6

Sampling, Weighting, and Variance Estimation in International Large-Scale Assessments

Keith Rust
Westat
Joint Program in Survey Methodology

CONTENTS

Introduction

Many educational assessment programs are conducted with the aim of obtaining results at the student, school, and administrative unit level. Such "high stakes" assessments may be used to make decisions about individual student progress through the education system, or as a tool used in the evaluation of teachers and schools. In such cases, generally every student in the population of interest is assessed, as results are desired for each student. In these circumstances there are no issues of sample design and selection involved, and no issues related to the need to provide analysis weights. In large-scale international assessments, however, almost invariably the goals of the study do not include the provision of individual student achievement results for all the individuals in the population. Rather, as is clearly demonstrated in the rest of this book, the purpose is to make inference about a whole population. This extends to interest in providing results for a wide variety of population subgroups, examining the distribution of achievement within and across these subgroups, and exploring the relationship of student and school characteristics to achievement.

Given these goals, it is not a specific requirement of the study to obtain achievement data for each student in the population. The inferences of interest can be obtained from a suitably designed and executed sample of students. This, of course, offers the potential to greatly reduce the cost and burden of these assessments, rendering them possible in cases where it would be infeasible to assess all students. This also permits such studies to simultaneously assess multiple subjects, or cognitive domains, while not unduly burdening any individual student.

While sampling methods provide the means to carry out assessments in an affordable manner, considerable attention to detail is required in designing and selecting samples from the participating countries. Correspondingly, care is required in the analysis of data collected from these samples. Analytic techniques are required that address the fact that, as a result of the way in which the sample is selected, the distribution of characteristics in the sample is likely to be different from that of the population from which the sample is drawn. The most obvious and best-known technique for addressing this issue is to apply sampling weights in the analysis of the data. This can be reasonably straightforward for some analyses, but is less so for others, particularly in cases where the model used to analyze the data has features that parallel the design of the sample. The estimation of sampling variances that account for the covariance structure of the data induced by the sample design is also a key feature of the analysis of data obtained from the sample designs typically used in large-scale assessments.

The use of probability-based, or scientific, methods for selecting samples for large-scale assessments has been practiced for many decades. Since the 1960s the International Association for the Evaluation of Educational

Achievement (IEA) has conducted many of these studies, and over the intervening period a number of other organizations have also conducted them. Over time those conducting these studies have increasingly seen the need to promote and, to the extent feasible, enforce a standardized approach to sample design and selection, with common standards applied. Perhaps these efforts can best be seen coming to fore in two studies conducted in the early 1990s. The IEA conducted the International Study of Reading Literacy from 1989 to 1991 (Elley 1992), while the Educational Testing Service conducted the International Assessment of Educational Progress (IAEP) in mathematics and science (Lapointe et al. 1988, 1992). These studies featured sampling manuals, with procedures described in detail, and considerable efforts to ensure the quality of the resulting samples in each country. Subsequent IEA studies in a variety of academic subjects, and the Programme for International Student Assessment (PISA), conducted by the Organisation for Economic Co-operation and Development in several cycles since 2000, have continued the promotion of adopting high-quality sampling procedures, well-documented and subject to substantial quality-control procedures.

Thus, the use of scientific sampling methods has become standard in these assessment programs. In this chapter we describe the most common sampling techniques that are applied in these studies, indicating the goals that the methods are designed to achieve. We then discuss the implications for analysis, specifically through the use of survey weights, and variance estimation procedures appropriate for the design. We also describe how missing data resulting from nonresponse can impact the analysis, with a discussion of the way in which the survey weights can be modified so as to reduce the potential for bias that is introduced by such nonresponse. We include a brief discussion of software available for analyzing data from these studies that addresses the need to incorporate weights and estimate sampling errors appropriately. However, experience has taught us that any detailed treatment of currently available software soon becomes dated, and so that will not be attempted. Users who undertake, or anticipate undertaking, regular analyses of complex survey data, whether from large-scale assessments or other surveys, are recommended to consult the book on this topic by Heeringa et al. (2010).

Sample Design

It has become routine over the past 25 years (and there were certainly earlier precedents) that the sample designs for international large-scale assessments follow the principles and practices of scientific, probability-based, sampling methods. A full introduction to such methods is given in Lohr (2010). In this section the key elements of scientific probability sampling that apply in the large-scale assessment context are described.

Population Definition

Strictly speaking, of course, the need to define the population of inference for an assessment study, and to ensure that this is done comparably across participating countries, is not related to whether one uses a sample or surveys the entire population. However, when a sample is selected in each country, it may not be as readily evident whether the population coverage is comparable across countries as might be the case if all students in the population were selected. For example, suppose that there is a school included in the sample that has 300 full-time students in the population, and 15 part-time students. If all the students in the school are to be assessed, it will be fairly readily apparent if those administering the assessment decide not to include any of the part-time students. But if a sample of 25 students is selected, and part-time students are omitted from the sampling frame, the fact that no part-time students end up being selected in the sample might not be noted. Issues with population definition and coverage tend to revolve around relatively small groups in the population, but ones whose distribution of achievement may be quite different from the rest of the population. Thus, on the one hand, their absence from the sample might not be noticed, while on the other hand, failure to cover them in the sampling procedure might induce a substantial bias in certain analyses of the data.

Large-scale assessments typically define some cohort of the student population as the population of interest. The two basic approaches are to define a single grade of interest, or a particular birth-cohort, typically a 12-month span. Each of these approaches has its advantages and limitations.

In the past two decades, the IEA studies have generally defined a common grade to be surveyed in each country. This has two main advantages: (1) within a given country, the population is meaningful to policy makers. There is interest in knowing "how our eighth-graders compare to the rest of the world"; and (2) by surveying a single grade, it is relatively straightforward to define the population, sample it, and administer the assessment. The major drawback is being able to define grades that are comparable across countries. "Eighth grade" is a very arbitrary label, affected by national policies about the age of starting school, and the labels given to early years of schools. Typically this is addressed by defining the population as, for example, the grade containing the largest proportion of 14-year-old students (i.e., students born during a specific 12-month period, with this same birth cohort used in every country). This is somewhat effective, but has three drawbacks. First, even if every country surveys the correct grade, there can legitimately be almost a 12-month spread in the mean age of students across the participating countries. Comparing students from a grade in one country where the mean age is 8.6 years with another where the mean age is 9.4 years can be problematic. Second, often when countries participate for the first time, they do not actually know which is the correct grade in cases where there are two grades each with a large proportion of the specified age cohort. As more

and more countries have come to participate in these studies, this issue has diminished as a problem. A pilot test can also indicate if there is a problem with age distribution of the chosen grade. The third problem is related to the first two; a slight change in the definition of the cohort can lead to a change in the target grade, especially if the age cohort definition is tied to the time of the assessment. That is, it is possible that in March the greatest number of 14-year-olds are in grade 9, but by May (shifting the cohort definition by 2 months) the greatest number are in grade 8. Problems can also arise when comparisons are made over time within a country, if there is a policy change about the age of schooling. In the late 1990s, Norway made such a policy change, and over a very short span of years the average age of students in each grade dropped by about 1 year. After this change, it was difficult to know which was the most appropriate grade to survey when an important goal of the study was to assess the trend over time. The IEA Trends in International Mathematics and Science Study was forced to confront this issue. A further complication is that different provinces within a country may have different policies relating to age of schooling. Thus, in the first TIMSS study in 1995, different grades were assessed in different states in Australia. In each state the two adjacent grades with the most 13-year-olds were surveyed. In some states the upper of these two grades was grade 9, but in others it was grade 8. In federal systems this type of complication seems inevitable.

PISA and the IAEP studies, on the other hand, have defined the population using a 12-month birth cohort. This approach too has its benefits and deficiencies; not surprisingly these complement those of the grade-based approach. One does not need to study, or understand in depth, the educational system of a country to define the target population, and it is relatively easy to conduct quality control procedures to ensure that the correct population has been surveyed. Diverse countries can be compared, but they will each have a population that is uniformly distributed in age over a common 1-year age span (setting aside the problem of students in the later teenage years dropping out of school differentially across countries). The use of an age definition also avoids issues that can arise because of artifactual distortions of a grade population. In addition to the presence of students in ungraded programs, in some cases there is a particular grade at the top of one level of school, where students may tend to disproportionately be "held back." Thus if there is a key exit exam at the end of grade 8, for example, the grade 8 cohort may contain relatively many students who are repeating the grade (and these will be among the lower achievers). Thus, a form of length-biased sampling may occur. This was a particular issue in the 1995 TIMSS study of grade 12 students, and the population was therefore defined not as students in grade 12, but as students who were in grade 12 for the first time. Such problems are avoided when an age-based population definition is used.

There are four main drawbacks to the use of an age-based definition. First, the population is often not very interesting at a national level. It might be of interest to know how U.S. 15-year-olds do compared with 15-year-olds

in Finland. But if 15-year-olds in the United States are half in grade 10 and half in grade 9, internal comparisons within the United States are not very enlightening, and the implications of the results for instructional policy are not as clear as when a grade-based approach is used. Second, administering the assessments is often somewhat unwieldy, since within a school students have to be pulled from classes in two or three different grades for the assessment. The third issue is that obtaining a frame of schools that comprehensively covers the population can be difficult. Often there are many schools in a given country that might have just a few eligible students enrolled, because the grades taught in the school generally have very few students of the required age. National study centers are loath to include such schools in the sampling frame, as they know that these schools will be a burden and relatively costly on a per-student basis to include. However, the students in such schools tend to be very much skewed toward one end or the other of the ability distribution. Consider the case of PISA in the United States. About 1–2% of the PISA age cohort is in eighth grade, with the rest in higher grades. Most schools that have grade 8 do not have higher grades, and there are many more schools with eighth grades than there are schools with higher grades (as middle schools tend to be smaller than high schools). This means that the sampling frame for the United States contains a large number of schools that have grade 8 as the highest grade. When such a school is sampled, the procedure is likely to be to include the handful of eligible students enrolled in the school. These will by definition be the oldest students in the school, which in many cases means that they are not the most able. Thus, there is a strong disincentive to include these schools, both on the part of the national center (because of the operational cost and burden), and also on the part of the school, which cannot understand the desire to assess these few atypical students, and question the representativeness of the resulting sample. Fourthly, the analysis and interpretation of school- and system-level factors, and their relationship to achievement, are problematic in cases where the population of students is divided across different levels of schooling.

There are other groups of students whose inclusion (or otherwise) must be carefully addressed whether an age- or grade-based population definition is used: special needs students, recent immigrants with a different first language from the test language, vocational students, apprenticeship, and other part-time students. Increasingly it seems likely that online students may be an issue at higher grades, although no study seems to have been as yet noticeably impacted by this phenomenon.

Multistage Sampling

It is entirely natural that one obtains a sample of students within a country by first selecting a sample of schools, and then selecting students within schools. Such a procedure is known in the survey sampling literature as two-stage sampling, and it is a commonly applied one. It is primarily used for

reasons of cost and operational feasibility. It is just not practicable to assess thousands of students within a country by testing one or two students from each of thousands of schools (although perhaps the rapid growth of computer-based testing may render this statement inaccurate within a matter of several years). In many cases also, there is no centralized data source listing all of the students in the population (and even if there is it is likely to be somewhat outdated). So the best approach to getting high-quality coverage of the student population is to sample schools and ask the schools directly for a current list of eligible students. However, in the case of achievement studies, there is an additional reason for the use of this type of design. Often analysis of the data involves consideration of school effects through a multilevel (mixed) model, and such models cannot be effectively analyzed if the sample size of students from any one school is very small.

Thus, the use of the two-stage approach means that there are two distinct elements to sample selection; the selection of schools and the selection of students. We will address each of these in the following. It is important to note that the two are closely related. In particular, the goal of the school sample design is primarily to lead to a sound sample of students, and the approach used to designing the school sample is driven by this consideration. Thus, the school sample that results does not necessarily have the characteristics that one would design for if the goal were solely to select a sample of schools to analyze school characteristics.

In certain countries a third, prior, stage of sampling may be called for. This is the selection of geographic domains, from within which schools will be sampled. There are two reasons for such an approach. The first is that there may be no up-to-date list of schools available centrally for the entire country. By selecting only a sample of a few dozen geographic regions, the construction of school lists of adequate quality may be made feasible. This is not required in many countries, as even developing countries generally have a centralized list of schools, or can easily compile one. The prime example that we are aware of where this was an issue was in Russia in the early 1990s, at the time of the first TIMSS study. A more common reason for using a geographic stage of sampling is to save on administrative costs, when the assessments are administered separately. Obviously this is an issue only for larger countries, and Russia and (at times, but not recently), the United States have adopted a three-stage design for this reason.

The use of a multistage design, and specifically a two-stage design of selecting schools and then students, can (and generally does) have profound implications for the precision of the resulting estimates. If one selects a simple random sample of schools, all of which have the same size enrollment of eligible students, and then draws a simple random sample of students from within each selected school, the sampling variance is given in formula 6.1. This formula ignores issues related to sampling from a finite population. This has been for simplicity of explication, but can often be justified on a combination of the following conditions: (1) the sample is selected using a with-replacement

procedure; (2) the sample sizes are small in comparison with the population size (this is likely to be true of the school sample, but not the student sample within each school); and (3) inferentially, strict reference to a finite population is not warranted (this is the case in most other areas of statistical analysis, and can particularly apply to the student sample in a large-scale assessment program, where we do not wish to limit our inference to the particular student population that happened to exist at the time the study was conducted).

$$Var\left(\bar{\bar{y}}\right) = \frac{1}{m}\left\{\sigma_1^2 + \frac{1}{M}\sum_{j=1}^{M}\sigma_{2j}^2/n_j\right\} \tag{6.1}$$

where

$\bar{\bar{y}}$ is the estimate of the per-student mean of y, $\bar{\bar{Y}}$;
m is the sample size of schools;
n_j is the sample size of students in school j;
M is the population size of schools;

$\sigma_1^2 = \frac{1}{M}\sum_{j=1}^{M}\left(\bar{Y}_j - \bar{\bar{Y}}\right)^2$ is the between-school variance of y;

$\sigma_{2j}^2 = \frac{1}{N}\sum_{i=1}^{N}\left(y_{ij} - \bar{Y}_j\right)^2$ is the within school variance of y;

N is the number of students in each school; and
$\bar{Y}_j$ is the mean of y for school j.

This compares with the variance of a simple random sample of students of the same size:

$$Var\left(\bar{\bar{y}}\right) = \left(\sigma_1^2 + \frac{1}{M}\sum_{j=1}^{M}\sigma_{2j}^2\right)/n_o \tag{6.2}$$

where $n_o = \sum_{j=1}^{m} n_j$ is the total sample size of students.

The ratio of these two quantities is referred to as the design effect for this particular design (a two-stage sample with simple random sampling at each stage):

$$Deff = \frac{n_o}{m}\frac{\left(\sigma_1^2 + \frac{1}{M}\sum_{j=1}^{M}\sigma_{2j}^2/n_j\right)}{\left(\sigma_1^2 + \frac{1}{M}\sum_{j=1}^{M}\sigma_{2j}^2\right)} \tag{6.3}$$

One can see that unless the ratio of the between-school variance to the average within-school variance is substantially smaller than the mean of the n_j, this design effect is likely to be noticeably greater than 1.0. This indicates that the level of sampling error from such a two-stage design is likely to be much higher than for a simple random sample. Under the simple case that all σ_{2j}^2 squared are equal, and all n_j are equal (to n say), this equation reduces to

$$Deff = \frac{n\sigma_1^2 + \sigma_2^2}{\sigma_1^2 + \sigma_2^2} \tag{6.4}$$

where $\sigma_2^2 = \frac{1}{M}\sum_{j=1}^{M} \sigma_{2j}^2$.

Defining the intraclass correlation, *Rho*, as

$$Rho = \sigma_1^2 / \left(\sigma_1^2 + \sigma_2^2\right) \tag{6.5}$$

the Design Effect can be expressed as

$$\begin{aligned} Deff &= \frac{n\sigma_1^2 + \sigma_2^2}{\left(\sigma_1^2 + \sigma_2^2\right)} \\ &= 1 + \left(n - 1\right) Rho \end{aligned} \tag{6.6}$$

To mitigate the negative effects of this design on sampling variances, we use techniques of stratification and systematic selection to effectively reduce the sizes of σ_1^2 and σ_2^2 for the two-stage design, thereby lowering the design effect. Also, it is important to keep in mind that the much lower per-student cost of administering the assessment to a two-stage sample means that while the amount of sampling variance per student is likely to be considerably higher for a two-stage design, the amount of sampling variance per unit cost will be considerably lower for a two-stage design, and this is the reason that two-stage designs are universally adopted in large-scale assessments of school students.

Sample Size Determination

Before turning to the sample design features of stratification, systematic selection, and probability-proportional-to-size selection, we will consider the question of sample size determination for international studies.

The topic is usually approached by considering what level of precision is required for the national student mean on an achievement scale, measured in standard deviation units, and what level of precision is required for an

estimate of a proportion of a certain size, again with the students as the units of analyses. IEA studies have specified the target to be that the standard error for a measure of student mean performance on an assessment should be 0.05 standard deviations, and that the standard error on an estimate proportion of about 0.5 should be 0.025. Both of these considerations dictate what is referred to as an effective sample size of 400 students. That is the sample size that would be required if, in fact, a simple random sample of students were selected. The actual sample size required for a given sample design is the product of the required effective sample size and the design effect.

Using the simplified model for the design effect above, this means that the student sample size required to meet the precision requirement is: $400(1 + (n–1)\ Rho)$. Thus for a value of *Rho* equal to 0.2 (a typical value in practice), with a sample of 30 students selected from each school (also not atypical), the required student sample size is 2720 assessed students, which would require a sample of about 90 schools.

However, other considerations also come into play in determining the desired sample size of students and schools. First, it is desirable to ensure that a certain minimum number of schools are included in the sample. This is to ensure that analyses involving school level variables have good precision, and also guards against undesirable consequences that can arise when one or two unexpectedly atypical schools are included in the sample. Also, there is generally interest in subgroup analyses for even moderately small population subgroups, and also the interest is not just in the mean student performance, but also in the overall distribution, including relatively extreme values, such as the 5th or 95th percentiles.

The combination of these considerations, and the establishment of historical precedent over the past 20 years, mean that a requirement to have a minimum of 150 participating schools, with 30 assessed students (or one classroom, in the case of IEA studies), has become an accepted norm. Of course individual countries often elect to adopt larger samples, almost always because they wish to obtain adequate precision for results at the region level—this is particularly the case for countries with federal governing systems, such as Australia, Canada, and Spain, where responsibility for education policy largely lies at the provincial level.

Stratification

Stratification involves forming a partition of the population, and selecting separate samples from each of the resulting groups (called strata). This has the effect of eliminating the variance between strata as a contributor to the sampling variance of any estimate, since all possible samples selected represent each stratum in exact proportion to the stratum population size (after weighting). Stratification also allows the sampling rates to differ across strata, so that some groups of particular interest can be sampled more heavily than others.

Stratification can be applied at both the school and student levels of sampling. In all large-scale assessment programs it is applied at the school level. It is also on occasion applied at the student level. This is generally only done when there are rare subgroups of special interest, which are not restricted to particular schools. Two examples from PISA are indigenous students in Australia (who constitute about 1–2% of the population) and recent immigrant students in Denmark.

As noted in Section "Multistage Sampling," when the intraclass correlation of schools is even moderately high, the design effect can be substantial, leading to considerable loss in the precision of estimates. Fortunately, in most countries there are school-level variables available in readily accessible sources, for all schools in the country, that are substantially correlated with school mean achievement. Thus, using these variables to form strata significantly reduces the impact of between-school variance on sampling variance, and thus effectively lowers the intraclass correlation and thus the design effect. An obvious variable that applies in many countries in the case of age-based studies is the grade level of the school. When surveying 15-year-olds, as in the case of PISA, in many countries some of these students are enrolled in upper-level schools, while the remainder are enrolled in lower-level schools. When surveying students in the upper levels of secondary education, in many countries good strata can be created by distinguishing between academic, vocational, and general high schools. In some countries the distinction between public and private schools is an important one. And in most countries some sort of geographic stratification is beneficial, as often schools in metropolitan areas have higher levels of achievement than those in more rural areas.

Table 6.1 shows the example of stratification for Sweden from PISA 2012. This illustrates the fact that a variety of school characteristics can be used to create school-level strata.

Under proportional allocation, in which the sample of schools allocated to each stratum is proportional to the population size of schools, Equation 6.4 for the design effect of the mean is replaced by

$$Deff = \frac{n\sigma_{s1}^2 + \sigma_2^2}{\left(\sigma_1^2 + \sigma_2^2\right)} \tag{6.7}$$

where

$$\sigma_{s1}^2 = \frac{1}{M}\sum_{l=1}^{S}\sum_{j\in l}\left(\overline{Y}_j - \overline{\overline{Y}}_l\right)^2$$

M_l denotes the school population size of stratum l;
$\overline{\overline{Y}}_l$ is the per student mean of y in stratum l;
S is the number of strata.

It can be seen that σ_{s1}^2 is generally less than σ_1^2, and often relatively considerably less, with the reduction being proportional to the squared correlation

TABLE 6.1

School Stratification and Sample Allocation for Sweden, PISA 2012

Stratification				Population Figures			Sample Sizes		
Type of Control	School Level	Community Type	Stratum ID	Number of Schools	Enrollment of 15-Year-Olds	Percentage of Total 15-Year-Old Enrollment	Number of Schools	Expected Number of Students	Percentage of Student Sample
Public	Lower secondary	Metropolitan	01	459	30,784	30.5	57	1710	30.0
		Other large city	02	331	22,166	22.0	42	1260	22.1
		Pop. at least 12,500	03	268	18,800	18.6	34	1020	17.9
		Manufacturing	04	85	6265	6.2	12	360	6.3
		Sparse population	05	117	5501	5.5	10	293	5.1
	Upper secondary		06	525	1953	1.9	19	82	1.4
Independent	Lower secondary	Metropolitan	07	234	7579	7.5	17	436	7.7
		Other large city	08	133	5199	5.2	14	371	6.5
		Pop. at least 12,500	09	82	1770	1.8	4	88	1.5
		Manufacturing	10	10	107	0.1	2	30	0.5
		Sparse population	11	12	124	0.1	2	25	0.4
	Upper secondary		12	488	613	0.6	12	19	0.3
			Total	2744	100,861	100.0	225	5694	100.0

between the stratification variables and the school mean achievement. Thus, the design effect is reduced in this case. However, a reduction in design effect may not result if, rather than using stratification with proportional allocation, stratification is used as a means to oversample certain types of schools. Since disproportionate allocation acts to increase the design effect, this may outweigh the gains that result from stratification and in aggregate lead to an overall increase in the design effect, but with the benefit of being able to produce reliable estimates for subgroups of interest.

Systematic Sampling

Systematic equal-probability sampling is used in place of simple random sampling. The technique involves ordering the population list (referred to as the sampling frame) in some systematic order. A sampling interval, I, is then obtained by dividing the population size, N, by the desired sample size, n. A random start, R, is then selected, uniformly between 0 and I. The selected units are those corresponding to the number R, $R + I$, $R + 2I$, and so on up until $R + (n - 1)I$. These numbers are in general not integers; rounding them up to the next exact integers identifies the sampled units.

Systematic selection has two advantages. First, it reduces sampling variance in much the same way that stratification does. The sampling variance from a systematic sample is very similar to that of a stratified design in which n equal-sized strata are created and one unit is sampled from each stratum. The second advantage is that it is easy to implement, and easy to check, on a large scale, such as is needed in international assessments in which many countries participate. Its disadvantages are that there is no unbiased variance estimator available, even for a simple mean, and, in certain rare instances, it can actually lead to an increase in design effect rather than a decrease as occurs with stratification.

Systematic sampling can be used in conjunction with stratification, and also with probability-proportional-to-size selection, discussed in the next section. When used with stratification, the typical application is to use one set of variables for stratification, and other variables for systematic selection within strata. The variables used for stratification would tend to be categorical variables (especially those with no natural ordering), and variables most strongly related to achievement, as well of course as variables defining groups to be oversampled. Variables used for sorting for systematic selection are often continuous in nature, and perhaps less highly correlated with achievement. Thus, for example, one might stratify a sample of schools by province, public/private, and level (upper/lower), and then within each resulting stratum, sort schools by size prior to systematic selection. This ensures that the resulting sample is well-balanced with respect to school size, even though that is not a stratification variable.

Systematic sampling can also be used effectively to select student samples in cases where no oversampling is required. For example, when selecting

students within each school, sorting the sample by grade (in the case of age-based samples) and gender, prior to systematic selection, can be very effective in reducing sampling errors. In this case it is σ_2^2 that is being effectively reduced, much in the way that school stratification reduces σ_1^2. It is generally much easier from a practical standpoint to obtain a sorted list of students from each sampled school and to select a systematic sample from the list, than it is to stratify the student list and select independent samples within each stratum.

Probability-Proportional-to-Size Selection

In the discussion to this point, it has been assumed that all units within a stratum receive the same probability of selection. However, in large-scale assessment surveys, almost always the school sample selection procedure selects larger schools with larger probabilities than smaller schools. This method provides a robust approach, when the primary goal of school sampling is to select a sample of students. The reason why this approach is effective is discussed later in this section.

In probability-proportional-to-size (PPS) selection, each unit is assigned a measure of size. In the case of assessment surveys, the measure of size for a school is generally a nondecreasing function of the estimated number of eligible students enrolled in the school. The probability that the unit (i.e., school) is included in the sample is directly proportional to this measure of size. There are many ways to select such a sample, although if more than two units are selected from a given stratum, all but one method are very complicated. Thus the approach that is generally adopted is to use systematic sampling to select a PPS sample. Let the measure of size for school j be denoted as Z_j, and let the sum of Z_j across all M schools be Z. Schools are sorted in the desired systematic order, and the Z_j values are cumulated. Let Q_j denoted the cumulated measure of size associated with school j. Thus $Q_1 = Z_1$, $Q_2 = Z_1 + Z_2$, and $Q_M = Z$. An interval I is calculated as $I = Z/M$, and a random start R is chosen, uniformly between 0 and I. The schools selected are those whose values of Q_j correspond to the values R, $R + I$, $R + 2I$, and so on up until $R + (n - 1)I$. That is, school j is selected if Q_j equals or exceeds one of the selection values, while $Q_{(j-1)}$ is less than that same selection value. The result is that school j is selected with probability equal to nZ_j/Z. If any school has a value for Z_j that exceeds mZ/M, then it is included with certainty.

The PPS selection of schools is paired with a student sampling procedure that selects an equal sample size of students from each school. That is (setting aside any oversampling at the student level), the sample size in each selected school is set equal to a constant, n, whenever $N_j > n$, and is set to N_j otherwise. Provided that $\max(N_j, n)$ is highly correlated with Z_j, this two-stage sampling procedure leads to a "self-representing" student sample, that is, one in which each student throughout the whole population has close to the same chance of selection. All else being equal, samples in which each student has close to the same selection probability have a lower design effect

(i.e., lower sampling variance for the same sample size), than ones in which the student sampling probabilities vary considerably.

A self-representing sample could also be obtained by selecting an equal-probability sample of schools, and then selecting a fixed fraction of students within each school. But while selecting, say, 30 students out of 150 eligible students in a school provides for an operationally feasible assessment, selecting 200 students from 1000 might be inconvenient, and selecting 2 students out of 10 seems like folly.

Furthermore, having variable sample sizes within schools acts to increase somewhat the design effect due to the clustering of the sample. If schools are of unequal size, and a simple random sample is selected at each stage, with students having equal probability of selection, the variance for the mean of a student characteristic, y, is given as

$$Var\left(\bar{y}\right) = \frac{1}{m}\left[\frac{1}{M}\sum_{j=1}^{M}\left(\frac{MN_j}{N_o}\bar{Y}_j - \bar{Y}\right)^2 + \frac{1}{N_o\bar{n}}\sum_{j=1}^{M}N_j\sigma_{2j}^2\right] \tag{6.8}$$

where

$$N_o = \sum_{j=1}^{M}N_j$$

and

$$\bar{n} = \frac{1}{m}\sum_{j=1}^{m}n_j$$

However, if the sample is chosen PPS with the selection probabilities exactly proportional to the true school size, N_j, with a constant sample size of n students selected in each school, the variance becomes

$$Var\left(\bar{y}\right) = \frac{1}{m}\left[\frac{1}{N_o}\sum_{j=1}^{M}N_j\left(\bar{Y}_j - \bar{Y}\right)^2 + \frac{1}{N_o\bar{n}}\sum_{j=1}^{M}N_j\sigma_{2j}^2\right] \tag{6.9}$$

Thus, the within-school variance contribution is the same in Equations 6.8 and 6.9, but the between-school variance is different. The between-school variance is likely to be higher in Equation 6.8, because not only the variation in school means for y contributes to this variance, but also the variation of the schools sizes, N_j, about the average school size, N_o/M.

Thus, for both practical and design efficiency reasons, the preferred design is one in which the school sample is selected with PPS, and a fixed number of

students is selected from each school (or all students if the school is smaller than the target student sample size).

It is important to note that the whole emphasis of this design discussion has been aimed at obtaining a student sample that will provide reliable estimates of student characteristics. It is sometimes also of interest to make statements about the population of schools. If each school in the *population* is given equal weight in such analyses, then a PPS sample of schools generally provides a rather poor basis for analysis, since the sample is heavily skewed toward larger schools. Population estimates can be rendered unbiased through weighting, but the resulting sampling errors will generally be much larger than would be the case if an equal probability sample of schools were selected.

However, even when analyzing school characteristics, it is often the case that, at least implicitly, inference is about the student population in the schools. Thus if one asks the question: "What proportion of schools have a principal who has a postgraduate degree?," it is likely that the real question of interest is: "What proportion of students attend a school where the principal has an advanced degree?" For this second question, a PPS sample of schools is substantially more efficient than an equal probability sample of schools. It is effective to include more large schools in the sample, because large schools each affect a greater number of students. Thus the use of a PPS school sample, with an equal probability student sample, is effective for most kinds of analyses that are likely to be of interest in large-scale assessment studies.

Classroom Samples

The above discussion of student sampling assumed that students are sampled directly from lists within each sampled school. This is the case in studies such as PISA and IAEP, for this is a very natural approach for age-based samples, where students may be drawn from several different grades, and in any one class in the school it is quite likely that not all students are eligible for inclusion in the population. However, as mentioned previously, IEA studies such as TIMSS select whole classrooms of students. This is a natural procedure when the population is grade based.

The selection of classrooms is a method of selecting students within a school, but by sampling clusters of students (classrooms) rather than selecting individual students via systematic sampling. However, similar principles apply as with direct student sampling. It is good to select schools with PPS, and then to select a fixed number of classrooms in each school (usually one or two), with equal-probability. Classroom characteristics can be used to sort the list of classrooms, and then systematic sampling used to select the sample.

Depending upon the type of classrooms used to define the list of classrooms to be provided for sampling, it is quite likely to be the case that a classroom-based sample has a much higher level of sampling error per student than does a direct student sample. This is because classrooms within a

school tend to have students of somewhat different mean ability. Especially in middle and upper levels of schooling, this can be an intentional part of school organization.

For example, the IEA TIMSS assessments select one or two mathematics classes from grade 8 (or equivalent). At that level of education, mathematics classes are very often arranged on the basis of mathematics proficiency of the students. Thus, students from a single mathematics classroom are much more likely to be similar to one another than students drawn at random from throughout the school. Thus, this intraclass correlation at the classroom level acts to increase the design effect of the resulting student sample, and often very substantially, since generally all students within the selected classrooms are included.

Classroom-based designs continue to be used despite their higher design effects. Not only do they facilitate the selection of the student sample, but this design also offers the possibility to analyze classroom effects, which would be unwieldy with a direct student sample (perhaps every teacher in the school needs to be surveyed, and carefully linked to individual students), and may greatly weaken the ability to analyze the data using multilevel models, using the classroom as the second level unit, with students as the first level.

The Overall Sample Design

Putting all of these design attributes together, we see that the typical sample design for a national sample in an international large-scale assessment most often has the following characteristics:

a. A two-stage sample of schools, and students within schools.
b. The sample size is at least 150 schools and 4000 students, and can often be much larger.
c. The school sample is stratified by a variety of characteristics that are related to mean school achievement, defined groups of special interest, or both.
d. The school sample is selected with probability proportional to a measure of size (PPS).
e. The student sample within each school is selected with equal probability, and an equal-size sample (of students or classrooms) is selected from within each school.
f. The school sample is selected using systematic sampling, sorted using additional characteristics likely to be related to school mean achievement.
g. The student sample is selected using systematic sampling, sorted using characteristics likely to be related to student achievement.

The use of designs such as these has implications for estimating population parameters from the sample, quantifying the precision of those estimates, and conducting multivariate analyses of the data. Compounding these issues is the presence of nonresponse, at the school level and at the student level. In the next section, we discuss the implications of the design for deriving estimates, which is addressed through weighting. Then we discuss the issues of nonresponse, and how the weights are modified to account for that, and finally how we estimate sampling variances appropriately for a wide variety of estimators, accounting for the effects of both the design and nonresponse adjustments.

Weighting

The Purpose of Weighting

The provision of survey weights on the data files for large-scale assessments and other surveys is intended to assist the user in obtaining estimates of population parameters that: (1) do not suffer from bias resulting from the use of a complex sample design, and (2) have minimal bias due to the presence of nonresponse, to the extent possible. In many applications there is a third objective: (3) to reduce sampling errors in estimates by utilizing auxiliary information about the population; but this objective is seldom, if ever, present for international large-scale assessments of student populations.

Weighting is, in effect, a part of the estimation procedure. The intention is to incorporate features that would ideally be included in many analyses into a single weight (or, on occasion, into several different weights) and place this on the data file. This has two benefits. It obviates the need of the analyst to repeat this part of the estimator for every analysis, and it ensures consistency across analyses, if all use the same weight variable.

Base Weights

The term "base weight" refers to the weight component that reflects the impact of the sample design, and specifically the inclusion probability. The base weight is also referred to as the design weight. The base weight is given as the reciprocal of the inclusion probability of the unit. It can be separated into two components. The first is the school-level base weight, which is the reciprocal of the inclusion probability of the school to which the student belongs. The second is the reciprocal of the inclusion probability of the student, given that the school is included in the sample. The overall student base weight is given as the product of these two components, and it is this weight that is required in order to obtain unbiased (or consistent, in the case

of more complex estimators) estimators of student population characteristics (in the absence of nonresponse). However, the two components are generally needed when conducting multilevel (hierarchical) linear model analyses, as the overall student base weight by itself is not sufficient for estimating parameters of such models correctly.

If we let d_i denote the overall base weight for student i (the "d" is used to denote "design weight" since this weight reflects the features of the sample design, but no other weight components), then for characteristic y, the mean of y for the student population is estimated from sample s as

$$\hat{\bar{y}} = \left(\sum_{i \in s} d_i y_i \right) \Big/ \left(\sum_{i \in s} d_i \right) \tag{6.10}$$

More generally, for a more complex statistic, such as regression coefficient, for example, each student's contribution to the estimator is multiplied by the base weight. Thus the estimator for a regression coefficient, where y is the dependent variable and x is the independent variable, is given as

$$\hat{B}_{yx} = \frac{\sum_{i \in s} d_i y_i \left(x_i - \hat{\bar{x}} \right)}{\sum_{i \in s} d_i \left(x_i - \hat{\bar{x}} \right)^2} \tag{6.11}$$

Nonresponse Adjustments

Weights are adjusted for nonresponse by creating nonresponse adjustment factors by which the base weights are multiplied. For nonresponse adjustments of this type to be effective in reducing nonresponse bias in estimates from the survey, they must be a function of a variable, available for both nonrespondents and respondents, that has two properties: It is correlated with the response status, and it is correlated with the outcome variables of interest. There are two basic approaches to establishing variables for use in creating nonresponse adjustments: The first is to find variables that are related to response, and hope that they are related to the outcome variables (perhaps looking for evidence that this is the case). The second is to find variables that are related to key outcome variables and response status. Whereas the first method generally uses data from the survey to be adjusted, the second method is generally based on historical data from a similar survey. Because over time the relationship with the outcome is likely to be more stable than the relationship to the response (which can be affected by policies to try to increase response rates, or current events that affect response rates at a given time), generally with this approach the emphasis is concentrated on finding variables that are correlated with the outcome, perhaps confirming that they are also correlated with response status.

Either approach is viable for international large-scale achievement studies. But in fact, for several reasons, the approach has generally been to focus on historical data with the aim of finding variables that are related to both the outcomes of interest, and response status. First, in these studies, unlike many surveys, there are generally a few outcome variables that are clearly the most important—namely, achievement on the assessments conducted. Second, as the weights for many countries must be processed in a short time, a fairly standard approach to creating nonresponse adjustments is needed. This is in contrast to the case with stratification, where time permits the tailoring of the choice of stratification variables for each individual country. Since these studies are often repeated over time, there is good opportunity to use historical data to determine which variables to use in creating nonresponse adjustments.

Generally nonresponse adjustments are implemented by creating groups of schools, and students, and applying a common weight-adjustment factor to all the units in the same group. Adjustments are made to the school base weight component to account for nonresponse by entire schools. Then adjustments are made to the student weights, to adjust for nonresponse by individual schools for which some students respond, but not all.

School Nonresponse Adjustments

Generally school nonresponse adjustment classes are created using stratification variables. These are a natural choice, since they are generally chosen for stratification for either or both of two reasons: (1) They are related to the outcomes of interest; a necessary condition for effective nonresponse adjustment, or (2) they represent subgroups of special interest, and it is important that such groups be represented in the correct proportions in the final weighted data. It is also the case that often the different strata experience different school response rates. This is particularly the case when strata represent different school systems (public versus private), different tracks (academic versus vocational), different levels (upper secondary schools versus lower secondary schools), or different provinces, which may differ in their level of support for the study. For schools grouped in school nonresponse class C, the nonresponse adjustment factor that is applied to the school component of the base weight, is given as

$$f_{1C} = \left(\sum_{j \in C} d_j e_j \right) \Bigg/ \left(\sum_{j \in C_R} d_j e_j \right) \tag{6.12}$$

where

C_R denotes the set of responding schools in class C; and

e_j denotes the anticipated enrollment of eligible students in school, based on frame data.

Note that the adjustment factor utilizes not only the base weight of each school, but also its enrollment of eligible students, as estimated from the data on the sampling frame. This estimated enrollment is used, rather than the actual enrollment, because the actual enrollment is not generally available for the nonresponding schools. If there is a systematic difference between the estimated and actual enrollments, and the actual enrollments were used for the respondents, this would lead to a bias being incorporated into the nonresponse adjustments.

The enrollment data are incorporated into the adjustment, because the purpose of the adjustment is to reduce, as much as possible, bias in estimates concerning the student population that is induced by the school nonresponse. The number of students in the population that are "represented" by each school in the sample is approximated by $d_j e_j$, and thus this is the contribution of each school to the school nonresponse adjustment. Therefore these adjustments are not suitable if one wishes to use the data to make inference about the population of schools (with each school in the population counted as a single unit, regardless of enrollment size), since in that case each school should be weighted just by d_j.

Recall that, with probability proportional to size sampling, except perhaps in the case of very small and very large schools, d_j is proportional to $1/e_j$, which means that often the school nonresponse adjustment is close in value to the ratio of the number of schools sampled in class C to the number of responding schools in class C_R. In some applications, in fact, this ratio can adequately serve as the nonresponse adjustment factor, without having to apply Equation 6.12. This would generally only be the case if each class C falls entirely within a single stratum.

It is typical in international large-scale assessments that school response rates vary considerable from country to country, or across strata within countries. Thus, it is not uncommon for a country to achieve a very high rate of response for public schools, but a lower rate for private schools. Because of these variations within and across countries, school nonresponse adjustments often serve an important role in bias reduction of survey estimates, but the importance of these adjustments varies considerably across countries.

Student Nonresponse Adjustments

Whereas school response varies considerably across countries, and typically many countries achieve 100% school participation or very close to it, it is usually more the case that student response rates are less variable across countries. There is almost always some student nonresponse, as a result of day-to-day absences of students from school, but the amount of student nonresponse is seldom high. High student nonresponse generally has one of two causes: (1) poor operational procedures in organizing the assessments within schools; and (2) a requirement that written parental consent be obtained prior to assessing the student (but situations where parents are notified of

the assessment in advance, and may exercise an option to opt their child out of the assessment, do not generally result in a high degree of nonresponse).

One might think that these circumstances would mean that student nonresponse is of less concern for the validity of the data. However, experience shows that, while school nonresponse is often quite high, it is also often not strongly related to the average achievement of students within the school. All sorts of local political factors can come into play in determining whether a school will participate, and often these have little to do with the level of student achievement within the school. But for student nonresponse, experience has shown that this is very often related to student achievement. Generally this is in the direction of a potential upward bias, as less capable students are more likely to be absent from the assessment. This relationship may be causal—less able students may seek to avoid the assessment—but it is also likely that the two share a common cause—lack of interest in school, or chronic poor health, for example. Whatever the case, the application of student nonresponse adjustments can substantially reduce student nonresponse bias.

In contrast to the case of school nonresponse adjustments, generally the information available at the student level for making student nonresponse adjustments is very similar across countries. Furthermore, available student characteristics are often limited—gender, age (relevant for grade-based studies), and grade (relevant for age-based studies) are often the only variables available for nonresponding students. However, often *school-level* variables are very strong predictors of *student* response. Thus, it is common to observe differential student response in upper- compared to lower-level schools, public compared to private schools, vocational compared to academic schools, and so on. Therefore, effective student nonresponse adjustments are generally accomplished by forming classes of students from similar schools (or even using a single school as the basis of a class), and perhaps adding gender and age or grade.

Student nonresponse adjustments are calculated using both the student base weight and the school nonresponse adjustment factor that applies to the student's school. Thus the student nonresponse adjustment factor is given as

$$f_{2C^*} = \left(\sum_{i \in C^*} d_{ij} f_{1C(j)} \right) \Bigg/ \left(\sum_{i \in C_R^*} d_{ij} f_{1C(j)} \right) \tag{6.13}$$

where

C_R^* denotes the set of responding students in class C^*;

d_{ij} denotes the base weight for student i in school j (reflecting both school and student-within-school inclusion probabilities); and

$f_{1C(j)}$ denotes the school nonresponse factor for the class of schools which contains the school (j) to which student i belongs.

While the absolute and relative importance of school, gender, and age/grade in determining student nonresponse bias obviously varies across studies, findings from research into an effective method of forming student nonresponse classes in PISA, reported in Rust et al. (2012), are informative.

In the 2003 cycle of PISA, each individual school formed a student nonresponse adjustment class. Further subclassification within school was done on the basis of grade where feasible. However, in most cases, the sample size did not permit this finer classification. It was noted that, in PISA, gender is strongly related to achievement (girls substantially outperform boys in reading in almost every country, while boys somewhat outperform girls in mathematics in almost every country). Furthermore, in many countries gender was related to response, with typically girls responding at a slightly higher rate than boys. Thus, there was evidence that the method used to create the student nonresponse adjustment classes for PISA in 2003 might not have been fully effective in removing nonresponse bias. In an attempt to improve the effectiveness of these adjustments, for the 2006 cycle of PISA an alternative approach was proposed. Instead of using the individual school as the highest level of classification for forming adjustment cells, school nonresponse adjustment classes were used. These classes were based on the school stratification variables. Such classes were formed even for countries with 100% school response. By creating classes based on school characteristics that were several-fold larger than those used in 2003, much scope was provided for further subdividing the cells based on gender and grade.

As an evaluation of the change in procedure, student nonresponse adjustments were applied to the 2006 PISA data using the method proposed for 2006, and that used in 2003. As expected, in those countries with either a sizeable response rate difference between boys and girls, or a response rate difference across grades, the two methods produced somewhat different results.

Table 6.2 shows the student response rates and the mean achievement in each of reading, mathematics, and science, for a selection of countries that participated in PISA in 2006. The countries listed are those for which the two methods of nonresponse adjustment led to the greatest changes in results, or where there were noticeable response rate differences, either by gender or grade. Many of them have over a 2-percentage point difference in response rates by gender. Australia is unusual in that the response rate was higher for males than for females. Note that in every country females had much higher achievement than males in reading, but noticeably lower achievement in mathematics. Thus, absent an adequate adjustment for nonresponse, results in many of these countries are likely to be biased upward for reading, and downward for mathematics (exceptions being Belgium, Ireland, and the United States, with little difference in the response rates by gender, and Australia where the directions of the biases are likely to be reversed).

Table 6.3 shows response rates by grade for the subset of countries listed in Table 6.2 for which a substantial proportion of the sample was drawn from more than one grade. Note that in every case, the response rate for the lowest

TABLE 6.2

Student Response Rates and Mean Achievement Scores, by Gender, for Selected Countries PISA 2006

	Females				Males				Difference			
Country	Response Rate (%)	Reading Mean	Mathematics Mean	Science Mean	Response Rate (%)	Reading Mean	Mathematics Mean	Science Mean	Response Rate (%)	Reading Mean	Mathematics Mean	Science Mean
Australia	85.2	532	513	527	86.4	495	527	527	–1.2	37	–14	0
Austria	92.9	513	494	507	88.2	468	517	515	4.8	45	–23	– 8
Belgium	93.1	522	517	510	92.9	482	524	511	0.2	40	–7	–1
Denmark	90.7	509	508	491	87.5	480	518	500	3.2	30	–10	–9
Iceland	84.4	509	508	494	81.8	460	503	488	2.6	48	4	6
Ireland	83.8	534	496	509	83.7	500	507	508	0.0	34	–11	0
Italy	93.2	489	453	474	90.9	448	470	477	2.2	41	–17	–3
Poland	92.8	528	491	496	90.6	487	500	500	2.1	40	–9	–3
Portugal	87.2	488	459	472	85.6	455	474	477	1.7	33	–15	–5
Spain	90.3	479	476	486	86.7	443	484	491	3.7	35	–9	–4
Tunisia	95.5	398	358	388	93.5	361	373	383	2.1	38	–15	5
USA	90.9	*	470	489	91.0	*	479	489	0.0	*	–9	–1

Note: * denotes reading results are not available for the United States.

TABLE 6.3

Student Response Rates, by Grade*, for Selected Countries PISA 2006

	Response Rate				
Country	**Grade 7**	**Grade 8**	**Grade 9**	**Grade 10**	**Grade 11**
Australia			81.6%	86.9%	83.9%
Belgium		81.8%	90.7%	95.5%	87.4%
Denmark		84.9%	89.8%		
Ireland		65.9%	85.8%	83.1%	81.4%
Italy		82.8%	83.1%	94.2%	87.8%
Poland		77.6%	92.5%	95.1%	
Portugal	76.0%	82.2%	86.8%	88.9%	
Spain		64.1%	80.7%	96.2%	
USA			84.1%	92.5%	88.2%

Note: * denotes only grades with atleast 100 students in sample are given.

grade shown is less than for the other grades. However, in many of these countries the response rate for the highest grade shown is below that of the next lowest grade. So absent an adequate adjustment, there is a potential for bias in the overall results for these countries, but the likely direction is not always clear.

Table 6.4 shows the results for this set of countries from each of the two methods of nonresponse adjustment. Note that in several countries there is little difference between the methods (Australia, Austria, Belgium, Italy, and Tunisia). This reflects the fact that either the 2003 method was adequate to remove the nonresponse bias, or the biases from differential response rates across grades have a canceling effect, since the lowest response rates occurred in the highest and lowest grades, with higher response rates in the middle grades (this is seen in Table 6.3 for Australia and Belgium in particular). However, in a few countries there was a noticeable difference in the results of the two methods, which strongly suggests that the 2006 method at least partially reduced nonresponse bias that the 2003 method failed to remove. The 2006 method gave higher scores in Ireland, and lower scores in Poland, Portugal, and Spain. In all cases the differences were less than one standard error of the mean (and less than half a standard error in all cases except that of Spain). The fact that in most cases the differences were the same for all subjects suggests that the differences in response rate by grade were contributing more bias than were the differences by gender.

This study, while hardly providing a definitive approach to performing nonresponse adjustments in international large-scale assessments, does illustrate the potential importance of these methods for reducing the biases that are very likely to be introduced through differential response to the assessments.

TABLE 6.4

Mean Achievement Scores, PISA 2006, Using Two Different Student Nonresponse Adjustment Methods, for Selected Countries

	Reading					Mathematics					Science				
	2003 Adjustment Method		2006 Adjustment Method			2003 Adjustment Method		2006 Adjustment Method			2003 Adjustment Method		2006 Adjustment Method		
Country	Mean	S.E.	Mean	S.E.	Difference	Mean	S.E.	Mean	S.E.	Difference	Mean	S.E.	Mean	S.E.	Difference
Australia	512.9	2.1	512.9	2.1	0.0	520.0	2.2	519.9	2.2	0.1	527.0	2.2	526.9	2.3	0.1
Austria	490.3	4.1	490.2	4.1	0.1	505.4	3.7	505.5	3.7	–0.1	510.9	3.9	510.8	3.9	0.0
Belgium	500.9	3.0	500.9	3.0	0.0	520.5	2.9	520.3	3.0	0.1	510.4	2.5	510.4	2.5	0.0
Denmark	495.0	3.1	494.5	3.2	0.5	513.1	2.6	513.0	2.6	0.1	496.0	3.1	495.9	3.1	0.1
Iceland	484.9	1.9	484.4	1.9	0.4	505.9	1.8	505.5	1.8	0.3	491.0	1.7	490.8	1.6	0.2
Ireland	516.4	3.6	517.3	3.5	–1.0	500.6	2.8	501.5	2.8	–0.9	507.4	3.2	508.3	3.2	–1.0
Italy	468.5	2.4	468.5	2.4	0.0	461.3	2.3	461.7	2.3	–0.4	475.2	2.0	475.4	2.0	–0.2
Poland	509.0	2.8	507.6	2.8	1.4	496.3	2.4	495.4	2.4	0.8	498.7	2.3	497.8	2.3	0.9
Portugal	473.5	3.6	472.3	3.6	1.2	467.2	3.1	466.2	3.1	1.0	475.5	3.0	474.3	3.0	1.2
Spain	462.4	2.2	460.8	2.2	1.6	481.5	2.3	480.0	2.3	1.6	490.0	2.6	488.4	2.6	1.5
Tunisia	380.5	4.0	380.3	4.0	0.2	365.3	4.0	365.5	4.0	–0.2	385.3	3.0	385.5	3.0	–0.2
USA	*	*	*	*	*	474.9	4.0	474.4	4.0	0.5	489.5	4.2	488.9	4.2	0.6

Note: * denotes reading results are not available for the United States.

Variance Estimation

Because large-scale survey assessments invariably involve complex sample designs, special procedures are needed to estimate the variances of the resulting estimates. An examination of the formula for the actual sampling variance of a mean achievement score, shown in Section "Probability-Proportional-to-Size Selection" (e.g., Equation 6.9) reveals that sampling variance from these designs is not the same as in the case of a simple random sample. To obtain unbiased (or consistent) estimates of sampling variance therefore requires the application of specialized variance estimation procedures.

As is well-documented in the literature (see, e.g., Wolter [2007]), there are two aspects to this problem. The first is to find a variance estimator that is unbiased (preferably) or consistent, for very simple estimators. By "very simple" we essentially mean estimated totals of population characteristics, calculated using the base weights. This estimator, known as the Horvitz-Thompson estimator, is given by the numerator in Equation 6.10. Formulae for variance of the Horvitz–Thompson estimator for a variety of complex designs are given in well-known sampling texts such as those by Särndal et al. (1992) and Lohr (2010). The second aspect arises from the fact that, especially in large-scale assessments, there is generally very little interest in anything that can be estimated via the Horvitz–Thompson estimator. As Equation 6.10 demonstrates, even something as simple as an estimator of mean achievement, obtained without any adjustments being made to the base weights, is not of this form.

There are two approaches to dealing with these joint issues of a complex sample and complex estimators (meaning estimators other than the Horvitz–Thompson estimator). Wolter (2007) provides a detailed treatment of both of these approaches. The first is to use Taylor series linearization to approximate the complex estimator by a linear combination of Horvitz–Thompson estimators, and then use the variance estimation formulae appropriate for the particular design involved to estimate the variance of this linear combination. The second general approach is to use a replication (or resampling) procedure, to estimate the complex variance via a fairly simple formula, but one that involves a large number of estimates of the parameter of interest, each obtained from a subsample of the full sample. In more general applications the best known of such method these days is the bootstrap (Efron, 1993), but in typical survey sampling applications such as large-scale assessment surveys, other methods such as the jackknife and balanced repeated replication are available and are more efficient than the bootstrap in most practical applications. For a discussion of the use of replicated variance estimation methods in surveys generally, see Rust and Rao (1996).

For a variety of reasons, both practical and historical, in large-scale assessment surveys the replication approach has tended to be used more frequently than the linearization approach for the estimation of sampling variance. This

has not been a result of any technical superiority of the resulting variance estimators. Studies of the asymptotic properties of linearization and replication estimators have shown them to be of similar consistency, and empirical simulations have supported these findings of similarity. The reasons for the popularity of replication methods can be traced to two practical aspects. First, many of those analyzing the data are not familiar with the special formulae needed to estimate the variances of Horvitz–Thompson estimators directly, and are not particularly anxious to master them. Second, many different types of complex estimators are employed in the analysis of assessment data. With the linearization approach a different variance formula is needed for each type of estimator. But for a given replication approach (the jackknife, say), a common formula can be used to estimate the variance of a wide variety of different kinds of parameter estimators, incorporating the effects of both the survey design and the survey weighting process in each case. Consequently, in the remainder of this chapter we focus on replication variance estimators.

Replication Variance Estimators in International Large-Scale Assessments

There are two replication methods used commonly in large-scale assessments: The jackknife, or jackknife repeated replication (JRR), and balanced repeated replication (BRR). In application they are quite similar, and have similar properties, with some slight differences. Traditionally IEA studies such as TIMSS and PIRLS have used the jackknife, while for PISA a modified version of BRR is used.

In most, if not all cases, the approach is to "model" the sample design as being a stratified (explicitly) two-stage sample, with two primary sampling units selected from each stratum with replacement (three units for some strata, in cases of an odd number of units). We use the expression "model the sample design," because generally the design does not actually fit the description. It is the case that two-stage sampling is generally involved, with schools as the first-stage units and students or classes as the second-stage units. However, most often more than two schools are selected per explicit stratum, and the schools are sorted within strata, based on some characteristic likely to be related to mean school achievement on the assessment in question. Following this the schools are sampled using systematic selection—a without replacement sampling method, rather than with replacement.

By approximating the design as a two-per-stratum design selected with replacement, the construction of appropriate replicate subsamples and the ensuing variance estimation formula is greatly simplified. Yet there is little bias introduced into the variance estimation. Systematic sampling very closely approximates a highly stratified design with independent selections of two (or three) units within each stratum. Treating the school sampling as being with-replacement when in fact it is not, leads to some overestimation of variance, but in most applications it is minor. In cases where the effect

of assuming without replacement sampling results in a noticeable overestimation of variance, more complex replication procedures are available (see Rizzo and Rust (2011), Lin et al. (2013)), and have been implemented in practice (Kali et al. 2011), but not in an international context.

We first consider the jackknife and BRR methods ignoring the complications of having an odd number of schools in the sample. We will also set aside the possibility that any school has a school-level selection probability of 1.0 (which can certainly happen in smaller countries). We also ignore in this discussion the effect of imputation error when estimating the variance of achievement scale scores and other latent characteristics—this is addressed in Chapter 7 (von Davier and Sinharay, 2013).

The Jackknife

Suppose that there are $2K$ schools in the sample for a given country, paired into K "strata" (hereafter called "variance strata") so as to reflect both the actual school stratification and the systematic sampling used within strata. The first replicate is formed by "dropping" from the sample one of the two schools in variance stratum 1, identified at random. For the dropped school, the design weights for this first replicate are set to zero. The design weights are doubled for all students in the remaining school for variance stratum 1. The design weights for all other students (i.e., those whose schools are not in variance stratum 1) are set equal to those used for the full sample. This process is repeated K times, once for each variance stratum. This results in each student on the data file having not only a full-sample design weight, d_i, but also a set of K replicate weights, where these are given by Equation 6.14.

$$\begin{aligned} d_i^{(k)} &= 0 \text{ if student } i \text{ is from the school in variance stratum } k \text{ that is} \\ &\qquad \text{dropped for replicate } k \\ &= 2d_i \text{ if student } i \text{ is from the other school in variance stratum } k \\ &= d_i \text{ otherwise} \\ k &= 1, \ldots, K. \end{aligned} \tag{6.14}$$

Ideally, nonresponse adjustments are applied to each set of replicate weights in turn, resulting in a set of K nonresponse adjusted replicate weights. Sometimes in practice this step is omitted, and the replication is applied to the nonresponse adjusted weights for the full sample. That is, nonresponse adjusted weights take the place of the design weights in Equation 6.14. This procedure fails to replicate the effect on sampling variance of the nonresponse adjustments, but in most applications these are relatively small.

To obtain the estimated variance of a particular estimate, a set of K replicate estimates is then generated by using each of the replicate weights in turn

in place of the original design weight. Thus, for example, for an estimate of mean achievement, we obtain the following replicate estimates:

$$\hat{\bar{y}}_{(k)} = \left(\sum_{i \in s} d_i^{(k)} y_i \right) \Big/ \left(\sum_{i \in s} d_i^{(k)} \right) \quad k = 1, \ldots, K. \tag{6.15}$$

The variance estimate for the mean achievement is then estimated via Equation 6.16 (again, ignoring the imputation variance, which must be added in the case of a mean scale score and other estimated latent characteristics).

$$var\left(\hat{\bar{y}}\right) = \sum_{k=1}^{K} \left(\hat{\bar{y}}_{(k)} - \hat{\bar{y}} \right)^2 \tag{6.16}$$

Note the simplicity of this formula. But not only is this formula simple, the same variance estimation formula applies to every parameter estimator for which the variance is estimated (e.g., the proportion of students above a cut score, or a regression coefficient in a multiple regression analysis).

With this approach of creating a set of jackknife replicate weights that are included in the data file, the analyst can derive valid inferences from the data without having to have any detailed understanding of either the sample design or the jackknife procedure. All that is needed is the application of formula 6.16. As discussed in Section "Software for Variance Estimation," many software packages available now utilize the replicate weights and this formula routinely, making this method even more straightforward for the analysts to apply. One cautionary note of application is that there some variants on the jackknife procedure that result in some variation of formula 6.16 being required. Thus, caution is needed to ensure that the correct form of variance estimator is being used, corresponding to the specific implementation of the jackknife replicate formation.

A theoretical limitation of the jackknife is that, in a single-stage sample, it has been demonstrated that the method does not provide consistent variance estimation for "nonsmooth" estimators such as the median and other population quantities (see Wolter 2007, Section 4.2.4). Roughly speaking, this means that, while the method may not result in a substantial bias, the variance estimator remains very unstable no matter how large the sample size or how many jackknife replicates are created. While the empirical evidence is strong that this problem is essentially completely mitigated in most applications with two-stage sampling, theoretical justification for this finding is lacking. Thus, when possible, it is probably best to avoid using the jackknife for estimating the variance of the median, and similar statistics, when analyzing large-scale assessment surveys.

Balanced Repeated Replication

Like the jackknife, BRR, or balanced half-sampling, proceeds by creating parallel sets of weights that vary systematically from the full sample weights. With BRR, the weights for a single replicate are created by increasing the weights for students in one school from each variance stratum, while simultaneously decreasing the weights of the remaining schools. In "classical" BRR, the increased weights are obtained by doubling the base weights, while the decreased weights are obtained by setting the replicate base weights to zero. However, it has become common to use a different approach to varying the base weights, for reasons discussed below.

If the choice of schools for which the weights were to be increased were both random and independent across variance strata and across replicates, this method would constitute a form of bootstrap variance estimator. Efficiency is gained by imposing a correlation structure on the assignment across variance strata and replicates, so that far fewer replicates are required than is generally the case with the bootstrap.

The determination as to which primary units will have their weights increased for a given replicate, and which decreased, is determined via an orthogonal array, known as a Hadamard matrix. A Hadamard matrix is a square matrix, with entries that have value +1 or –1, with the property that the vector product of any row, with any column of a different number is zero, while the product of a row and column of the same number is equal to the rank of the matrix. Hadamard matrices can be readily constructed of arbitrarily large size, and exist for most dimensions that are a multiple of 4. To construct the BRR replicates for a set of H pairs of primary units, one uses the Hadamard matrix of size H^*, where H^* is the smallest multiple of 4 that exceeds H. For further discussion of Hadamard matrices and their use with BRR, see Wolter (2007).

Let a_{ij} denote the (i, j) entry of Hadamard matrix A, of dimension K^*. Each value of a_{ij} is either +1 or –1. The properties of the matrix are such that

$$\begin{aligned} \sum_{k=1}^{K^*} a_{hk}a_{kl} &= 0 \quad \text{for all } h \neq l \\ \sum_{k=1}^{K^*} a_{lk}a_{kl} &= K^* \quad \text{for all } l \end{aligned} \tag{6.17}$$

We create one set of replicate weights corresponding to each row of H, for a total of K^* replicate weights. In each primary unit pair, one is arbitrarily numbered 1 and the other 2, with the numbering assignment being independent across pairs. Then for pair l in replicate k, if $a_{lk} = +1$, the weight for unit 1 is increased and the weight for unit 2 is decreased, and vice versa when $a_{lk} = -1$.

Thus

$$
\begin{aligned}
d_i^{(k)} &= (1+\delta)d_i \text{ if student } i \text{ is from school labelled } (1.5-0.5a_{lk}) \text{ in pair } l \\
&= (1-\delta)d_i \text{ if student } i \text{ is from school labelled } (1.5+0.5a_{lk}) \text{ in pair } l \\
k &= 1,\ldots,K^*
\end{aligned}
\tag{6.18}
$$

where δ is a factor greater than zero, but no greater than 1. In classical BRR, $\delta = 1$, but in PISA and many other surveys, $\delta = 0.5$.

The variance estimator is somewhat similar to that for the jackknife but with sufficient crucial differences that it is disastrous if one uses the BRR formula with jackknife replicate weights, and vice versa (only if $K^* = \delta^{-2}$ do the jackknife and BRR variance formulae coincide).

$$
var_{BRR}\left(\hat{\bar{y}}\right) = \frac{1}{K^*\delta^2}\sum_{k=1}^{K^*}\left(\hat{\bar{y}}_{(k)} - \hat{\bar{y}}\right)^2 \tag{6.19}
$$

As with the jackknife, the same variance estimation formula can be used for a wide variety of complex estimators. Nonresponse weighting adjustments are applied to each set of replicate base weights. The BRR method is consistent for "nonsmooth" estimators such as the median and other quantiles, unlike the jackknife, and so is a preferable method when these quantities are of particular interest.

Some Practical Considerations in the Use of Replication Variance Estimators

Odd Number of Primary Units (Schools)

The discussion in section "Replication Variance Estimators in International Large-Scale Assessments" assumed that there is an even sample size of primary units (i.e., schools in most large-scale assessment surveys), but this is not always the case. To deal with an odd number of schools, we form one (or a few) triples of schools, rather than pairs. The approach then used with BRR is given in Adams and Wu (2002), Appendix 12. For two possible approaches when using the jackknife, see Rizzo and Rust (2011), and Martin and Kelly (1997, Chapter 5).

Large Numbers of Primary Units

Large numbers of primary units can arise in two situations in large-scale assessments. The first case is when a participating country wishes to obtain reliable estimates at a region level. Thus, for example, in PISA Canada typically has a

sample of about 1000 schools, since they wish to obtain reliable estimates for each language by province combination. The second case occurs in small countries, where individual schools may be included with certainty. In these cases the students are the primary units. Thus in PISA Iceland typically includes all schools, and the resulting sample of over 3000 students is thus composed of over 3000 primary units. In these cases, some kind of grouping of primary units is required, since it is not feasible to produce a set of replicate weights that number in the same order as the number of primary units. However, considerable care is required in the way in which units are grouped. An inappropriate approach can result in biased variance estimators (either positively or negatively biased), unstable variance estimators, or both. For a detailed discussion of the best approaches to reducing the number of replicate weights required, while maintaining the precision of variance estimation and introducing little bias, see Lu, Brick, and Sitter (2006), Rust (1986), and Wolter (2007).

Software for Variance Estimation

Software suitable for analyzing complex survey data is widely available, and is frequently being updated. For this reason it is of little value to give a detailed treatment of the currently available software in a volume such as this, since such a discussion will be outdated shortly after publication.

There are a few key features that analysts of large-scale assessment surveys should be alert to when determining which software to use that should be kept in mind when considering whether to use a particular program or package.

A. *The ability of the program to handle the kind of sample design that is used in large-scale surveys.* Not all software is suitable for the multistage designs used in large-scale assessments, although increasingly these designs are covered. The software must be able to deal with, for example, an analysis that compares the United States, with a two-stage sample of schools and students, with Iceland, where all schools are included, and perhaps all students.

B. *The ability to use sampling weights appropriately.* Generally software that is capable of dealing with complex sample designs can also deal with weights appropriately. But, for example, if one uses the default option with SAS to obtain a weighted estimate of a population standard deviation, unless the weights are normalized in advance the result will be completely spurious. See Chapter 17 (Stapleton, 2013) for a discussion of weight normalization.

C. The ability to take advantage of the replicate weights, and other variables related to variance estimation, that are included in the data files. It is much more foolproof if the user is able to utilize replicate weights that are already provided, rather than attempting to use software to create replicate weights from scratch (sometimes

confidentiality provisions and other considerations mean that the data needed to estimate variances, without using the replicate weights provided, are not available).

D. *The ability of the software to do other kinds of analyses where the complex design can reasonably be ignored*. It is a nuisance to have to use one type of software to estimate the variance of a mean appropriately, and something else to implement a complex modeling procedure. This is the strength of systems such as STATA, SAS, and R, which provide such "one-stop shopping."

E. *The ability to include measurement error as well as sampling error.* Generally when analyzing large-scale achievement data, valid inferences can be drawn only if the analyst incorporates a measure of uncertainty related to the individual estimates of the latent constructs—that is, the student's scale score. This can be achieved either through the multiple imputation ("plausible value") approach, or through an explicit maximum likelihood-based approach. It is very convenient if the same software can be used to incorporate both sources of error simultaneously, rather than requiring the user to estimate the components separately and then write additional software to combine them.

F. *Price, user support, and documentation*. Some suitable software is freely available, and this can be very attractive, especially for a one-off application. The user should beware, however, as to the availability of sound documentation and user support in cases where the software is free.

To see a comprehensive comparative review of current software for analyzing complex survey data, the reader is encouraged to visit the Survey Research Methods Section of the American Statistical Association website devoted to this issue: www.hcp.med.harvard.edu/statistics/survey-soft/.

Without specifically endorsing any particular set of software, listed below are some candidates that warrant consideration by anyone wishing to analyze large-scale assessment data. Indicated for each are those aspects from the list above that are particular strengths of the software. First-time users are likely to want to pay particular attention to those programs that support the analysis of latent variables (E), since student achievement scales are latent. All packages listed below have the properties A and B. Again, the reader is reminded that both the capabilities and availability of specific software packages vary over time, often changing rapidly.

i. AM Software (E, F)
ii. R (C, D, F)
iii. Mplus (C, E, F)
iv. SAS (D)
v. SPSS (D)

vi. Stata (D)
vii. SUDAAN (C)
viii. WesVar (C, E, F)

Conclusion

International Large Scale Assessments invariably involve the collection of data on student achievement via the conduct of a sample survey with a complex design. A variety of standard survey sampling techniques can be applied so that data can be collected at a reasonable cost, yet can provide estimates of adequate reliability for useful inference. The estimation procedures applied generally also call for adjustments to account for nonresponse at both the school and student levels.

Because the data are collected via a complex design, methods of inference must be used that reflect the sampling variance that actually results from the design, rather than relying on methods that assume a much simpler design. This applies both to parameter estimators, and estimators of sampling variance. In the case of parameter estimators, unbiased estimates will result only if survey weights are incorporated into the analyses. For variance estimation, techniques must be employed that give unbiased (or nearly unbiased) and reliable estimators of the true sampling variance engendered by the design. This is further complicated by the fact that this approach needs to be incorporated with methods used to reflect the measurement uncertainty of key latent characteristics measured in the survey, most notably student achievement placed on a scale.

Despite these challenges, well-established methods exist for the appropriate analysis of ILSA data and are routinely used both by those who produce the initial reports from each study, but also by secondary analysts of the data. Research continues into the best methods to analyze large-scale assessment data obtained from surveys.

In the future it seems likely that there will be interest in incorporating auxiliary data into the estimation procedures for these studies. It is fair to say that, with the exception of adjustments for nonresponse, current practice is that auxiliary data are frequently used at the design stage, but seldom at the analysis stage, at least for initial reporting. There are several reasons for this. Because the design information varies in nature and quality across countries, it is difficult to meet the time requirements for reporting the results if additional steps are added to the estimation, especially when these must be tailored to each country. A second consideration is that, generally speaking, errors in auxiliary data used at the design phase lead to some increase in variance, but errors in auxiliary data used in analysis can often lead to bias in the parameter estimates. Nevertheless, it seems likely that

there will be interest in the future in modifying estimation procedures to use external data to increase the reliability and usefulness of the results, as this is happening in other fields of survey research, and statistical analysis more generally.

References

Adams, R. and Wu, M., eds., 2002. *PISA 2000 Technical Report*. OECD: Paris.

Efron, B. 1993. *Introduction to the Bootstrap*. Chapman & Hall: New York.

Elley, W.B. 1992. *How in the World do Students Read? IEA Study of Reading Literacy*. IEA: The Hague, Netherlands.

Heeringa, S.G., West, B.T., and Berglund, P.A. 2010. *Applied Survey Data Analysis*. CRC Press: Boca Raton, FL.

Kali, J., Burke, J., Hicks, L., Rizzo, L., and Rust, K.F. 2011. Incorporating a first-stage finite population correction (FPC) in variance estimation for a two-stage design in the National Assessment of Educational Progress (NAEP). *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 2576–2583.

Lapointe, A.E., Mead, N.A., and Phillips, G.W. 1988. *A World of Differences: An International Assessment of Mathematics and Science*. Educational Testing Service: Princeton, NJ.

Lapointe, A.E., Mead, N.A., and Askew, J.A. 1992. *Learning Mathematics*. Educational Testing Service: Princeton, NJ.

Lin, C.D., Lu, W., Rust, K., and Sitter, R.R. 2013. Replication variance estimation in unequal probability sampling without replacement: One-stage and two-stage. *Canadian Journal of Statistics*. To appear.

Lohr, S. 2010. *Sampling: Design and Analysis*, 2nd Edition. Duxbury Press: Pacific Grove, CA.

Lu, W.W., Brick, J.M., and Sitter, R.R. 2006. Algorithms for constructing combined strata variance estimators. *Journal of the American Statistical Association* **101**, 1680–1692.

Martin, M.O. and Kelly, D.L., eds., 1997. *Third International Mathematics and Science Study Technical Report. Volume II: Implementation and Analysis*. Boston College: Chestnut Hill, MA.

Rizzo, L. and Rust, K. 2011. Finite population correction (FPC) for NAEP variance estimation. *Proceedings of the Section on Survey Research Methods,* American Statistical Association, 2501–2515.

Rust, K., Krawchuk, S., and Monseur, C. 2013. PISA nonresponse adjustment procedures. In Prenzel, M., Kobarg, M., Schöps, K., and Rönnebeck, S., eds., *Research on PISA*. Springer: Heidelberg, Germany.

Rust, K.F. 1986. Efficient replicated variance estimation. *Proceedings of the American Statistical Association, Survey Research Methods Section,* 81–87.

Rust, K.F. and Rao, J.N.K. 1996. Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research* **5**, 281–310.

Särndal, C-E., Swensson, B., and Wretman, J. 1992. *Model Assisted Survey Sampling*. Springer-Verlag: New York.

Stapleton, L.M. 2013. Incorporating sampling weights into linear and non-linear models. Ch. XII in Rutkowski, L., von Davier, M., and Rutkowski, D., eds., *Handbook of International Large-Scale Assessment Data Analysis.* CRC Press: Boca Raton, FL.

von Davier, M. and Sinharay, S. 2013. Item response theory extensions and population models. Ch. VII in Rutkowski, L., von Davier, M., and Rutkowski, D., eds., *Handbook of International Large-Scale Assessment Data Analysis.* CRC Press: Boca Raton, FL.

Wolter, K.M. 2007. *Introduction to Variance Estimation*, 2nd Edition. Springer: New York.

This page intentionally left blank