Chapter 4

PIRLS 2006 Sample Design

Marc Joncas

4.1 Overview

This chapter describes the PIRLS 2006 sample design, which consists of a set of specifications for the target and survey populations, sampling frames, survey units, sample selection methods, sampling precision, and sample sizes. The sample design is intended to ensure that the PIRLS 2006 survey data provide accurate and economical estimates of national student populations. Since measuring trends is a central goal of PIRLS, the sample design also aims to provide accurate measures of changes in student achievement from 2001 to 2006. In addition to the sample design, the PIRLS 2006 sampling activities also include estimation procedures for sample statistics and procedures for measuring sampling error. These other components are described in Chapters 9 and 12, respectively. The basic PIRLS sample design has two stages: schools are sampled with probability proportional to size at the first stage, and one or two intact classes of students from the target grade are sampled at the second stage.

All participants followed the uniform sampling approach specified by the PIRLS 2006 sample design, with minimum deviations. This ensured that high quality standards were maintained for all participants, avoiding the possibility that differences between countries in survey results could be attributable to the use of different sampling methodologies. This uniform approach also facilitated an efficient approval process of the national designs by the international project team.

The PIRLS National Research Coordinator (NRC) of each participating country was responsible for implementing the sample design, including documenting every step of the sampling procedure for approval by the TIMSS & PIRLS International Study Center and Statistics Canada prior to implementation. To support NRCs in their sampling activities, a series of manuals (the *School Sampling Manual (*PIRLS, 2004*),* the *Survey Operations Procedures* (PIRLS*,* 2005b*)*, and the *School Coordinator Manual* (PIRLS, 2005a) and sampling software (IEA, 2005)) were provided. In addition to these materials, Statistics Canada consulted with each country throughout the process.

4.2 PIRLS 2006 Target Population

PIRLS is a study of student achievement in reading comprehension in primary school, and is targeted at the grade level in which students are at the transition from learning to read to reading to learn, which is the fourth grade in most countries. The formal definition of the PIRLS target population makes use of UNESCO's International Standard Classification of Education (ISCED) in identifying the appropriate target grade:

> …all students enrolled in the grade that represents four years of schooling, counting from the first year of ISCED Level 1, providing the mean age at the time of testing is at least 9.5 years. For most countries, the target grade should be the fourth grade, or its national equivalent.

ISCED Level 1 corresponds to primary education or the first stage of basic education, and should mark the beginning of "systematic apprenticeship of reading, writing, and mathematics" (UNESCO, 1999). By the fourth year of Level 1, students have had 4 years of formal instruction in reading, and are in the process of becoming independent readers.

In IEA studies, the above definition corresponds to what is known as the *international desired target population*. Each participating country was expected to define its *national desired population* to correspond as closely as possible to this definition (i.e., its fourth grade of primary school). In order to measure trends, it was critical that countries that participated in PIRLS 2001, the previous cycle of PIRLS, choose the same target grade for PIRLS 2006 that was used in

PIRLS 2001. Information about the target grade in each country is provided in Chapter 9.

Although countries were expected to include all students in the target grade in their definition of the population, sometimes it was not possible to include all students who fell under the definition of the international desired target population. Consequently, occasionally a country's *national desired target population* excluded some section of the population, based on geographic or linguistic constraints. For example, Lithuania's national desired target population included only students in Lithuanian-speaking schools, representing approximately 93 percent of the international desired population of students in the country.

Working from the national desired population, each country had to operationalize the definition of its population for sampling purposes and define their *national defined population.* While this national defined target population should ideally coincide with the national desired target population, in reality, there may be some regions or school types that cannot be included. All students in the desired population who were not included in the defined population are referred to as the excluded population.

PIRLS participants were expected to ensure that the national defined population included at least 95 percent of the national desired population of students. Exclusions (which had to be kept to a minimum) could occur at the school level, within the sampled schools, or both. Although countries were expected to do everything possible to maximize coverage of the national desired population, *school-level exclusions* sometimes were necessary. Keeping within the 95 percent limit, school-level exclusions could include schools that:

- were geographically remote,
- had very few students,
- had a curriculum or structure different from the mainstream education system, or
- were specifically for students with special needs.

The difference between these school-level exclusions and those at the previous level is that these schools were included as part of the sampling frame (i.e., the list of schools to be sampled). They then were eliminated on an individual basis if it was not feasible to include them in the testing.

In many education systems, students with special educational needs are included in ordinary classes. Due to this fact, another level of exclusions is necessary to reach an effective target population—the population of students who ultimately will be tested. These are called *within-school exclusions* and pertain to students who are unable to be tested for a particular reason but are part of a regular classroom. There are three types of within-school exclusions, which are explained below.

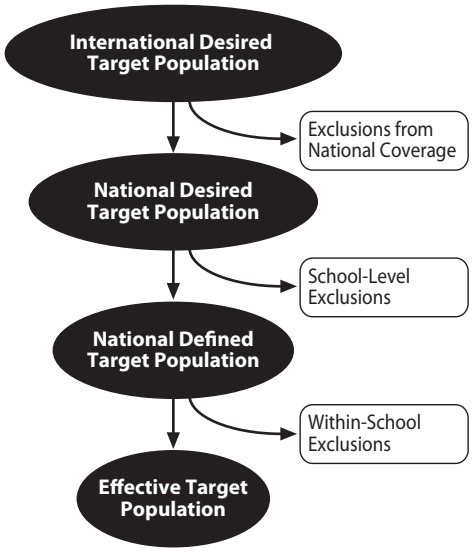
- **Intellectually disabled students:** These are students who are considered in the professional opinion of the school principal, or by other qualified staff members, to be intellectually disabled or who have been tested psychologically as such. This includes students who are emotionally or mentally unable to follow even general test instructions. Students should not be excluded solely because of poor academic performance or normal disciplinary problems.
- **Functionally disabled students:** These are students who are permanently, physically disabled in such a way that they cannot perform in the PIRLS testing situation. Functionally disabled students who are able to respond should be included in the testing.
- **Non-native language speakers:** These are students who are unable to read or speak the language(s) of the test and would be unable to overcome the language barrier of the test. Typically, a student who has received less than 1 year of instruction in the language(s) of the test should be excluded, but this definition may need to be adapted in different countries.

Students eligible for within-school exclusion were identified by staff at the schools and could still be administered the test if the school did not want the student to feel out of place during the assessment (though the data from these students were not included in any analyses). Again, it was important to ensure that this population was as close to the national desired target population as possible.

If combined, school-level and within-school exclusions exceeded 5 percent of the national desired target population, results were annotated in the *PIRLS 2006 International Report* (Mullis, Martin, Kennedy, & Foy, 2007). Target population coverage and exclusion rates are displayed for each country in Chapter 9. Descriptions of the countries' school-level and within-school exclusions can be found in Appendix B.

In any study that utilizes sampling, the population that ultimately participates usually differs slightly from the target population, with some portion of the target population being excluded from the study. A major objective of the PIRLS sampling strategy was to ensure that the effective target population, the population actually sampled by PIRLS, was as close as possible to the international desired population, and to document clearly all excluded populations. Exhibit 4.1 illustrates the relationship between successively more refined definitions of the target population and the excluded populations at each stage.

Exhibit 4.1 Relationship Between the Desired Populations and Exclusions



4.3 Sample Design

Once the survey population was defined, the next step involved building a sampling frame in which all sampling units within the national defined target population have a known probability of being sampled. In PIRLS 2006, however, it is important to note that in addition to gathering data on sampled students, a large amount of information also was gathered about their classes and schools, which required other types of sampling units. The intrinsic, hierarchical nature of these nested units necessitated the creation of a sampling frame by stages.

Therefore, a two-stage stratified cluster sample design was used, with schools as the first stage and intact classes as the second.[1]

4.3.1 Sampling Precision and Sample Size

Because PIRLS is fundamentally a study of reading comprehension among fourth-grade students, the precision of survey estimates of student characteristics was of primary importance. However, PIRLS reports extensively on school, teacher, and classroom characteristics also, so it is necessary to have sufficiently large samples of schools and classes. The PIRLS standard for sampling precision requires that all student samples should have an effective sample size of at least 400 students for the main criterion variable, which is reading achievement. In other words, all student samples should yield sampling errors that are no greater than would be obtained from a simple random sample of 400 students.

Given that sampling error, when using simple random sampling, can be expressed as $SE_{SRS} = S/\sqrt{n}$ where S gives the population standard deviation and n the sample size, a simple random sample of 400 students would yield a 95 percent confidence interval for an estimate of a student-level mean of plus and minus 10 percent of its standard deviation (1.96 times $1/\sqrt{400}$ times S). Because the PIRLS achievement scale has a standard deviation of 100 points, this translates into a ±10 point confidence interval (or a standard error estimate of approximately 5 points). Similarly, sample estimates of student-level percentages would have confidence intervals of approximately ±5 percentage points.

Notwithstanding these precision requirements, PIRLS required that all student sample sizes should not be less than 4,000 students. This was necessary to ensure adequate sample sizes for analyses where the student population was broken down into many subgroups. Furthermore, since PIRLS planned to conduct analyses at the school and classroom level in addition to the student level, all school sample sizes were required to be not less than 150 schools, unless a complete census fails to reach this minimum. Under simple random sampling assumptions, a sample of 150 schools yields a 95 percent confidence interval for an estimate of a school-level mean that is plus and minus 16 percent of a standard deviation.

Although the PIRLS sampling precision requirements are such that they would be satisfied by a simple random sample of 400 students, student samples chosen using multi-stage cluster designs, such as the PIRLS 2006 school-and-class design, typically require much larger student samples to achieve the same

1 Because their large population size, it was necessary to include a preliminary sampling stage in the United States and the Russian Federation, where regions were sampled first, and then schools.

level of precision. Because students in the same school, and even more so in the same class, tend to be more like each other than like other students in the population, sampling a single class of 30 students will yield less information per student than a random sample of students drawn from across all students in the population. PIRLS uses the intra-class correlation, a statistic indicating how much students in a group are similar on an outcome measure, and a related measure known as the design effect, to adjust for this "clustering" effect in planning sample sizes.

For countries taking part in PIRLS for the first time in 2006, we used the following mathematical formulas to estimate how many schools should be sampled to achieve an acceptable level of sampling precision.

$$Var_{PPS} = Deff * Var_{SRS} = \frac{Deff * S^2}{n} = \frac{\left[1+\rho(mcs-1)\right] * S^2}{n} = \frac{\left[1+\rho(mcs-1)\right] * S^2}{a * mcs}$$

Where *Deff* is a compensation factor for using a sample selection method that differs from a simple random sample (also called design effect). S^2 gives the variance of the population, ρ measures the intra-class correlation between clusters, *mcs* corresponds to the average number of sampled students per class (assuming one class per school), and *a* gives the number of schools to sample. Incorporating the precision requirements into this equation gives the number of schools required as:

(1)
$$a = 400 * \frac{\left[1+\rho(mcs-1)\right]}{mcs}$$

For planning purposes, the intra-class correlation coefficient was usually set to 0.3 if no other information was available. For example, with a MCS of 20 students and a ρ of 0.3, equation (1) gives 134 schools.

Equation (1) is a model for determining how many schools would be required for the PIRLS 2006 sample under the assumption that the standard error of the criterion variable (student reading achievement) reflects only sampling variance—the usual situation in sample surveys. However, because of its complex matrix-sampling assessment design, standard errors in PIRLS include an imputation error component in addition to the usual sampling error component (see Chapter 11). To keep the standard error within the prescribed

precision limits, the number of schools determined by equation (1) have to be increased, as shown in equation (2):

(2)

$$a_{imp} = (400 * 0.5) / mcs$$

Continuing the example for a country with a MCS of 20 students, according to equation (2), 10 schools would have to be added to the 134 schools from equation (1), for a total of 144 schools.

For PIRLS 2006 countries that also had participated in PIRLS 2001, the standard error estimates computed from the 2001 data were reviewed to ensure that the student samples had been large enough to meet the precision requirements in 2001 and would be sufficiently precise to measure trends to 2006. For the several countries falling somewhat short of the sampling requirements not met in 2001, the school sample size for 2006 was increased, using as a rule of thumb that sampling error is inversely proportional to the square root of the sample size. For example, if the sample size in 2001 yielded a standard error of 7 points for an estimate of a mean, the sample size in 2006 was increased by a factor of 2 to provide a standard error of 5 points $((7/5)^2=2)$. Intra-class correlation coefficients also were calculated for countries that participated in PIRLS 2001. These coefficients were presented in the *PIRLS 2006 School Sampling Manual* (PIRLS, 2004).

4.3.2 Stratification

Stratification is the grouping of sampling units into smaller sampling frames according to information found on the initial sampling frame prior to sampling, and may be employed to improve the efficiency of the sample design, to sample sections of the population at different rates, or to ensure adequate representation of specific groups in the sample. The stratification by itself can take two forms: explicit or implicit.

Explicit stratification physically creates smaller sampling frames from which samples of schools and classes will ultimately be drawn. In PIRLS, this type of stratification is used when the usual proportional allocation (i.e., students in certain regions or types of school are represented in the sample in proportion to their distribution in the population) may not result in adequate representation of some groups in the sample. For example, if a country wanted to make generalizations regarding the reading achievement of private sector

students, the sampling frame could be split into two strata—public and private sector schools. The sample could then be allocated between the two strata to achieve the desired level of precision in each. In most countries in PIRLS 2006, the sample allocation among strata was proportional to the number of students found in each stratum. However, it could be noted in passing that, even without any stratification, the PIRLS samples represented the different groups found in the population, on average.

Implicit stratification only requires that the sampling frame is sorted according to some variable(s) prior to sampling and can be nested within explicit stratification. By combining the sorting of the frame with a systematic sampling of the units, we get a sample where units are in the same proportions as those found at the population level. When schools from the same implicit stratum tend to have similar behavior, in terms of reading achievement, implicit stratification will produce more reliable estimates.

In the basic PIRLS 2006 sample design, all schools in the sampling frame for a country were sorted according to some measure of their size (MOS—see next section). If implicit stratification was used, then the sorting by MOS was done within each stratum using a serpentine approach—high to low for the first stratum, followed by low to high for the next, and so on (see example in Exhibit 4.2).

Exhibit 4.2 MOS Sort Order Across Implicit Strata

Implicit Stratum	Sort Order of MOS
1. Rural – Public	High to Low
2. Rural – Private	Low to High
3. Urban – Public	High to Low
4. Urban – Private	Low to High

This way of sorting sampling units optimizes the chances of choosing replacement schools with a MOS close to the original sampled schools they are meant to replace.

4.3.3 Replacement Schools

Ideally, response rates to study samples should always be 100 percent, and although the PIRLS 2006 participants worked hard to achieve this goal, it was anticipated that a 100 percent participation rate would not be possible in all

countries. To avoid sample size losses, the PIRLS sampling plan identified, *a priori*, replacement schools for each sampled school. Therefore, if an originally selected school refused to participate in the study, it was possible to replace it with a school that already was identified prior to school sampling. Each originally selected school had up to two pre-assigned replacement schools. In general, the school immediately following the originally selected school on the ordered sampling frame and the one immediately preceding it were designated as replacement schools. Replacement schools always belonged to the same explicit stratum, although they could come from different implicit strata if the originally selected school was either the first or last school of an implicit stratum.

The main objective for having replacement schools in PIRLS 2006 was to ensure adequate sample sizes for analysis of sub-population differences. Although the use of replacement schools did not eliminate the risk of bias due to nonresponse, employing implicit stratification and ordering the school sampling frame by size increased the chances that any sampled school's replacements would have similar characteristics. This approach maintains the desired sample size while restricting replacement schools to strata where nonresponse occurred. Since the school frame is ordered by school size, replacement schools also tended to be of the same size as the school they were meant to replace. For the field test, replacement schools were used to make sure sample sizes were large enough to validate new items, and no more than one replacement school was assigned per originally selected school.

4.4 Sample Selection

The school sampling selection method used in PIRLS 2006 is a classic approach that can be found in most sampling textbooks (e.g., Cochran, 1997). The method is usually referred to as a systematic *probability proportional-to-size* (PPS) technique. This sampling method is a natural match with the hierarchical nature of the sampling units, with classes of students nested within schools. Even if a country had a list from which students could be selected directly, the sampling frame for most of the countries participating in PIRLS was first made of schools. From these sampled schools, lists of classes were created and sampled. For each sampled class, a list of students was created.

4.4.1 Sampling Schools

In order to draw school samples representative of the student population, NRCs were asked to provide vital information about the schools within the sampling frame. The following data were required for each school:

- A *measure of size* (MOS) (e.g., the average student enrollment in the fourth grade, the number of classrooms in the fourth grade, or the total student enrollment in the school);
- The expected number of sampled students per class, also called *minimum cluster size* (MCS). This was required if the number of classrooms in the fourth grade couldn't be provided and was calculated as the ratio of the total number of students to the total number of classes for schools having more than one class in the fourth grade; and
- Any variables describing school characteristics to be used for stratification purposes, such as type of school, degree of urbanization, or sex of students served by the school.

Schools were sampled using systematic random sampling with probability proportional to their MOS. For example, if school A had a MOS value twice as large as school B, then School A had twice the chance of being in the sample compared to school B. Similarly, if region A had a MOS value twice as large as region B, then region A had twice the chance of being in the sample.

To implement the school sampling, schools in each explicit stratum were sorted in order by the implicit stratification variables and within these by the MOS. The measures of size are accumulated from school to school, and a running total, the cumulative measure of size, is recorded next to each school. The cumulative MOS is an indicator of the size of the population of sampling elements (students). Dividing the cumulative MOS by the number of schools to sample gives the sampling interval.

With systematic PPS sampling, it is possible for a large sampling unit to be selected more than once if its size is greater than the sampling interval. To avoid this situation, all such units were automatically selected by changing their MOS to the sampling interval of the associated explicit stratum.

Some schools have so few students that their selection using probability proportional to their size (MOS) becomes problematic. Since the selection of these schools depends on their size, a difference between the number of expected students when drawing the sample and the number of students actually

found in the field can substantially contribute to the sampling error. To lessen the impact of this eventuality, any schools with fewer expected students than the average minimum cluster size (MCS) for the explicit stratum were sampled with equal probabilities. For example, if the MCS was 30 students and there were 28 schools with less than 30 students for a total of 476 students, the MOS of these small schools was changed to 476/28 = 17. By doing this, the overall size of the explicit stratum stayed the same but all small schools had an equal chance of being selected.

The MCS also was used to define very small schools. Whenever a school had an expected number of students less than one quarter of the average MCS, the school was labeled as a very small school. These schools could be excluded, as long as they did not exceed 2 percent of the national desired target population and the overall exclusion rate did not exceed 5 percent.

4.4.1 Sampling Classes

For all participants to PIRLS 2006 but two (Morocco and Singapore),[2] intact student classes were the second and final sampling stage, with no student subsampling. This means that all students within sampled classes participated in PIRLS 2006, with the exception of excluded students and students absent the day of the assessment. Classes were selected with equal probability of selection using systematic random sampling. Within each sampled school, all fourth-grade classes were listed, and one or two classes were sampled, using a random start (different in each sampled school). This method, combined with the PPS sampling method for schools, results in a self-weighted student sample under the following conditions: a) there is a perfect correlation between the school MOS reported in the sampling frame and the actual school size; b) the same number of classes is selected in each school and c) the MCS is the same for all schools. Given that these conditions were never totally met, student sampling weights varied somewhat from school to school (see Chapter 9 for details about sampling weights).

Within sampled schools, some classes have so few students that it was unreasonable to go through the sampling process and end up with these small classes. Furthermore, small classes tend to increase the risk of unreliable survey estimates. To avoid these problems, a class smaller than half the specified MCS was combined with another class from the same school prior to class sampling.

2 Two classes per school were selected using systematic PPS sampling in Singapore, and then 19 students were sampled within each class. One class per school was selected using PPS sampling in Morocco, with 25 students (all student if less than 25 students in the class) were sampled within each class.

4.5 Selecting Field-test Samples

Prior to the main data collection, which was conducted October–November 2005 in Southern Hemisphere countries and April–May 2006 in Northern Hemisphere countries, PIRLS 2006 conducted a full-scale field test in April 2005 in all participating countries. The field-test sample size was approximately 30 schools in each country. Countries were required to draw their field-test samples using the same random sampling procedures that they employed for the main sample. This ensured that field-test samples approximated closely the main samples, while reducing the burden on schools, the field-test and main data collection samples were drawn simultaneously, so that a school could be selected for either the field test or the main data collection, but not both. For example, if 150 schools were needed for the main data collection and another 30 schools needed for the field test, a larger sample of 180 schools was selected using the sampling method described earlier. A systematic subsample of 30 schools then was selected from the 180 schools and assigned to the field test, leaving 150 schools for data collection.[3]

References

Cochran, W. G. (1997). *Sampling techniques*. New York: John Wiley.

IEA. (2005). *WinW3S: Within-school sampling software manual*. Hamburg: IEA Data Processing and Research Center.

Mullis, I.V.S., Martin, M.O., Kennedy, A.M., & Foy, P. (2007). *PIRLS 2006 international report: IEA's progress in international reading literacy study in primary schools in 40 countries.* Chestnut Hill, MA: Boston College, PIRLS.

TIMSS & PIRLS International Study Center. (2004). *PIRLS 2006 school sampling manual*. Chestnut Hill, MA: Boston College.

TIMSS & PIRLS International Study Center. (2005a). *PIRLS 2006 school coordinator manual*. Chestnut Hill, MA: Boston College.

TIMSS & PIRLS International Study Center. (2005b). *PIRLS 2006 survey operations procedures unit 1: Contacting schools and sampling classes*. Chestnut Hill, MA: Boston College.

UNESCO Institute for Statistics. (1999). *Operational manual for ISCED-1997 : International standard classification of education*.

3 In countries where it was necessary to conduct a complete census of all schools, or where the NRC believed that the sampling frame used to draw the combined sample was not appropriate for the data collection, separate sampling frames were provided for the field test and main data collection. In such situations, no attempt was made to minimize the overlap. This issue is discussed in more detail in Appendix B.