Annex A3. Technical notes on analyses in this volume

Standard errors, confidence intervals, significance test and p-values

The statistics in this report represent estimates based on samples of students, rather than values that could be calculated if every student in every country had answered every question. Consequently, it is important to measure the degree of uncertainty of the estimates. In PISA, each estimate has an associated degree of uncertainty, which is expressed through a standard error. The use of confidence intervals provides a way to make inferences about the population parameters (e.g. means and proportions) in a manner that reflects the uncertainty associated with the sample estimates. If numerous different samples were drawn from the same population, according to the same procedures as the original sample, then in 95 out of 100 samples the calculated confidence interval would encompass the true population parameter. For many parameters, sample estimators follow a normal distribution and the 95% confidence interval can be constructed as the estimated parameter, plus or minus 1.96 times the associated standard error.

In many cases, readers are primarily interested in whether a given value in a particular country is different from a second value in the same or another country, e.g. whether girls in a country perform better than boys in the same country. In the tables and figures used in this report, differences are labelled as statistically significant when a difference of that size or larger, in either direction, would be observed less than 5% of the time in samples, if there were actually no difference in corresponding population values. Throughout the report, significance tests were undertaken to assess the statistical significance of the comparisons made.

Some analyses in this volume explicitly report p-values (e.g. Table I.B1.5.4). P-values represent the probability, under a specified model, that a statistical summary of the data would be equal to or more extreme than its observed value (Wasserstein and Lazar, 2016[1]). For example, in Table I.B1.5.4, the p-value represents the likelihood of observing, in PISA samples, a trend equal to or more extreme (in either direction) than what is reported, when in fact the true trend for the country is flat (equal to 0).

Statistical significance of differences between subgroup means, after accounting for other variables

For many tables, subgroup comparisons were performed both on the observed difference (“before accounting for other variables”) and after accounting for other variables, such as the PISA index of economic, social and cultural status of students. The adjusted differences were estimated using linear regression and tested for significance at the 95% confidence level. Significant differences are marked in bold.

Range of ranks (confidence interval for rankings of countries)

An estimate of the rank of a country mean, across all country means, can be derived from the estimates of the country means from student samples. However, because mean estimates have some degree of uncertainty, this uncertainty should also be reflected in the estimate of the rank. While mean estimates from samples follow a normal distribution, this is not the case of the rank estimates derived from these. Therefore, in order to construct a confidence interval for ranks, simulation methods were used.

Data are simulated assuming that alternative mean estimates for each relevant country follow a normal distribution around the estimated mean, with a standard deviation equal to the standard error of the mean. Some 1 000

simulations are carried out and, based on the alternative mean estimates in each of these simulations, 1 000 possible estimates for each country are produced.

There are two steps to estimating the confidence sets of ranks. For each country, all possible differences in score estimates are considered between the reference country and all other participating countries. Then for every country, confidence sets of ranks are computed with respect to all other participating countries (with respect to all other OECD countries in the case of the OECD country ranking). Using these individual confidence sets, a *simultaneous* confidence set is computed, covering all possible differences of the reference country with all other countries with a confidence level of 95%. Given this, the simultaneous confidence sets that are fully above or fully below zero (i.e. where differences are significantly different from zero) are used to determine confidence sets for the ranking of a country.

The ranking that results from these simultaneous confidence sets is obtained using a stepwise multiple testing procedure. This implies that first, some countries will be ranked higher or lower compared to the reference country as described above. In the following steps, the rank of the remaining countries accounts for the countries that were ranked higher or lower in previous steps, until all countries are ranked with respect to the reference country. These are the ranks reported in Tables I.2.4, I.2.5 and I.2.6, see Chapter 2. For further details on this procedure, see (Mogstad et al., 2023[2]).

The main difference between the range of ranks (e.g. Table I.2.4) and the comparison of countries' mean performance (e.g. Table I.2.1) is that the former takes into account the multiple comparisons involved in ranking countries/economies, while the latter does not. Therefore, sometimes there is a slight difference between the range of ranks and counting the number of countries above a given country, based on pairwise comparisons of the selected countries' performance. For instance, OECD countries Hungary, Portugal and Spain have similar mean performance and the same set of countries whose mean score is not statistically different from theirs, based on Table I.2.1; but the range of ranks amongst OECD countries for Hungary and Portugal can be restricted to be with 97.5% confidence between 16th and 30th for Hungary and between 17th and 30th for Portugal, while the range of ranks for Spain is narrower (between 18th and 29th) (Table I.2.4). When interest lies in examining countries' rankings, this range of ranks should be used.

Statistics based on multilevel models

Statistics based on multilevel models include variance components (between- and within-school variance) and the index of inclusion derived from these components (i.e. by index of inclusion we refer here to the index of academic inclusion [see Tables I.B1.2.12 and I.B1.2.13] and to the index of social inclusion [see Tables I.B1.4.40 and I.B1.4.41]). Multilevel models are generally specified as two-level regression models (student and school levels), with normally distributed residuals, and estimated with maximum likelihood estimation. Where the dependent variable is mathematics performance, the estimation uses 10 plausible values for each student's performance on the mathematics scale. Models were estimated using the Stata (version 17) "mixed" module.

The index of inclusion is defined and estimated as:

$$100 * \frac{\sigma_W^2}{\sigma_W^2 + \sigma_B^2} \qquad \text{Equation I.A3.1}$$

where σ_W^2 and σ_B^2, respectively, represent the within- and between-variance estimates.

For statistics based on multilevel models (such as the estimates of variance components) the standard errors are not estimated with the usual replication method, which accounts for stratification and sampling rates from finite populations. Instead, standard errors are "model-based": their computation assumes that schools, and students

within schools, are sampled at random (with sampling probabilities reflected in school and student weights) from a theoretical, infinite population of schools and students, which complies with the model's parametric assumptions. The standard error for the estimated index of inclusion is calculated by deriving an approximate distribution for it from the (model-based) standard errors for the variance components, using the delta method.

Parity index

The parity index for an indicator is used by the UNESCO Institute of Statistics to report on Target 4.5 of the Sustainable Development Goals. It is defined as the ratio of the indicator value for one group to the value for another group. Typically, the group more likely to be disadvantaged is in the numerator, and the parity index takes values between 0 and 1 (with 1 indicating perfect parity).

However, in some cases the group in the numerator has a higher value on the indicator. To restrict the range of the parity index between 0 and 2, and to make its distribution symmetrical around 1, an adjusted parity index is defined in these cases. For example, the gender parity index for the share of students reaching Level 2 proficiency on the PISA scale is computed from the share of boys (p_b) and the share of girls (p_g) reaching Level 2 proficiency as follows:

$$PI_{b,g} = \begin{cases} \frac{p_b}{p_g} & \text{if } p_b \geq p_g \\ 2 - \frac{p_g}{p_b} & \text{if } p_b < p_g \end{cases}$$ Equation I.A3.2

The "parity index" reported in Table I.B1.3.13 corresponds to the adjusted parity.

Odds ratios

The odds ratio is a measure of the relative likelihood of a particular outcome across two groups. The odds ratio for observing the outcome when an antecedent is present is simply

$$\text{OR} = {(p_{11}/p_{12})}\Big/{(p_{21}/p_{22})}$$ Equation I.A3.3

where p_{11}/p_{12}represents the "odds" of observing the outcome when the antecedent is present, and p_{21}/p_{22} represents the "odds" of observing the outcome when the antecedent is not present.

Logistic regression can be used to estimate the log ratio: the exponentiated logit coefficient for a binary variable is equivalent to the odds ratio. A "generalised" odds ratio, after accounting for other differences across groups, can be estimated by introducing control variables in the logistic regression.

Figures in bold in the data tables presented in Annex B1 of this report indicate that the odds ratio is statistically significantly different from 1 at the 95% confidence level. To construct a 95% confidence interval for the odds ratio, the estimator is assumed to follow a log-normal distribution, rather than a normal distribution.

References

Mogstad, M. et al. (2023), “Inference for Ranks with Applications to Mobility across Neighbourhoods and Academic Achievement across Countries”, *Review of Economic Studies*, https://doi.org/10.1093/restud/rdad006. [2]

Wasserstein, R. and N. Lazar (2016), “The ASA Statement on p-Values: Context, Process, and Purpose”, *The American Statistician*, Vol. 70/2, pp. 129-133, https://doi.org/10.1080/00031305.2016.1154108. [1]