2

Exploratory Analysis Procedures

INTRODUCTION

PISA surveys use complex methodologies that condition the way data should be analysed. As this is not yet included in standard procedures included in the statistical software packages such as SAS® or SPSS®, this manual describes the methodologies in detail and also presents syntax and macros developed specially for analysing the PISA data.

First of all, PISA does not draw simple random samples of students from exhaustive lists of 15-year-olds. The sampling design applied in PISA, its rationale and its consequences on how data should be analysed are mainly presented in Chapters 3 and 4. Briefly, PISA usually samples students in two stages: schools are first sampled and then students are sampled in the participating schools. Such sampling design increases the standard errors of any population estimates. As most of the statistical packages assume the data were collected on a simple random sample, analysing the PISA data with such software would systematically underestimate the standard errors and therefore lead to reporting non-significant results as significant. This would jeopardise the credibility of the programme.

Secondly, PISA uses imputation methods, denoted plausible values, for reporting student performance. From a theoretical point of view, any analysis that involves student performance estimates should be analysed five times and results should be aggregated to obtain: *(i)* the final estimate; and *(ii)* the imputation error that will be combined with the sampling error in order to reflect the test unreliability on the standard error. The detailed description of plausible values and its use are presented in Chapters 6 and 8.

All results published in the OECD initial and thematic reports have been computed accordingly to these methodologies, which means that the reporting of a country mean estimate and its respective standard error requires the computation of 405 means as described in detail in the next sections.

This chapter discusses the importance and usefulness of applying these recommended procedures, depending on the circumstances and on the stage of the data analysis process. Alternatives that shorten the procedures will be also presented, as well as the potential bias associated with such shortcuts.

The chapter is structured according to the three methodological issues that affect the way data should be analysed:

- weights,
- replicates for computing the standard errors,
- plausible values.

WEIGHTS

Weights are associated to each student and to each school because:

- students and schools in a particular country did not necessarily have the same probability of selection;
- differential participation rates according to certain types of school or student characteristics required various non-response adjustments;
- some explicit strata were over-sampled for national reporting purposes.

Weighting data is a straightforward process in SAS®. Most of the SAS statistical procedures include a WEIGHT statement. Box 2.1 presents the weight statement in the proc means procedure, while w_fstuwt is the variable name of the student final weights.

Box 2.1 **WEIGHT statement in the proc means procedure**

```
proc means data=temp1;
var pv1scie;
weight w_fstuwt;
run;
```

The syntax of Box 2.1 will provide unbiased estimates of some statistics such as mean and percentile. However, it will return biased estimates of the variance and consequently all related statistics such as standard deviation and standard error. For example, in order to compute weighted variance, SAS® firstly computes a weighted sum of square according to the following formulae:

$$SS = \sum_{i=1}^{n} w_i (x_i - \bar{x})^2$$

with w_i the weight for student i, X_i the value of student i for variable X and $\bar{X}$ the weighted mean estimate of variable X.

Then, as the default setting in SAS®, the weighted sum of square is divided by degree of freedom, which is equal to *N-1* for the variance. Consequently, the results largely overestimate the variance and its related statistics.

To overcome this problem, VARDEF must be specified in the proc means procedure. It indicates the divisor that will be used in the computation of the variance. Four divisors are available:

- *N*, *i.e.* the number of valid observations;
- DF for Degree of Freedom, which is equal to *N-1* for the variance;
- WGT, *i.e.* the sum of the weights for the valid observations;
- WDF for the Weighted Degree of Freedom, which corresponds to the sum of the weights minus 1.

As the default divisor is DF, the sum of squares, with weighted or without weighted, will be divided by *N-1* when the VARDEF option is not included. The DF divisor will largely overestimate the variance and its related statistics. Realistic estimates of the variance can be obtained with the WGT or WDF divisors by adding VARDEF=WGT or VARDEF=WDF.

It is, however, worth noting that SAS® does not compute standard errors with these two divisors of WGT and WDF. This is not an issue for computing the final estimates for reporting, since replicates are used for computing standard errors, as described in the following section. But, when analysts are interested in computing rough estimates of standard errors for provisional exploratory analysis, this becomes an issue. One way of obtaining realistic rough estimates of a standard error,[1] without using replicates, is to normalise the weights. The weight included in the database should be multiplied by a ratio of the number of observations to the sum of the weights. In other words, the weights should be multiplied by the total number of students and divided by the weighted total number of students. This linear transformation will ensure that the sum of the weights is equal to the number of observations. In this context, the VARDEF option does not need to be specified in SAS®.

Can analyses be conducted without weighting the data? Figure 2.1 represents the unweighted and weighted mean proficiency estimates in science for OECD countries in PISA 2006. In most countries, the difference is negligible. However, for some countries, the difference is quite substantial. Large differences between weighted and unweighted means usually result from over-sampling some strata in the population for national reporting purposes.

Figure 2.1
Science mean performance in OECD countries (PISA 2006)

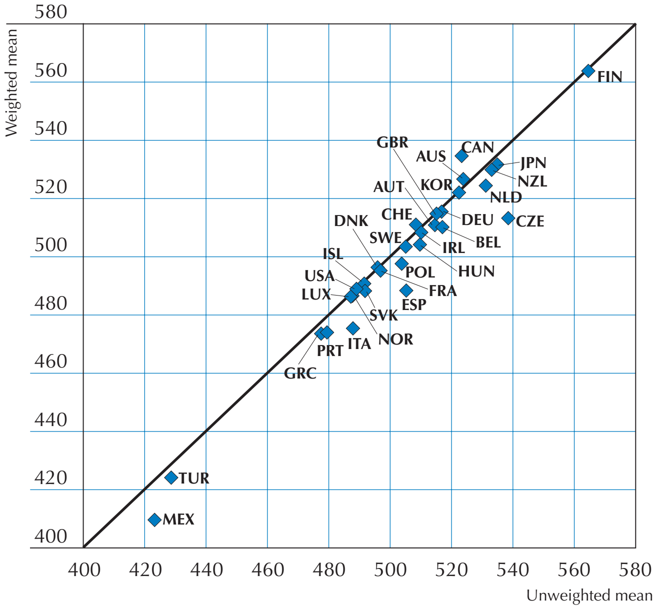
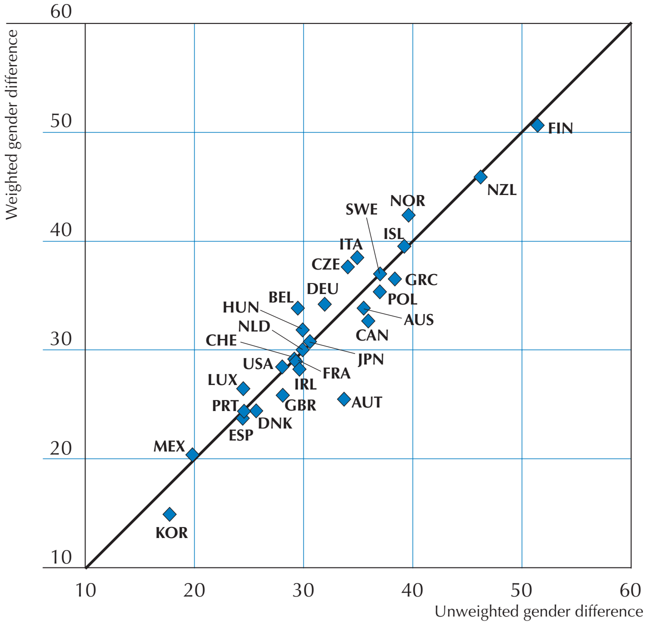


Figure 2.2
Gender differences in reading in OECD countries (PISA 2000)



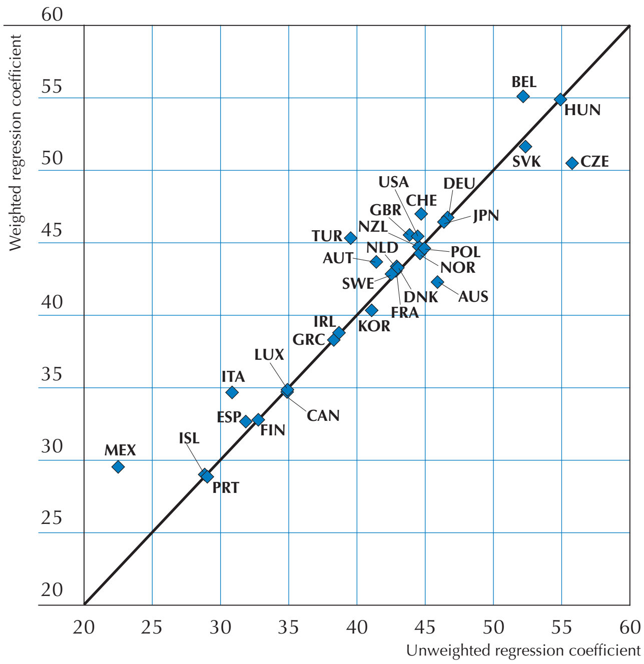
PISA DATA ANALYSIS MANUAL: SAS® SECOND EDITION – ISBN 978-92-64-05624-4 –

Figure 2.2 compares the unweighted and weighted gender differences in reading in PISA 2000. In most countries, the difference is negligible, but in Austria, for instance, the unweighted and weighted gender differences are equal to 33.5 and 25.4 respectively. Uneven distribution of males and females per type of schools (vocational versus academic) and differential participation rates per type of schools might explain such gaps between unweighted and weighted gender differences.

Finally, Figure 2.3 presents the unweighted and weighted regression coefficient of student socio-economic background (ESCS) on mathematic performance in PISA 2003. As shown by the figure, differences between unweighted and weighted coefficient are sometimes not negligible.

Figure 2.3

Regression coefficient of ESCS on mathematic performance in OECD countries (PISA 2003)



These three examples clearly demonstrate the impact of the weights on population parameter estimates. The bias of unweighted estimates could be substantial.

In conclusion, the weighting process does not make the analysis procedures more complex and guarantees that population estimates will be unbiased. Analyses should therefore always be weighted, at any stage of the process, whether it is the provisional exploration of the data or the final analyses before reporting.

REPLICATES FOR COMPUTING THE STANDARD ERROR

PISA applies two-stage sampling instead of simple random sampling. Chapter 3 describes the sampling design of the PISA surveys in detail and why such a design is implemented. This section, however, briefly describes the differences between these two sampling designs in order to provide rationale for using replicate weights. As previously indicated , statistical packages such as SAS® or SPSS® make the assumption that data are collected on a simple random sample of individuals.

One of the differences between simple random sampling and two-stage sampling is that for the latter, selected students attending the same school cannot be considered as independent observations. This is because students within a school usually have more common characteristics than students from different schools. For instance, they would have access to the same school resources, have the same teachers, be taught a common curriculum, and so on. Differences between students from different schools are also greater if different educational programmes are not available in all schools. For example, it would be expected that differences between students from a vocational school and students from an academic school would be bigger than differences between students from two vocational schools.

Furthermore, it is likely that within a country, within subnational entities, and within cities, people tend to live in areas according to their financial resources. As most children tend to attend schools close to their homes, it is assumed that students attending the same school come from similar socio-economic backgrounds.

A simple random sample of 4 000 students is therefore likely to cover the diversity of the population better than a sample of 100 schools with 40 students observed within each school. It follows that the uncertainty associated with any population parameter estimate (*i.e.* standard error) will be greater for a two-stage sample than for a simple random sample of the same size.

Reporting accurate and unbiased standard error estimates is of prime importance, since these estimates could be used for reporting differences that are statistically significant between countries or within countries. Reporting gender differences, for example, might lead to educational reforms aimed to reduce the gap between males and females. It is therefore essential to assure that these differences are indeed statistically significant.

Earlier student assessment surveys used to increase the simple random sample standard errors by the design effect (usually denoted in the statistical literature as DEFF) were roughly estimated on a few key variables for some population estimators, such as means, correlation and regression coefficients. For instance, in the First International Mathematics Study (FIMS) (Husen, 1967):

> "four subsamples of each subpopulation were obtained – this meant that instead of having only one sample representing a population, there were four. The purpose of doing this was twofold: *(i)* the standard errors of sampling could be obtained from the comparison of subsamples and, *(ii)* the answer sheets for each subsample could be shipped separately; thus if one was lost, three still remained."[2]

The International Association for the Evaluation of Educational Achievement (IEA) Six Subject Survey extended the FIMS procedure for the estimation of standard errors by integrating the scientific development of John Tukey on the Jackknife replication method. The whole sample for each country was divided into ten subsamples, following the sampling structure, and then ten complementary samples were obtained by leaving out, from the whole sample, each subsample in turn. Population estimates were then computed on each complementary subsample. The variability of these population estimates was used to estimate the standard errors and their respective design effect. The comparison between these design effects and their respective theoretical design effects based on the school variance and the average within school sample size showed quite consistent results, which allowed using the theoretical design effect.

As noted by Peaker (1975), "this evidence was combined with the evidence from the Mathematics Study in 1967, and suggested that appropriate values of DEFF were 2.4 for criterion means, 1.6 for correlations and 1.4 for regression coefficients."

In the late 1980s, the power of computers allowed the systematic use of replication methods. Standard errors were estimated for the Second International Science Study (SISS) by the Jackknife method for unstratified sample which consists of creating as many complementary samples as the number of schools in the whole sample. Each complementary sample was created by dropping one school at a time. The IEA Reading Literacy Study also used this replication method as well.

This manual presents how these replicates are computed in detail (Chapter 4) and how to estimate a standard error with these replicates (Chapter 7).

This section discusses the consequences of not using the replicates for estimating the standard errors and the appropriateness of using them in all phases of the data analysis process.

The PISA Technical Reports (OECD, 2002c, 2005, 2009) describe the sampling design effects for the performance country mean estimates in the chapter devoted to the sampling outcomes. Mathematically, the design effect corresponds to the ratio between the unbiased estimate of the sampling variance for a particular parameter and the sampling variance for that parameter if the observed sample was considered as a simple random sample. In PISA 2000, the sampling design effect for the country mean estimate on the combined reading literacy scale ranged from 2.32 to 19.92. This means that the actual standard error is from 1.5 to 4.4 times larger than the simple random sample standard error. In PISA 2003 and PISA 2006, countries requesting an adjudication of their data at a subnational level had to over-sample. The sampling design was, therefore, less effective and the design effect was higher. For instance, the design effect for the country performance mean estimate of Mexico was higher than 50 in PISA 2003.

Table 2.1 presents the type I error depending on the design effect. For instance, with a design effect of 4, a data analyst using the standard error returned by statistical packages assuming simple random sample will be working with the type I error of 0.33. As 0.01, 0.05 or 0.1 are normally used for the criteria of the significance level, this is a very important difference. Let us suppose an analysis estimates gender difference in science performance. When the gender difference is significantly different from 0 at the significance level of 0.33, the analysis has a 33% chance of being wrong in saying that there is a significant gender difference.

Table 2.1
Design effect and type I errors

Design effect (coefficient of increase)	Type I error	Design effect (coefficient of increase)	Type I error
1.5	0.11	11.0	0.55
2.0	0.17	11.5	0.56
2.5	0.22	12.0	0.57
3.0	0.26	12.5	0.58
3.5	0.29	13.0	0.59
4.0	0.33	13.5	0.59
4.5	0.36	14.0	0.60
5.0	0.38	14.5	0.61
5.5	0.40	15.0	0.61
6.0	0.42	15.5	0.62
6.5	0.44	16.0	0.62
7.0	0.46	16.5	0.63
7.5	0.47	17.0	0.63
8.0	0.49	17.5	0.64
8.5	0.50	18.0	0.64
9.0	0.51	18.5	0.65
9.5	0.52	19.0	0.65
10.0	0.54	19.5	0.66
10.5	0.55	20.0	0.66

The design effect varies from one country to another, but it also varies from one variable to another within a particular country. Figure 2.4 compares the design effect on the country mean estimates for the science performance and for the student socio-economic background (ESCS) in PISA 2006. The design effect on the mean estimate for the student socio-economic background is usually smaller than the design effect for science performance, since grouping students into different schools is usually based on their academic performance and, to a lesser extent, based on student socio-economic background.

Figure 2.4

Design effect on the country mean estimates for science performance and for ESCS in OECD countries (PISA 2006)

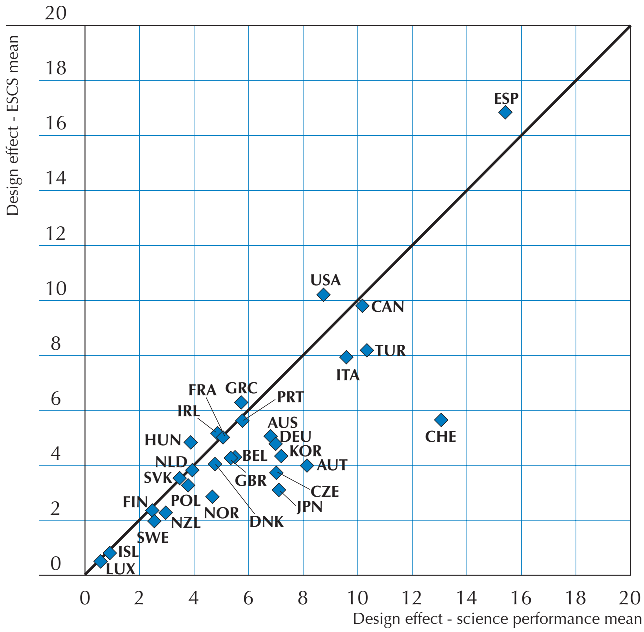


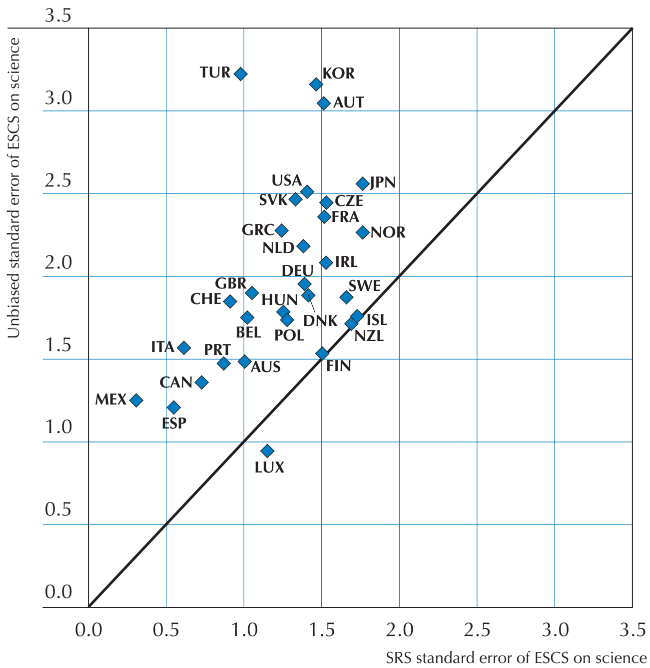
Figure 2.5 compares two different types of standard errors of the regression coefficient of ESCS on science performance: one is computed just as simple random sample (SRS) and the other is computed with replicates (unbiased). In Figure 2.5, the following can be observed:

- For most countries unbiased standard errors are bigger than SRS standard errors (*i.e.* dots are above the diagonal line),[3] but unbiased standard errors are not twice as big as SRS standard errors. This means that design effects are not as big as two in most countries. This result, therefore, supports the notion that design effects for regression coefficients (Figure 2.5) are smaller than design effects for mean estimates (Figure 2.4), as already noted by Peaker (1975).

- No specific patterns between SRS and unbiased standard errors are observed in Figure 2.5. This means that the design effect for regression coefficients varies from one country to another.

As illustrated by these few examples, the design effect depends on: *(i)* the population parameter that needs to be estimated; *(ii)* the sampling design of the country; *(iii)* the variables involved in the analyses (in particular the importance of the between-school variance relative to the within-school variance). Therefore, it would be inappropriate to suggest a single design effect for a particular parameter to be used for all countries to obtain a rough estimate of the actual standard error, based on the simple random sample standard error – especially given the increasing number of countries implementing a study design for regional adjudications and the large number of countries implementing international or national options.

Figure 2.5
Simple random sample and unbiased standard errors of ESCS on science performance in OECD countries (PISA 2006)



In sum, the results that will be reported have to be computed according to the recommended procedures, *i.e.* standard errors have to be estimated by using the replicates. During the exploratory phase, analysts might skip the replicate computations to save time. Instead, analysts could use the normalised weights and apply design effects. But, it is advised not to wait until the last stage of the process to compute unbiased estimates of the standard errors. Indeed, it might change a major outcome that would require rewriting some section of the reports. It is also important to note that analysis with the PISA data for only one country might inflate the standard error by using some fixed design effect values. This would require starting by estimating sensitive values of design effects for parameters such as mean, correlation, regression coefficient and so on. With a little practice, the procedures developed for analysing PISA data are not a constraint anymore. Moreover, with standard computers, these procedures do not take more than a couple of minutes.

PLAUSIBLE VALUES

This section briefly presents the rationale for using plausible values. The detailed description of plausible values and its use are presented in Chapters 6 and 8.

Since the Third International Mathematics and Science Survey conducted by the IEA in 1995, student proficiency estimates are returned through *plausible values*.

"The simplest way to describe plausible values is to say that plausible values are a representation of the range of abilities that a student might reasonably have. (…). Instead of directly estimating a student's ability θ, a probability distribution for a student's θ, is estimated. That is, instead of obtaining a point estimate for θ, (like a WLE[4]), a range of possible values for a student's θ, with an associated probability for each of these values is estimated. Plausible values are random draws from this (estimated) distribution for a student's θ." (Wu and Adams, 2002)

As will be described in Chapter 6, plausible values present several methodological advantages in comparison with classical Item Response Theory (IRT) estimates such as the Maximum Likelihood Estimates or Weighted Maximum Likelihood Estimates. Indeed, plausible values return unbiased estimates of:

- population performance parameters, such as mean, standard deviation or decomposition of the variance;
- percentages of students per proficiency level as they are on a continuous scale, unlike classical estimates which are on a non-continuous scale;
- bivariate or multivariate indices of relations between performance and background variables as this information is included in the psychometric model.

Usually, five plausible values are allocated to each student on each performance scale. Statistical analyses should be performed independently on each of these five plausible values and results should be aggregated to obtain the final estimates of the statistics and their respective standard errors. It is worth noting that these standard errors will consist of sampling uncertainty and test unreliability.

The plausible value methodology, combined with the replicates, requires that the parameter, such as a mean, a standard deviation, a percentage or a correlation, has to be computed 405 times (*i.e.* 5 plausible values by one student final weights and 80 replicates) to obtain the final estimate of the parameter and its standard error. Chapter 8 describes an unbiased shortcut that requires only 85 computations.

Working with one plausible value instead of five will provide unbiased estimate of population parameters but will not estimate the imputation error that reflects the influence of test unreliability for the parameter estimation. With a large dataset, this imputation error is relatively small. However, the smaller the sample size, the greater the imputation error.

Table 2.2 to Table 2.5 present the differences for four population parameters (*i.e.* mean, standard deviation, correlation and regression coefficient) between the estimates based on one plausible value and the same estimates based on five plausible values. These analyses were computed on the PISA 2006 science performance data in Belgium. Simple random samples of various sizes were selected. Each table shows:

- the estimated statistic based on one plausible value,
- the estimated standard error based on one plausible value,
- the estimated statistic based on five plausible values,
- the estimated standard error based on five plausible values,
- the sampling error based on five plausible values,
- the imputation error based on five plausible values.

With a sample size of 6 400 students, using one plausible value or five plausible values does not make any substantial difference in the two mean estimates (510.56 versus 510.79) as well as in the two standard error estimates (2.64 versus 2.69). In term of type I error, that would correspond to a shift from 0.050 to 0.052.

Table 2.2
Mean estimates and standard errors

Number of cases	Estimate on 1 PV	S.E. on 1 PV	Estimate on 5 PVs	S.E. on 5 PVs	Sampling error	Imputation error
25	500.05	19.47	493.87	21.16	20.57	4.55
50	510.66	17.70	511.48	16.93	16.76	2.18
100	524.63	12.25	518.00	12.42	11.70	3.81
200	509.78	7.52	509.46	7.79	7.56	1.72
400	507.91	6.34	508.31	6.52	6.46	0.86
800	507.92	4.55	508.69	4.58	4.50	0.79
1 600	506.52	3.54	507.25	3.44	3.39	0.52
3 200	511.03	2.77	511.48	2.76	2.70	0.49
6 400	510.56	2.64	510.79	2.69	2.67	0.23

Notes: PV = plausible value; S.E. = standard error.

Table 2.2 also illustrates how the imputation error increases as the sample size decreases. With a sample of 25 students, the imputation error is as big as the sampling error with a sample of 800 students. However, even if the imputation error is quite large with a sample of 25 students, working with one plausible value instead of five would correspond to a small shift in type I error from 0.05 to 0.072.

Under normal assumptions, the imputation error implies that the average, *i.e.* 493.87 for a sample of 25 students, can vary from 485 to 503. Using one plausible value instead of five for a very small sample may therefore have a considerable impact on the parameter estimates.

Table 2.3
Standard deviation estimates and standard errors

Number of cases	Estimate on 1 PV	S.E. on 1 PV	Estimate on 5 PVs	S.E. on 5 PVs	Sampling error	Imputation error
25	116.86	14.87	114.99	13.62	11.95	5.97
50	106.53	17.05	104.38	15.32	15.00	2.88
100	90.36	8.79	90.73	8.75	8.19	2.81
200	101.66	6.49	101.18	6.75	6.50	1.65
400	97.52	3.63	97.67	4.39	3.83	1.95
800	100.03	2.66	99.97	3.65	2.92	2.00
1 600	96.82	2.51	96.36	2.41	2.35	0.48
3 200	100.66	2.09	100.29	2.19	2.14	0.42
6 400	98.66	1.97	99.09	2.01	1.94	0.48

Notes: PV = plausible value; S.E. = standard error.

Table 2.4
Correlation estimates and standard errors

Number of cases	Estimate on 1 PV	S.E. on 1 PV	Estimate on 5 PVs	S.E. on 5 PVs	Sampling error	Imputation error
25	0.57	0.13	0.65	0.13	0.11	0.07
50	0.58	0.12	0.58	0.13	0.12	0.05
100	0.47	0.09	0.49	0.09	0.09	0.03
200	0.54	0.05	0.54	0.05	0.04	0.02
400	0.40	0.05	0.40	0.05	0.05	0.01
800	0.39	0.04	0.39	0.04	0.04	0.00
1 600	0.45	0.02	0.45	0.03	0.02	0.01
3 200	0.43	0.02	0.43	0.02	0.02	0.00
6 400	0.43	0.01	0.44	0.02	0.01	0.00

Notes: PV = plausible value; S.E. = standard error.

Table 2.5
ESCS regression coefficient estimates and standard errors

Number of cases	Estimate on 1 PV	S.E. on 1 PV	Estimate on 5 PVs	S.E. on 5 PVs	Sampling error	Imputation error
25	57.76	24.99	51.43	28.32	27.34	6.73
50	34.19	11.20	31.64	11.67	10.90	3.80
100	37.44	12.33	41.19	12.43	11.90	3.28
200	36.43	7.60	41.60	8.65	7.92	3.17
400	53.27	5.43	53.89	5.79	5.61	1.31
800	47.83	4.20	47.98	4.62	4.26	1.64
1 600	47.26	3.12	47.86	3.56	3.17	1.48
3 200	47.98	2.45	48.22	2.54	2.53	0.25
6 400	46.91	1.92	47.23	2.08	1.97	0.63

Notes: PV = plausible value; S.E. = standard error.

Similar conclusions can be drawn from the three tables above that refer respectively to standard deviation, correlation and ESCS regression coefficient.

CONCLUSION

This chapter briefly described the three methodological components of PISA that condition the data analysis process: weights, replicates and plausible values. It also discussed the consequences of not applying the recommended statistical procedures according to the data analysis phase.

In summary, the recommendations are:

- At any stage of the data analysis process, data should always be weighted. Unweighted data will return biased estimates. The importance of weighting the data is reinforced by the increasing number of countries that request a data adjudication at a subnational level, since such a request requires oversampling in almost all cases. As weighting data does not slow down the data analysis process and can easily be implemented in statistical packages, there is no valid reason for skipping this process.

- Use of replicates for estimating the standard error is certainly the methodological component that slows down the data analysis process the most. During the exploratory phase of the data, it is not of prime importance to estimate the standard error with the replicates. Standard errors returned by statistical software with normalised weight, and inflated by a rough estimate of the design effect, can provide the data analyst with an acceptable indication of the statistical significance of hypotheses. However, any results that will be published or communicated to the scientific community and to policy makers should be computed with replicates.

- Finally, using one plausible value or five plausible values does not really make a substantial difference on large samples. During the exploratory phase of the data, statistical analyses can be based on a single plausible value. It is, however, recommended to base the reported results on five plausible values, even on large samples. This will guarantee consistencies between results published by the OECD and results published in scientific journals or national reports. Further, results based on five plausible values are, from a theoretical point of view, incontestable.

Notes

1. This rough estimate of standard error is based on the assumption of a simple random sample.

2. In the IEA Six Subject Survey, a box containing answer sheets from Belgium fell out of a boat into the sea.

3. PISA in Luxembourg is not a sample survey but a census. SRS does not take into account the school stratification variables, while PISA does. Therefore, in Luxembourg, SRS standard errors are bigger than unbiased standard errors.

4. Weighted Likelihood Estimates.

References

Beaton, A.E. (1987), *The NAEP 1983-1984 Technical Report*, Educational Testing Service, Princeton.

Beaton, A.E., *et al.* (1996), *Mathematics Achievement in the Middle School Years, IEA's Third International Mathematics and Science Study*, Boston College, Chestnut Hill, MA.

Bloom, B.S. (1979), *Caractéristiques individuelles et apprentissage scolaire*, Éditions Labor, Brussels.

Bressoux, P. (2008), *Modélisation statistique appliquée aux sciences sociales*, De Boek, Brussels.

Bryk, A.S. and **S.W. Raudenbush** (1992), *Hierarchical Linear Models for Social and Behavioural Research: Applications and Data Analysis Methods*, Sage Publications, Newbury Park, CA.

Buchmann, C. (2000), *Family structure, parental perceptions and child labor in Kenya: What factors determine who is enrolled in school?* a*Soc. Forces,* No. 78, pp. 1349-79.

Cochran, W.G. (1977), *Sampling Techniques*, J. Wiley and Sons, Inc., New York.

Dunn, O.J. (1961), "Multilple Comparisons among Menas", *Journal of the American Statistical Association*, Vol. 56, American Statistical Association, Alexandria, pp. 52-64.

Kish, L. (1995), *Survey Sampling*, J. Wiley and Sons, Inc., New York.

Knighton, T. and **P. Bussière** (2006), "Educational Outcomes at Age 19 Associated with Reading Ability at Age 15", Statistics Canada, Ottawa.

Gonzalez, E. and **A. Kennedy** (2003), *PIRLS 2001 User Guide for the International Database*, Boston College, Chestnut Hill, MA.

Ganzeboom, H.B.G., P.M. De Graaf and **D.J. Treiman** (1992), "A Standard International Socio-economic Index of Occupation Status", *Social Science Research* 21(1), Elsevier Ltd, pp 1-56.

Goldstein, H. (1995), *Multilevel Statistical Models,* 2nd Edition, Edward Arnold, London.

Goldstein, H. (1997), "Methods in School Effectiveness Research", *School Effectiveness and School Improvement* 8, Swets and Zeitlinger, Lisse, Netherlands, pp. 369-395.

Hubin, J.P. (ed.) (2007), *Les indicateurs de l'enseignement*, 2nd Edition, Ministère de la Communauté française, Brussels.

Husen, T. (1967), *International Study of Achievement in Mathematics: A Comparison of Twelve Countries,* Almqvist and Wiksells, Uppsala.

International Labour Organisation (ILO) (1990), *International Standard Classification of Occupations: ISCO-88.* Geneva: International Labour Office.

Lafontaine, D. and **C. Monseur** (forthcoming), "Impact of Test Characteristics on Gender Equity Indicators in the Assessment of Reading Comprehension", *European Educational Research Journal*, Special Issue on PISA and Gender.

Lietz, P. (2006), "A Meta-Analysis of Gender Differences in Reading Achievement at the Secondary Level", *Studies in Educational Evaluation* 32, pp. 317-344.

Monseur, C. and **M. Crahay** (forthcoming), "Composition académique et sociale des établissements, efficacité et inégalités scolaires : une comparaison internationale – Analyse secondaire des données PISA 2006", *Revue française de pédagogie*.

OECD (1998), *Education at a Glance – OECD Indicators,* OECD, Paris.

OECD (1999a), *Measuring Student Knowledge and Skills – A New Framework for Assessment,* OECD, Paris.

OECD (1999b), *Classifying Educational Programmes – Manual for ISCED-97 Implementation in OECD Countries,* OECD, Paris.

OECD (2001), *Knowledge and Skills for Life – First Results from PISA 2000,* OECD, Paris.

OECD (2002a), *Programme for International Student Assessment – Manual for the PISA 2000 Database,* OECD, Paris.

OECD (2002b), *Sample Tasks from the PISA 2000 Assessment – Reading, Mathematical and Scientific Literacy,* OECD, Paris.

OECD (2002c), *Programme for International Student Assessment – PISA 2000 Technical Report,* OECD, Paris.

OECD (2002d), *Reading for Change: Performance and Engagement across Countries – Results from PISA 2000,* OECD, Paris.

OECD (2003a), *Literacy Skills for the World of Tomorrow – Further Results from PISA 2000,* OECD, Paris.

OECD (2003b), *The PISA 2003 Assessment Framework – Mathematics, Reading, Science and Problem Solving Knowledge and Skills,* OECD, Paris.

OECD (2004a), *Learning for Tomorrow's World – First Results from PISA 2003,* OECD, Paris.

OECD (2004b), *Problem Solving for Tomorrow's World – First Measures of Cross-Curricular Competencies from PISA 2003,* OECD, Paris.

OECD (2005a), *PISA 2003 Technical Report,* OECD, Paris.

OECD (2005b), *PISA 2003 Data Analysis Manual,* OECD, Paris.

OECD (2006), *Assessing Scientific, Reading and Mathematical Literacy: A Framework for PISA 2006,* OECD, Paris.

OECD (2007), *PISA 2006: Science Competencies for Tomorrow's World*, OECD, Paris.

OECD (2009), *PISA 2006 Technical Report,* OECD, Paris.

Peaker, G.F. (1975), *An Empirical Study of Education in Twenty-One Countries: A Technical report. International Studies in Evaluation VIII*, Wiley, New York and Almqvist and Wiksell, Stockholm.

Rust, K.F. and **J.N.K. Rao** (1996), "Variance Estimation for Complex Surveys Using Replication Techniques", *Statistical Methods in Medical Research,* Vol. 5, Hodder Arnold, London, pp. 283-310.

Rutter, M., *et al.* (2004), "Gender Differences in Reading Difficulties: Findings from Four Epidemiology Studies", *Journal of the American Medical Association* 291, pp. 2007-2012.

Schulz, W. (2006), *Measuring the socio-economic background of students and its effect on achievement in PISA 2000 and PISA 2003*, Paper presented at the Annual Meetings of the American Educational Research Association (AERA) in San Francisco, 7-11 April.

Wagemaker, H. (1996), *Are Girls Better Readers. Gender Differences in Reading Literacy in 32 Countries*, IEA, The Hague.

Warm, T.A. (1989), "Weighted Likelihood Estimation of Ability in Item Response Theory", *Psychometrika,* Vol. 54(3), Psychometric Society, Williamsburg, VA., pp. 427-450.

Wright, B.D. and **M.H. Stone** (1979), *Best Test Design: Rasch Measurement,* MESA Press, Chicago.

Table of contents

PISA DATA ANALYSIS MANUAL: SAS® SECOND EDITION – ISBN 978-92-64-05624-4 – © OECD 2009

LIST OF BOXES

LIST OF FIGURES

LIST OF TABLES

PISA DATA ANALYSIS MANUAL: SAS® SECOND EDITION – ISBN 978-92-64-05624-4 – © OECD 2009

PISA DATA ANALYSIS MANUAL: SAS® SECOND EDITION – ISBN 978-92-64-05624-4 – © OECD 2009

User's Guide

Preparation of data files

All data files (in text format) and the SAS® control files are available on the PISA website (*www.pisa.oecd.org*).

SAS® users

By running the SAS® control files, the PISA data files are created in the SAS® format. Before starting analysis, assigning the folder in which the data files are saved as a SAS® library.

For example, if the PISA 2000 data files are saved in the folder of "c:\pisa2000\data\", the PISA 2003 data files are in "c:\pisa2003\data\", and the PISA 2006 data files are in "c:\pisa2006\data\", the following commands need to be run to create SAS® libraries:

```
libname PISA2000 "c:\pisa2000\data\";
libname PISA2003 "c:\pisa2003\data\";
libname PISA2006 "c:\pisa2006\data\";
run;
```

SAS® syntax and macros

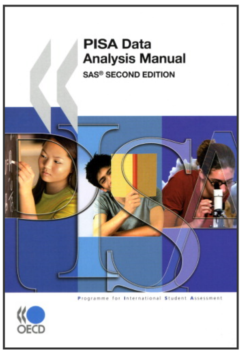
All syntaxes and macros in this manual can be copied from the PISA website (*www.pisa.oecd.org*). The 17 SAS® macros presented in Chapter 17 need to be saved under "c:\pisa\macro\", before staring analysis. Each chapter of the manual contains a complete set of syntaxes, which must be done sequentially, for all of them to run correctly, within the chapter.

Rounding of figures

In the tables and formulas, figures were rounded to a convenient number of decimal places, although calculations were always made with the full number of decimal places.

Country abbreviations used in this manual

AUS	Australia	FRA	France	MEX	Mexico
AUT	Austria	GBR	United Kingdom	NLD	Netherlands
BEL	Belgium	GRC	Greece	NOR	Norway
CAN	Canada	HUN	Hungary	NZL	New Zealand
CHE	Switzerland	IRL	Ireland	POL	Poland
CZE	Czech Republic	ISL	Iceland	PRT	Portugal
DEU	Germany	ITA	Italy	SVK	Slovak Republic
DNK	Denmark	JPN	Japan	SWE	Sweden
ESP	Spain	KOR	Korea	TUR	Turkey
FIN	Finland	LUX	Luxembourg	USA	United States



From:

PISA Data Analysis Manual: SAS, Second Edition

Access the complete publication at:

https://doi.org/10.1787/9789264056251-en

Please cite this chapter as:

OECD (2009), “Exploratory Analysis Procedures”, in *PISA Data Analysis Manual: SAS, Second Edition*, OECD Publishing, Paris.

DOI: https://doi.org/10.1787/9789264056251-3-en