SSC Annual Meeting, May 2013
Proceedings of the Survey Methods Section

STANDARDISATION OF SAMPLING PLANS AND QUALITY ASSURANCE IN COMPARATIVE SURVEYS

Jean Dumais, Marc Joncas, Sylvie LaRoche[1]

ABSTRACT

International comparative studies in education have standardised their statistical methods, in particular, the sampling design they use. Benefits and challenges associated with such standardised designs will be examined. Lessons learned during more than fifteen years of practice will be given in closing remarks.

KEY WORDS: International studies; Standards; Sampling.

RÉSUMÉ

Les études comparatives internationales en éducation ont éprouvé le besoin de normaliser leurs méthodes statistiques, en particulier, les plans de sondage. Nous examinons les bénéfices et défis liés à la mise en place d'un plan de sondage normalisé pour ces études. Nous terminons en discutant des leçons apprises au cours de plus de quinze ans de participation à différentes enquêtes internationales.

MOTS CLÉS : Études internationales; Normes; Échantillonnage

1. Introduction

International comparative surveys in the field of education have been conducted for many years. They look to measure the effectiveness of education systems as a whole and are generally administered to students and/or teachers. A number of these surveys assess the skills acquired by students and thus provide data for comparing the performance of participating countries' education systems. Among the best known such surveys are the Programme for International Student Assessment (PISA)[2], the Progress in International Reading Literacy Study (PIRLS)[3] and the Trends in International Mathematics and Science Study (TIMSS)[4]. In addition, since 2008, the Organisation for Economic Co-operation and Development (OECD) has administered the Teaching and Learning International Survey (TALIS), which focuses on teachers' beliefs, practices and attitudes, their working conditions and pedagogical environment.

Since most of these studies rank participating countries' education systems on the basis of student performance, their findings receive heavy media coverage and are often politically sensitive. Because of their comparative nature, international studies tend to be controversial. Consequently, their managers face many challenges in ensuring that every aspect of the surveys is checked and tested so that the results are comparable and credible.

[1] Jean Dumais, Statistique Canada, Ottawa (ON), Canada, K1A 0T6 (jean.dumais@statcan.gc.ca); Marc Joncas, Statistique Canada, Ottawa (ON), Canada, K1A 0T6 (marc.joncas@statcan.gc.ca) ; Sylvie LaRoche, Statistique Canada, Ottawa (ON), Canada, K1A 0T6 (sylvie.laroche@statcan.gc.ca).)
[2]Conducted by the Organisation for Economic Co-operation and Development (OECD).
[3]Carried out by the TIMSS and PIRLS International Study Center and funded by the International Association for the Evaluation of Educational Achievement (IEA).
[4]Also carried out by the International Study Center and funded by the IEA.

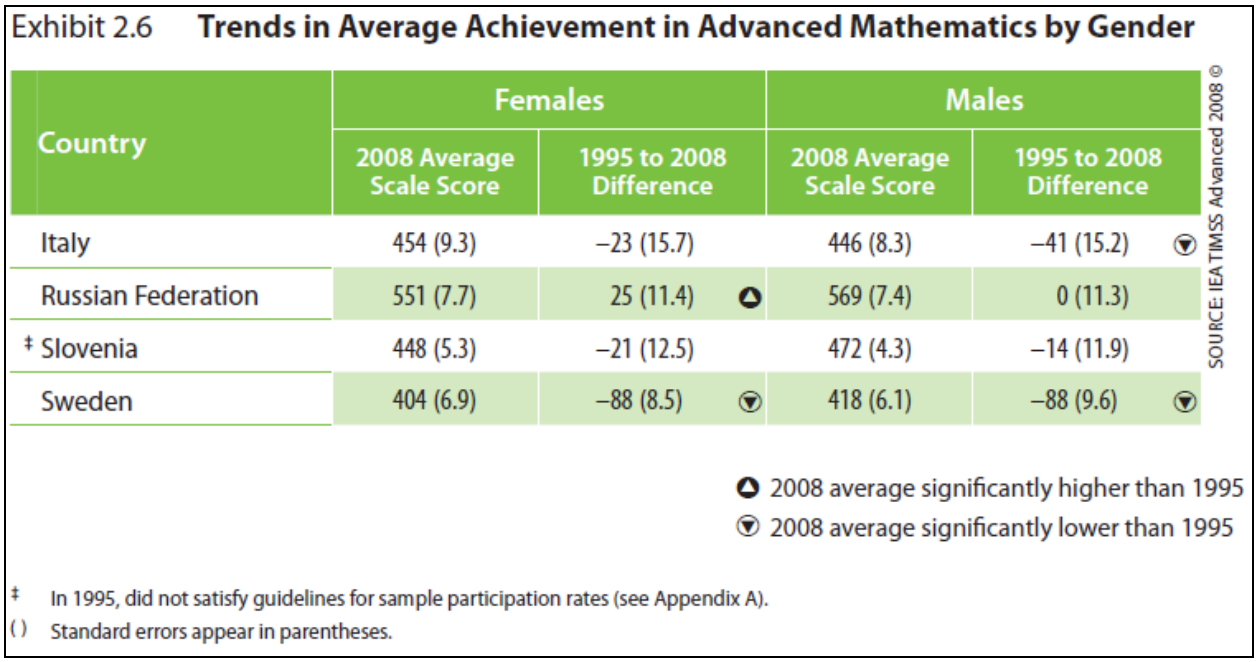

Exhibit 2.6 **Trends in Average Achievement in Advanced Mathematics by Gender**

Country	Females		Males	
	2008 Average Scale Score	1995 to 2008 Difference	2008 Average Scale Score	1995 to 2008 Difference
Italy	454 (9.3)	–23 (15.7)	446 (8.3)	–41 (15.2) ▼
Russian Federation	551 (7.7)	25 (11.4) ▲	569 (7.4)	0 (11.3)
‡ Slovenia	448 (5.3)	–21 (12.5)	472 (4.3)	–14 (11.9)
Sweden	404 (6.9)	–88 (8.5) ▼	418 (6.1)	–88 (9.6) ▼

SOURCE: IEA TIMSS Advanced 2008 ©

▲ 2008 average significantly higher than 1995
▼ 2008 average significantly lower than 1995

‡ In 1995, did not satisfy guidelines for sample participation rates (see Appendix A).
( ) Standard errors appear in parentheses.

Figure 1: Example of a table publisehed by *TIMSS Advanced 2008*

Sampling is certainly an important component. The sampling methodology used in previous surveys has often been criticized. We can always ask ourselves the following questions: Is a country's sample representative of the target population? Are the exclusion levels comparable between countries? Were the samples properly selected? Are the participation rates acceptable? These questions and many others led to the need for a series of controls and standards specifically for comparative education surveys; these controls and standards are reviewed in the sections below. Note that our review is limited to elements that relate to the sampling plan in particular. More specifically, we first describe the conditions for an ideal context leading to an appropriate sampling plan and then move on to the difficulties of implementing such a plan in the field. That section is followed by a brief discussion of the reasons for standardization in an international context. Section 4 presents a number of examples of the most widely used controls and standards in this type of survey, and section 5, the conclusion, contains a list of the lessons learned from many years of experience working on international comparative education surveys.

2. CONTEXTS

2.1 Ideal context

In an ideal context, a sampling plan should lead to unbiased, accurate and internationally comparable results. Consequently, three conditions appear to be essential in designing international surveys.

First, the survey population has to be the same as the target population. This condition is crucial because the inferences relate to the survey population. Following collection, it is not always possible to make adjustments in the estimation process to remedy coverage flaws. It is often easier to convince people of the importance of fully covering the population of interest when you are conducting a census. What good would a census be with just 80% coverage?

In the case of a sample survey, however, people tend to think that this condition is not as critical, though it actually is. It is even more important in an international context, since it is the first thing that can undermine the survey's credibility. Coverage rates that differ between countries inevitably invite discussions that cast doubt on the validity of the comparisons and can limit the potential for analysis. When it comes to education surveys, this element is particularly important because it can leave the impression that the results will be biased if the portion of a country's population that is not covered includes the worst students.

The second essential condition is a valid sampling plan. It is our connection with the survey population. Every unit in the survey population must have a chance of being selected (a non-zero inclusion probability), or the exclusion level will increase (and coverage will be incomplete). The probability of inclusion must be known and calculable to eliminate a risk of bias. It is also important to ensure that the sampling plan is properly implemented (because there is no point in developing good sampling plans if they are not executed correctly in the field). Without a valid sampling plan, it is risky to use the survey results to make inferences about the survey population.

The third and final condition is that sampling errors should be as small as possible. This will ensure that the survey data are as close as possible to the values that would have been obtained with a census.

In an ideal context, where all these conditions are met, with perfect implementation of the sampling plan and the collection procedures, and with 100% response rates, there would be no need for standards such as those set out in section 4.1.

2.2 International context (in the field)

In survey projects involving a number of countries, it is rare to find an ideal context in which the conditions described in the previous section are fully satisfied. Just about everything differs from one country to another, and education is no exception. No matter what attribute one examines—the education systems themselves, geographic or cultural characteristics, the availability of and access to pertinent administrative data, the capacity to conduct surveys, the response burden or the survey culture,[5] to name only a few—it is impossible to find two perfectly comparable countries. Invariably, coverage rates, response rates, the quality of the implementation—in fact, everything that relates to the sampling plan—is affected to differing degrees when a survey goes from one country to another. This observation does not mean that all attempts at international comparison are doomed to failure, but rather that such projects require the establishment of minimum standards and norms below which comparisons become suspect.

3. WHY STANDARDISE

The use of standards and controls in this type of study validates comparisons of results and increases their credibility for users. The aim is to be able to attribute a significant difference (statistically speaking) in the results to a real difference in the populations being compared, and not to a combination of uncontrolled errors (inadequate response rate, poor coverage, inferior implementation of field procedures, measurement errors, *etc*.). Users' motivation to analyze the results is heavy influenced by the quality of the data and the relevance of results comparisons.

Establishing standards also makes it much easier to implement the sampling plan. In the particular domain of international education surveys, standardization often means unification of procedures. For example, in prestigious surveys such as PISA or TIMSS, the sampling plan, the sample size, the collection method or even the wording of questionnaires is nearly the same for all participants (examples of standards will be provided in the next section). Such unification ensures more balanced workloads among the participating countries and makes it possible to introduce effective, uniform minimum measures for data quality control.

In summary, the establishment of standards and controls pays off for all parties involved: the sponsors are reassured by valid, comparable results that their investment was worthwhile; the survey managers can guarantee the quality of the procedures and the validity of the results; and the participating countries obtain results of assured quality for what is essentially an equivalent amount of work regardless of their individual conditions, constraints and environment.

4. EXAMPLES OF STANDARDS AND CONTROLS

4.1 Introduction

We distinguish between standards and controls as described below. Standards are norms established to ensure data quality. Those norms are described and documented. For countries that fail to comply with standards, the consequences usually range from notes in published tables to relegation of all their results to an appendix.[6] Controls are less formal and often take the form of more flexible criteria. They are more commonly referred to as quality control procedures. Such control procedures were defined and applied on the basis of years of experience. However, some procedures that were controls early in the history of international comparative studies are now established standards.

4.2 Target and survey populations

The first examples of standards relate to the target population and the survey population. Each participating country is required to provide a description of its national population. At the very least, it must describe its education system (including the age at which

[5]Survey culture refers to a country's openness to surveys. In some countries, participation in surveys may be mandatory, while in others, surveys are voluntary and are sometimes perceived as an invasion of privacy.

[6] See example in Introduction. The results of non-compliant countries are omitted from the main tables and placed in an appendix at the end of the reports.

children start school, the school structure, and the ISCED level[7]). The survey population must cover at least 95% of the target population (this standard is probably the most widely recognized by the education community). Any divergence between survey population and target population must be documented (type and magnitude of exclusions). If a country has less than 95% coverage of the target population, this will automatically be noted in international publications. A 95% standard may seem high, but it is important to keep in mind that in international publications, there is one row per country in the results tables. The user naturally assumes that the results are representative of the country as a whole, and comparisons between countries are made on that basis. As mentioned previously with regard to education surveys, there is often a strong presumption of correlation between coverage and measured performance. It would be difficult to lower those requirements and still claim that the comparisons are credible.

Certain controls are commonly used to ensure that the definition of the national target population matches the definition of the international target population. Additional checks of the information provided by the country (such as age, years of schooling, and school attendance[8]) are performed against external data sources. For repeated surveys, we check that the definition of populations is comparable between cycles so that estimates reflect valid trends. If the target population changes from one cycle to the next, the portion of the population that is common to both cycles must be identified so that a trend analysis can be carried out. The latter situation is not so unlikely. For example, it may arise following a change in a country's education system that pushes the school year representing four years of formal education (often used as a basis for defining a population) from the fourth year to the fifth year at the start of the new cycle (which generally extends over three to five years). A reform may also change the age at which children start school. To our knowledge, there are currently no norms that require a specific percentage in those situations (such as a minimum percentage of the population common to the two cycles).

4.3 Sampling plan

This subsection covers standards applied to the sampling frame, the sample selection method, the sample size and the implementation of the sampling plan.

4.3.1 Sampling frame

In education surveys, the sampling frame used is often composed of a list of schools, and the size measure is the number of units in the target population (teachers or students). To our knowledge, there are no established, published standards for checking the frame's quality. However, good practices (controls) are followed that are similar to the checks usually performed on all sampling frames. First, we check that the frame is as up to date as possible. Second, we make sure that the frame supplied by participants provides complete coverage of the survey population and that it contains no erroneous data, duplicates or elements extraneous to the survey/target population. In addition, wherever possible, an up-to-date size measure for each unit in the frame is required. We also insist that the sampling frames supplied by the countries provide access to the entire target population. This makes it possible to estimate and document exclusions more effectively. Moreover, we use tools such as the Web, information from previous cycles, and information from other countries to validate the information supplied by country representatives. For example, a number of countries have international schools. A country may have omitted these schools because they are not considered part of the education system.

4.3.2 Selection method

With regard to the standards associated with sample selection, the norm is to require a single selection method for all participating countries. Adaptations and/or deviations are permitted, but they must be approved by the survey managers before implementation, and they must be documented. Having a single selection method allows us to develop and use generalized sample selection and weighting programs, thereby minimizing the risks of error. This approach facilitates the equitable distribution of work among the participants, resulting in much more uniform sample sizes across the various countries. It also facilitates the validation of selections and minimizes the number of control programs required for implementation. In addition, with the adoption of a single method, collection operations can follow uniform procedures, which limits the number of operations manuals required. This reduces the risks of error, preventing differences in instructions from affecting data quality and comparability. The use of a single selection method also helps reassure participants of the comparability of the results: non-sampling errors are expected to be comparable. Sampling errors too are expected to be of similar magnitude, which is not necessarily the case; nevertheless, the perception remains. Note that there are consequences to not meeting standards. The risk that the data will not be published or will be annotated in the tables increases substantially if the plan is not approved or irregularities are observed.

[7]The International Standard Classification of Education (ISCED) is a UNESCO standard for classifying education systems.
[8] School attendance is defined as the percentage of the age cohort in a given school year that is attending school.

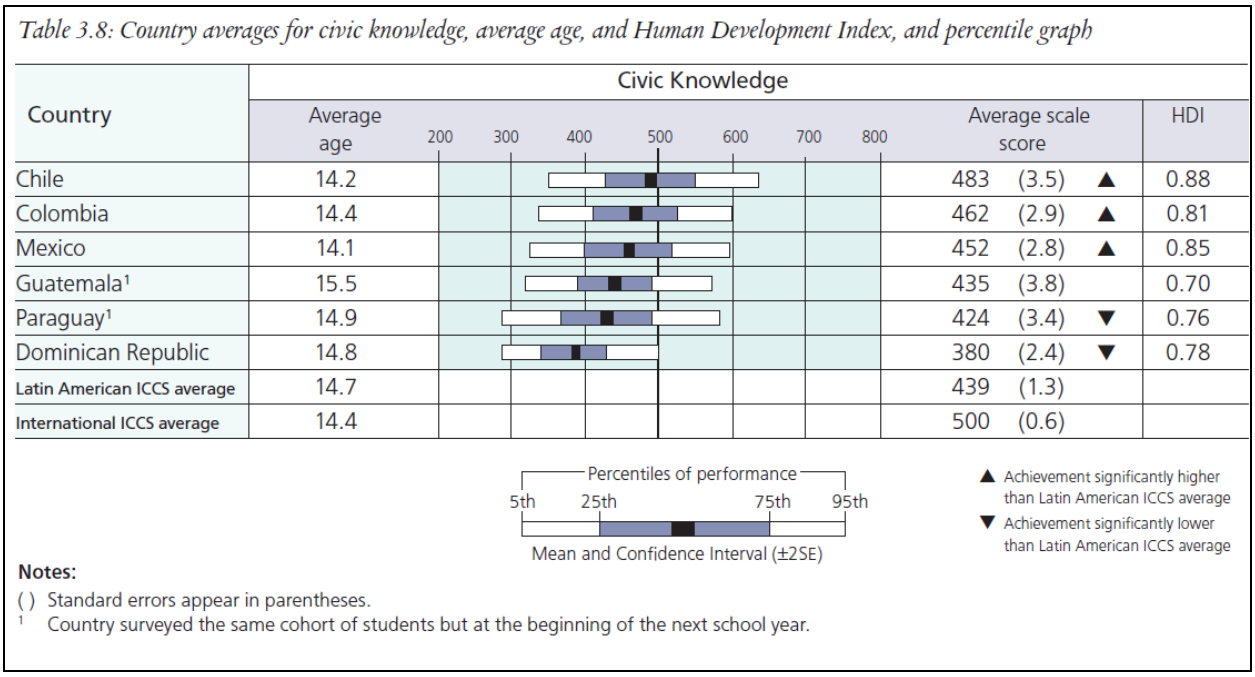

Table 3.8: Country averages for civic knowledge, average age, and Human Development Index, and percentile graph

Country	Civic Knowledge			
	Average age	Average scale score		HDI
Chile	14.2	483	(3.5) ▲	0.88
Colombia	14.4	462	(2.9) ▲	0.81
Mexico	14.1	452	(2.8) ▲	0.85
Guatemala[1]	15.5	435	(3.8)	0.70
Paraguay[1]	14.9	424	(3.4) ▼	0.76
Dominican Republic	14.8	380	(2.4) ▼	0.78
Latin American ICCS average	14.7	439	(1.3)	
International ICCS average	14.4	500	(0.6)	

Notes:
( ) Standard errors appear in parentheses.
[1] Country surveyed the same cohort of students but at the beginning of the next school year.

Figure 2: Example of table published by the *International Civic And Citizenship Education Study, 2009*

4.3.3 Sample size

Invariably in this type of survey, there is a standard for the minimum size of the sample. Most of the time, the minimum size is based on the desired margins of error and the study's inherent constraints. Another commonly used standard is the advance identification of so-called replacement schools. In general, there is a maximum of two replacement schools for each school originally selected. Again, any deviation must be documented and approved. Note that replacement schools cannot be used to replace eligible schools that refuse to participate. The use of replacements satisfies requirements concerning the sample size and may help minimize the risk of bias. Nevertheless, we maintain strict requirements regarding the minimum participation rate of originally selected schools (see the next section).

4.3.4 Implementation

The most widely recognized standard in this area is the requirement that all countries take part in a trial. The purpose of the trial is to test procedures in the field and, in particular, take corrective measures (which are often needed) before the survey begins. In principle, participation in the trial is compulsory; non-compliant countries will have their data omitted from the international publications.

Another recognized standard is the minimum response rate. In general, we can define three zones:

(1) There is the absolute minimum zone, or red zone. If a country fails to achieve these minimum rates, its data will simply be excluded from all international publications.

(2) At the other end of the spectrum is the green zone. A country is in the green zone if its participation rates are above a certain threshold, without the use of replacement schools. In this case, the risk of bias in the statistics derived from that country's data is considered negligible. If the threshold is attained only after the replacement schools are brought in, the country's results are included in the international publications, but they are annotated to alert users to the increased risk of bias.

(3) Then there is the grey zone. If a country's participation rates are above the threshold for the red zone but below the threshold for the green zone, even after the replacement schools are used, a decision is usually made on a case-by-case basis. The results may be placed at the bottom of the tables or in appendices, or they may not be published.

It is essential that response rates (sometimes referred to as participation rates) be documented so that analysts can assess the quality of the inferences and analyses based on the data.

With regard to controls, it is worth noting that during implementation, countries are often instructed to contact the survey managers when they encounter an unusual situation. This makes it possible to check and take action before data collection is complete and no further corrective measures can be taken. The presence of a measure of school size in the sampling frame provides another control: comparison of that size with the size observed in the field. It is possible to request explanations and more detailed documentation in cases where the differences are substantial (omission of classes, an error in identifying the school, a change in the school's structure, *etc.*). In addition, it is not uncommon to validate the status of non-participating schools following collection to determine whether a more appropriate status should be considered (for example, classify some refusals as exclusions). Standard errors are calculated in part to detect outliers, influential values, and influential or abnormal weights on the basis of the key variable of interest. It is also possible to compare observed and expected estimates (for example, exclusion rates in schools compared with rates in previous cycles, population totals compared with known totals from previous cycles). All these controls help detect potential violations of the rules set out in the sampling plan. Participants are usually required to provide written explanations for any abnormalities detected.

Lastly, we would like to point out the importance of conducting an evaluation of the implementation. Sampling plans and their implementation are usually reviewed in the presence of an expert from outside the survey's management circle. This independent evaluation and approval lend important credibility to the survey. In addition, it is essential to wrap up the project by preparing a technical report describing all the procedures affecting the sampling plan and its implementation.

5. CONCLUSION

From our experience with international comparative surveys, we have learned the following:

(1) It is difficult to have standards that meet every need and retain some flexibility. In every survey or cycle, we have to deal with unusual situations. It is therefore important to have a technical team responsible for supporting the participants and to invite them to consult the team before and during the survey's implementation to address the various unforeseen problems.

(2) The establishment of standards is necessary and critical to dispel any doubts about the relevance of the analyses based on the survey.

(3) It is important to check, at a reasonable cost, all the procedures followed to improve the quality of the data collected.

(4) It is also important to quantify and document the actions taken to ensure the quality and comparability of the data and build confidence in those responsible for implementing the survey.

It is safe to say, of course, that there is always room for improvement. Some of the above-mentioned controls could easily be beefed up and turned into standards. As survey managers, however, we have to exercise caution and maintain a degree of flexibility, always with a view to guaranteeing an acceptable level of quality. Standards are a constraint for the participating countries. Should we therefore aim for more standards to the detriment of flexibility and accommodation in the field, or should we take the opposite course, with the risks of possible abuse? The debate is still open.

REFERENCES

International Association for the Evaluation of Educational Achievement (2011) *ICCS 2009 Latin American Report : Civic knowledge and attitudes among lower-secondary students in six Latin American countries*, Amsterdam: IEA.

OECD (2012) *PISA 2009 Technical Report*, PISA, OECD Publishing, http://dx.doi.org/10.1787/9789264167872-en

TIMSS & PIRLS International Study Center (2009), *TIMSS Advanced 2008 International Report*, Chestnut Hill, MA; TIMSS & PIRLS International Study Center, Boston College.

TIMSS & PIRLS International Study Center (2007), *TIMSS 2007 Technical Report*, Chestnut Hill, MA; TIMSS & PIRLS International Study Center, Boston College.

TIMSS & PIRLS International Study Center (2005), *TIMSS 2007 school sampling manual*, Chestnut Hill, MA; TIMSS & PIRLS International Study Center, Boston College.