CHAPTER 11. SAMPLING OUTCOMES

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

POPULATION COVERAGE

Tables 11.1 and 11.2 (by adjudicated regions) show the quality indicators for population coverage and the various pieces of information used to derive them. The following notes explain the meaning of each coverage index and how the data in each column of the table were used.

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 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.

Coverage Index 1: Coverage of the national population, calculated by $P/(P+E) \times$(ST7b_3/ST7b_1):

- Coverage Index 1 shows the extent to which the weighted participants covered the final target population after all school exclusions. The following bullet points give details of its computation.
- In the preceding expression $P/(P+E)$ broadly represents the coverage proportion due to within school exclusion, and (ST7b_3/ST7b_1) the coverage proportion due to school level exclusion.
- The national population (NP) value, defined by Sampling Task 7b response box [1] and denoted here as ST7b_1 (and in Table 11.1 as the target population) is the population that includes all enrolled 15-year-old students in grades 7 and above in each participating country (with the possibility of small levels of exclusions), based on national statistics. However, the final NP value reflected for each country's school sampling frame might have had some school-level exclusions. The value that represents the population of enrolled 15-year-old students minus those in excluded schools is represented initially by response box [3] on Sampling Task 7b. It is denoted here as ST7b_3. As in PISA 2012, the procedure for PISA 2015 was that small schools having only one or two PISA-eligible students could not be excluded from the school frame but could be excluded in the field if the school still had only one or two PISA-eligible students at the time of data collection. Therefore, what is noted in Coverage Index 1 as ST7b_3 (and in Table 11.1 as target minus school level exclusions) was a number after accounting for all school level

exclusions, which means a number that omits schools excluded from the sampling frame in addition to those schools excluded in the field. Thus, the term (ST7b_3/ST7b_1) provides the proportion of the NP covered in each country based on national statistics.

- The value (*P*+*E*) provides the weighted estimate from the student sample of all PISA-eligible 15-year-olds in each participating country, where *P* is the weighted estimate of PISA-eligible non-excluded 15-year-old students and *E* is the weighted estimate of PISA-eligible 15-year-old students that were excluded within schools. Therefore, the term *P*/(*P*+*E*) provides an estimate, based on the student sample, of the proportion of the PISA-eligible 15-year-old population represented by the non-excluded PISA-eligible 15-year-old students.

- The result of multiplying these two proportions together (*P*/(*P*+*E*) and (ST7b_3/ST7b_1) indicates the overall proportion of the NP covered by the non-excluded portion of the student sample.

Coverage 1Index 2: Coverage of the national enrolled population, calculated by *P*/(*P*+*E*) × (ST7b_3/ST7a_2.1):

- Coverage Index 2 shows the extent to which the weighted participants covered the target population of all enrolled students in grades 7 and above.

- The national enrolled population (NEP), defined by Sampling Task 7a response box [2.1] and denoted here as ST7a_2.1 (and as enrolled 15-year-old students in Table 11.1), is the population that includes all enrolled 15-year-old students in grades 7 and above in each participating country, based on national statistics. The final NP, denoted here as ST7b_3 as described above for Coverage Index 1, reflects the 15- year-old population after school-level and other small exclusions. This value represents the population of enrolled 15-year-old students less those in excluded schools;

- The value (*P*+*E*) provides the weighted estimate from the student sample of all eligible 15-year-olds in each country, where *P* is the weighted estimate of PISA-eligible non-excluded 15-year-old students and *E* is the weighted estimate of PISA-eligible 15-year-old students that were excluded within schools. Therefore, the term *P*/(*P*+*E*) provides an estimate based on the student sample of the proportion of the PISA-eligible 15-year-old population that is represented by the non-excluded PISA-eligible 15-year-old students;

- Multiplying these two proportions together (*P*/(*P*+*E*) and (ST7b_3/ST7a_2.1)) gives the overall proportion of the NEP that was covered by the non-excluded portion of the student sample.

Coverage Index 1 and Coverage Index 2 will differ when countries have excluded geographical areas or language groups apart from other school level exclusions. In these cases Coverage Index 2 will be less than Coverage Index 1.

Coverage Index 3: Coverage of the national 15-year-old population, calculated by *P*/ST7a_1:

- The national population of 15-year-olds, defined by Sampling Task 7a response box [1] and denoted here as ST7a_1 (and called all 15-year-olds in Table 11.1), is the entire population of 15-year-olds in each country (enrolled and not enrolled), based on national statistics. The value *P* is the weighted estimate of PISA-eligible non-excluded 15-year-old students from the student sample. Thus (*P*/ST7a_1) indicates the proportion of the national population of 15-year-olds covered by the non-excluded portion of the student sample. It therefore also reflects the proportion of 15-year-olds excluded or not at school.

Coverage Index 4: Coverage of the estimated school population, calculated by (*P*+*E*)/*S:*

- The value (*P*+*E*) provides the weighted estimate from the student sample of all PISA-eligible 15-year-old students in each country, where *P* is the weighted estimate of PISA-eligible non-excluded 15-year-old students and *E* is the weighted estimate of PISA-eligible 15-year-old students who were excluded within schools.

- The value *S* is an estimate of the 15-year-old school population in each participating country (called estimate of enrolled students from frame in Table 11.1). This is based on the actual or (more often) approximate number of 15-year-old students enrolled in each school in the sample, prior to contacting the school to conduct the assessment. The *S* value is 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 (*ENR*) as recorded on the school sampling frame.

- Thus, (*P+E)/S* is the proportion of the estimated school 15-year-old population that is represented by the weighted estimate from the student sample of all PISA-eligible 15-year-old students. It will be influenced by the accuracy of the school sample frame, fluctuations in the target population size and the accuracy of the within school sampling process. Its purpose is to check whether the student sampling has been carried out correctly, and to assess whether the value of *S* is a reliable measure of the number of enrolled 15-year-olds. This is important for interpreting Coverage Index 5.

Coverage Index 5: Coverage of the school sampling frame population, calculated by *S*/ST7b_3:

- The value (S/ST7b_3) is the ratio of the enrolled 15-year-old population, as estimated from data on the school sampling frame, to the size of the enrolled student population, as reported on Sampling Task 7b and adjusted by removing any additional excluded schools in the field. In some cases, this provided a check as to whether the data on the sampling frame gave a reliable estimate of the number of 15-year-old students in each school. In other cases, however, it was evident that ST7b_3 had been derived using data from the sampling frame by the NPM , so that this ratio may have been close to 1.0 even if enrolment data on the school sampling frame were poor. Under such circumstances, Coverage Index 4 would differ noticeably from 1.0, and the figure for ST7b_3 would also be inaccurate.

Table 11.1: PISA target populations and samples

Table 11.2: PISA target populations and samples, by adjudicated regions

SCHOOL AND STUDENT RESPONSE RATES

Tables 11.3 to 11.8 present school and student-level response rates at the national and regional levels.

- Tables 11.3 and 11.4 (by adjudicated regions) indicate the rates calculated by using only original schools and no replacement schools.
- Tables 11.5 and 11.6 (by adjudicated regions) indicate the improved response rates when first and second replacement schools were accounted for in the rates.
- Tables 11.7 and 11.8 (by adjudicated regions) indicate the student response rates among the full set of participating schools.

For 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.

Table 11.3: Response rates before school replacement

Table 11.4: Response rates before school replacement, by adjudicated regions

For 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.

In 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 with few certainty school selections and no over-sampling or under-sampling of any explicit strata, weighted and unweighted rates are very similar. The weighted school response rate before replacement is given by the formula:

$$\begin{matrix} weighted\ school\ response\ rate \\ before\ replacement \end{matrix} = \frac{\sum_{i \in Y} W_i E_i}{\sum_{i \in (Y \cup N)} W_i E_i} \quad 12\backslash * \text{ MERGEFORMAT ()}$$

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.

Table 11.5: Response rates after school replacement

Table 11.6: Response rates after school replacement, by adjudicated regions

The weighted school response rate, after replacement, is given by the formula:

$$\begin{matrix} weighted\ school\ response\ rate \\ after\ replacement \end{matrix} = \frac{\sum_{i \in (Y \cup R)} W_i E_i}{\sum_{i \in (Y \cup R \cup N)} W_i E_i} \quad 34\backslash * \text{ MERGEFORMAT ()}$$

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 less 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.

For weighted student response rates, the same number of students appears in the numerator and denominator as for unweighted rates, but each student was 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 with no over-sampling of any explicit strata, weighted and unweighted student participation 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.

Table 11.7: Response rates, students within schools after school replacement

Table 11.8: Response rates, students within schools after school replacement, by adjudicated regions

TEACHER RESPONSE RATES

Unweighted response rates for both science and non-science 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 science teachers and in Table 11.10 for the non-science teachers.

In addition to these rates, unweighted response rates were calculated also for each sampled school in each country 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, 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.

A simple random sample of 4 000 students 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:

$$\sigma_{(\hat{\mu})} = \sqrt{\frac{\sigma^2}{n}} \qquad 56\backslash* \text{ MERGEFORMAT ()}$$

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:

$$\sigma_{(\hat{\mu})} = \sqrt{\frac{\sigma^2_{schools}}{n_{schools}} + \frac{\sigma^2_{within}}{n_{schools} n_{students}}} \qquad 78\backslash* \text{ MERGEFORMAT ()}$$

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:

$$Rho = \frac{\sigma^2_{schools}}{\sigma^2_{schools} + \sigma^2_{within}} \qquad 910\backslash* \text{ MERGEFORMAT ()}$$

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.

Table 11.9 provides the standard errors on the PISA 2015 main domain scales, calculated as if the participating country sample was selected according to (i) a simple random sample; (ii) a multistage procedure without using complementary information (unstratified multistage sampling, with sampling weights ignored) and (iii) the unbiased BRR estimate for the actual PISA 2015 design, using Fay's method. It should be mentioned that the plausible value imputation variance was not included in these computations, which thus only reflect sampling error.

Note that the values in Table 11.9 for the standard errors for the unstratified multistage design are overestimates for countries that had a school census (Cyprus, Iceland, Luxembourg, Macao-China, Malta, Trinidad and Tobago, and Qatar) since these standard error estimates assume a sample of schools was collected.

Also note that in some of the countries where the BRR estimates in Table 11.9 are greater than the values for the unstratified multistage sample, this is because of regional or other oversampling (The countries with oversampling were: Australia, Argentina, Belgium, Brazil, Canada, China, Colombia Czech Republic, Denmark, Italy, Malaysia, Portugal, United Arab Emirates, United Kingdom).

The BRR estimates in Table 11.9 are also greater than the values for the unstratified multistage sample for almost all countries since nearly every country undersamples very small schools. As described in the sampling design chapter, some countries have a substantial proportion of students attending schools that have fewer students than the target cluster size (*TCS*). When small school undersampling was done, very small schools were undersampled while all other schools were slightly oversampled in compensation. In such cases, very small schools with at most 0, 1, or 2 age-eligible PISA students expected to be enrolled were typically undersampled by a factor of 4 while the very small schools with between 3 and TCS/2 age-eligible PISA students expected to be enrolled were undersampled by a factor of 2. This takes the allocation of schools to strata slightly away from proportional allocation, which can add slightly to weight variability and therefore to sampling variance. This is done though, to help countries minimize the operational burden of having too many small schools in their sample.

For the other instances of countries in Table 11.9 that have BRR estimates that are somewhat greater than estimates based on an unstratified multistage design it is unclear why the BRR variance should be larger, though it is possible that the stratification undertaken possibly did not explain enough between-school variance in these countries.

Table 11.9: Standard errors for the PISA 2015 main domain scales

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:

$$Deff(t) = \frac{Var_{BRR}(t)}{Var_{SRS}(t)} \qquad (6)$$

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 Table 11.9, the unbiased BRR standard error on the mean estimate in science in Australia (for example) is equal to 1.46 (rounded from *1.45568*). As the standard deviation of the

science performance is equal to *102.29735*, the design effect in Australia for the mean estimate in science is therefore equal to:

$$Deff(t) = \frac{Var_{BRR}(t)}{Var_{SRS}(t)} = \frac{(1.45568)^2}{\left[102.29735^2 / 14530\right]} = 2.942195 \quad (7)$$

The sampling variance on the science performance mean in Australia is about *2.94* times larger than it would have been with a simple random sample of the same sample size. Note that the participating students are 14 530 as this number were assessed for science.

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:

$$Effn(t) = \frac{n}{Deff(t)} = \frac{n \times Var_{SRS}(t)}{Var_{Brr}(t)} \quad (8)$$

where n is equal to the actual number of units in the sample. The effective sample size in Australia for the science performance mean is equal to:

$$Effn(t) = \frac{n}{Deff(t)} = \frac{14530}{2.942195} = 4938.4898 \quad (9)$$

In other words, a simple random sample of *4938* students in Australia would have been as precise as the actual PISA 2015 sample for the national estimate of mean science proficiency.

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, 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 is 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 the international PISA initial report, *PISA 2015 Results* (OECD, 2016) that involves performance variables (scale scores) consist of two components: sampling variance and measurement variance. The standard error of proficiency estimates in PISA is inflated because the students were not sampled according to a simple random sample and also because the estimation of student proficiency includes some amount of measurement error.

For any statistic *r*, the population estimate and the sampling variance are computed for each plausible value and then combined as described in Chapter 9.

The five design effects and their respective effective sample sizes are defined as follows:

- Design Effect 1

$$Deff_1(r) = \frac{Var_{SRS}(r) + MVar(r)}{Var_{SRS}(r)} \tag{10}$$

  where *MVar(r)* is the measurement variance for the statistic *r*. 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.

- Design Effect 2

$$Deff_2(r) = \frac{Var_{BRR}(r) + MVar(r)}{Var_{SRS}(r) + MVar(r)} \tag{11}$$

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

- Design Effect 3

$$Deff_3(r) = \frac{Var_{BRR}(r)}{Var_{SRS}(r)} \tag{12}$$

shows the inflation of the sampling variance due to the use of a complex design.

- Design Effect 4

$$Deff_4(r) = \frac{Var_{BRR}(r) + MVar(r)}{Var_{BRR}(r)} \tag{13}$$

shows the inflation of the total variance due to measurement variance.

- Design Effect 5

$$Deff_5(r) = \frac{Var_{BRR}(r) + MVar(r)}{Var_{SRS}(r)} \tag{14}$$

shows the inflation of the *total* variance due to the measurement variance and due to the complex sampling design.

The product of the first and second design effects equals the product of the third and fourth design effects, and both products are equal to the fifth design effect.

Tables 11.12 through 11.16 present for each of the major domains the values of the different design effects and the corresponding effective sample sizes.

Table 11.12: Design effects and Effective Sample Sizes for Scientific Literacy

Table 11.13: Design effects and Effective Sample Sizes for Mathematical Literacy

Table 11.14: Design effects and Effective Sample Sizes for Reading Literacy

Table 11.15: Design effects and Effective Sample Sizes for Collaborative Problem Solving

Table 11.16: Design effects and Effective Sample Sizes for Financial Literacy

To better understand the design effect for a particular country, some information related to the design effects and their respective effective sample sizes, are presented in Annex C. In particular, the design effect and the effective sample size depend on:

- ***The sample size***, the number of participating schools, the number of participating students and the average within-school sample size, which are provided in Table C.2 (Annex C);
- ***The school variance***, school variance estimates and the intraclass correlation, which are provided respectively in Tables C.3 and C.4 (Annex C);
- ***The stratification variables***, the intraclass correlation coefficient within explicit strata and the percentage of school variance explained by explicit stratification variables, which are provided respectively in Tables C.5 and C.6 (Annex C).

Finally, the standard errors on the mean performance estimates are provided in Table C.1 (Annex C).