Computing Sampling Weights in Large-Scale Assessments in Education

Survey Methods: Insights from the Field, Special issue: Weighting: Practical Issues and 'How to' Approach

Sabine Meinck

How to cite this article : Meinck, S. (2015). Computing Sampling Weights in Large-Scale Assessments in Education. Survey Insights: Methods from the Field, Weighting: Practical Issues and 'How to' Approach. Retrieved from https://surveyinsights.org/?p=5353

DOI : 10.13094/SMIF-2015-00004

Copyright : © the authors 2015. This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0)

Abstract : Sampling weights are a reflection of sampling design; they allow us to draw valid conclusions about population features from sample data. This paper explains the fundamentals of computing sampling weights for large-scale assessments in educational research. The relationship between the nature of complex samples and best practices in developing a set of weights to enable computation of unbiased population estimates is described. Effects of sampling weights on estimates are shown, as well as potential consequences of not using weights when analysing data from complex samples. Illustrative examples are provided in order to make it easy to understand the rationale behind the mathematical foundations.

1. Introduction

There are a number of technical standards that are universally applied to all modern large-scale assessments in education (LSA) today, including TIMSS (Martin et al., 2012), PIRLS (Mullis et al., 2012) and PISA (OECD, 2014), to name just a few of the most important current educational studies. One important standard is the call for unbiased random samples. Only such samples can guarantee that valid inferences from sample data on population features can be made, an essential criterion for international comparability as well as comparability between sub-national entities.

The sample designs used in LSA, however, are not only random, but also complex. The features of complex sample designs need to be taken into account when estimating population parameters and standard errors. In order to understand the relation between sampling weights and sample design, this paper will first introduce the different types of weights and then review the specific features of a complex sampling design, their effects on estimates, and the use of weights to statistically offset these effects.

The general approaches to deriving sampling weights described here were applied in all current LSA, including all studies by the International Association for the Evaluation of Educational Achievement (IEA)[1], as well as those conducted by the OECD[2], and many other important national educational studies.[3]

2. Types of weights

Weights are values that are assigned to every unit sampled; they are usually stored as variables in the public-use data files of the various studies. If a study targets more than one population, the weights are computed independently for each target population. Franklin & Walker (2003), Groves et al. (2004) and Rust (2014) are recommended for further reading on the topic.

2.1 Design weights

The design weight *d* of a sampled unit *i* at a particular sampling stage is computed as:

$$d_i = \frac{1}{p}$$

The so-called “design” or “base” weights reflect the selection probability *p* of the sampled unit at each sampling stage and thus account for multiple-stage sampling procedures. For example, if 1 out of 10 schools was selected and all schools had the same selection probability, then the probability of any school being sampled is 1/10 and the design weight of the sampled school is 10. As another example, if a bowl contains 100 balls with individual student names on them and a sample of 20 is selected, each student’s selection probability is 1/5 and the weight of the sampled students is 5. This selection approach is called “simple random sampling” (SRS). By definition, in SRS sampling each possible sample of size *n* has the same selection probability (Lohr, 1999). LSA generally rely on other selection methods (see section 3.3), but for any random sampling method, the selection probabilities will be known.

There are two approaches to think about the meaning of the design weight values: In a sense, the value of the design weight of a sampled unit refers to the number of other frame units represented by this sampled unit. For example, if the sample weight for a student is 100, this could be seen as representing 100 students in the population. But there is a methodologically more accurate interpretation of the weight value. Much like the famous “throw of the dice” experiment that is undoubtedly familiar to most students around the world, if, for example, we select as a sample 2 out of 10 possible schools a million times over, each of the schools will be represented in 20% of the samples. In some types of sampling designs, explained later in this paper, selection probabilities can vary for the sampling units. Design weights thus must compensate for the fact that some units are part of a greater number of samples than others.

2.2. Non-response adjustments

A non-response adjustment factor reflects the degree of success in survey implementation as it accounts for the possibility of a non-response, which occurs when a sample unit refuses to participate in the survey.[4] As with weights, adjustment factors are computed separately for each sampling stage. The main principle at work here is that the weight of the non-respondents within a specific adjustment cell must be re-distributed among the responding units in that cell. Such an “adjustment cell” contains sampling units that share specific features. For example, all private schools in a given region could

comprise a “stratum” of schools, from which a sample of schools can be selected. If some of the sampled schools refuse to participate, then the remaining (participating) schools in this stratum will carry the weight of the non-participating schools. In other words, we try to model the non-response, assuming that it is related to some features known for responding and non-responding units (see also Lohr, 1999).

In SRS sampling (all units carry the same weight), if n_i is the number of sample units within the adjustment cell and r_i is the number of responding units in that cell, the non-response adjustment f for each responding unit i is computed as

$$f_i = \frac{n_i}{r_i}$$

within each adjustment cell. If, for example, 8 out of 10 sampled schools in a particular stratum agreed to participate (the other 2 schools declined), then the non-response adjustment is 10/8 = 1.25. The sum of all non-response adjustment values for the 8 responding schools is 10 – hence, we do not “lose” the weight of a non-responding school or a part of the population to be represented in the sample. To give another example, if out of 20 students sampled in a class 4 are absent, the non-response adjustment for the 16 participating students is computed as 20/16 = 1.25 (the non-response adjustment cell is usually the class).

In those cases where the selection probabilities vary for sampling units this is taken into account, in some studies, when computing the non-response adjustment (for details see Rust, 2014).

This approach of adjusting for non-response assumes a non-informative response model. In other words, it presumes that non-response occurs completely at random within the adjustment cell – the non-responding units do not vary systematically from the responding units. We can also look at this from a more theoretical perspective: over all possible samples, all sampled units within an adjustment cell are assumed to respond (or not respond) equally often. If this assumption is violated, population inferences may be biased. This actually poses a big challenge on any assessment, as the features of the non-respondents usually remain unknown; thus there is no proof that the assumption holds. This is why LSA adopt very high standards for participation rates (see, for example, Gregory et al., 2001; OECD, 2012): Bias risk due to systematic non-response should clearly be kept at a minimum.

The determination of the adjustment cell should be carefully considered in any LSA. For example, in LSA with class sampling, the class-level non-response adjustment is usually applied within the stratum rather than the school. If, for instance, in a stratum containing 10 sample schools 2 classes are selected, and a single class in a school rejects participation, the non-response adjustment is computed as 20/19 (the adjustment is applied to all schools in the stratum). Why? If the school were the non-response adjustment cell, the participating class would double its contribution to any estimate (non-response adjustment = 2/1). This approach would increase the bias if classes with specific features more frequently refuse to participate than others (non-response occurs not completely at random) – in practice, a very likely scenario, e.g., when school principals, fearing negative consequences from poor results, decline participation for the lower-achieving of the two selected classes. Of course, if there is evidence that a non-responding class is more similar to other classes in its school than to classes in other schools in the stratum, retaining the school as non-response adjustment cell would be the better choice.

In fact, if non-response depends on a known feature of the sampled units, this feature should be used for the determination of the adjustment cells. If, for example, boys participate less often than girls, and there is evidence that the main objectives of the studies are gender-dependent, the most appropriate adjustment cell for student non-response may be not the class but gender groups within classes. It should

be noted that this approach is limited because adjustment cells that are too small are disadvantageous.[5]

Careful non-response analysis always provides valuable information on how adjustment cells are to be determined and on the reliability of the survey results and therefore should be a standard part of data analysis in any assessment.

2.3 Estimation weight

Once the design weights and non-response adjustment factors have been computed for each sampling stage, the “final” or “estimation” weight can be derived by multiplying all these factors. For example, if schools were sampled in a first sampling stage and then students from the sampled schools were sampled in a second stage, two design weights and two non-response adjustments would have to be computed. The estimation weight *e* of each sampled school *j* is then given as

$$e_j = d_j \times f_j$$

while the estimation weight *e* of each sampled student *i* has to consider both sampling stages and is therefore given by

$$e_i = d_j \times f_j \times d_{ji} \times f_{ji}$$

When adding further sampling stages, the respective design weights and adjustment factors have to be included as further factors in the above formula. The formulas also show that in multi-stage sampling designs, all stages carried out to sample a specific unit contribute to its estimation weight.

The estimation weight is used for estimating any population statistic relying on sample data, such as means, percentages, correlation or single-level regression coefficients, etc.[6] It also constitutes a tool for checking whether the survey sampling procedures were implemented correctly. The sum of the estimation weight values is an estimate of the population size, which can then be compared with available official statistics. For example, the sum of all student estimation weight values in PIRLS 2011 in country X is an estimator of the total number of 4^{th} grade students in the country that were eligible for the assessment in 2011. Significant deviations between this estimate and the country-level statistics could be a sign that mistakes were made during survey implementation. For example, omitting classes on class lists prior to class sampling or omitting students within sampled classes would lead to unduly small population size estimates.

3. Features of complex sample designs and their effects on selection probabilities, weights, and estimates

Comprehensive introductions to sampling techniques are to be found, for example, in Cochran (1977),

Lohr (1999), Franklin & Walker (2003) and Rust (2014). The following sections focus on the relation between sampling design features and sampling weights.

3.1 Stratification

Stratification is part of many LSA sampling designs and comprises the grouping of sampling frame units by common characteristics. Examples would include the allocation of schools by geographic region or school type if school is the frame unit, or by gender if individuals are units on the sampling frame. IEA technical reports refer to “implicit” and “explicit” stratification. Implicit stratification implies simply sorting the sampling frame by stratification variables prior to sampling – a straightforward method to achieve a fairly proportional sample allocation across all strata, but with no direct influence on sampling weights. Explicit stratification is generally used for the following reasons:

– To improve the efficiency of the sample design, thereby making survey estimates more reliable;

– To apply disproportional sample allocations[7] to specific groups of schools to ensure adequate representation of specific groups of interest (domains) of the target population in the sample; this is crucial when estimates per domain are required by the national analysis plan.

Disproportional sample allocation is directly related to sampling weights: the bigger the skew in the sample allocation, the more the selection probabilities – and consequently the design weights – will vary for the units in the different strata. As already noted, strata are often used to determine adjustment cells for school-level non-response adjustments. If the non-response patterns vary between strata, then the non-response adjustments and, subsequently, the estimation weights will also vary. The following example should illustrate the effect (see also Table 1).

Table 1: Example: Stratification and weights

		School type A (Stratum 1)	School type B (Stratum 2)	Total
Population size		1,000	10,000	11,000
Sample size		100	100	200
Design weight		10	100	
Respondents		100	50	150
Non-response adjustment		1	2	
Estimation weight		10	200	
Mean score	unweighted	500	600	**533**
	weighted	500	600	**591**

Let us suppose there are two school types, A and B, in a given country. Type A schools are attended by 1,000 students, while the greater majority of pupils (10,000) attend Type B schools; hence, the total school population in the country is 11,000 students. The students are tested in their mathematics achievement and a mathematical score is derived from the assessment for each student. The objective is to compare the mathematics achievement of the two school types and to estimate the overall average achievement across the school types, i.e., in the whole population. In order to derive estimates for both school types with similar precision levels, the same sample size is applied and 100 students from each school type are selected using SRS. This is a very disproportional sample allocation – a proportional allocation would have resulted in a sample of roughly 18 type A and 182 type B students if the total sample size is held constant. To account for the disproportion, the design weights for students attending the two school types differ by factor of 10.

To add a little complexity, let us assume the non-response patterns differ between school types. While all sampled students of type A participate, half of the type B students refuse and therefore the non-response adjustment varies accordingly for students in the different strata. As a result, the estimation weight varies by a factor of 20. Finally, let us assume the students of school type A have an average mathematics score of 500, while those of type B vary around a score of 600. As long as each school type is considered separately, the weights do not matter when estimating the average scores in this example. However, when estimating the mean achievement of the whole population, the unweighted score is obviously heavily biased towards the mean achievement of type A students, a result of their overrepresentation in the sample. Once this overrepresentation has been accounted for (by applying the estimation weight), the estimated average population score is much closer to the average score of the majority population group (type B students) and presumably much closer to the true population mean.

Of course, this is an extreme example that serves to illustrate the general concept. Let us also look at a real-world example, an analysis of ICCS 2009 student data from Chile. Table 2 shows an estimated percentage distribution of highest achieved educational level of the students' fathers. If the sampling weight is neglected (right column), one would infer that almost 18% of the fathers of 8th grade students (who constituted the ICCS target population) achieved a university degree (ISCED level 5A or 6). In fact, the unbiased estimated percentage is just 13.3%. The reason for the discrepancy between the weighted and unweighted estimates is that an oversample of private schools was selected; fathers of students in private schools more frequently reported a high education level. The mean civics and citizenship knowledge score drops for the same reason from 494 (unweighted score) to 483 when (correctly) applying the weights.

Table 2: Example: Unbiased and biased estimates (in %) of education levels of students' fathers (ICCS 2009, Chile)

Father's highest level of education	Correct estimate (unbiased, weighted)	Incorrect estimate (biased, unweighted)
<ISCED level 5A or 6>	13.3	17.7
<ISCED level 4or 5B>	12.1	12.6
<ISCED level 3>	44.6	40.4
<ISCED level 2>	21.5	17.2
<ISCED level 1 or none>	8.5	6.7

3.2 Multiple sampling stages and cluster sampling

All previously cited LSA use multiple-stage sampling. The major reason for choosing this approach is that comprehensive sampling frames for the in-scope individuals are usually not available, e.g., a list of all 4th grade students eligible for PIRLS cannot be delivered by most, if not all, countries participating in this study; however, a complete list of all schools offering 4th grade education is available. Therefore, a sample of such schools is selected first and then, in a second stage, a sample of classes within those schools is selected. In other LSA, students comprise the second-stage sampling unit (e.g., PISA, ICILS); three-stage cluster sampling is also applied in specific circumstances.[8]

This multiple-stage sampling approach usually yields less efficient samples because of the cluster nature of the samples, along with the fact that individuals within a cluster tend to be more similar to each other than individuals between clusters. Students in the same class, for example, tend to share similar social backgrounds, a common school environment, and the same teachers. Hence, the gain in information when the sample size within a cluster is increased is usually not as large as when the number of sampled clusters is increased, even when the total sample size of individuals remains constant. This is the so-called “design effect”, defined as the ratio between the sampling variance of a complex design divided by the sampling variance of an SRS sample of the same size. The cluster sampling approach, however, also has valuable advantages. Student-related variables can be put into the classroom and school context, and various target populations can be linked together at low costs. Because of the similarity of individuals within a cluster, clusters are often used as non-response adjustment cells.

As already stated, the design weights and non-response adjustments have to be computed separately for each sampling stage. This is crucial as the multi-stage procedure can lead to greatly varying weights. An extreme (yet actual) scenario will possibly enhance understanding of this concept.

Let us assume we would like to sample every second school from a list of schools[9] with equal probabilities (a quite likely scenario under certain circumstances in LSA) and, in a second stage, select one classroom in each school (see Table 3). This design is referred to as a two-stage cluster design. Within the sampled schools, one class (and all its students) is to be sampled. Obviously, the likelihood of being part of the sample will be much larger for a student in a small school, i.e., a school with few classes. If there is only one class in a sampled school, all eligible students will be sampled; in other words, their selection probability equals 1. In a larger school, the likelihood of any one student being in the sample depends on the number of classes – the more classes the school has, the smaller the likelihood; as a consequence, in this design students of large schools will be underrepresented in the sample.[10] If they differ systematically from students in small schools, estimates will be incorrect unless we account for this underrepresentation by using the correct estimation weights as computed in Table 3.

Table 3: Example: Multiple sampling stages and weights

	Small schools (1 class)	Large schools (5 classes)
Design weight 1 (school weight)	2	2

Design weight 2 (class weight)	1	5
Estimation weight	2	10

3.3 Sampling with selection probabilities proportional to size (PPS)

If the goal of a study is to collect information from more than one target population, a decision must be made when determining the sampling design for the study – the design can only be optimized for one particular target population; this is a problem that applies to most LSA. Data is not only collected from students, but also from schools; in some studies, teachers themselves comprise a target population (ICCS, ICILS).[11] Mostly, though, students constitute the core target population, which is why the sampling design is usually optimized for student population estimates. A so-called “self-weighting” design is considered optimal for this purpose. In a self-weighting design, the sampling units have the same (or at least similar) selection probabilities and hence similar weights. If one looks at the example in Table 3, it becomes clear that a simple random sample of schools will not yield a self-weighted student sample (unless all classes are selected within the sampled schools). One method for achieving self-weighted samples is sampling with probabilities proportional to size (PPS sampling; Lohr, 1999). In this method, the selection probabilities depend on the size of the primary sampling unit: large schools have large selection probabilities and, vice versa, small schools have small selection probabilities. The measure of size (MOS) used for each school is usually equal to the expected number of eligible individuals (e.g., students) MOS_j in that school. The selection probabilities in PPS are then computed for school j as:

$$d_i = \frac{n \times MOS_j}{\sum MOS}$$

with n being the number of schools selected and the denominator comprising the total measure of size (e.g., the total number of all students) in the stratum (Brewer & Hanif, 1983).

In general, in order to achieve a self-weighted sample, PPS must be used at all sampling stages except the final one, where a sample of fixed size is selected by means of SRS. In a two-stage design, if a subsample of the same size is selected within each primary sampling unit, then all subsampled units have the same total selection probabilities and consequently the same final weights. This is illustrated by means of a simplified example in Table 4. Let us assume we have a population with 6 schools, of which 2 are large schools with 100 students each, and 4 are small schools with 50 students each. The total student population size is thus 400. We wish to sample 2 schools and 25 students per school. Finally, let us assume that all sampled schools and students participate in the survey. As shown in the table, the estimation weights in this example are equal for all students (one could thus ignore them when analysing the data), but they are far from equal for the schools – thus acknowledging the fact that large schools are overrepresented in the sample.

Table 4: Example: Self-weighting sampling design

	Small schools (50 students)	Large schools (100 students)
Selection probability (school)	2 x 50/400 = 1/4	2 x 100/400 = 1/2

Selection probability (student within the school)	25/50 = 1/2	25/100 = 1/4
Total selection probability (student)	1/4 x 1/2 = 1/8	1/2 x 1/4 = 1/8
Estimation weight (students)	1/(1/8) = 8	1/(1/8) = 8
Estimation weight (school)	1/(1/4) = 4	1/(1/2) = 2

PPS sampling is used in the vast majority of countries in all LSA. Thus, one might be tempted to argue that as the data are derived from self-weighting designs, the weights will not vary (and therefore matter) much and can safely be ignored. This, it should be emphasized, would be a big mistake! The example above is an ideal one and has little in common with actual practice. The following gives a list of scenarios that will lead to varying weights even though the general sample design is intended to yield self-weighted samples.

First, let us recognize that even our simplified example from Table 4 will most certainly result in misleading estimates when unweighted school level data are being analysed and the variables of interest are related to the size of the school. A simple analysis of the grade 4 TIMSS 2011 school-level data file for Germany shows this easily. When estimating the average number of computers per school (neglecting weights), we would (incorrectly) infer that primary schools in Germany have 15 computers per school on average. When the school estimation weight is applied, we find that the correct average is 13. Obviously, large schools have more computers and are overrepresented in the sample because it was sampled using the PPS method.

Second, quite many studies apply class sampling as second sampling step. All students in the sampled class are asked then to participate. This sampling design yields hardly ever “pure” self-weighted sample; it is – at best – close to a self-weighting design. Moreover, countries are often asked to sample two classes instead of one in large schools. This is done to increase the total student sample size without increasing the number of schools to be sampled (a cost issue) – the situation, for example, for about half of all countries participating in PIRLS 2011. In another case, in order to reduce sampling variance, countries may be requested to sample two classes in all schools that group students in classrooms by ability.[12] In both cases, the class selection probabilities will differ from the ideal self-weighting scenario.

Third, the magnitude of non-response often varies by stratum or school. As already shown in section 2.2 and in the example in Table 1, this can also lead to varying weights. Imagine, for example, a country where participation is mandatory for schools supervised by the ministry of education, but private schools are left free to make their own decision: the non-response adjustment factor for public schools will be 1, but most likely > 1 for private schools, which, of course, will result in a difference in final student weights.

Finally, oversampling of schools in particular strata occurs is a common occurrence, but remains invisible to the users of published databases.

4. Conclusions and recommendations

This paper shows evidence of the importance of sampling weights in LSA and the consequences of not applying such weights when analysing the data. It should now be evident that using the weights when estimating population features from sample data is absolutely critical.

In general terms, the concepts and procedures introduced in this paper hold for any study that attempts to draw inferences from sample data on populations – the concepts are not limited to educational studies, but can be translated to any social survey. Determining an appropriate sample design for such studies constitutes a crucial step in the study-planning phase. Standards for response rates and strategies on how to handle non-response also need to be defined at an early stage. While selecting the samples, researchers need to track the selection probabilities at all sampling stages as well as non-response. With this information in hand, it will be possible to derive appropriate sampling weights in order to account for complex sampling designs and achieve unbiased population estimates.

---

[1] TIMSS and PIRLS (Martin & Mullis, 2013), TIMSS-Advanced (Arora et al., 2009), CIVED (Schulz & Sibberns, 2011), SITES (Carstens et al., 2009)

[2] PISA (OECD, 2012), TALIS (OECD, 2010), PIACC (OECD 2013)

[3] For example, NAEP in the United States (Gorman, 1994) or the German National Educational Standards (Stanat et al, 2012).

[4] It should be noted that partial non-response (i.e., where items of a questionnaire remain unanswered) is usually not accounted for by non-response adjustments; it leads to missing data in the public-use data files and can be treated, for example, by imputation techniques (see, for example, Shin, 2014).

[5] Too small in terms of the number of respondents who carry the weight of the non-respondents.

[6] It should be noted that estimation weights in multilevel models are derived differently; please see, for example, Rutkowski et al. (2010).

[7] I.e., the sample does not reflect the true proportions of the population.

[8] For example, the Russian Federation, in most assessments, first samples by region, then school, then class.

[9] with an even number of schools

[10] For the sake of simplicity, we assume 100% participation of all sampled units for this example.

[11] Even though data is also collected from parents and teachers in TIMSS and PIRLS, they do not

constitute a separate target population; their data is directly linked to the student data and thus considered a feature of the surveyed students.

[12] Such schools are expected to have a high variance between classes.

References

1. Arora, A., Foy, P., Mullis, I. V. S., Martin, M.O. (2009). TIMSS Advanced 2008 Technical Report. Amsterdam: International Association for the Evaluation of Educational Achievement.
2. Brewer K.R.W. and M. Hanif. 1983. Sampling with Unequal Probabilities. New York: Springer.
3. Carstens, R., Ed, Pelgrum, W. J. (eds) (2009). Second Information Technology in Education Study: SITES 2006 Technical Report. Amsterdam: International Association for the Evaluation of Educational Achievement.
4. Cochran, W.G. 1977. Sampling Techniques. New York: John Wiley and Sons,.
5. Fraillon, J., Schulz, W., Ainley, J. (2013). International Computer and Information Literacy Study Assessment Framework. Amsterdam: International Association for the Evaluation of Educational Achievement.
6. Franklin S, Walker C (eds) (2003) Survey methods and practices. Statistics Canada. Social Survey Methods Division.
7. Lohr, S. L. (1999). Sampling: Design and analysis. Pacific Grove, CA: Duxbury Press.
8. Gorman, S. (1994). The 1992 NAEP Technical Report for the National Assessment. U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics.
9. Gregory, K. D., Martin, M. O., Wagemaker, P. (2001). Technical standards for IEA studies: an annotated bibliography. Amsterdam: International Association for the Evaluation of Educational Achievement.
10. Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E., Tourangeau, R. (2004). Survey Methodology, Wiley.
11. Martin, M. O., Mullis, I. V. S. (eds) (2013). Methods and procedures in TIMSS and PIRLS 2011. Chestnut Hill, MA: Lynch School of Education, Boston College.
12. Martin, M. O., Mullis, I. V. S., Foy, P., Stanco, G. M. (2012). TIMSS 2011 International Results in Science. Chestnut Hill, MA: Lynch School of Education, Boston College.
13. Mullis, I. V. S., Martin, M.O., Foy, P., Drucker, K. (2012). PIRLS 2011 International results in reading. Chestnut Hill, MA: Lynch School of Education, Boston College.
14. OECD (2010). TALIS 2008 Technical Report. TALIS, OECD Publishing.
15. OECD (2012). PISA 2009 Technical Report. PISA, OECD Publishing.
16. OECD (2013), OECD Skills Outlook 2013: First Results from the Survey of Adult Skills, OECD Publishing.
17. OECD (2014). PISA 2012 Results: What Students Know and Can Do (Volume I, Revised edition, February 2014) – Student Performance in Mathematics, Reading and Science. Retrieved July 10, 2014 from http://www.oecd-ilibrary.org/education/pisa-2012-results-what-students-know-and-can-do-volume-i-revised-edition-february-2014_9789264208780-en
18. Rutkowski, L., Gonzalez, E., Joncas, M., von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. Educational Researcher, 39(2), 142-151.
19. Rust, K. (2014): Sampling, weighting, and variance estimation in international large-scale assessments. In: Rutkowski, L., von Davier, M., Rutkowski D. (eds) (2014). Handbook of international large-scale assessment. New York: CRC Press.

20. Schulz, W., Sibberns, H. (eds) (2004). IEA Civic Education Study Technical Report. Amsterdam: The International Association for the Evaluation of Educational Achievement.
21. Schulz, W., Ainley, J., Fraillon, J. (eds) (2011). ICCS 2009 Technical Report. Amsterdam: The International Association for the Evaluation of Educational Achievement.
22. Shin, Y. (2014): Efficient handling of predictors and outcomes having missing values. In: Rutkowski, L., Von Davier, M., Rutkowski D. (eds) (2014). Handbook of international large-scale assessment. New York: CRC Press.
23. Stanat, P., Pant, H.A., Böhme, K., Richter, D. (2012). Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik – Ergebnisse des IQB-Ländervergleichs 2011. Münster: Waxmann
24. Tatto, M. T., Bankov, K. (eds) (2013). The teacher education and development study in mathematics (TEDS-M): policy, practice, and readiness to teach primary and secondary mathematics in 17 countries : Technical report. Amsterdam: International Association for the Evaluation of Educational Achievement.