ANNEX A6
Are PISA reading 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 the assessment instruments and the 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 reading experts conducted qualitative reviews of the full set of items included in the PISA 2018 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 2018, the PISA consortium asked national experts to confirm or revise their original ratings, with respect to the final instruments. Sixty-five national centres submitted ratings of the relevance of PISA 2018 reading items to measure students' "preparedness for life" – a key aspect of the validity of PISA (response options were: "not at all 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, 76% of items were rated as "highly relevant for students' preparedness for life" (the highest possible rating); only 3% received a low rating on this dimension (rating equal to 1). Thirty-five out of 65 countries/economies did not rate any item as being "not relevant" to students' preparedness for life.

On the other hand, many national experts indicated less overlap between national curricula and the PISA reading item set. On average, 63% of items were rated as "standard curriculum material", and 9% of items were identified as "not in curriculum". National experts from six countries – Australia, Costa Rica, Estonia, Finland, Iceland and the Republic of Moldova – 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 2018 set of reading items.

NATIONAL ITEM DELETIONS, ITEM MISFIT, AND ITEM-BY-COUNTRY INTERACTIONS

PISA reporting scales in reading, mathematics 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 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 the 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 reading for PISA 2018 is given in Figure I.A6.1 and Figure I.A6.2; the corresponding figures for other domains can be found in the *PISA 2018 Technical Report* (OECD, forthcoming[1]). Each set of stacked bars in these figures represents a country or economy (for countries and economies with multiple language groups, a weighted average of the scaling groups is presented).

Table I.A6.1 [1/2] **How national experts rated PISA reading 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	Australia	0.0	0.0	100.0	0.0	0.0	100.0
	Austria	0.4	20.0	79.6	2.0	33.9	64.1
	Belgium (Flemish Community)	0.0	9.0	91.0	0.0	2.0	98.0
	Belgium (French Community)	0.4	5.0	94.6	0.0	5.0	95.0
	Canada	0.0	26.9	73.1	0.0	15.9	84.1
	Chile	0.8	28.6	70.6	5.3	14.3	80.4
	Colombia	1.3	14.4	84.3	1.3	3.4	95.3
	Czech Republic	2.9	45.7	51.4	0.4	39.2	60.4
	Denmark	0.0	45.7	54.3	0.0	29.8	70.2
	Estonia	0.0	0.0	100.0	0.0	0.0	100.0
	Finland	0.0	0.0	100.0	0.0	0.0	100.0
	France	22.9	28.6	48.6	3.7	14.3	82.0
	Germany	0.0	9.0	91.0	0.0	0.8	99.2
	Greece	9.0	28.6	62.4	4.9	2.0	93.1
	Hungary	20.4	52.7	26.9	0.0	23.7	76.3
	Iceland	0.0	0.0	100.0	0.0	3.7	96.3
	Israel	10.2	26.1	63.7	9.0	44.5	46.5
	Italy	5.3	28.3	66.4	5.7	4.1	90.2
	Japan	1.2	0.4	98.4	1.2	0.4	98.4
	Korea	0.0	13.1	86.9	0.0	0.4	99.6
	Latvia	0.0	7.8	92.2	0.0	3.7	96.3
	Luxembourg	0.0	11.8	88.2	0.0	0.0	100.0
	Mexico	0.0	15.7	84.3	0.0	0.0	100.0
	Netherlands	0.8	46.5	52.7	0.0	14.7	85.3
	New Zealand	0.0	18.8	81.2	0.0	11.4	88.6
	Norway	8.6	14.3	77.1	6.5	5.7	87.8
	Poland	0.4	14.3	85.3	0.0	0.8	99.2
	Portugal	53.9	24.1	22.0	20.0	31.0	49.0
	Slovak Republic	0.0	85.3	14.7	0.4	35.5	64.1
	Slovenia	27.3	20.0	52.7	8.2	46.5	45.3
	Sweden	0.8	19.7	79.5	0.0	11.6	88.4
	Switzerland	0.0	31.8	68.2	0.0	0.4	99.6
	United States	m	m	m	m	m	m

Note: Percentages may not add up 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. In Switzerland, three experts from distinct language regions reviewed the items. For the few items where their ratings differed, a national rating was determined as follows: for relevance to "preparedness for life", the modal rating was considered; for curriculum overlap, the rating "in some curricula" was used unless all three experts agreed on one of the two other options. For Belgium, ratings are reported separately for the Flemish Community and for the French Community. For Denmark, the category "in some curricula" should be interpreted as "partly relevant to" the (single) national learning standards. Ratings for the United States are reported as missing; the education system in the United States is highly decentralised, with over 13 600 school districts that make curriculum decisions based on state recommendations. This makes it difficult to determine curriculum coverage in relation to assessment items.

StatLink https://doi.org/10.1787/888934028881

The bars represent the items used in the country. A colour-code indicates whether international item parameters were used in scaling (the same as in PISA 2015), or whether, due to misfit when using international parameters, national item parameters were used.[1] For items where international equality constraints were released, a distinction is made between two groups:

- 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)
- items for which the "non-invariant" item parameters used in 2018 could be constrained to the same values used in 2015 for the particular country/language group (these items contribute to measurement invariance over time, but not across groups).

Table I.A6.1 [2/2] **How national experts rated PISA reading 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 (%)
Partners	Albania	23.7	19.2	57.1	11.0	31.8	57.1
	Argentina	26.4	20.8	52.8	12.5	19.4	68.1
	Baku (Azerbaijan)	0.4	96.7	2.9	0.0	10.7	89.3
	Belarus	0.0	13.1	86.9	0.0	41.2	58.8
	Brazil	0.0	3.7	96.3	1.2	4.1	94.7
	Brunei Darussalam	21.2	63.3	15.5	22.4	58.0	19.6
	B-S-J-Z (China)	1.2	13.1	85.7	0.4	6.1	93.5
	Bulgaria	0.0	22.9	77.1	0.0	31.0	69.0
	Costa Rica	0.0	0.0	100.0	0.0	0.0	100.0
	Croatia	21.6	48.2	30.2	0.0	17.6	82.4
	Cyprus	0.0	33.9	66.1	0.0	5.7	94.3
	Hong Kong (China)	5.7	46.9	47.3	0.8	41.2	58.0
	Jordan	11.1	25.0	63.9	6.9	8.3	84.7
	Kazakhstan	0.0	82.9	17.1	0.0	29.8	70.2
	Macao (China)	58.8	41.2	0.0	20.8	70.6	8.6
	Malaysia	6.5	51.4	42.0	0.4	42.9	56.7
	Malta	2.4	40.4	57.1	0.4	49.0	50.6
	Moldova	0.0	0.0	100.0	2.8	5.6	91.7
	Montenegro	2.9	4.5	92.7	5.7	17.1	77.1
	Morocco	24.9	47.8	27.3	3.3	40.0	56.7
	Panama	0.0	59.2	40.8	0.0	95.5	4.5
	Peru	0.0	18.4	81.6	0.0	3.7	96.3
	Qatar	2.5	50.4	47.1	0.0	9.4	90.6
	Romania	0.0	5.6	94.4	1.4	6.9	91.7
	Russia	17.2	20.9	61.9	0.0	55.3	44.7
	Serbia	68.6	18.8	12.7	0.0	1.6	98.4
	Singapore	0.8	0.4	98.8	0.0	6.5	93.5
	Chinese Taipei	0.0	86.9	13.1	0.0	75.9	24.1
	Thailand	0.0	18.4	81.6	0.0	7.3	92.7
	Ukraine	18.1	11.1	70.8	0.0	1.4	98.6
	United Arab Emirates	46.1	18.8	35.1	14.7	43.3	42.0
	Uruguay	9.4	36.5	54.1	7.3	36.1	56.7
	Viet Nam	45.8	51.4	2.8	45.8	51.4	2.8

Note: Percentages may not add up 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. In Switzerland, three experts from distinct language regions reviewed the items. For the few items where their ratings differed, a national rating was determined as follows: for relevance to "preparedness for life", the modal rating was considered; for curriculum overlap, the rating "in some curricula" was used unless all three experts agreed on one of the two other options. For Belgium, ratings are reported separately for the Flemish Community and for the French Community. For Denmark, the category "in some curricula" should be interpreted as "partly relevant to" the (single) national learning standards. Ratings for the United States are reported as missing; the education system in the United States is highly decentralised, with over 13 600 school districts that make curriculum decisions based on state recommendations. This makes it difficult to determine curriculum coverage in relation to assessment items.

StatLink https://doi.org/10.1787/888934028881

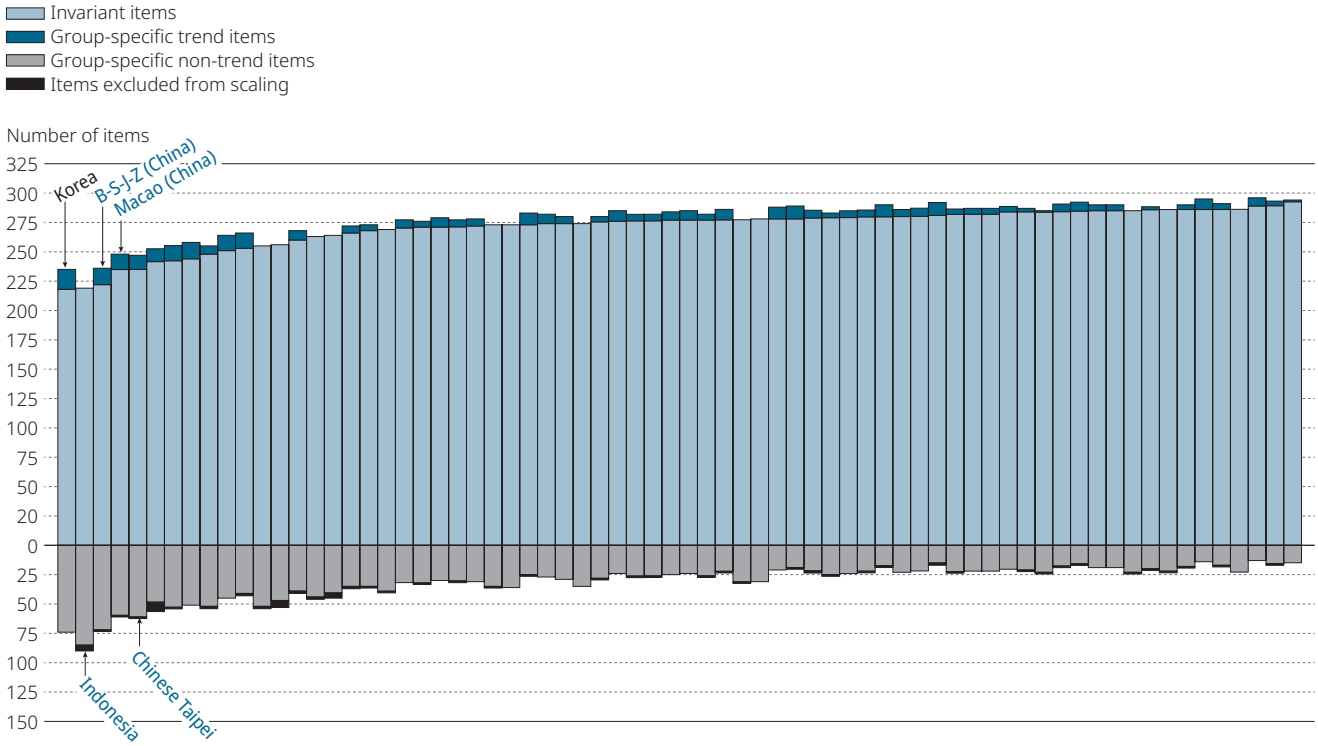
For any pair of countries/economies, the larger the number and share of common item parameters, the more comparable the PISA scores. As the figures show, comparisons between most countries' results are supported by strong links involving many items (in 58 of 79 countries/economies, over 85% of the items 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 2018 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 (the same was found in mathematics and science too). The proportion of international trend items is between 50% and 60% in each subject. A similar level of misfit was also found in PISA 2015.

The possible reasons why the item-response theory model that fits all other countries well is not a good fit for Viet Nam's data are still being investigated. Initial analyses explored, at the item level, the direction of misfit (using mean deviation statistics), the characteristics of misfit items, and any potential sign of data manipulation or coder bias. For example, students' booklets were inspected, and the answers were compared to the codes included in the database. The analysis also involved comparisons of booklets and response patterns in PISA 2018 with the PISA 2015 main study and with the PISA 2015 and 2018 field trials.

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

Analyses based on 309 items (including reading-fluency tasks)



Notes: Each set of stacked columns corresponds to a distinct country/economy. For countries/economies with more than one scaling group, a weighted average of invariant and non-invariant items across scaling groups is reported.

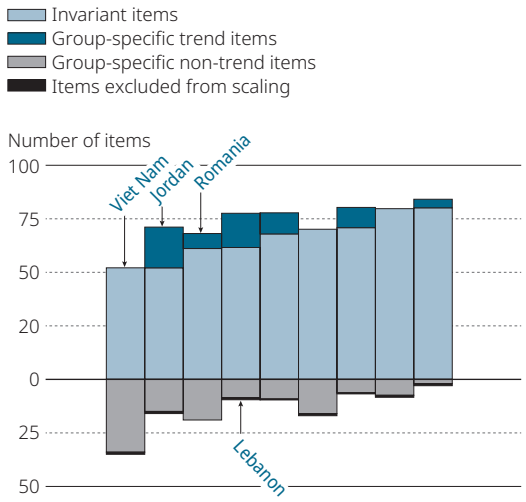
Item CR563Q12 was excluded from scaling in all countries and is not included among the 309 items considered for this figure.

Source: OECD, PISA 2018 Database; *PISA 2018 Technical Report* (OECD, forthcoming[1]).

StatLink https://doi.org/10.1787/888934028900

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

Analyses based on 88 items ("A" booklets) or 87 items ("B" booklets)



Note: Each set of stacked columns corresponds to a distinct country. For countries with more than one scaling group, a weighted average of invariant and non-invariant items across scaling groups is reported.

Source: OECD, PISA 2018 Database; *PISA 2018 Technical Report* (OECD, forthcoming[1]).

StatLink https://doi.org/10.1787/888934028919

Indeed, while overall performance can vary across PISA administrations (and particularly between the field trial and the main study), the item-response patterns, conditional on overall performance, should remain relatively stable across administrations, unless the patterns are strongly influenced by test conditions, such as the print quality.

This initial investigation did not find any evidence of data manipulation or coder bias. Initial findings indicate that a significant amount of misfit could be modelled as a country-specific response-format effect, meaning that selected-response questions, as a group, appeared to be significantly easier for students in Viet Nam than expected, given the usual relationship between open-ended and selected-response questions reflected in the international model parameters. The initial investigation also found that for a number of selected-response items, response patterns were not consistent across field-trial and main study administrations. This inconsistency over time within the same country cannot be explained by familiarity, curriculum or cultural differences. After reviewing the data for Viet Nam, the PISA Adjudication Group concluded that targeted training and coaching on PISA-like items (and occasional errors induced by training or coaching) constitutes the most plausible explanation for the differences between student-response patterns observed in Viet Nam in 2018 and those observed in other countries or in previous cycles.

Whatever its causes, the statistical uniqueness of Viet Nam's response data implies that performance in Viet Nam cannot be validly reported on the same PISA scale as performance in other countries. It may still be possible to estimate an item-response-theory model for Viet Nam and report performance on a scale that retains some level of within-country trend comparability, but this scale could not be used to compare Viet Nam with other countries and could not be interpreted in terms of international proficiency levels.

In addition, Beijing, Shanghai, Jiangsu and Zhejiang (China) (hereafter "B-S-J-Z [China]"), Indonesia, Korea, Macao (China) and Chinese Taipei (as well as, amongst countries that delivered the PISA test in the paper-based format, Jordan, Lebanon and Romania) show a relatively large number of patterns that are unexpected, based on international item parameters and given the overall performance level observed in these countries/economies. In all of these countries/economies, except Jordan, items with group-specific parameters and items excluded from scaling represent between 23% and 30% of all items in reading (in Jordan, they represent 40% of items in reading, 39% in science and 13% in mathematics). This mirrors earlier findings that differential item functioning in the PISA reading test is higher in Asian countries and in countries using non-Indoeuropean languages (Grisay and Monseur, 2007[2]; Grisay, Gonzalez and Monseur, 2009[3]). Another tentative pattern that can be established is that item misfit is higher, in reading, in countries where the language of assessment is not the language spoken outside of school by many of their students; this is the case in Indonesia and Lebanon. In these cases, the target construct for reading items may be confounded by language proficiency.

While the number of items affected is relatively large, the nature and extent of misfit are unlikely to affect the validity and comparability of PISA results in these cases. For example, in each of the countries/economies that delivered PISA on computer, including B-S-J-Z (China), Indonesia, Korea, Macao (China) and Chinese Taipei, comparisons of reading scores across countries are supported by at least 218 items with common, invariant parameters. In the case of Jordan and Lebanon, while international comparability is lower (also because these countries used paper-based instruments; see Annex A5), trend comparability is strong: for a majority of the items receiving country-specific item parameters, the observed response patterns are consistent with what had already been observed in 2015.

ARE PISA RANKINGS DETERMINED BY THE SELECTION OF ITEMS FOR THE TEST?

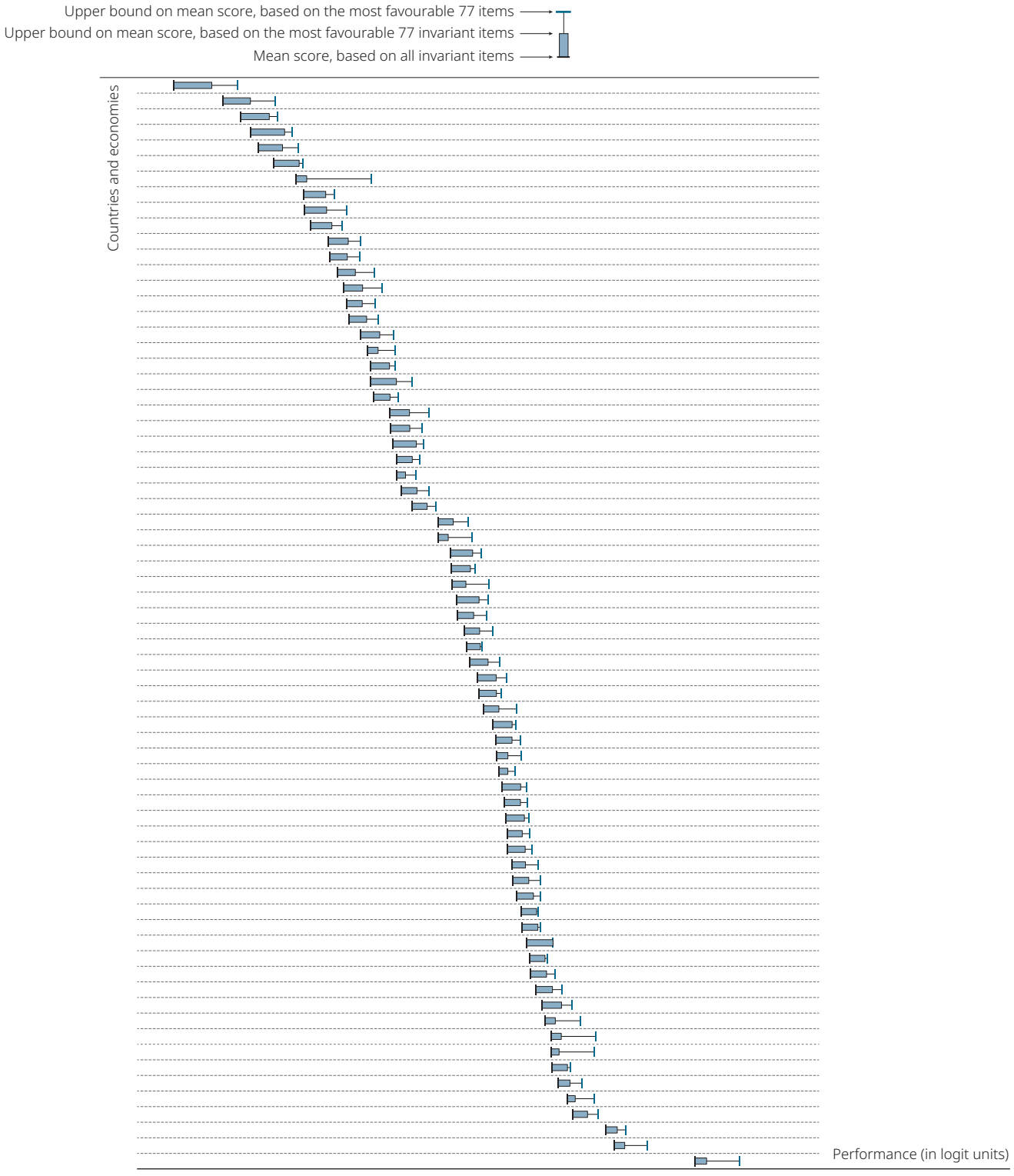
A key assumption of a fully invariant "international" item-response model is that a single model can describe the relationship between student proficiency and (international) item characteristics for all countries and economies. This would imply, for example, that any sufficiently large subset of items would result in the same performance estimate for the country/economy, up to a small "measurement error". In practice, the assumption of full invariance is relaxed in PISA, which estimates a "partial" invariance model that allows some items to have country/language-specific characteristics (see above). This strongly limits the impact of item selection on performance scores.

This section analyses the impact of item selection on mean-score rankings for countries that delivered the PISA 2018 test on computer. It does so both in the context of a hypothetical fully invariant item-response model and in the context of the partial-invariance model used in PISA. In both situations, the analysis asks: to what extent could a country improve its ranking simply through a more favourable selection of test items (i.e. without changing students' behaviour)?

In particular, for each country, three approximate measures of mean performance are computed: one based on the full set of invariant items, which is used as a reference, and two "upper bound" estimates based on more favourable sets of items. These upper bound estimates are based on two-thirds of the items only. In the "strong invariance" case, all items are considered when selecting the most-favourable 77 items (out of 115 items available in total); in the "partial invariance" case, only items that are scaled using international trend items are considered when selecting the most-favourable 77 items for each country/economy.

Figure I.A6.3 **Robustness of country mean scores in science**

Mean performance and upper bound on mean performance based on most favourable selection of 77 items



Note: Mean performance is computed based on invariant items only as the mean of logit-transformed percentages of correct answers, centred around the international mean and divided by the median absolute deviation. The value of 0 corresponds to the international mean for computer-based countries. To compute the upper bound on mean performance, only the most favourable 77 items (i.e. about two-thirds of the overall set of items) are considered for each country. The high mark selects these 77 items among all 115 items, assuming that they are invariant and can be used to compare countries; the more narrow range assumes that only the science items that are scaled with international item parameters are comparable across countries and economies.

Source: OECD, PISA 2018 Database.

StatLink https://doi.org/10.1787/888934028938

To avoid embedding other model assumptions in the comparison, country mean scores are not computed through an item-response model, but as simple averages of logit-transformed percent-correct statistics, centred around the international mean for each item.[2] The average score for a country whose students succeed at the international mean level on each item is therefore 0. Positive scores indicate that the country has, on average across items, higher success rates than the international mean; negative scores indicate that the country has, on average, lower success rates than the international mean.

The analysis in this section is based on the science test, because item-level statistics, including the percentage of correct answers or its logit-transformed values, are not directly comparable across countries for the reading test, which was delivered in adaptive fashion. The analysis intends to illustrate what the observed level of misfit implies for the substantive conclusions that are drawn from PISA, both before any country- and language-specific parameters are assigned, and after the set of invariant items is tailored to each country. Because the amount of model misfit is similar in every domain, the qualitative conclusions are expected to generalise to reading too.

The analysis shows that the selection of items only minimally affects the most important comparative conclusions – for example, whether a country scores above or below another country, on average – and that the influence of item selection on country rankings is reduced particularly when the partial-invariance model that PISA applies to student responses is duly considered. This means that the potential for improving a country's mean performance in PISA through a more favourable selection of items, indicated by the blue segments in Figure I.A6.3, is small in comparison to the overall variation in performance across countries.

ARE MEASURES OF READING FLUENCY COMPARABLE ACROSS COUNTRIES AND LANGUAGES?

Reading-fluency tasks required test-takers to decide as quickly as possible whether a simple sentence made sense (see Annex C).

Student scores on reading-fluency tasks (i.e. whether they correctly affirmed that a meaningful sentence made sense and rejected meaningless sentences) were considered together with the remaining reading tasks during scaling. These tasks amount to very simple literal understanding tasks. The analysis of country-by-item effects (DIF) did not highlight particular issues related to this group of items.

Timing information, however, was not used during scaling.[3] An initial analysis of item completion time for reading-fluency tasks indeed showed considerable country differences and, most important, item-by-country effects. For this reason, the Reading Expert Group that guided the development of the reading test does not recommend the use of time data at the item level as part of the international PISA reports, nor the construction of a simple international timing-based measure of fluency. At the same time, the Reading Expert Group supports the use of timing-based measures of fluency in national analyses, and encourages further research into the modelling of timing and accuracy data at national and international levels. Simple, descriptive measures of the total time spent by students on reading-fluency tasks are provided in Table I.A8.19 (available on line).

Data about response time and score (correct/incorrect) are available for all items, including reading fluency items, and for all students, as part of the public-use cognitive database. Interested researchers can access these data through the PISA website at www.oecd.org/pisa.

Notes

1. For countries that distributed the paper-based test, group invariance is assessed with respect to international paper-based item parameters. When comparing countries using the paper-based test to countries using the computer-based test, the number and share of items for which the difficulty parameter differs (metric invariant items; see Table I.A5.3) should also be considered.
2. The approximate mean scores used in Figure I.A6.3, based on logit-transformed and centred percent-correct statistics for invariant items, correlate at $r = 0.998$ (N = 70) with the mean scores based on plausible values reported in Table I.B1.6.
3. Timing information is collected and reported in databases for all items in the computer-based test, but is not considered, in general, part of the construct that is being assessed by these items. In contrast, in the case of reading-fluency items, both "speed" and "accuracy" are important aspects of the target construct, and students were explicitly told that their completion time would be considered, along with their answers ("You will have three minutes to read and respond to as many sentences as you can"). For this reason, the question whether timing information should be included in scaling was considered.

References

Grisay, A., E. Gonzalez and **C. Monseur** (2009), *Equivalence of item difficulties across national versions of the PIRLS and PISA reading assessements*, http://www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/IERI_Monograph_Volume_02.pdf#page=63 (accessed on 16 July 2019). [3]

Grisay, A. and **C. Monseur** (2007), "Measuring the equivalence of item difficulty in the various versions of an international test", *Studies in Educational Evaluation*, Vol. 33/1, pp. 69-86, http://dx.doi.org/10.1016/j.stueduc.2007.01.006. [2]

OECD (forthcoming), *PISA 2018 Technical Report*, OECD Publishing, Paris. [1]