Foy, P., Rust, K., and Schleicher, A. (1996) “Sample Design” in M.O. Martin and D.L. Kelly (eds.), *Third International Mathematics and Science Study (TIMSS) Technical Report, Volume I: Design and Development*. Chestnut Hill, MA: Boston College.

# 4. Sample Design

Pierre Foy
Keith Rust
Andreas Schleicher

## 4.1 OVERVIEW

This chapter describes the procedures developed to ensure that the student populations that were the focus of the study were properly sampled in each participating country. To be acceptable for TIMSS, national sample designs had to result in probability samples which give accurate weighted estimates of population parameters, and for which estimates of sampling variance could be computed. An effort was made in designing TIMSS to strike a balance between the analytical requirements and operational constraints, while keeping the survey design simple enough for all participants to implement it. The selection of valid and efficient samples was crucial to the success of the project. The accuracy of the survey results are dependent on the quality of the sampling information available at the design stage, and particularly on the implementation of the sampling procedures.

The National Research Coordinators (NRCs) were aware that in a study as ambitious as TIMSS, the sample design and sampling procedures would be complex, and that the gathering of the required information about the national education systems would place considerable demands on resources and expertise. At the same time, those directing and coordinating the project realized that the national centers had only limited numbers of qualified sampling personnel. Simplifying the sampling procedures to the extent possible, especially the sample selection within schools, was thus a major consideration in developing the sample design. Sometimes simplicity and practicality had to be given a higher priority than optimizing the sample design in terms of precision and cost.

NRCs were allowed to adapt the sample design for their educational system, using more sampling information and more sophisticated sample designs and procedures than the base design provided. However, these solutions had to be approved and monitored by the international project management (the International Coordinating Center at the University of British Columbia, Vancouver, until August 1993, and the International Study Center at Boston College thereafter).

The international project management provided manuals and expert advice to help NRCs to adapt the TIMSS sample design to their national system, and to guide them through the phases of sampling. The *Sampling Plan* (TIMSS, 1992) provided an overview of the sample design and described the survey design options offered. The *Sampling Manual* (TIMSS, 1994a) described how to implement the sampling plan and offered advice on initial planning, working within constraints, establishing appropriate sample selection procedures, and fieldwork. The *Survey Operations Manuals* (TIMSS, 1994d, 1994e) and *School Coordinator Manuals* (1994b, 1994c) provided information on sample selection and execution within schools, the assignment of rotated test instruments to selected students, and administration and monitoring procedures used to identify and track respondents and nonrespondents. NRCs also received software designed to automate the sometimes complex within-school sampling procedures.

NRCs also had several sources of expert support. Statistics Canada, in consultation with the TIMSS sampling referee and the TIMSS Technical Advisory Committee (TAC), reviewed and approved the national sampling plans, sampling data, and sampling frames, and the sample execution. In addition, Statistics Canada provided advice and support to NRCs at all stages of the sampling process.

## 4.2 TARGET POPULATIONS AND EXCLUSIONS

In IEA studies, the target population for all countries is known as the *International Desired Population*. TIMSS chose to study student achievement in three such populations in each country. The international desired populations for TIMSS were as follows:

- **Population 1.** All students enrolled in the two adjacent grades that contain the largest proportion of 9-year-olds at the time of testing.
- **Population 2.** All students enrolled in the two adjacent grades that contain the largest proportion of 13-year-olds at the time of testing.
- **Population 3.** Students enrolled in their final year of secondary education. Population 3 had two optional subpopulations:
    - Students taking advanced courses in mathematics
    - Students taking advanced courses in physics.

### 4.2.1 POPULATIONS 1 AND 2

In defining populations for international comparisons of student achievement it is usually necessary to choose between age and grade level as the basis of comparison. An age-based definition focuses on a specific age cohort, for example all 13-year-old students in an education system. A grade-based definition focuses on a specific grade, for example the eighth grade in an education system, counting from the beginning of primary schooling. Since TIMSS is mainly a survey of mathematics and science instruction, with the classrooms functioning as units of analysis as well as sampling units, a grade-based definition was chosen. It was difficult, however, to identify internationally comparable grades, for lack of standard international grade definitions. It was therefore decided to identify the target grades on the basis of an age cohort.

The Population 1 and Population 2 target populations are thus defined as the two adjacent grades that will maximize coverage of a specific age cohort (9-year-olds for Population 1, and 13-year-olds for Population 2). Two adjacent grades were chosen to ensure extensive coverage of the age cohort for most countries–thereby increasing the likelihood of producing useful age-based comparisons also. Furthermore, two grades allow the measurement of growth between grades.

### 4.2.2 POPULATION 3

The intention in surveying Population 3 was to try to measure what might be considered the “yield” of the elementary and secondary education systems of a country with regard to mathematics and science. Thus the definition of the population is student-oriented; it is the body of students who are in *their* last year of school. For many students, this does not represent the highest level of education, especially mathematics and science education, available in the country.

For each secondary-education track in the country, the final grade of the track was identified as being part of Population 3. This allowed substantial coverage of students in their final year of schooling. For example, grade 10 might be the final year of a vocational program, and grade 12 the final year of an academic program. Both of these grade/track combinations are considered to be part of Population 3 (but grade 10 in the academic track is not).

There are two further difficulties in defining the international desired population for Population 3. The first is that many students drop out before the final year of any track. This is addressed in the TIMSS Population 3 assessment by the calculation of a Secondary Education Coverage Index which quantifies the proportion of the general population that reaches the final year. The Secondary Education Coverage Index (SECI) was defined as follows:

$$SECI = \frac{5 * Total\ Enrollment\ in\ Population\ 3\ in\ 1995}{Total\ National\ Population\ Aged\ 15-19\ in\ 1995}$$

This definition reflected the fact that Population 3 is likely to be almost entirely a subset of the population of 15- to 19-year-olds, and that, by age 19, someone who has never belonged to Population 3 during any of the five most recent years is very unlikely to ever belong to Population 3. The SECI represents a kind of moving average measure of the proportion of the general population that undertakes the final year of a track of the secondary education system.

The second issue is that some students repeat the final year of a track, or take the final year in more than one of the tracks at two different times. That is, some students who are in the final year of a track are not in fact completing their secondary education that year. At the time of the TIMSS testing, these students would generally not have been aware (or at least certain) whether this was to be their final year. If this occurs within a country to any great extent, sampling students from the final grade may bias the estimate of the educational "yield." On the one hand, students who in fact are not completing their education still have the potential to gain further knowledge in additional years of schooling, and thus will not have attained their full yield at the time of the TIMSS assessment. On the other hand, and of more serious concern, the presence both of students who are repeating the final track, and of those who will repeat that track can contribute a substantial downward bias to the estimated achievement of the population. Repeating students are represented twice in the population, and are likely to be lower-achieving on average than those who do not repeat. The only practical way for TIMSS to deal with this problem was to exclude students who were repeating the final year. Thus Population 3 is formally defined as those students taking the final year of one track of the secondary system for the first time.

The International Study Center tried to maximize standardization across countries for the definition of Population 3. However, the precise definitions of the mathematics and physics subpopulations was necessarily a consultative process. Each country identified the group of students that it wished to compare internationally, based on a consideration of the general contents of the tests and practical considerations in sampling and administration. The analysis of Population 3 will include for each country a measure of the proportion of the total test population who were included in the advanced mathematics subpopulation, and the proportion who were included in the physics subpopulation.

The interest in measuring mathematics and science literacy levels extended to the whole of Population 3, not just the nonspecialist students. This means that the comparability of countries with regard to the literacy assessment is not affected by how the countries chose to define their mathematics and physics subpopulations. It also means that the sample design for Population 3 had to ensure that a representative sample of the advanced course-taking students took the literacy assessment, in addition to those taking the specialist tests.

### 4.2.3 SCHOOL AND WITHIN-SAMPLE EXCLUSIONS

TIMSS expected all participating countries to define their national desired populations to correspond as closely as possible to its definition of the international desired populations. However, sometimes NRCs had to restrict their coverage. For example, some countries had to restrict geographical coverage by excluding remote regions; or by excluding a segment of its education system. The international reports will document any deviations from the international definition of the TIMSS target populations. Significant differences in terms of number of students excluded would mean that the survey results will be deemed not representative of the whole national school system.

Using their national desired populations as a basis, participating countries had to operationally define their populations for sampling purposes. This operational definition, known in IEA terminology as the *National Defined Population*, is essentially the sampling frame from which the first stage of sampling takes place. The national defined populations could be subsets of the national desired populations. All schools and students from the former excluded from the latter are referred to as excluded populations.

TIMSS participants were expected to keep such exclusions to no more than 10% of the national desired populations. Exclusions could occur at the school level, within schools, or both. Because national desired populations were restricted to schools that contain the required grades, schools not containing any of the target grades were not considered as excluded. In general, practical reasons were invoked for excluding schools or students, such as increased survey costs, increased complexity in the sample design, and difficult test conditions. The size of the excluded populations were documented and serve as an index of the coverage and representativeness of the selected samples.

Participants could exclude schools from the sampling frame for the following reasons:

- They are in geographically remote regions
- They are of extremely small size
- They offer a curriculum, or school structure, that is different from the mainstream educational system(s)
- They provide instruction only to students in the exclusion categories defined under “within-school exclusions.”

Within-school exclusions were limited to students who, because of some disability, were unable to take the TIMSS tests. TIMSS participants were asked to define anticipated within-school exclusions. Because these definitions can vary internationally, they were also asked to follow certain rules, adapted to their jurisdictions. In addition, they were to estimate the size of such exclusions so that their compliance with the 10% rule could be gauged.

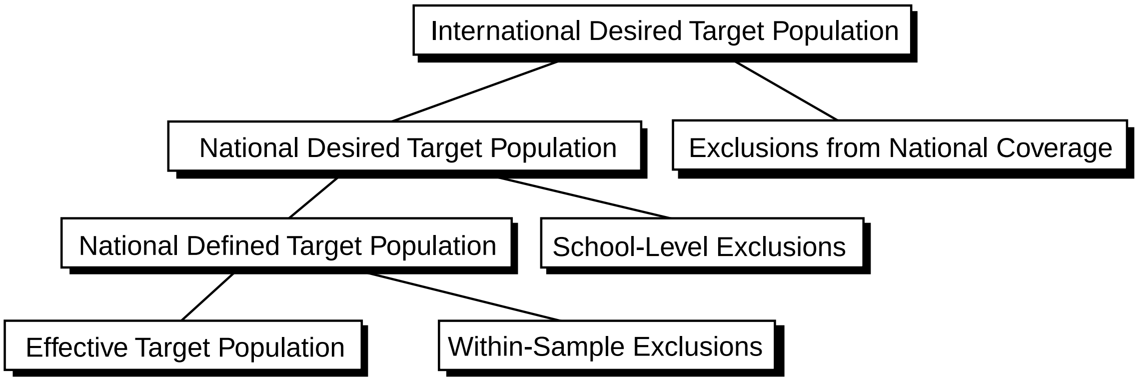
The general TIMSS rules for defining within-school exclusions are the following.

- **Educable mentally disabled students.** These are students who are considered, in the professional opinion of the school principal or other qualified staff members, to be educable mentally disabled, or who have been so diagnosed in psychological tests. This includes students who are emotionally or mentally unable to follow even the general instructions of the TIMSS test. It does not include students who merely exhibit poor academic performance or discipline problems.
- **Functionally disabled students.** These are students who are permanently physically disabled in such a way that they cannot perform in the TIMSS tests. Functionally disabled students who can perform should be included in the testing.
- **Non-native-language speakers.** These are students who cannot read or speak the language of the test and so could not overcome the language barrier of testing.

> Typically, a student who has received less than one year of instruction in the language of the test should be excluded, but this definition should be adapted in different countries.

The stated objective in TIMSS was that the effective population, the population actually sampled by TIMSS, be as close as possible to the international desired population. Figure 4.1 illustrates the relationship between the desired populations and the excluded populations. Any exclusion of eligible students from the international desired population had to be accounted for. This applies to school-level exclusions as well as within-sample exclusions.

**Figure 4.1 Relationship Between the Desired Populations and Exclusions**



## 4.3 SAMPLE DESIGN

The basic sample design proposed for TIMSS is generally referred to as a two-stage stratified cluster sample design. The first stage consists of a sample of schools[1], which may be stratified; the second stage consists of samples of classrooms from each eligible target grade in sampled schools. In some countries a third stage was added, in which students were sampled within classrooms. This design lends itself to the many analytical requirements of TIMSS. Survey estimates were required for students, teachers, classrooms, and schools.

### 4.3.1 UNITS OF ANALYSIS AND SAMPLING UNITS

The TIMSS analytical focus is both on the cumulative learning of students and on instructional characteristics affecting learning. The sample design, therefore, had to address both the measurement of explanatory characteristics thought to influence cumulative learning and the measurement of specific characteristics of the instructional settings. The first focus included characteristics of system organization, school organization and differentiation, national cross-grade curriculum specifications, resource allocations, national goals, and the like. The second focus included the measurement of teacher characteristics, classroom composition, teaching practices, implemented curriculum, and measurements of pupils' experiences and exposure to instruction. As a consequence, schools, classrooms,

---

[1] In some very large countries, it was necessary to include an extra preliminary stage, where school districts were sampled first, and then schools.

and students would all be potential units of analysis. They therefore had to be considered as sampling units in the sample design in order to meet specific requirements for data quality and sampling precision at all levels.

Although in the second sampling stage the sampling units were intact classrooms, the ultimate sampling elements were students, and so it was important that each student from the target grades be a member of one, and only one, of the classes in a school from which the sampled classes would be selected. Ideally, from a sampling perspective, the student should belong to the same class for both mathematics and science instruction. In most education systems, the mathematics class coincided with a student homeroom or science class, especially in Population 1. However, in some systems, mathematics and science classes did not coincide; students formed different groups for mathematics and for science instruction. In that case, participating countries were asked to define the classrooms on the basis of mathematics instruction. If not all students in the national desired population belonged to a mathematics class, then an alternative definition of the classroom was required for ensuring that the nonmathematics students had an opportunity to be selected.

The analytical objectives for Population 3 focused on the achievement of students in their final year of secondary schooling, rather than on the instructional context. In fact, there was no teacher questionnaire for Population 3, which meant that classrooms need not be a sampling unit. In practical terms, however, many education systems define classrooms by curriculum tracks. This made classrooms a useful sampling unit in those systems, especially when separate samples were selected for the advanced students. In education systems where the advanced course-taking students were not conveniently clustered in classrooms, student samples were selected at random within selected schools, using specified procedures.

### 4.3.2 SAMPLING PRECISION AND SAMPLE SIZE

Sample sizes for TIMSS had to be specified so as to meet the analytic requirements of the study. Since students were the principal units of analysis, the emphasis for data reliability was placed on the ability to produce reliable estimates of student characteristics. The TIMSS standard for sampling precision requires that all population samples have an effective sample size of at least 400 students for the main criterion variables. In other words, all population samples should yield sampling errors that are no greater than those that would be obtained from a simple random sample of 400 students.

Furthermore, since TIMSS planned to conduct analyses at the school and classroom levels, at least 150 schools were to be selected per target population. A sample of 150 schools yields 95% confidence limits for school- and classroom-level mean estimates that are precise to within ± 16% of their standard deviations. To ensure sufficient sample precision for these units of analysis, some participants had to sample more schools than they would have selected otherwise.

An effective sample size of 400 students results in the following approximate 95% confidence limits for sample estimates of population means, percentages, and correlation coefficients.

- Means: $m \pm 0.1s$ (where $m$ is the mean estimate and $s$ is the estimated standard deviation for students)
- Percentages: $p \pm 5.0\%$ (where $p$ is a percentage estimate)
- Correlations: $r \pm 0.1$ (where $r$ is a correlation estimate).

Multistage cluster sample designs are generally affected by what is called the clustering effect. A classroom as a sampling unit constitutes a cluster of students who tend to be more like each other than like other members of the population. The *intraclass correlation* is a measure of this within-class similarity. Sampling 30 students from a single classroom, when the intraclass correlation is positive, will yield less information than a random sample of 30 students spread across all classrooms in a school. Such sample designs are less efficient, in terms of sampling precision, than a simple random sample of the same size. This clustering effect was a factor to be considered in determining the overall sample size for TIMSS.

The magnitude of the clustering effect is determined by the size of the cluster (classroom) and the size of the intraclass correlation. For TIMSS the intraclass correlation for each country was estimated from past studies or national assessments. In the absence of these sources, an intraclass correlation of 0.3 was assumed.

To allow the planning of sample sizes, each participant had to specify a cluster size, known as the minimum cluster size for that country. Since most participants chose to test intact classrooms, the minimum cluster size was in fact the average classroom size. For participants who chose to subsample students from selected classrooms, the minimum cluster size was the number of students subsampled per classroom. The specification of the minimum cluster size not only affected the number of schools to sample, but also affected how small schools and small classrooms would be treated.

Sample-design tables were produced and included in the *Sampling Manual* (TIMSS, 1994a) (see Table 4.1 for an example). These tables illustrated the number of schools to sample for a range of intraclass correlations and minimum cluster size values. TIMSS participants could refer to these tables to determine how many schools they should sample given their intraclass correlation and minimum cluster size. A participant whose intraclass correlation was expected to be 0.6 and whose average classroom size was 30 would need to sample a minimum of 186 schools. Whenever the estimated number of schools to sample fell below 150, participants were asked to sample at least 150 schools.

The sample-design tables could be used also to determine sample sizes for more complex designs. For example, a stratum of small schools could be constructed where a smaller minimum cluster size could be specified, thereby avoiding the administrative complexity of defining pseudo-schools. (See section 4.4.1 Small Schools).

## Table 4.1 Sample-Design Table, Populations 1 and 2

| 95% Confidence Limits For Means ± 0.1s/Percentages ± 5.0% | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Minimum Cluster Size | | Intraclass Correlation | | | | | | | | |
| | | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
| 5 | a | 94 | 118 | 142 | 166 | 190 | 214 | 238 | 262 | 286 |
| | n1 | 470 | 590 | 710 | 830 | 950 | 1,070 | 1,190 | 1,310 | 1,430 |
| | n2 | 470 | 590 | 710 | 830 | 950 | 1,070 | 1,190 | 1,310 | 1,430 |
| 10 | a | 62 | 89 | 116 | 143 | 170 | 197 | 224 | 251 | 278 |
| | n1 | 620 | 890 | 1,160 | 1,430 | 1,700 | 1,970 | 2,240 | 2,510 | 2,780 |
| | n2 | 620 | 890 | 1,160 | 1,430 | 1,700 | 1,970 | 2,240 | 2,510 | 2,780 |
| 15 | a | 52 | 80 | 108 | 136 | 164 | 192 | 220 | 248 | 276 |
| | n1 | 780 | 1,200 | 1,620 | 2,040 | 2,460 | 2,880 | 3,300 | 3,720 | 4,140 |
| | n2 | 780 | 1,200 | 1,620 | 2,040 | 2,460 | 2,880 | 3,300 | 3,720 | 4,140 |
| 20 | a | 46 | 75 | 103 | 132 | 160 | 189 | 217 | 246 | 274 |
| | n1 | 920 | 1,500 | 2,060 | 2,640 | 3,200 | 3,780 | 4,340 | 4,920 | 5,480 |
| | n2 | 920 | 1,500 | 2,060 | 2,640 | 3,200 | 3,780 | 4,340 | 4,920 | 5,480 |
| 25 | a | 43 | 72 | 101 | 130 | 158 | 187 | 216 | 245 | 274 |
| | n1 | 1,075 | 1,800 | 2,525 | 3,250 | 3,950 | 4,675 | 5,400 | 6,125 | 6,850 |
| | n2 | 1,075 | 1,800 | 2,525 | 3,250 | 3,950 | 4,675 | 5,400 | 6,125 | 6,850 |
| 30 | a | 41 | 70 | 99 | 128 | 157 | 186 | 215 | 244 | 273 |
| | n1 | 1,230 | 2,100 | 2,970 | 3,840 | 4,710 | 5,580 | 6,450 | 7,320 | 8,190 |
| | n2 | 1,230 | 2,100 | 2,970 | 3,840 | 4,710 | 5,580 | 6,450 | 7,320 | 8,190 |
| 35 | a | 40 | 69 | 98 | 127 | 156 | 185 | 214 | 244 | 273 |
| | n1 | 1,400 | 2,415 | 3,430 | 4,445 | 5,460 | 6,475 | 7,490 | 8,540 | 9,555 |
| | n2 | 1,400 | 2,415 | 3,430 | 4,445 | 5,460 | 6,475 | 7,490 | 8,540 | 9,555 |
| 40 | a | 38 | 68 | 97 | 126 | 155 | 185 | 214 | 243 | 272 |
| | n1 | 1,520 | 2,720 | 3,880 | 5,040 | 6,200 | 7,400 | 8,560 | 9,720 | 10,880 |
| | n2 | 1,520 | 2,720 | 3,880 | 5,040 | 6,200 | 7,400 | 8,560 | 9,720 | 10,880 |

a= number of sampled schools
n1 = number of sampled students in upper grade
n2 = number of sampled students in lower grade

### 4.3.3 STRATIFICATION

Stratification is the grouping of schools according to some attribute or variable. It is generally used for the following reasons:

- To improve the efficiency of the sample design, thereby making survey estimates more reliable
- To apply different sample designs, or disproportionate sample-size allocations, to specific groups of schools (such as those within certain states or provinces)
- To ensure adequate representation in the sample of specific groups from the target population.

Examples of stratification variables for school samples are geography (such as states or provinces, school type (such as public and private schools), and level of urbanization (such as rural and urban). Stratification variables in the TIMSS sample design could be used explicitly, implicitly, or both.

*Explicit stratification* consists of building separate school lists, or sampling frames, according to the stratification variables under consideration. If, for example, geographic regions were an explicit stratification variable, then separate school sampling frames would be constructed for each region. Possibly different sample designs, or different sampling fractions, would then be applied to each school-sampling frame to select the sample of schools. In practice, the major reason for considering explicit stratification in the context of TIMSS was disproportionate allocation of the school sample to the strata. For example, the same number of schools might have been required from each stratum, regardless of the relative size of each stratum.

*Implicit stratification* makes use of a single school-sampling frame, but sorts the schools in this frame by a set of implicit stratification variables. This type of stratification is a simple way of ensuring proportional sample allocation without the complexity of explicit stratification. It can also improve the reliability of survey estimates, provided the implicit stratification variables are related to school mean student achievement in mathematics and science.

## 4.4 FIRST SAMPLING STAGE

The sample-selection method proposed for first-stage sampling in TIMSS makes use of a systematic probability-proportional-to-size (PPS) technique. In order to use this method it is necessary to have some measure of size (MOS) of the sampling units. Ideally this should be the number of sampling elements within the unit (e.g. number of students in the target grades in the school). If this is unavailable, some other, highly correlated measure, such as total school enrollment, may be used.

The schools in each explicit stratum are listed in order of the implicit stratification variables, together with the MOS for each school. They are further sorted by MOS within implicit stratification variable. The measures of size are accumulated from school to school, and the running total (the cumulative MOS) is listed next to each school (see Table 4.2). The total cumulative MOS is a measure of the size of the population of sampling elements.

Dividing the total cumulative MOS by the number of schools to be sampled gives the *sampling interval*.

The first school is sampled by choosing a random number in the range between 1 and the sampling interval. The school whose cumulative MOS contains the random number is the sampled school. By adding the sampling interval to that first random number, a second school is identified. This process of consistently adding the sampling interval to the previous selection number results in a PPS sample of the required size.

If an implicit stratification is in effect, then the resulting school sample will be allocated proportionately to the sizes of the implicit strata. Furthermore, if the implicit stratification variables used act to reduce sampling variance, then this sample selection method will reap that benefit, resulting in more reliable estimates than would otherwise be achieved.

Of the many benefits of this sample-selection method, the main reasons for its use in TIMSS, are that it is easy to implement, and it is easy to verify that it was implemented properly. The latter is critical since one of TIMSS' major objectives was to ensure that a sound sampling methodology be used.

Table 4.2 illustrates the PPS systematic sampling method applied to a fictitious sampling frame. The first three sampled schools are shown, as well as their preselected replacement schools should the originally selected schools not participate (see Section 4.4.3).

## Table 4.2 Application of the PPS Systematic Sampling Method

Total MOS: 392,154 Sampling Interval: 23,614.3600

School Sample: 150 Random Start: 1,135.1551

| School Code | School MOS | Cumulative MOS | Sample |
|---|---|---|---|
| 917740 | 532 | 532 | |
| 875870 | 517 | 1049 | |
| 924942 | 487 | 1536 | √ |
| 893204 | 461 | 1997 | R1 |
| 952774 | 459 | 2456 | R2 |
| 806290 | 437 | 2893 | |
| 161758 | 406 | 3299 | |
| 357056 | 385 | 3684 | |
| 997650 | 350 | 4034 | √ |
| 778732 | 341 | 4375 | R1 |
| 216873 | 328 | 4703 | R2 |
| 336426 | 311 | 5014 | |
| 97015 | 299 | 5313 | |
| 486237 | 275 | 5588 | |
| 221573 | 266 | 5854 | |
| 696152 | 247 | 6101 | |
| 645538 | 215 | 6316 | |
| 540649 | 195 | 6511 | √ |
| 330383 | 174 | 6685 | R1 |
| 914017 | 152 | 6837 | R2 |
| 76874 | 133 | 6970 | |
| 406509 | 121 | 7091 | |
| 66513 | 107 | 7198 | |
| 429291 | 103 | 7301 | |
| 88501 | 97 | 7398 | |

√ = Sampled School

R1, R2 = Replacement Schools

### 4.4.1 SMALL SCHOOLS

Small schools tend to be problematic in PPS samples because students sampled from these schools get disproportionately large sampling weights, and when the school size falls below the minimum cluster size, they reduce the overall student sample size. A school was deemed to be small for TIMSS' purposes if it could not yield an adequate sample of students per grade, as specified by the minimum cluster size. For example, if the minimum cluster size was set at 20, then a school with fewer than 20 students in each target grade was considered a small school.

In TIMSS, small schools were handled either through explicit stratification or through the use of pseudo-schools. In the first case, an explicit stratum of small schools was created for which a smaller number of students were required. The second approach consisted of creating clusters of small schools, called pseudo-schools, that would be sampled as a single unit. Any sampled cluster, or pseudo-school, would then be able to provide the required number of students.

The construction of pseudo-schools complicates data collection. Therefore, they were used only when absolutely necessary. In TIMSS, pseudo-schools were required whenever student enrollment in small schools exceeded 5% of total student enrollment. Also, participants who proposed sample designs with suitable explicit stratification for small schools were not required to construct pseudo-schools.

### 4.4.2 OPTIONAL PRELIMINARY SAMPLING STAGE

Some very large countries chose to introduce a preliminary sampling stage before sampling schools. This consisted of a PPS sample of geographic regions. A sample of schools was then be selected from each sampled region. This design was used mostly as a cost-reduction measure. The construction of a comprehensive list of schools would have been either impossible or prohibitively expensive. Also, this additional sampling stage reduces the dispersion of the school sample, thereby potentially reducing travel costs.

Sampling guidelines were put in place to ensure that an adequate number of sampling units would be sampled from this preliminary stage. The sampling frame had to consist of at least 100 primary sampling units, of which at least 50 had to be sampled at this stage.

### 4.4.3 REPLACEMENT SCHOOLS

A high participation rate among sampled schools is not always possible. To avoid sample-size losses, a mechanism was instituted to identify, a priori, replacement schools for each sampled school. For each sampled school the next school on the ordered school-sampling frame was identified as its replacement; and the one after that as a second replacement, should it be necessary.

The use of implicit stratification variables and the subsequent ordering of the school-sampling frame by size ensured that any sampled school's replacement would have similar characteristics. Although this approach was not guaranteed to avoid response bias, it would tend to minimize the potential for bias. Furthermore, it was deemed more acceptable than oversampling to accommodate a low response rate.

## 4.5 SECOND SAMPLING STAGE

For Populations 1 and 2, the second sampling stage consisted of selecting classrooms within sampled schools. As a rule, one classroom per target grade was sampled, although some participants opted to sample two classrooms per grade.

Classrooms were selected either with equal probabilities or with probabilities proportional to their size. Participants who opted to test all students in selected classrooms sampled classrooms with equal probabilities. This was the method of choice for most participants. Participants who chose to subsample students within selected classrooms sampled classrooms with PPS.

### 4.5.1 SMALL CLASSROOMS

Generally, classrooms in an education system tend to be of roughly equal size. Frequently, however, small classrooms are devoted to special situations, such as remedial or accelerated programs. These classrooms can become problematic since they can lead to a shortfall in sample size, and thus introduce some instability in the resulting sampling weights when classrooms are selected with equal probabilities.

In order to avoid these problems, it was suggested that any classroom smaller than half the specified minimum cluster size be combined with another classroom from the same grade and school. For example, if the minimum cluster size was set at 30, then any classroom with fewer than 15 students should be combined with another. The resulting pseudo-classroom would then constitute a sampling unit. If a pseudo-classroom was sampled, then all of its component classrooms would fall in the sample.

### 4.5.2 POPULATION 3

For Population 3, the second sampling stage consisted either of sampling classrooms or of sampling students directly from the target grades, depending on how students taking advanced courses in mathematics or physics were organized into schools and classes. Chapter 9 describes the within-school sampling at Population 3, for systems where students could be selected in intact classes and for systems where students in each subpopulation were sampled from across the entire grade level in a school.

## 4.6 OPTIONAL THIRD SAMPLING STAGE

An optional third sampling stage consisted of selecting students within sampled classrooms. Generally, all students in selected classrooms were included in the TIMSS sample. Participants with particularly large classrooms in their education system could opt to subsample a fixed number of students per selected classroom. This was done using a simple random sampling method whereby all students in a sampled classroom were assigned equal selection probabilities.

## 4.7 RESPONSE RATES

Weighted and unweighted response rates were computed for each participant by grade, at the school level and at the student level. Specific criteria were put in place to determine acceptable response rates at each level.

### 4.7.1 SCHOOL-LEVEL RESPONSE RATES

The minimum acceptable school-level response rate, before the use of replacement schools, was set at 85%. This criterion was applied to the unweighted school-level response rate. School-level response rates will be computed and reported by grade weighted and unweighted, with and without replacement schools. The general formula for computing weighted school-level response rates is shown in the following equation:

$$R_{wgt}(sch) = \frac{\sum_{part} MOS_i / \pi_i}{\sum_{elig} MOS_i / \pi_i}$$

For each sampled school, the ratio of its MOS to its selection probability ($\pi_i$) is computed. The weighted school-level response rate is the sum of the ratios for all participating schools divided by the sum of the ratios for all eligible schools. The unweighted school-level response rates are computed in a similar way, where all school ratios are set to unity. This becomes simply the number of participating schools in the sample divided by the number of eligible schools in the sample. Since in most cases, in selecting the sample, the value of $\pi_i$ was set proportional to $MOS_i$ within each explicit stratum, it is generally the case that weighted and unweighted rates are similar.

### 4.7.2 STUDENT-LEVEL RESPONSE RATES

Like the school-level response rate, the minimum acceptable student-level response rate was set at 85%. This criterion was applied to the unweighted student-level response rate. Student-level response rates will be computed and reported by grade, weighted and unweighted. The general formula for computing student-level response rates is shown in the following equation:

$$R_{wgt}(stu) = \frac{\sum_{part} 1/p_j}{\sum_{elig} 1/p_j}$$

where $p_j$ denotes the probability of selection of the student, incorporating all stages of selection. Thus the weighted student-level response rate is the sum of the inverse of the selection probabilities for all participating students divided by the sum of the inverse of the selection probabilities for all eligible students. The unweighted student response rates will be computed in a similar way, but with each student contributing equal weight.

Student-level response rates in Population 3 will be calculated separately by subpopulation. There will therefore be separate student-level response rates for the general population, and for students taking courses in advanced mathematics, and for students taking courses in physics.

### 4.7.3 OVERALL RESPONSE RATES

The minimum acceptable overall response rate was set at 75% for the upper grade. This overall response rate for each grade was calculated as the product of the weighted school-level response rate at the grade without replacement schools and the weighted student-level response rate at the grade.

Weighted overall response rates will be computed and reported by grade, both with and without replacement schools.

## REFERENCES

Third International Mathematics and Science Study (TIMSS). (1992). *Sampling Plan* (Doc. Ref.: ICC438/NPC116). Prepared by Richard Wolfe and David Wiley.

Third International Mathematics and Science Study (TIMSS). (1994a). *Sampling Manual–Version 4* (Doc. Ref.: ICC 439/NPC117). Prepared by Pierre Foy and Andreas Schleicher. Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1994b). *School Coordinator Manual–Populations 1 and 2* (Doc. Ref.: ICC891/NRC427). Prepared by the IEA Data Processing Center. Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1994c). *School Coordinator Manual–Population 3* (Doc. Ref.: ICC907/NRC440). Prepared by the IEA Data Processing Center. Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1994d). *Survey Operations Manual–Populations 1 and 2* (Doc. Ref.: ICC889/NRC425). Prepared by the IEA Data Processing Center. Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1994e). *Survey Operations Manual–Population 3* (Doc. Ref.: ICC 906/NRC439). Prepared by the IEA Data Processing Center. Chestnut Hill, MA: Boston College.