computer-based test of mathematics across countries/economies and over time

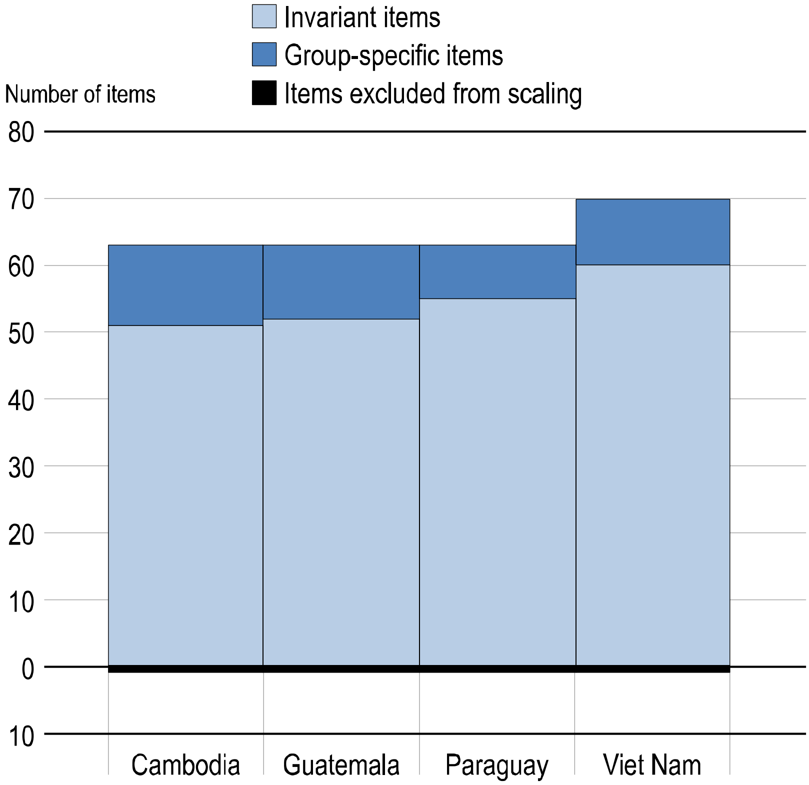
Analyses based on 234 items



Note: Each set of stacked columns corresponds to a distinct country-by-language group.
Source: OECD, PISA 2022 Database; PISA 2022 Technical Report (OECD, Forthcoming[1]).

Figure I.A6.2. Invariance of items in the paper-based test of mathematics across countries and over time

Analyses based on 64 (“new” paper-based assessment) or 71 items (“old” paper-based assessment)



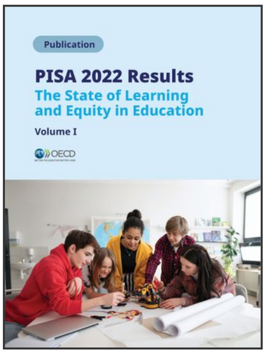
Note: Each set of stacked columns corresponds to a distinct country.

In PISA 2022, a paper-based version of the assessment that included only trend units was implemented in Cambodia, Guatemala and Paraguay (“new” PBA). Viet Nam used the same paper-based materials as in the 2015 and 2018 cycles (based on items that were first used in PISA 2012 or earlier) (“old” PBA). See Annex A5 for more details on paper-based assessments in PISA 2022.

Source: OECD, PISA 2022 Database; PISA 2022 Technical Report (OECD, Forthcoming[1]).

References

OECD (Forthcoming), *PISA 2022 Technical Report*, PISA, OECD Publishing, Paris. [1]



From:

PISA 2022 Results (Volume I)

The State of Learning and Equity in Education

Access the complete publication at:

https://doi.org/10.1787/53f23881-en

Please cite this chapter as:

OECD (2023), “Are PISA mathematics scores comparable across countries and languages?”, in *PISA 2022 Results (Volume I): The State of Learning and Equity in Education*, OECD Publishing, Paris.

DOI: https://doi.org/10.1787/21bfadca-en

This document, as well as any data and map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area. Extracts from publications may be subject to additional disclaimers, which are set out in the complete version of the publication, available at the link provided.

The use of this work, whether digital or print, is governed by the Terms and Conditions to be found at http://www.oecd.org/termsandconditions.