Chapter 11
Sampling Outcomes

This chapter reports on PISA sampling outcomes. Details of the sample design and selection are provided in Chapter 4.

POPULATION COVERAGE

Quality indicators for population coverage and the information used to develop them are presented in Table 11.1 and Table 11.2, for participating countries/economies and adjudicated regions, respectively. The following notes explain the meaning of each coverage index and how the data in each column of the table were used.

Table 11.1: Population characteristics, sample characteristics, exclusions and coverage indices for participating countries/economies

Table 11.2: Population characteristics, sample characteristics, exclusions and coverage indices for participating adjudicated regions

Coverage indices 1, 2 and 3 are intended to measure PISA population coverage. Coverage indices 4 and 5 are intended to be diagnostic in cases where indices 1, 2 or 3 have unexpected values. Many references are made in this chapter to the various sampling tasks on which National Project Managers (NPMs) documented statistics and other information needed in undertaking the sampling of schools and students. Note that although no comparison is made between the total population of 15-year-olds and the enrolled population of 15-year-old students, generally the enrolled population was expected to be less than or equal to the total population. Occasionally this was not the case due to differing data sources for these two values.

The components used for the coverage indices are the following:

- ST7a_1: National population of all 15-year-olds based on national statistics.
- ST7a_2.1: Enrolled 15-year-old students in grades 7 and above based on national statistics
- ST7b_1: Target population that includes all enrolled 15-year-old students in grades 7 and above that omits schools based on national statistics such as schools located in unsafe areas
- ST7b_3: Target population that includes all enrolled 15-year-old students in grades 7 and above, minus school-level exclusions, based on national statistics
- P: Weighted number of participating students calculated from the PISA sample
- E: Weighted estimate of within school excluded students calculated from the PISA sample
- S: Estimate of enrolled students from school sampling frame calculated as the sum over all sampled schools of the product of each school's sampling weight and its number of 15-year-old students

Coverage Index 1: Coverage of the national *target* population, $P/(P+E)$ × (ST7b_3/ST7b_1). This estimates the extent to which the weighted participants covered the final *target* population after all exclusions. It indicates the overall proportion of the *target* population covered by the non-excluded portion of the student sample.

Coverage Index 2: Coverage of the national *enrolled* population, $P/(P+E)$ × (ST7b_3/ST7a_2.1). This estimates the extent to which the weighted participants covered the population of all *enrolled* students in grades 7 and above. Thus, this index may be somewhat lower than Index 1.

Coverage Index 3: Coverage of the national *15-year-old* population, *P*/ST7a_1. This estimates the proportion of the national population of 15-year-olds covered by the non-excluded portion of the student sample. It is below 1.0 to the extent that 15-year-olds were excluded, or not enrolled in grade 7 or higher.

Coverage Index 4: Coverage of the estimated school population, $(P+E)/S$. This estimates the proportion of the estimated school 15-year-old population that is represented by the weighted student sample of all PISA-eligible 15-year-old students. Its purpose is to assess whether the enrolment data on the sampling frame is a reliable measure of the number of enrolled 15-year-olds. As the enrolment data on the frame was often inaccurate, this index usually differed noticeably from 1.0. In such cases, Indexes 1 and 2 may be suspect, as they rely on national enrolment data for their denominators, often derived from the same source as the school-level enrolment data.

Coverage Index 5: Coverage of the school sampling frame population, *S*/ST7b_3. This estimate provides a check as to whether the data on enrolment obtained from national statistics is consistent with the enrolment on the sampling frame. However, in most cases for PISA, the enrolment data based on national statistics were derived using data from the sampling frame by the NPM, and so this ratio was close to 1.0 for most countries/economies, even when the enrolment data on the school sampling frame were poor.

SCHOOL AND STUDENT RESPONSE RATES

Tables 11.3 to 11.8 present school and student-level response rates at the national and regional levels. Response rates are all presented separately by participating country/economy, and by adjudicated regions.

Table 11.3: Response rates for participating countries/economies calculated by using only original schools and no replacement schools.

Table 11.4 Response rates for adjudicated regions calculated by using only original schools and no replacement schools.

Tables 11.5: Response rates for participating countries/economies when first and second replacement schools were accounted for in the rates.

Tables 11. 6 : Response rates for adjudicated regions when first and second replacement schools were accounted for in the rates.

Tables 11.7 Student response rates among the full set of participating schools in participating countries/regions.

Tables 11.8 Student response rates among the full set of participating schools in adjudicated regions.

When calculating school response rates before replacement, the numerator consisted of all original sample schools with enrolled age-eligible students who participated (i.e., assessed a sample of PISA-eligible students, and obtained a student response rate of at least 50%). The denominator consisted of all the schools in the numerator, plus those original sample schools with enrolled age-eligible students that either did not participate or failed to assess at least 50% of PISA-eligible sample students. Schools that were included in the sampling frame, but were found to have no age-eligible students, or which were excluded in the field were omitted from the calculation of response rates. Replacement schools do not figure in these calculations.

When calculating school response rates after replacement, the numerator consisted of all sampled schools (original plus replacement) with enrolled age-eligible students that participated (i.e., assessed a sample of PISA-eligible students and obtained a student response rate of at least 50%). The denominator consisted of all the schools in the numerator, plus those original sample schools that had age-eligible students enrolled, but that failed to assess at least 50% of PISA-eligible sample students and for which no replacement school participated. Schools that were included in the sampling frame, but were found to contain no age-eligible students, were omitted from the calculation of response rates. Replacement schools were included in rates only when they participated, and were replacing a refusing school that had age-eligible students.

When calculating weighted school response rates, each school received a weight equal to the product of its base weight (the reciprocal of its selection probability) and the number of age-eligible students enrolled in the school, as indicated on the school sampling frame.

With the use of probability proportional to size sampling, where there are no certainty or small schools, the product of the initial weight and the enrolment will be a constant, so in participating countries/economies with few certainty school selections and no oversampling or undersampling of any explicit strata, weighted and unweighted rates are very similar. The weighted school response rate before replacement is given by the formula:

Formula 11.1

$$\text{weighted school response rate before replacement} = \frac{\sum_{i \in Y} W_i E_i}{\sum_{i \in (Y \cup N)} W_i E_i}$$

where Y denotes the set of responding original sample schools with age-eligible students, N denotes the set of eligible non-responding original sample schools, W_i denotes the base weight for school i, $W_i = 1/P_i$ where P_i denotes the school selection probability for school i, and E_i denotes the enrolment size of age-eligible students, as indicated on the sampling frame. The weighted school response rate, after replacement, is given by the formula:

Formula 11.2

$$\text{weighted school response rate after replacement} = \frac{\sum_{i \in (Y \cup R)} W_i E_i}{\sum_{i \in (Y \cup R \cup N)} W_i E_i}$$

where Y denotes the set of responding original sample schools, R denotes the set of responding replacement schools, for which the corresponding original sample school was eligible but was non-responding, N denotes the set of eligible refusing original sample schools, W_i denotes the base weight for school i, $W_i = 1/P_i$, where P_i denotes the school selection probability for school i, and for weighted rates, E_i denotes the enrolment size of age-eligible students, as indicated on the sampling frame.

For unweighted student response rates, the numerator is the number of students for whom assessment data were included in the results, omitting those in schools with between 25 and 50% student participation. The denominator is the number of sampled students who were age-eligible, and not explicitly excluded as student exclusions, also omitting those in schools with between 25 and 50% student participation.

For weighted student response rates, the same students appear in the numerator and denominator as for unweighted rates, but each student is weighted by its student base weight. This is given as the product of the school base weight – for the school in which the student was enrolled – and the reciprocal of the student selection probability within the school.

In countries/economies with no oversampling of any explicit strata, weighted and unweighted student response rates are very similar.

Overall response rates are calculated as the product of school and student response rates. Although overall weighted and unweighted rates can be calculated, there is little value in presenting overall unweighted rates. The weighted rates indicate the proportion of the student population represented by the sample prior to making the school and student non-response adjustments.

TEACHER RESPONSE RATES

Unweighted response rates for both reading/language arts and non-reading/language arts teachers were created using similar methods to those for unweighted student and school response rates – that is, ineligible teachers are not used in the denominator for the rate calculation.

These rates are presented in Table 11.9 for reading/language arts teachers and in Table 11.10 for the non-reading/language arts teachers.

Table 11.9: Unweighted Teacher response rates for reading/language arts teachers

Table 11.10: Unweighted Teacher response rates for non-reading/language arts teachers

In addition to these rates, unweighted response rates were calculated also for each sampled school in each country/economy which implemented the Teacher Questionnaire. These rates were created as quality indicators for the questionnaire team who would use the Teacher Questionnaire data to create derived variables to help provide context about PISA students.

DESIGN EFFECTS AND EFFECTIVE SAMPLE SIZES

Surveys in education and especially international surveys rarely sample students by simply selecting a random sample of students (known as a simple random sample, or SRS). Rather, a sampling design is used where schools are first selected and, within each selected school, classes or students are randomly sampled. Sometimes, geographic areas are first selected before sampling schools and students. This sampling design is usually referred to as a cluster sample or a multi-stage sample.

Selected students attending the same school cannot be considered as independent observations as assumed with a simple random sample because they are usually more similar to one another than to students attending other schools. For instance, the students are offered the same school resources, may have the same teachers and therefore are taught a common implemented curriculum, and so on. School differences are also larger if different educational programmes are not available in all schools. One expects to observe greater differences between a vocational school and an academic school than between two comprehensive schools.

Furthermore, it is well known that within a country/economy, within sub-national entities and within a city, people tend to live in areas according to their financial resources. As children usually attend schools close to their home, it is likely that students attending the same school come from similar social and economic backgrounds.

Therefore, a simple random sample of 4 000 students within a country/economy is thus likely to cover the diversity of the population better than a sample of 100 schools with 40 students observed within each school. It follows that the uncertainty associated with any population parameter estimate (i.e., standard error) will be larger for a clustered sample estimate than for a simple random sample estimate of the same size.

In the case of a simple random sample, the standard error of a mean estimate is equal to:

Formula 11.3

$$s_{(\hat{m})} = \sqrt{\frac{s^2}{n}}$$

where σ^2 denotes the variance of the whole student population and n is the student sample size.

For an infinite population of schools and infinite populations of students within schools, the standard error of a mean estimate from a cluster sample is equal to:

Formula 11.4

$$s_{(\hat{m})} = \sqrt{\frac{s^2_{schools}}{n_{schools}} + \frac{s^2_{within}}{n_{schools} n_{students}}}$$

where $\sigma^2_{schools}$ denotes the variance of the school means, σ^2_{within} denotes the variances of students within schools, $n_{schools}$ denotes the sample size of schools, and $n_{students}$ denotes the sample size of students within each school.

The standard error for the mean from a simple random sample is inversely proportional to the square root of the number of selected students. The standard error for the mean from a cluster sample is proportional to the variance that lies between clusters (i.e. schools) and within clusters, and inversely proportional to the square root of the number of selected schools and is also a function of the number of students selected per school.

It is usual to express the decomposition of the total variance into the between-school variance and the within-school variance by the coefficient of intraclass correlation, also denoted *Rho*. Mathematically, this index is equal to:

Formula 11.5

$$Rho = \frac{s^2_{schools}}{s^2_{schools} + s^2_{within}}$$

This index provides an indication of the percentage of variance that lies between schools. A low intraclass correlation indicates that schools are performing similarly while higher values point towards large differences between school performance.

To limit the reduction of precision in the population parameter estimate, multi-stage sample designs usually use supplementary information to improve coverage of the population diversity. In PISA the following techniques were implemented to limit the increase in the standard error: (i) explicit and implicit stratification of the school sampling frame and (ii) selection of schools with probabilities proportional to their size. Complementary information generally cannot compensate totally for the increase in the standard error due to the multi-stage design however but will greatly reduce it.

It is usual to express the effect of the sampling design on the standard errors by a statistic referred to as the design effect. This corresponds to the ratio of the variance of the estimate obtained from

the (more complex) sample to the variance of the estimate that would be obtained from a simple random sample of the same number of sampling units. The design effect has two primary uses – in sample size estimation and in appraising the efficiency of more complex sampling plans (Cochran, 1977).

In PISA, as sampling variance has to be estimated by using the 80 *BRR* replicates, a design effect can be computed for a statistic *t* using:

Formula 11.6

$$Deff(t) = \frac{Var_{BRR}(t)}{Var_{SRS}(t)}$$

where $Var_{BRR}(t)$ is the sampling variance for the statistic *t* computed by the *BRR* replication method, and $Var_{SRS}(t)$ is the sampling variance for the same statistic *t* on the same data but considering the sample as a simple random sample.

Based on a hypothetical country/economy, where the unbiased BRR standard error on the mean proficiency estimate is equal to 1.46, and the standard deviation is equal to *102.29*, on a sample of 14,530 students, the design effect for the mean proficiency estimate is therefore calculated as:

Formula 11.7

$$Deff(t) = \frac{Var_{BRR}(t)}{Var_{SRS}(t)} = \frac{(1.46)^2}{[102.29^2/14\,530]} = 2.96$$

This means the sampling variance on the proficiency estimate is about *2.96* times larger than it would have been with a simple random sample of the same sample size.

Another way to express the reduction of precision due to the complex sampling design is through the effective sample size, which expresses the simple random sample size that would give the same sampling variance as the one obtained from the actual complex sample design. The effective sample size for a statistic *t* is equal to:

Formula 11.8

$$Effn(t) = \frac{n}{Deff(t)} = \frac{n' \, Var_{SRS}(t)}{Var_{BRR}(t)}$$

where *n* is equal to the actual number of units in the sample. The effective sample size in our example would then be equal to:

Formula 11.9

$$Effn(t) = \frac{n}{Deff(t)} = \frac{14530}{2.96} = 4909$$

In other words, a simple random sample of about *4 909* students in this hypothetical country/economy would have been as precise as the actual sample for the national proficiency estimate.

VARIABILITY OF THE DESIGN EFFECT

Neither the design effect nor the effective sample size is a definitive characteristic of a sample. Both the design effect and the effective sample size vary with the variable and statistic of interest.

As previously stated, the sampling variance for estimates of the mean from a cluster sample is proportional to the intraclass correlation. In some countries/economies, student performance varies between schools. Students in academic schools usually tend to perform well while on average student performance in vocational schools is lower. Let us now suppose that the height of the students was also measured, and there are no reasons why students in academic schools should be of different height than students in vocational schools. For this particular variable, the expected value of the between-school variance should be equal to zero and therefore, the design effect should tend to one. As the segregation effect differs according to the variable, the design effect will also differ according to the variable.

The second factor that influences the size of the design effect is the choice of requested statistics. It tends to be large for means, proportions, and sums but substantially smaller for bivariate or multivariate statistics such as correlation and regression coefficients.

Design effects in PISA for performance variables

The notion of design effect as given earlier can be extended and gives rise to five different design effect formulae to describe the influence of the sampling and test designs on the standard errors for statistics.

The total errors computed for population estimates based on performance variables (scale scores) in the international PISA reports (OECD, 2019) consist of two components: sampling variance (Var_{BRR}) and measurement variance. The measurement variance is approximated by means of the imputation variance (*MVar*) which is calculated form the statistics calculated from imputed plausible values assigned to the participating students.

The standard error of proficiency estimates in PISA are inflated because the students were not sampled according to a simple random sample and because the estimation of student proficiency includes some amount of measurement error.

Therefore, the variance of a statistic calculated using plausible values is then calculated as the sum of the sampling and the imputation variances, or $Var_{BRR} + MVar$.

The five design effects and their respective effective sample sizes can then be defined as follows:

Design Effect 1: This design effect shows the inflation of the total variance that would have occurred due to measurement error if in fact the samples were considered as a simple random sample.

Formula 11.10

$$Deff_1(r) = \frac{Var_{SRS}(r) + MVar(r)}{Var_{SRS}(r)}$$

Design Effect 2: shows the inflation of the *total* variance due only to the use of a complex sampling design.

Formula 11.11

$$Deff_2(r) = \frac{Var_{BRR}(r) + MVar(r)}{Var_{SRS}(r) + MVar(r)}$$

Design Effect 3: shows the inflation of the sampling variance due to the use of a complex design. This is the same as Formula 11.7 introduced above.

Formula 11.12

$$Deff_3(r) = \frac{Var_{BRR}(r)}{Var_{SRS}(r)}$$

Design Effect 4: shows the inflation of the total variance due to measurement variance.

Formula 11.13

$$Deff_4(r) = \frac{Var_{BRR}(r) + MVar(r)}{Var_{BRR}(r)}$$

Design Effect 5: shows the inflation of the total variance due to the measurement variance and due to the complex sampling design.

Formula 11.14

$$Deff_5(r) = \frac{Var_{BRR}(r) + MVar(r)}{Var_{SRS}(r)}$$

Tables 11.11 through 11.15 present the values of the different design effects and the effective sample size using $Deff_5$, for each of the main PISA domains.

Table 11.11: Standard errors and related statistics for the average reading proficiency

Table 11.12: Standard errors and related statistics for the average mathematics proficiency

Table 11.13: Standard errors and related statistics for the average science proficiency

Table 11.14: Standard errors and related statistics for the average global competence proficiency (Forthcoming)

Table 11.15: Standard errors and related statistics for the average financial literacy proficiency

Table 11.16: Standard errors and related statistics for the average reading proficiency (Financial Literacy sample)

Table 11.17: Standard errors and related statistics for the average mathematics proficiency (Financial Literacy sample)

To better understand the design effect for a country/economy, some information related to the design effects and their respective effective sample sizes are presented in Annex C.

REFERENCES

Cochran, W. (1977), *Sampling Techniques (3rd ed.)*, John Wiley and Sons.

OECD (2019), *PISA 2018 Results (Volume I): What Students Know and Can Do*, PISA, OECD Publishing, Paris, *https://doi.org/10.1787/5f07c754-en*.