Chapter 7

Sampling, Weighting, and Variance Estimation

Sabine Meinck

Abstract The International Association for the Evaluation of Educational Achievement (IEA) undertakes international large-scale assessments (ILSAs) to provide reliable measures of student achievement in the context of learning. These ILSAs can be used to assess the quality and equity of education systems and enable countries to make informed decisions for improving education. Surveying all individuals belonging to the target population (e.g., students, classes, or teachers) would be a costly enterprise. The ILSA design instead focuses on measuring high-quality representative samples, ensuring the results are close to the true values in the populations and can be usefully compared across cultures and countries. Sampling is thus a key element in establishing the validity and reliability of the results obtained from cross-national large-scale assessments. This chapter reviews IEA's sampling strategies, from defining the target populations, to compiling sampling frames, applying complex sampling methods, and accounting for the methodology when analyzing data. These concepts are explained for a non-technical audience in an illustrative way. Common methodologies used when selecting random samples for IEA studies include stratification, multiple-stage and cluster sampling, and unequal selection probabilities. The samples yielded by implementing any one of these methods call for specific methods of data analysis that take account of the sampling design. Sampling weights have to be applied to arrive at accurate conclusions on population features when using sample data; specific methods are needed to compute standard errors and confidence intervals correctly. This comprehensive guide to IEA's methodology illustrates the vital importance of sampling in providing high-quality, reliable data that are generalizable to the targeted populations in an unbiased manner.

Keywords Cluster sampling · Complex samples · Multiple-stage sampling · Nonresponse · Standard errors · Target population · Variance estimation · Weighting

S. Meinck (✉)

International Association for the Evaluation of Educational Achievement (IEA),
Hamburg, Germany

e-mail: sabine.meinck@iea-hamburg.de

7.1 Introduction

The studies of IEA, like any other international large-scale assessments (ILSAs) of education, aim to provide a precise (and valid) picture of the state of education systems at a defined point in time for a particular target group and domain (e.g., mathematics achievement of grade 4 students). In doing so, IEA needs to satisfy the twin imperatives of international comparability (fair, valid, and reliable) and national relevance.

Assessing all individuals belonging to the target population would generally be too costly, which is why ILSAs are applied to selected representative samples instead. Based on these samples, researchers approximate the population features of interest. A sample helps in reducing the workload, respondent burden, and costs while providing estimates that are close enough to values from a complete census to meet the intended purposes. This means that the sample must be selected in a way that ensures that it will represent the targeted population in a precise and undistorted manner. The expressions “design-unbiasedness” and “sampling precision” are often used to summarize those characteristics of a statistical sampling design (Dumais and Gough 2012a). Therefore, sampling is key to ensuring the validity and reliability of ILSAs. Sampling strategies developed for IEA studies ensure that: (1) cost and quality requirements are balanced; (2) population estimates are close to what a complete census would have given and precise enough for the intended purposes; (3) they can be applied in a variety of educational systems; and (4) they allow for valid cross-national comparison of the results.

All samples selected for IEA studies are random samples. Common features applied are stratification, multiple-stage sampling, cluster sampling, and sampling with unequal selection probabilities. Samples yielded by implementing any one of these methods are called “complex samples.” In most ILSAs, these methods are part of the international sampling design, applied in most, but not all countries, and more than one method can be used. Typically, the international sampling design is optimized with respect to the specific circumstances of a particular country, while complying with the international objectives of design-unbiasedness and sampling precision. Importantly, samples are selected in a way that replicates the structure of the educational system, allowing data to be linked across schools, classes, students, and teachers.

Like any other survey based on random sampling, IEA studies estimate two parameters for each characteristic of interest: its point value (e.g., the years teachers spent on average in the profession, the proportion of students who have access to a school library, the average science score of grade 8 students, the difference in achievement between boys and girls, regression and correlation coefficients) and its precision (e.g., a margin of error for the estimated science score of grade 8 students). Both measures are affected by the complex sampling design, and addressed by specific weighting and variance estimation procedures. After reviewing how target populations are defined in IEA studies, this chapter illustrates the IEA’s sampling strategies and weighting procedures, and introduces approaches for estimating population

characteristics and their standard errors. The chapter focuses on how these strategies are related to the reliability and validity of the survey results, and also covers related quality control measures.

7.2 Defining Target Populations

When looking at the study results, readers will intuitively assume they describe features of the whole target population. For example, readers of the Progress in International Reading Literacy Study (PIRLS; see IEA 2020) reports will suppose a particular average pertains to all grade 4 students in a given country. Furthermore, to compare features across different nations or educational systems, equivalent items have to be compared. IEA studies make great efforts to ensure valid comparisons are possible.

At first glance, it may seem a simple task to define the target population of a study. Looking at some exemplary IEA target population definitions, however, it becomes obvious that the details are significant. For instance, the population definitions of students often rely on the internationally accepted International Standard Classification of Education (ISCED) scheme to describe levels of schooling across countries (UIS [UNESCO Institute for Statistics] 2012) and determine the correct target grade in each country. For example, IEA's PIRLS assesses:

All students enrolled in the grade that represents four years of schooling counting from the first year of ISCED Level 1, providing the mean age at the time of testing is at least 9.5 years. (Martin et al. 2017, p. 3.3)

Reference to the ISCED scheme overcomes the challenge of differently structured educational systems, with different names for grades and different school entry rules (for example, to distinguish pre-primary from primary classes). PIRLS specifies a minimum average age for the cohort of interest, acknowledging that school entry occurs at different ages in different countries and the assessment may pose excessive cognitive demands on very young children. Some educational systems (for example, Northern Ireland and England) have a very early school entry age and therefore test students at grade 5 instead of grade 4 so that their national definition of the target population deviates from the international definition. Compromises are unavoidable when accommodating cross-national comparisons, but also pose a threat to the validity of such comparisons. Not all targeted children have received the same number of years of education, and/or the average age of children within the target population may differ by country, which may affect the maturity of test takers. Nonetheless, the national cohort that is closest to the internationally defined target population represents the optimum choice for cross-country comparisons.

Defining other than student populations is often even more difficult. For example, determining a cross-nationally comparable population of teachers poses significant challenges. This entails detailed wording and operationalized definitions to enable

national teams to identify the individuals targeted by the survey correctly. The International Civic and Citizenship Education Study (ICCS; see IEA 2020) and the International Computer and Information Literacy Study (ICILS; see IEA 2020) both target teachers of grade 8 students. Each study defines precisely which target teachers are to be studied and provides detailed criteria for establishing whether the target groups are teaching grade 8 students (Meinck 2015b; Weber 2018). The definitions are substantiated by practical examples. IEA's sampling experts are available throughout recruitment periods to help national teams interpret the definitions correctly and advise on particular situations.

Another challenge regarding comparability is that not all eligible individuals¹ can be covered by the studies in all countries. Reasons to remove parts of the internationally defined target population usually relate to undue collection costs and/or a lack of fit between the assessment and the abilities of the test takers. Relatively high collection costs occur, for example, for remote or small schools, which is why countries are usually allowed to exclude such schools from the assessment. If clearly distinguishable parts of target populations are removed, such as, for example, minority language groups, this is usually referred to as “reduced coverage.” In addition, students with disabilities or those lacking the requisite language skills can be released from the test. As a general guideline, students with physical disabilities, such that they cannot perform in the study test situation,² or students with intellectual disabilities, unable to follow even the general instructions of the test, can be excluded (Martin et al. 2016, 2017; Meinck 2015b; Weber 2018). It is advisable that students who are non-native language speakers (i.e., students who are unable to read or speak the language(s) of the test and would be unable to overcome the language barrier in the test situation) are also excluded. Typically, this affects students who have received less than one year of instruction in the language(s) of the test. In all IEA studies, the reasons and scope of exclusions and under-coverage are meticulously examined and documented. To minimize the potential risk of bias, exclusions from the core target population (mostly students) must not exceed five percent of the target population. Countries surpassing this threshold are annotated in every single table in the international report, clearly signaling potential doubts on the comparability of their results to the readers. If the national target population definitions are altered between cycles, trends over time are not reported, or are reported based on homogeneous population parts. An issue of concern is the increasing number of students being excluded because of disabilities in recent study cycles in a number of countries, as can be seen from comparing respective technical documentation over time. For example, the average total exclusion rate in PIRLS increased from 3.8% in 2001 to 5.8% in 2016, including all countries in the analysis who participated in both cycles (Martin et al. 2003, 2017). This may be related to reforms initiated after the United Nations (2019) enacted the Convention on the Rights of Persons with Disabilities, which entered into force in

¹Eligible individuals are all those belonging to the internationally defined target population.

²Students that cannot be accommodated by the test situation or where accommodation cannot be provided as part of normal practice can be excluded.

2008, instigating new procedures on diagnosing and treating students with disabilities. IEA continually reviews this and related developments to ensure comparability is not jeopardized.

Given such challenges, readers of IEA study results are strongly advised to consult the encyclopedias and technical documentation that describe the school systems of each participating country, the chosen target grades and average ages, and information on excluded or not covered parts of the populations. This contextual information needs to be carefully considered when interpreting the results.

To ensure countries identify the suitable target populations in their country, and specify their exclusions correctly, national research coordinators (NRCs), who are responsible for the implementation of a study in a particular country, complete a set of forms. Among other tasks, NRCs identify the target grade, the country's name for the grade, the average age of students in that grade at the time of data collection, and the type and scope of exclusions or reduced coverage. This information is validated by the IEA sampling experts, using information from earlier cycles or other studies, or/and education statistics from reliable online sources, such as the World Bank (World Bank 2020), UNESCO Institute for Statistics (UIS 2020), the Organisation for Economic Co-operation and Development (OECD 2020), Eurydice (European Commission 2020), and national statistical agencies.

7.3 Preparing Valid Sampling Frames for Each Sampling Stage

A sampling frame is a list of units from which the sample is selected. Samples can only be representative of the units enlisted on the frame. Therefore, sampling frames in IEA studies must be comprehensive lists of all eligible units belonging to the target population. As described in Sect. 7.4.1, many IEA studies involve multiple sampling stages, which implies that valid sampling frames have to be compiled for each selection stage. It is critical that all NRCs are able to identify the correct units and list them on the sampling frame. Therefore, detailed definitions are not only needed for the target populations but also for every step of the selection process.

Most often, the first selection step entails a school sample, in which case all schools offering education to the targeted students must be listed on the sampling frame. This list must include a national identification number, allowing staff at the national center to identify the sampled schools correctly. If sampling with probabilities proportional to size (PPS; see Sect. 7.4.3) is to be employed, a measure of school size, such as the number of students or classes, has to be included. Further, variables grouping homogeneous schools together (a process called stratification; see Sect. 7.4.2) need to be added if this design feature is deemed useful for increasing sampling precision or addressing research questions of particular policy relevance. Variables such as addresses and contact information are not needed for sampling, but could be helpful for validating the frame or simplifying the correct identification of schools later on in the process. A school sampling frame is rarely fully up-to-date, as school statistics are usually available only with some delay. However, significant efforts

must be made to make them as accurate as possible. Missing out schools on the frame will lead to an undercoverage bias in the population (i.e., the sample is not representative of, for example, newly opened schools). Carrying ineligible schools in the frame, however (e.g., those that recently closed or no longer teach students in the target grade), can lead to decreasing sample sizes, as they have to be dropped from the sample without replacement when being sampled. Sampling experts check all school sampling frames rigorously for consistency and plausibility. Checking routines include the search for duplicates, and comparison of frame information with information on sampling forms, estimated population sizes from earlier cycles or other studies, and officially available data, such as enrollment figures and birth statistics from reliable online sources (as mentioned in Sect. 7.2).

Deciding upon the unit to be listed on the school frame can be sometimes more difficult than expected. Options can be to list physical buildings, administrative units, or even tracks within units. In some countries, schools as administrative units are not tied to a building, but rather comprise a set of buildings or even locations (sometimes called main and satellite schools). In such a case, it is methodologically sound to list school buildings separately or as complete administrative units. Importantly, all persons involved in the study need to be informed about this decision, as this affects not only the compilation of valid sampling frames for the second sampling stage but also the validity of questionnaire responses. The school coordinator must list the respective target grade classes to prepare for proper class sampling, whether this is just one school building or the complete administrative set. In addition, the principal of a sampled school who is asked to provide information about the school needs to answer the school questionnaire with respect to the sampled unit. Incomplete or imprecise information can lead to noisy data or even severe bias. For example, if school coordinators incorrectly listed only classes based in the school's main building, even though the whole administrative unit was sampled, then classes from satellite buildings (often located in rural areas) would be omitted. The latter would then have a zero chance of being selected and thus would not be represented by any sample selected from the frame. Alternatively, if the satellite of a school, located on a remote island, is sampled and the principal of this school (i.e., the person completing the school questionnaire) has their office at the main school building, their response may instead reflect the whole administrative unit if they are not informed about the location of the sampled unit. This may lead to potentially inaccurate or misleading answers regarding many of the sampled school's features, such as socioeconomic intake, resources, or location. A last pertinent example relates to schools organized in shifts, a situation that is often present in developing countries. In such schools, a school building is used for two, or even three different shifts (morning, afternoon and evening shifts), each with completely distinct student and (often) teacher populations. Again, a decision needs to be made on whether shifts or actual school buildings are listed on the sampling frame, and all people involved in the study (i.e., school coordinators, principals, and test administrators) must be accordingly informed to avoid biased results and thereby threats to the validity of the collected data.

Similar constraints can occur when compiling sampling frames for subsequent selection stages. When sampling classes, a key requirement is that every eligible

student in a school must belong to exactly one, and only one class. In other words, classes must contain mutually exclusive and exhaustive groups of students. If this requirement is violated, students may have zero or multiple selection chances, which would lead to biased results. Further, it is key that all eligible classes (i.e., those containing eligible students) must be listed. Classes exclusively dedicated to students that should be excluded from the test must be marked as such, and will not be sampled. Finally, a complete list of students is needed for each sampled class. With this list, a sample of students within the class can be selected as a last sampling step, or all students in the class invited to participate in the survey.

Listing and sampling classes and students in IEA studies is performed by national staff, using a dedicated software made available by IEA to support participating countries, IEA's Windows Within School Sampling Software (WinW3S). Sampling experts provide comprehensive documentation on listing procedures and train national staff on correct implementation of class and student listing and sampling. IEA's sampling experts are available for help whenever questions arise, a measure that has proved to be highly effective at ensuring the high quality of class and student sampling frames and procedures.

It should be noted that IEA ensures full confidentiality of the information received with the sampling frames. School sampling frames are kept fully confidential; sensitive information about classes, students, and teachers, such as names or addresses, are automatically removed from the WinW3S database when it is submitted by the national center to IEA. Schools, classes, and individuals cannot be identified in the publicly available databases after the survey.

7.4 Sampling Strategies and Sampling Precision

IEA studies rely on sampling strategies recognized as state-of-the art by the scientific community, which are now implemented in most contemporary ILSAs (Rutkowski et al. 2014). The chosen sampling strategies are well described in the technical documentation accompanying all ILSAs and comprehensive introductions to the topic are readily available (for example, Dumais and Gough 2012a; Lohr 1999; Rust 2014; Statistics Canada 2003). Hence, rather than reviewing the sampling approaches used in IEA studies, this chapter focuses on certain aspects directly related to the reliability and validity of the results.

IEA studies implement mostly so-called complex samples, that is, samples that employ at least one of the following features: multiple stage sampling, cluster sampling, stratification, and sampling with unequal selection probabilities. All of these features have a direct or indirect effect on sampling precision (i.e., how close the estimated values are to the true values in the population) or, in other words, how reliable and valid the statements are, based on survey data. The concept of “sampling precision” was neatly illustrated by Meinck (2015b), using the famous photograph of

Fig. 7.1 Illustration of sampling precision: simple random sampling. *Source* Meinck (2015b) Copyright © 2015 International Association for the Evaluation of Educational Achievement (IEA)

Einstein taken by Arthur Sasse in 1951. Selecting pixels from a picture and reassembling the picture based only on the sampled pixels shows that the picture gets more precise as the sample size (here the number of pixels) increases (Fig. 7.1).

In educational studies, with students or schools as sampling units, it is not as easy to visualize the effect of sampling precision. Furthermore, most ILSA do not employ simple random sampling (SRS) methods. IEA studies have established precision requirements that allow highly precise estimates for large populations and large population subsets. ICCS and ICILS employ sampling strategies yielding samples equivalent to a simple random sample of 400 sampling units, a precision level deemed sufficient in most ILSAs. Samples with this precision yield percentage point estimates with a confidence interval of ± 5 percentage points (Meinck 2015b; Weber 2018). The Trends in International Mathematics and Science Study (TIMSS; see IEA 2020) and PIRLS sample size requirements are equivalent to a simple random sample of 800 sampling units to permit precise trend measurement (Martin et al. 2016, 2017). However, why and how do sampling strategies affect sampling precision?

7.4.1 Multiple Stage Sampling and Cluster Sampling

In most IEA studies, samples are selected in multiple stages. Schools are usually selected first; they comprise the so-called “primary sampling units” (PSUs). In a second step, one or multiple classes of the target grade are sampled in TIMSS, PIRLS, and ICCS, while students from across all classrooms of the target grade are selected in ICILS. The approach of selecting multiple units within a previously selected group is called cluster sampling. The method has several advantages: collection costs are

low and the method enables researchers to connect different target populations (e.g., students with their peers, teachers, and schools), and thereby examine the contexts of learning, yielding additional analytical power and possibilities. Comprehensive lists of students are not available in many countries, while lists of schools exist; hence, the process of compiling sampling frames is simplified. However, a significant downside to this approach is that sampling precision is reduced. The effect occurs if elements within a cluster are more similar than elements from different clusters. For example, students attending the same school share the same learning environment, or, similarly, students within a class are instructed by the same teacher, and these students will therefore be more likely to exhibit similar learning outcomes. Studies combining school and class sampling thus suffer from a double cluster effect. A simple comparison of SRS with cluster sampling can be made by using the different sampling methods to select same number of pixels from a picture. This illustrates neatly that the sample derived by SRS provides a far better representation of the original picture (Fig. 7.1a), even though the same number of pixels is displayed (Fig. 7.2).

The cluster effect can differ for any variable of interest, and vary among countries and over time. Its statistical measure is the intraclass correlation coefficient (ICC; Kish 1965). As a very prominent example, the socioeconomic composition of students in schools has a relatively large effect on the ICC of achievement in many countries. This is, variance between schools in student achievement is strongly associated with the “average” socioeconomic status of the students within schools (e.g., Sirin 2005).

The effect of ICC in ILSA is often so large that the required sample sizes actually need to be approximately ten times greater than required for SRS to achieve the same

Fig. 7.2 Sampling precision with equal sample sizes of 50,000 pixels. The results of **a** simple random sampling, and **b** cluster sampling *Source* Meinck (2015b) Copyright © 2015 International Association for the Evaluation of Educational Achievement (IEA)

level of sampling precision. This is why IEA studies that aim for a sample equivalent to an SRS of 400 (ICCS and ICILS) or 800 (TIMSS and PIRLS) units, in fact require about 3000 to 4000 units. While a minimum sample size is specified, countries are able to adjust this upwards to meet the requirements for generating estimates for subpopulations that may be of particular policy interest.

7.4.2 *Stratification*

Stratification is another feature regularly used in ILSA, including IEA studies. Unlike cluster sampling, stratification can increase sampling precision (Cochran 1977). This effect occurs if units within a stratum have similar survey outcome variables. For example, suppose students' abilities in operating electronic devices are the variable of interest in a survey. If students in rural areas lack access to such devices while students in urban schools use them in lessons on a regular basis, students in rural areas are likely to have systematically lower abilities to operate them compared to students in urban areas. In such a case, stratification by urbanization can increase sampling precision, given a constant sample size. This is because it is enough to select a few students from a larger pool of students that are all alike, making sure that a few are selected from every distinct group of students. Applying this rationale to the Einstein picture (Fig. 7.1a), six strata can be distinguished: the dark background, the mouth, both eyes, the nose, and the hair. Selecting some pixels from each of those parts has a high probability of providing a rather precise representation of the original picture.

Stratification can be employed at any sampling stage. If schools separate their students into high and low ability classes, it would be reasonable to stratify classes by ability before selecting a class sample. Equally, when selecting teachers within schools, it may be useful to stratify by subject, gender, or age group, depending on the variables of interest and expected group differences. IEA studies employ stratification at all stages whenever reasonable, and distinguish between different approaches of stratification (please refer to, e.g., Martin et al. 2016 for details on these methods).

Stratification can also be useful if countries are interested in comparing specific subgroups. In such cases, disproportional sample allocations to strata are possible, for example increasing the sample size for small subgroups within the population, thereby increasing the precision (and reliability) of results for small subgroups (Dumais and Gough 2012a). Computing sampling weights (see Sect. 5) will ensure that unbiased features of the subgroups are estimated, but allows for unbiased estimation of the whole population at the same time.

Stratification and cluster sampling may seem similar to a layperson, as both methods imply assembling units of the target population into homogeneous groups. However, they have opposing effects on sampling precision. The essential difference between these methods is that cluster sampling selects only some clusters from a large pool of clusters, while stratification selects some units within each of the strata (Fig. 7.3).

(a) Cluster sampling: Selecting some groups but all or nearly all units within groups

(b) Stratification: Selecting few units from each cluster

Fig. 7.3 A schematic illustration of the distinction between **a** cluster sampling and **b** stratification. *Note* Large bubbles represent homogeneous groups; shaded bubbles are sampled. Small white bubbles within selected clusters illustrate the fact that clusters are not necessarily sampled in full. Sampling precision will be higher with stratification, even though the number of sampled units is the same in both scenarios

7.4.3 Sampling with Probabilities Proportional to Size

Sampling with probabilities proportional to size (PPS), is also commonly used when selecting samples for IEA studies. This method is again described in detail in many textbooks (e.g., Lohr 1999). With PPS, selection probabilities depend on a “measure of size” of the sampling unit; large schools (e.g., those with many students) have high selection probabilities, and, vice versa, classes or students within large schools have low selection probabilities balancing out the two selection stages. Applying PPS aims for similar estimation weights in the core target population (a so-called “self-weighting design”), a measure that can increase sampling precision for this population but will likely decrease sampling precision for secondary target populations, such as schools, as their estimation weights are more variable (Solon et al. 2015). The effect occurs only if the chosen measure of size is related to the outcome variable. Moreover, PPS is a simple way to ensure selecting an approximately equal number of students in each sampled school while getting roughly similar final sampling weights.

7.4.4 *Estimating Sampling Precision*

Sampling precision is documented by reporting estimates of standard error³ or confidence intervals along with any presented population feature. To account for the complex sampling design, standard errors must be estimated by respective statistical methods. In most IEA studies, jackknife repeated replication (JRR; Wolter 2007) is used to achieve unbiased estimates of sampling error, a method not available in statistical standard packages, such as for example, the base module of SAS (SAS Institute Inc. 2013) or SPSS (IBM Corporation 2016). Standard statistical packages use a default method applicable only to simple random samples and thus underestimate standard errors in complex sampling designs. To address this problem, IEA data sets are prepared for proper estimation of standard errors, and IEA offers with its IDB Analyzer a tool that applies the correct method automatically (see Chap. 13 for more information). Applying this method is of utmost importance regarding the reliability and validity of the study results: failure will likely lead to considerably underestimated standard errors, implying overly high precision levels in the results (i.e., confidence intervals that are too small), and can lead to falsely detecting significant group differences.

7.5 **Weighting and Nonresponse Adjustment**

Sampling weights are a reflection of the sampling design; they enable those analyzing the data to draw valid conclusions about population features from sample data. There are two reasons why it is necessary to compute weights and use them for data analysis, namely varying selection probabilities and nonresponse. Readers should refer to the technical documentation of IEA studies for detailed information on the exact computation algorithms for sampling weights and nonresponse adjustments and how to apply them for analysis; general introductions on weighting in ILSA can be found for example in Dumais and Gough (2012b), or Meinck (2015a). Rather than reiterating these procedures, this section focuses on the effects of weights for unbiased estimation in IEA studies, or, in other words, how weights help to retrieve valid and reliable statements about population features.

IEA studies account for the selection probabilities, varying due to multiple selection stages and the application of PPS, by computing design weights. In thinking about the meaning of the design weight values, the value of the design weight of a sampled unit refers to the number of other frame units represented by the sampled unit; a methodologically more accurate interpretation of the weight value is that the design weights must compensate for the fact that some units are part of a greater

³For

estimates of achievement, standard errors are composites of measurement and sampling error in IEA studies. The standard error is equivalent to the sampling error for all other variables (see, e.g., Martin et al. 2017).

number of samples than others (Meinck 2015a). For example, if school A has twice the chances of being sampled (due to stratification, oversampling, and PPS effects) compared to school B, the latter school weight must be doubled to compensate for school A being selected twice as often as school B over all possible samples of schools with a given design. By doing so, over all possible samples, the weighted contribution of schools A and B will be identical to a census of all schools.

There is generally no doubt in the research community that weights are needed to achieve unbiased estimates of population features. The following simple example illustrates this effect (Table 7.1). Suppose that a researcher wishes to compare the percentages of female and male teachers teaching in private schools with the respective percentages in public schools. To achieve estimates with the same precision, the sample size for both private and public schools has to be identical, even though private schools compose only 10% of schools in the country.⁴ Investigators seek to establish the average percentage of female teachers in each type of school, and overall. In this example, private schools are nine times more likely to be selected and, if the researcher does not apply a weighting factor, the percentage of female teachers in the total population would be drastically overestimated, inflating the average percentage of female teachers in all sampled schools (70%). However, if the appropriate design weights are applied, the estimate becomes design-unbiased and the average percentage of female teachers drops to 62%, a figure that is much closer to the expected average value for the larger part of the population, namely the public schools.

Design weights are computed separately for each sampling stage, and multiplied for sampling units of subsequent stages. If studies aim for multiple target populations, weights are computed separately for each of them.

The validity and reliability of ILSA studies and any other similar surveys can be negatively affected by failure to complete the survey; this is known as nonresponse.

Table 7.1 Example illustrating the effect of disproportional sample allocation on design weights and estimated population features

Factors		Private schools	Public schools	Total
Number of	Schools in population	1000	9000	10,000
	Schools in sample	100	100	200
Design weight		10	90	
Proportion of female teachers (%)	In population and sample	80	60	
	Unweighted sample (biased, descriptive statistic)			70
	Weighted sample (unbiased, inference statistic)			62

⁴The example assumes SRS of teachers in each stratum.

Nonresponse bias can be substantial when three conditions hold: (1) the response rate to the survey is relatively low; (2) there are significant differences between the characteristics of respondents and nonrespondents; and (3) nonresponse is highly correlated with survey outcomes. As opposed to item-level nonresponse (i.e., failure of respondents to answer individual single items), unit-level nonresponse creates higher risk to the validity of the ILSA because usually very little is known about the nonresponding units. Therefore, conditions (2) and (3) are difficult to assess. This applies especially to cross-sectional surveys and hence to ILSA surveys, and is the reason why sophisticated nonresponse models known from, for example, longitudinal studies, cannot be applied in ILSAs.⁵ Instead, IEA specifies very strict requirements for participation rates. The requirements for participation rates in IEA studies can be obtained from the technical reports; generally they translate into minimums of 85% participation of sampled schools, 95% of sampled classes, and 85% of sampled students within participating schools/classes. Further, sampling weights are adjusted for nonresponse within supposedly homogenous adjustment cells (e.g., sampling strata, classes), assuming a non-informative response model (i.e., that nonresponse occurs at random) within these cells. For example, if all schools in urban areas participated in a study but not all schools in rural areas, the weight of the rural schools has to be increased to compensate for their loss in the sample. Thus, participating rural schools represent nonresponding rural schools, while urban schools only represent the urban schools they were sampled from.

Nonresponse can occur at every stage of sampling, hence, nonresponse adjustments are computed separately for each stage. All nonresponse adjustment and design weight factors are multiplied to achieve the “final” or “estimation” weight for every unit in the sample. Again, if studies aim for multiple target populations, estimation weights are computed separately for each of them and are stored with the respective data sets. The individual technical reports for each IEA study provide considerable detail on the procedures used to calculate the weights and nonresponse adjustments in each instance.

Computing sampling weights is an elaborative process requiring accuracy and attention to detail. To ensure the quality of the final weights, multiple checks for consistency and validity are conducted. Using exported WinW3S databases, information on classes and school sizes is compared with information on the sampling frame, and significant deviations are probed and clarified. The estimated sizes of total populations and the various subpopulations are compared to the sampling frames, official statistics, and previous survey data. Large deviations may indicate systematic errors in sampling and listing procedures, and are consequently carefully investigated. Variation of the weights is checked, and outliers are identified and handled depending on the outcome of these investigations.

⁵Longitudinal studies can base their nonresponse models on responses of the participants from earlier cycles. It should be noted that nonresponse bias analysis is feasible for countries that can provide the relevant data.

7.6 Sampling Adjudication

A final important quality control measure implemented in all IEA studies is sampling adjudication, mostly conducted as a face-to-face meeting at the end of a study cycle attended by the study directors, the sampling experts, and the sampling referee. The sampling referee assumes a key role regarding sampling quality, acting as an objective external expert who is consulted over the whole cycle of a study on particular cases and challenging situations regarding sampling matters. During the adjudication phase, all debatable issues are brought to the referee's attention, and are resolved in consultation with the adjudication committee, usually aiming for a consensus, but, if in doubt, following the referee's opinion. Violations of the sampling procedures for single schools are usually handled by dropping them from the sample; exceeding limits for exclusion and coverage rates, or moderate shortfalls in participation rate requirements⁶ are annotated in all IEA reports; more severe transgressions may lead to separate reporting of the affected participating educational systems as a way to emphasize to readers that there are issues with data validity and due care is required when interpreting this data. In cases of extreme concern, the affected education system may be entirely excluded from the reported data.

Variables affected by high item-level nonresponse are also annotated in IEA reports. Users of complementary questionnaire data (e.g., those provided by parents or principals) are encouraged to pay close attention to completion rates, and to analyze the frequencies of missing values. When undertaking multivariate analyses, listwise deletion (i.e., removing all responses of a person from the analysis) may lead to exponential dropouts of cases, potentially resulting in biased analyses. In ILSA studies, this may often be the case when including information on socioeconomic features collected from parents in the analysis. Students with low achievement have higher missing rates for such features, possibly because parents do not want to reveal their (potentially low) levels of education or occupation, and thus the analysis results may be not representative for affected students.

7.7 Conclusions

The issues related to sampling are complex. Accounting for the sampling methods used is critical for understanding and interpreting study outcomes and correct analysis of the data. Users of IEA data and readers of the international reports should always refer to the extensive technical documentation that accompanies each study release and pay particular attention to the table annotations, which provide critical information designed to ensure valid data interpretation.

⁶That is, meeting participation rate requirements only after including replacement schools.

References

- Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York, NY: Wiley.
- Dumais, J., & Gough, H. (2012a). School sampling methodology. In V. Greaney & T. Kellaghan (Eds.), *Implementing a national assessment of educational achievement* (Vol. 3). Washington, DC: The World Bank.
- Dumais, J., & Gough, H. (2012b). Weighting, estimation and sampling error. In V. Greaney & T. Kellaghan (Eds.), *Implementing a national assessment of educational achievement* (Vol. 3). Washington, DC: The World Bank.
- European Commission. (2020). *Eurydice* [webpage]. Brussels, Belgium: European Commission. https://eacea.ec.europa.eu/national-policies/eurydice/home_en.
- IBM Corporation. (2016). *IBM SPSS Statistics for Windows* (Version 24.0) [Computer software]. Armonk, NY: IBM Corp. <https://www.ibm.com/analytics/spss-statistics-software>.
- IEA. (2020). *IEA studies* [webpage]. Amsterdam, the Netherlands: IEA. <https://www.iea.nl/studies/ieastudies>.
- Kish, L. (1965). *Survey sampling*. New York, NY: Wiley.
- Lohr, S. (1999). *Sampling: Design and analysis*. New York, NY: Duxbury Press.
- Martin, M. O., Mullis, I. V. S., & Kennedy, A. (Eds.). (2003). *PIRLS 2001 technical report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. <https://www.iea.nl/publications/technical-reports/pirls-2001-technical-report>.
- Martin, M. O., Mullis, I. V. S., & Hooper, M. (Eds.). (2016). *Methods and procedures in TIMSS 2015*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. <https://www.iea.nl/publications/technical-reports/methods-and-procedures-timss-2015>.
- Martin, M. O., Mullis, I. V. S., & Hooper, M. (Eds.). (2017). *Methods and procedures in PIRLS 2016*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. <https://www.iea.nl/publications/study-reports/international-reports-iea-studies/methods-and-procedures-pirls-2016>.
- Meinck, S. (2015a). Computing sampling weights in large-scale assessments in education. *Survey insights: Methods from the field, weighting: Practical issues and 'how to' approach*. <http://surveyinsights.org/?p=5353>.
- Meinck, S. (2015b). Sampling design and implementation. In J. Fraillon, W. Schulz, T. Friedman, J. Ainley, E. Gebhardt (Eds.), *ICILS 2013 technical report* (pp. 67–87). Amsterdam, the Netherlands: IEA. <https://www.iea.nl/publications/technical-reports/icils-2013-technical-report>.
- OECD. (2020). *OECD data*. Education [webpage]. Paris, France: OECD. <https://data.oecd.org/education.htm>.
- Rutkowski, L., Von Davier, M., & Rutkowski, D. (Eds.). (2014). *Handbook of international large-scale assessment*. New York, NY: Chapman and Hall/CRC. <https://doi.org/10.1201/b16061>.
- Rust, K. (2014). Sampling, weighting and variance estimation in international large-scale assessments. In L. Rutkowski, M. Von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment*. New York, NY: Chapman and Hall/CRC. <https://doi.org/10.1201/b16061>.
- SAS Institute Inc. (2013). *SAS university edition 9.4* [Computer software]. Cary, NC: SAS Institute. https://www.sas.com/en_us/software/university-edition.html.
- Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research*, 75(3), 417–453.
- Solon, G., Haider, S. J., & Wooldridge, J. M. (2015). What are we weighting for? *Journal of Human Resources*, 50(2), 301–316.
- Statistics Canada. (2003). *Survey methods and practices*. Ottawa, Canada: Statistics Canada. <http://www.statcan.gc.ca/pub/12-587-x/12-587-x2003001-eng.pdf>.
- UIS. (2012). *International standard classification of education. ISCED 2011*. Montreal, Canada: UNESCO Institute for Statistics. <http://uis.unesco.org/sites/default/files/documents/international-standard-classification-of-education-isced-2011-en.pdf>.
- UIS. (2020). *Education* [webpage]. Montreal, Canada: UNESCO Institute for Statistics. http://data.uis.unesco.org/Index.aspx?DataSetCode=edulit_ds.

- United Nations. (2019). *Convention on the rights of persons with disabilities* (CRPD). New York, NY: United Nations. <https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities.html>.
- Weber, S. (2018). Sample design and implementation. In W. Schulz, B. Losito, R. Carstens, & J. Fraillon (Eds.), *ICCS 2016 technical report* (pp. 43–51). Amsterdam, the Netherlands: IEA. <https://www.iea.nl/publications/technical-reports/iccs-2016-technical-report>.
- Wolter, K. (2007). *Introduction to variance estimation* (2nd ed.). New York, NY: Springer Verlag.
- World Bank. (2020). *Education statistics* (EdStats) [webpage]. Washington, DC: The World Bank. <https://datatopics.worldbank.org/education/>.

Sabine Meinck works for the IEA in Hamburg, Germany, being head of both its Research and Analysis Unit and Sampling Unit. Since 2006, she has been involved with the sampling, weighting, variance estimation, and analysis of nearly all contemporary large-scale assessments in education. Her experience as a member of the project management teams for IEA's TIMSS and PIRLS, with the consortia of the IELS and TALIS Starting Strong Surveys, and on the joint management committees of IEA's ICILS, ICCS, ECES, and TEDS-M, have enabled her to develop a diverse knowledge and expertise; she also serves on the board of the IERI Institute.

Dr. Meinck coordinates, guides and supports all research activities within the IEA. Her main research interest lies with the science of cross-national large-scale assessments, and the methodological challenges of complex survey data.

In support of the IEA's enduring commitment to knowledge dissemination, Dr. Meinck has conducted multiple workshops for international audiences designed to share her experiences, and teach best practices and methodologies in educational research. Topics taught range from basic to advanced statistical methods, survey design, and publication and dissemination strategies for diverse audiences. Further, she teaches a Masters Course at the University of Hamburg on "Quantitative methods in educational research." Dr. Meinck is associate editor of the Springer journal *Large-scale Assessments in Education*. She is honored to serve as a peer reviewer for several scientific journals on educational research, and many educational research networks (such as AERA and CIES).

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.