ANNEX A3
Technical notes on analyses in this volume

STANDARD ERRORS, CONFIDENCE INTERVALS AND SIGNIFICANCE TESTS

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, if there were actually no difference in corresponding population values. Similarly, the risk of reporting an association as significant if there is, in fact, no correlation between two measures, is contained at 5%.

Throughout the report, significance tests were undertaken to assess the statistical significance of the comparisons made.

Statistical significance of gender differences and differences between subgroup means

Gender differences in student performance or other indices were tested for statistical significance. Positive differences indicate higher scores for girls while negative differences indicate higher scores for boys. Generally, differences marked in bold in the tables in this volume are statistically significant at the 95% confidence level.

Similarly, differences between other groups of students (e.g. non-immigrant students and students with an immigrant background, or socio-economically advantaged and disadvantaged students) were tested for statistical significance. The definitions of the subgroups can, in general, be found in the tables and the text accompanying the analysis. All differences marked in bold in the tables presented in Annex B of this report are statistically significant at the 95% level.

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.

Statistical significance of performance differences between the top and bottom quartiles of PISA indices and scales

Differences in average performance between the top and bottom quarters of the PISA indices and scales were tested for statistical significance. Figures marked in bold indicate that performance between the top and bottom quarters of students on the respective index is statistically significantly different at the 95% confidence level.

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

$$\mathrm{OR} = \frac{(p_{11} / p_{12})}{(p_{21} / p_{22})} \qquad \text{Equation II.A3.2}$$

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.

Statistical significance of odds ratios

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.

In many tables, odds ratios after accounting for other variables are also presented. These odds ratios were estimated using logistic regression and tested for significance against the null hypothesis of an odds ratio equal to 1 (i.e. equal likelihoods, after accounting for other variables).

OVERALL RATIOS AND AVERAGE RATIOS

In this report, the comparisons of ratios related to teachers, such as student-teacher ratio or the proportion of certified teachers, are made using overall ratios. This means, for instance, that the student-teacher ratio is obtained by dividing the total number of students in the target population by the total number of teachers in the target population. The overall ratios are computed by first computing the numerator and denominator as the (weighted) sum of school-level totals, then dividing the numerator by the denominator. Similar estimations are made for the proportion of novice teachers, the proportion of teachers with at least a master's degree, the proportion of fully certified teachers, etc. In most cases (i.e. unless all schools are exactly the same size) this overall ratio differs from the average of school-level ratios.

SOCIAL AND ACADEMIC SEGREGATION INDICES

Statistics based on multilevel models

Statistics based on multilevel models include variance components (between- and within-school variance), the index of inclusion derived from these components, and regression coefficients where this has been indicated. Multilevel models are generally specified as two-level regression models (the student and school levels), with normally distributed residuals, and estimated with maximum likelihood estimation. Where the dependent variable is reading performance, the estimation uses ten plausible values for each student's performance on the reading scale. Models were estimated using the Stata (version 15.1) "mixed" module.

The index of inclusion is defined and estimated as:

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

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

Standard errors in statistics estimated from multilevel models

For statistics based on multilevel models (such as the estimates of variance components and regression coefficients from two-level regression models) 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.

The isolation index and the exposure index

The isolation index used in the report corresponds to the normalised exposure indicator (Frankel and Volij, 2011[1]),

$$\mathrm{I} = 1 - \frac{\sum_{j=1}^{J} \frac{n_j^a}{N^a} \frac{(1 - n_j^a)}{n_j}}{1 - p^a} \qquad \text{Equation II.A3.2}$$

where n_j^a (respectively N^a) stands for the number of students of type a (for instance, those with an immigrant background) in school j (respectively, in the country), n_j the total number of students in this school j and with $p^a = \frac{n^a}{N}$ the proportion of the group a in the population. This index ranges from 0 (no segregation) to 1 (full segregation), meaning that the index increases with the concentration of the students of the group a in a limited number of schools.

In the report, this index is also used for measuring the concentration of students in schools of socio-economically advantaged and disadvantaged students (defined as those in the first and the fourth quarters, respectively, of the national distribution of the ESCS index) and of low and high performers (defined as those in the first and the fourth quarters, respectively, of the national distribution of reading performance).

A related index, the exposure index, represents the probability E that an average student from one of these groups is in contact at school with students who do not belong to the same group (who represent three-quarters of the population). The exposure index can be computed as

$$E = \sum_{j=1}^{J} \frac{n_j^{disad}}{N^{disad}} \frac{n_j^{highperf}}{n_j} \approx 0.75 * (1 - I) \qquad \text{Equation II.A3.3}$$

This probability is (i.e. equal to the proportion of the other group in the population) when the allocation of students across schools does not depend on group membership (student type), and lower if the group the students belong to matters in the allocation of students to schools.

A derived version of the isolation index, the isolation of disadvantaged students (defined as those in the first quarter of the national distribution of the ESCS index) from high achievers (defined as those in the fourth quarter of the national distribution of reading performance) is also used in the report. It may be written formally as:

$$I_{disad}^{highperf} = 1 - \frac{\sum_{j=1}^{J} \frac{n_j^{disad}}{N^{disad}} \frac{n_j^{highperf}}{n_j}}{\frac{N^{highperf}}{N^{highperf} + N^{disad}}} \qquad \text{Equation II.A3.4}$$

The lowest value (0) is observed when the two subgroups are clustered in the same schools; the highest value (1) is observed when they are clustered in different schools. Medium values are observed when the two populations are randomly mixed within schools. Again, one may derive from this indicator the probability that an average disadvantaged student is in contact at school with a high performer, corresponding to:

$$E_{disad}^{highperf} = \sum_{j=1}^{J} \frac{n_j^{disad}}{N^{disad}} \frac{n_j^{highperf}}{n_j} \approx 0.5 * \left(1 - I_{disad}^{highperf}\right) \qquad \text{Equation II.A3.5}$$

The no social diversity index

The no social diversity index is a multi-group index, meaning that it provides a more accurate description of the social diversity in schools – comparing not only a group (such as disadvantaged students) with all other students, but all groups of students. This index is often referred to in the literature as the entropy index, or mutual information index (Frankel and Volij, 2011[1]; Reardon and Firebaugh, 2002[2]). The no social diversity index is computed as:

$$H = \sum_{j=1}^{J} \frac{n_j}{N} \frac{h\left(q_j\right) - h(q)}{h(q)} \qquad \text{Equation II.A3.6}$$

where $h(q) = -\sum_{k=1}^{4} q^k ln(q^k)$ is a measure of the diversity in the population, depending on the proportions of the four socio-economic groups in the population (defined by the quarter of ESCS index, meaning that $q^k = 0.25$), and $h(q_j)$ is its counterpart measured at the school level with $q_j = \left(q_j^1, q_j^2, q_j^3, q_j^4\right)$ the proportion of the four groups of the students amongst the students in school (and the total number of students). The no-diversity index goes from 0 (no segregation) to 1 (full segregation).

The no-social diversity index is additively decomposable. If one aggregates schools at a higher level, typically comparing private schools to public schools, the no-diversity index can be decomposed into three components. The first component corresponds to

the social segregation within private schools, the second to the segregation within public schools, and the third to the additional segregation that reflects the fact that the social composition in the public sector could be distinct from that of the private sector.

Formally, this can be written as:

$$H = H^{Priv\,/\,Pub} + \theta^{Pub} H^{Pub} + \theta^{Private} H^{Private}$$ Equation II.A3.7

With $H^{Priv\,/\,Pub}$ interpreted as the segregation due specifically to the coexistence of private and public sectors.

Modal grade schools

The segregation measures, such as between-school variations or the isolation indices, depend on how schools are defined and organised within countries and by the units that were chosen for sampling purposes. For example, in some countries, some of the schools in the PISA sample were defined as administrative units (even if they spanned several geographically separate institutions, as in Italy); in others, they were defined as those parts of larger educational institutions that serve 15-year-olds; in still others they were defined as physical school buildings; and in others they were defined from a management perspective (e.g. entities having a principal).

The *PISA 2018 Technical Report* (OECD, forthcoming) and Annex A2 provide an overview of how schools are defined. In Slovenia, for example, the primary sampling unit is defined as a group of students who follow the same study programme within a school (an education track within a school). In this case, the segregation indices between schools actually estimate the segregation between the distinct tracks in these schools. The use of stratification variables in the selection of schools may also affect the estimate of the between-school variation, particularly if stratification variables are associated with between-school differences.

In PISA 2018 the estimation of the segregation indices was restricted to schools with the "modal ISCED level" for 15-year-old students. The "modal ISCED level" is defined here as the level attended by at least one-third of the PISA sample. As PISA students are sampled to represent all 15-year-old students, whatever type of schools they are enrolled in, they may not be representative of their schools. Restricting the sampling to schools with the modal ISCED level for 15-year-old students ensures that the characteristics of students sampled for PISA represent the profile of the typical student attending the school. Modal grade may be either lower secondary (ISCED level 2), either upper secondary (ISCED level 3), or both (as in Albania, Argentina, Baku [Azerbaijan], Beijing, Shanghai, Jiangsu and Zhejiang [China], Belarus, Colombia, Costa Rica, the Czech Republic, the Dominican Republic, Indonesia, Ireland, Kazakhstan, Luxembourg, Macao [China], Morocco, the Slovak Republic, Chinese Taipei and Uruguay). In all other countries, analyses are restricted to either lower secondary or upper secondary schools. In several countries, lower and upper secondary education are provided in the same school. As the restriction is made at the school level, some students from a grade other than the modal grade in the country may also be used in the analysis. Table II.C.1 in Annex C shows the type of ISCED used for every country and economy, as well as the respective proportions of schools and students in the sample used in the analysis.

INDEX OF SOCIO-ECONOMIC INEQUALITY IN THE PROBABILITY OF BEING A HIGH PERFORMER

The index of socio-economic inequalities in high achievement quantifies the relative socio-economic inequalities in the probability of attaining Level 5 or 6 in reading proficiency. It calculates the cumulative number of high achievers concentrated in a cumulative percentage of the population of 15-year-olds ranked by the PISA index of economic, social and cultural status (ESCS), as described for instance in (O'Donnell et al., 2008[3]). This index may be related to the concentration line that would plot the cumulative numbers of high achievers (y-axis) against the cumulative percentage of the population of 15-year-olds, ranked by ESCS, beginning with the students with the lowest socio-economic status, and ending with those with the highest value (x-axis). If everyone, irrespective of his or her living standards, had exactly the same probability of being high achievers, the concentration curve would be a 45-degree line (hereafter, the line of equality), running from the bottom left-hand corner to the top right-hand corner. However, if being a high achiever is much less likely amongst students with the highest values in the ESCS index, the concentration curve would lie below the line of equality; conversely, if high achievers are more concentrated amongst students with the lowest values in the ESCS index, the concentration curve would be above the line of equality.

The farther the curve is below the line of equality, the more concentrated are high achievers amongst the most-advantaged students (similarly, the farther the line is above the line of equality, the more concentrated are high achievers amongst the least-advantaged students). The concentration index is then defined as twice the area between the concentration curve and the line of equality. When there is no socio-economic-related inequality, the concentration index is zero. By convention, the index takes a positive value when the curve lies below the line of equality, indicating a disproportionate concentration of high achievers amongst advantaged students; it takes a negative value when it lies above the line of equality. As the variable of interest (being high achievers) is binary, one should use a factor of normalisation. The calculation is made using the Stata

(version 15.1) procedure "conindex", using the f normalisation for bounded variable proposed by (Wagstaff, 2011[4]). This corresponds to the calculation:

$$C = \frac{2}{\pi_h * (1-\pi_h) * n^2} \sum_{i=1}^{n} \left(\frac{n+1}{2} - r_i \right) * h_i \qquad \text{Equation II.A3.4}$$

Where h_i is a binary variable that takes the value 1 if student *i* is high performer and 0 instead, $r_i = i/n$ is the relative rank of student *i*, n is the total number of students and is the proportion of high performers amongst the population of 15-year-old students. As emphasised by (Kjellsson and Gerdtham, 2013[5]), this means that the index will take the maximum value, 1, only when the students at the top of the ESCS index are high performers.

USE OF STUDENT, SCHOOL AND TEACHER WEIGHTS

The target population in PISA is 15-year-old students, but a two-stage sampling procedure was used. After the population was defined, school samples were selected with a probability proportional to the expected number of eligible students in each school. Only in a second sampling stage were students drawn from amongst the eligible students in each selected school.

Although the student samples were drawn from within a sample of schools, the school sample was designed to optimise the resulting sample of students, rather than to give an optimal sample of schools. It is therefore preferable to analyse the school-level variables as attributes of students (e.g. in terms of the share of 15-year-old students affected), rather than as elements in their own right.

Most analyses of student and school characteristics are therefore weighted by student final weights (or their sum, in the case of school characteristics), and use student replicate weights for estimating standard errors.

As an exception, estimates of "overall ratios" in which the denominator corresponds to the population of teachers (student-teacher ratios; proportions of fully certified teachers and proportion of teachers with at least a master's degree) use school weights, which correspond to the inverse of the prior probability of selection for each selected school. Replicate school weights were generated for these analyses in analogy with the student replicate weights in the database, by applying the replicate factors observed for student weights within the school (one value among 0.2929, 0.5, 0.6464,1, 1.3536, 1.5 or 1.7071) to the base school weights (OECD, Forthcoming[6]).

In PISA 2018, as in PISA 2012 and 2015, multilevel models weights are used at both the student and school levels. The purpose of these weights is to account for differences in the probabilities of students being selected in the sample. Since PISA applies a two-stage sampling procedure, these differences are due to factors at both the school and the student levels. For the multilevel models, student final weights (W_FSTUWT) were used. Within-school weights correspond to student final weights, rescaled to amount to the sample size within each school. Between-school weights correspond to the sum of final student weights (W_FSTUWT) within each school.

Analyses based on teacher responses to the teacher questionnaires are weighted by student weights. In particular, in order to compute averages and shares based on teacher responses, final teacher weights were generated so that the sum of teacher weights within each school was equal to the sum of student weights within the same school. The same procedure was used to generate replicate teacher weights in analogy with the student replicate weights in the database. All teachers within a school have the same weight. For the computation of means, this is equivalent to aggregating teacher responses to the school level through simple, unweighted means, and then applying student weights to these school-level aggregates.

References

Frankel, D. and **O. Volij** (2011), "Measuring school segregation", *Journal of Economic Theory*, http://dx.doi.org/10.1016/j.jet.2010.10.008. [1]

Kjellsson, G. and **U. Gerdtham** (2013), "On correcting the concentration index for binary variables", *Journal of Health Economics*, Vol. 32/3, pp. 659-670, http://dx.doi.org/10.1016/J.JHEALECO.2012.10.012. [5]

O'Donnell, O. et al. (2008), *Analyzing Health Equity Using Household Survey Data*, World Bank. [3]

OECD (Forthcoming), *PISA 2018 Technical Report*. [6]

Reardon, S. and **G. Firebaugh** (2002), "Measures of multigroup segregation", *Sociological Methodology*, Vol. 32, pp. 33-67, http://dx.doi.org/10.1111/1467-9531.00110. [2]

Wagstaff, A. (2011), "The concentration index of a binary outcome revisited", *Health Economics*, Vol. 20/10, pp. 1155-1160, http://dx.doi.org/10.1002/hec.1752. [4]