



2

Exploratory Analysis Procedures

Introduction.....	36
Weights.....	36
Replicates for computing the standard error.....	39
Plausible values.....	43
Conclusion.....	46



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 pvlscie;
weight wfstuwt;
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)

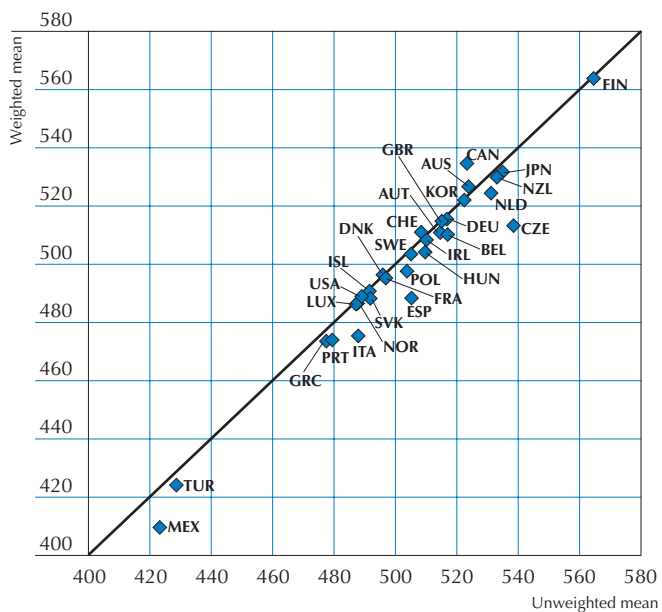


Figure 2.2
Gender differences in reading in OECD countries (PISA 2000)

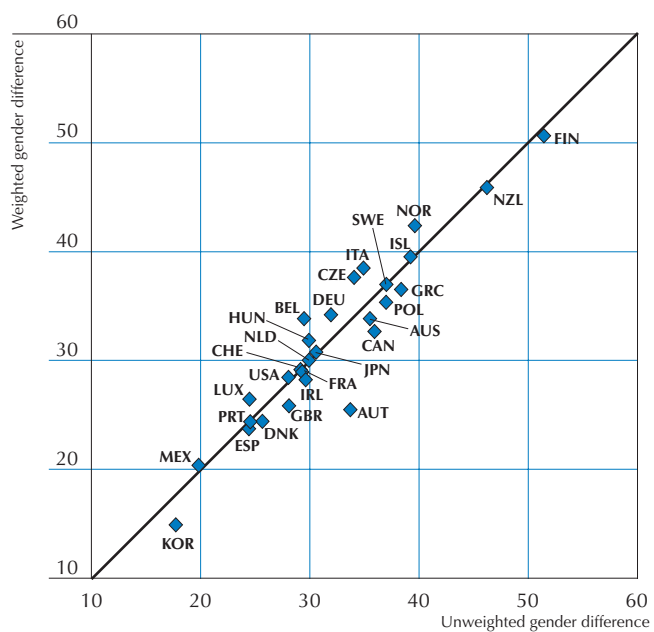


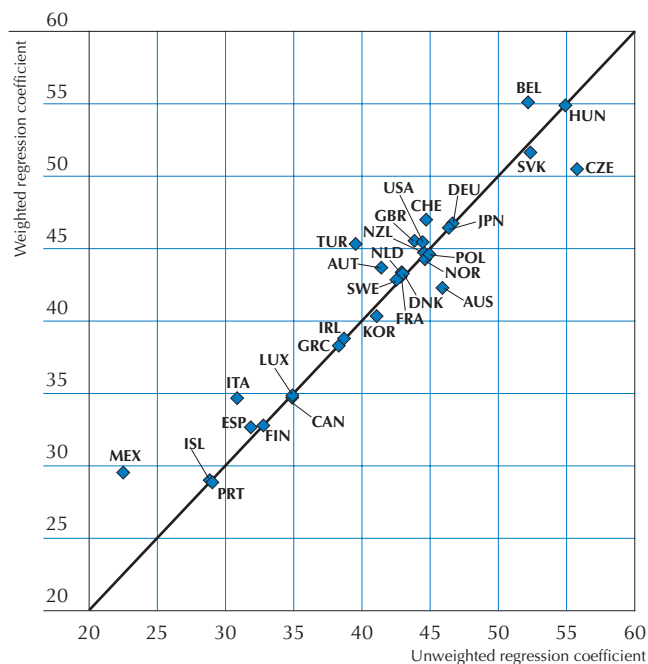


Figure 2.2 compares the unweighted and weighted gender differences in reading in PISA 2000. In most countries, the difference is negligible, but in Austria, for instance, the unweighted and weighted gender differences are equal to 33.5 and 25.4 respectively. Uneven distribution of males and females per 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)

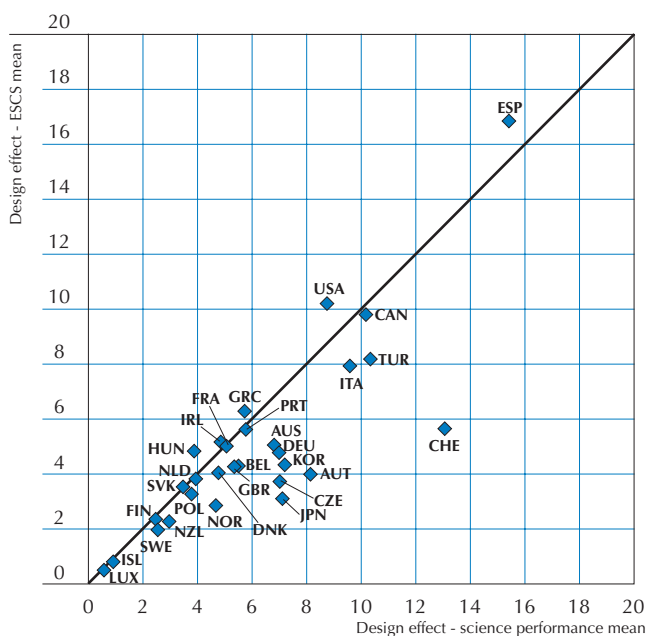


Figure 2.5 compares two different types of standard errors of the regression coefficient of ESCS on science performance: one is computed just as simple random sample (SRS) and the other is computed with replicates (unbiased). In Figure 2.5, the following can be observed:

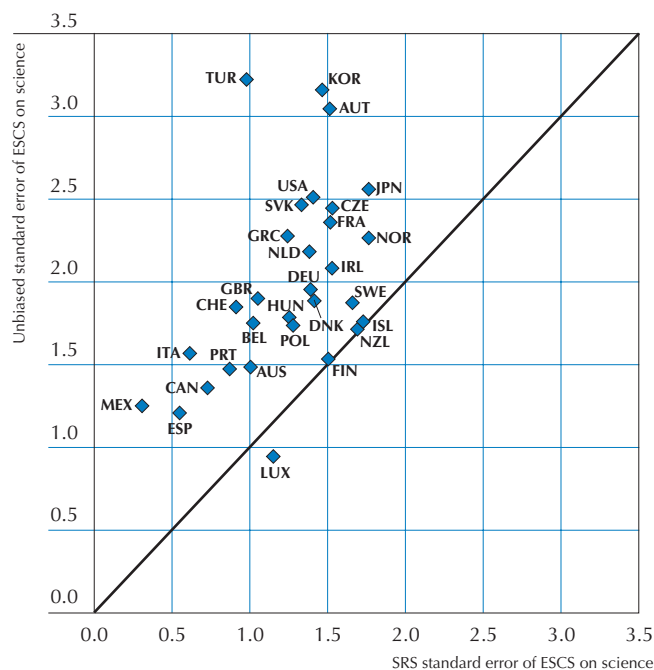
- For most countries unbiased standard errors are bigger than SRS standard errors (*i.e.* dots are above the diagonal line),³ but unbiased standard errors are not twice as big as SRS standard errors. This means that design effects are not as big as two in most countries. This result, therefore, supports 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 WLE⁴), 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 Boeck, 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?* *aSoc. Forces*, No. 78, pp. 1349-79.
- Cochran, W.G.** (1977), *Sampling Techniques*, J. Wiley and Sons, Inc., New York.
- Dunn, O.J.** (1961), "Multiple 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

FOREWORD	3
USER'S GUIDE	17
CHAPTER 1 THE USEFULNESS OF PISA DATA FOR POLICY MAKERS, RESEARCHERS AND EXPERTS ON METHODOLOGY	19
PISA – an overview	20
▪ The PISA surveys.....	20
How can PISA contribute to educational policy, practice and research?	22
▪ Key results from PISA 2000, PISA 2003 and PISA 2006.....	23
Further analyses of PISA datasets	25
▪ Contextual framework of PISA 2006.....	28
▪ Influence of the methodology on outcomes.....	31
CHAPTER 2 EXPLORATORY ANALYSIS PROCEDURES	35
Introduction	36
Weights	36
Replicates for computing the standard error	39
Plausible values	43
Conclusion	46
CHAPTER 3 SAMPLE WEIGHTS	49
Introduction	50
Weights for simple random samples	51
Sampling designs for education surveys	53
Why do the PISA weights vary?	57
Conclusion	58
CHAPTER 4 REPLICATE WEIGHTS	59
Introduction	60
Sampling variance for simple random sampling	60
Sampling variance for two-stage sampling	65
Replication methods for simple random samples	70
Replication methods for two-stage samples	72
▪ The Jackknife for unstratified two-stage sample designs.....	72
▪ The Jackknife for stratified two-stage sample designs.....	73
▪ The Balanced Repeated Replication method.....	74
Other procedures for accounting for clustered samples	76
Conclusion	76



CHAPTER 5 THE RASCH MODEL	79
Introduction	80
How can the information be summarised?	80
The Rasch Model for dichotomous items	81
▪ Introduction to the Rasch Model.....	81
▪ Item calibration.....	85
▪ Computation of a student's score.....	87
▪ Computation of a student's score for incomplete designs.....	91
▪ Optimal conditions for linking items.....	92
▪ Extension of the Rasch Model.....	93
Other item response theory models	94
Conclusion	94
 CHAPTER 6 PLAUSIBLE VALUES	 95
Individual estimates versus population estimates	96
The meaning of plausible values (PVs)	96
Comparison of the efficiency of WLEs, EAP estimates and PVs for the estimation of some population statistics	99
How to perform analyses with plausible values	102
Conclusion	103
 CHAPTER 7 COMPUTATION OF STANDARD ERRORS	 105
Introduction	106
The standard error on univariate statistics for numerical variables	106
The SAS® macro for computing the standard error on a mean	109
The standard error on percentages	112
The standard error on regression coefficients	115
The standard error on correlation coefficients	117
Conclusion	117
 CHAPTER 8 ANALYSES WITH PLAUSIBLE VALUES	 119
Introduction	120
Univariate statistics on plausible values	120
The standard error on percentages with PVs	123
The standard error on regression coefficients with PVs	123
The standard error on correlation coefficients with PVs	126
Correlation between two sets of plausible values	126
A fatal error shortcut	130
An unbiased shortcut	131
Conclusion	133
 CHAPTER 9 USE OF PROFICIENCY LEVELS	 135
Introduction	136
Generation of the proficiency levels	136
Other analyses with proficiency levels	141
Conclusion	143



CHAPTER 10 ANALYSES WITH SCHOOL-LEVEL VARIABLES	145
Introduction	146
Limits of the PISA school samples	147
Merging the school and student data files	148
Analyses of the school variables	148
Conclusion	150
CHAPTER 11 STANDARD ERROR ON A DIFFERENCE	151
Introduction	152
Statistical issues and computing standard errors on differences	152
The standard error on a difference without plausible values	154
The standard error on a difference with plausible values	159
Multiple comparisons	163
Conclusion	164
CHAPTER 12 OECD TOTAL AND OECD AVERAGE	167
Introduction	168
Recoding of the database to estimate the pooled OECD total and the pooled OECD average	170
Duplication of the data to avoid running the procedure three times	172
Comparisons between the pooled OECD total or pooled OECD average estimates and a country estimate	173
Comparisons between the arithmetic OECD total or arithmetic OECD average estimates and a country estimate	175
Conclusion	175
CHAPTER 13 TRENDS	177
Introduction	178
The computation of the standard error for trend indicators on variables other than performance	179
The computation of the standard error for trend indicators on performance variables	181
Conclusion	185
CHAPTER 14 STUDYING THE RELATIONSHIP BETWEEN STUDENT PERFORMANCE AND INDICES DERIVED FROM CONTEXTUAL QUESTIONNAIRES	187
Introduction	188
Analyses by quarters	188
The concept of relative risk	190
▪ Instability of the relative risk	191
▪ Computation of the relative risk	192
Effect size	195
Linear regression and residual analysis	197
▪ Independence of errors	197
Statistical procedure	200
Conclusion	201



CHAPTER 15 MULTILEVEL ANALYSES	203
Introduction	204
Two-level modelling with SAS®	206
▪ Decomposition of the variance in the empty model.....	206
▪ Models with only random intercepts.....	209
▪ Shrinkage factor.....	213
▪ Models with random intercepts and fixed slopes.....	213
▪ Models with random intercepts and random slopes.....	215
▪ Models with Level 2 independent variables.....	220
▪ Computation of final estimates and their respective standard errors.....	223
Three-level modelling	225
Limitations of the multilevel model in the PISA context	227
Conclusion	228
CHAPTER 16 PISA AND POLICY RELEVANCE – THREE EXAMPLES OF ANALYSES	231
Introduction	232
Example 1: Gender differences in performance	232
Example 2: Promoting socio-economic diversity within school?	236
Example 3: The influence of an educational system on the expected occupational status of students at age 30	242
Conclusion	246
CHAPTER 17 SAS® MACRO	247
Introduction	248
Structure of the SAS® Macro	248
REFERENCES	313
APPENDICES	315
Appendix 1 Three-level regression analysis.....	316
Appendix 2 PISA 2006 International database.....	324
Appendix 3 PISA 2006 Student questionnaire.....	333
Appendix 4 PISA 2006 Information communication technology (ICT) Questionnaire.....	342
Appendix 5 PISA 2006 School questionnaire.....	344
Appendix 6 PISA 2006 Parent questionnaire.....	351
Appendix 7 Codebook for PISA 2006 student questionnaire data file.....	355
Appendix 8 Codebook for PISA 2006 non-scored cognitive and embedded attitude items.....	399
Appendix 9 Codebook for PISA 2006 scored cognitive and embedded attitude items.....	419
Appendix 10 Codebook for PISA 2006 school questionnaire data file.....	431
Appendix 11 Codebook for PISA 2006 parents questionnaire data file.....	442
Appendix 12 PISA 2006 questionnaire indices.....	448



LIST OF BOXES

Box 2.1	WEIGHT statement in the proc means procedure	37
<hr/>		
Box 7.1	SAS® syntax for computing 81 means (e.g. PISA 2003).....	106
Box 7.2	SAS® syntax for computing the mean of HISEI and its standard error (e.g. PISA 2003).....	109
Box 7.3	SAS® syntax for computing the standard deviation of HISEI and its standard error by gender (e.g. PISA 2003).....	112
Box 7.4	SAS® syntax for computing the percentages and their standard errors for gender (e.g. PISA 2003).....	112
Box 7.5	SAS® syntax for computing the percentages and its standard errors for grades by gender (e.g. PISA 2003).....	114
Box 7.6	SAS® syntax for computing regression coefficients, R ² and its respective standard errors: Model 1 (e.g. PISA 2003).....	115
Box 7.7	SAS® syntax for computing regression coefficients, R ² and its respective standard errors: Model 2 (e.g. PISA 2003).....	116
Box 7.8	SAS® syntax for computing correlation coefficients and its standard errors (e.g. PISA 2003).....	117
<hr/>		
Box 8.1	SAS® syntax for computing the mean on the science scale by using the PROC_MEANS_NO_PV macro (e.g. PISA 2006).....	121
Box 8.2	SAS® syntax for computing the mean and its standard error on PVs (e.g. PISA 2006).....	122
Box 8.3	SAS® syntax for computing the standard deviation and its standard error on PVs by gender (e.g. PISA 2006).....	123
Box 8.4	SAS® syntax for computing regression coefficients and their standard errors on PVs by using the PROC_REG_NO_PV macro (e.g. PISA 2006).....	124
Box 8.5	SAS® syntax for running the simple linear regression macro with PVs (e.g. PISA 2006).....	125
Box 8.6	SAS® syntax for running the correlation macro with PVs (e.g. PISA 2006).....	126
Box 8.7	SAS® syntax for the computation of the correlation between mathematics/quantity and mathematics/space and shape by using the PROC_CORR_NO_PV macro (e.g. PISA 2003).....	129
<hr/>		
Box 9.1	SAS® syntax for generating the proficiency levels in science (e.g. PISA 2006).....	137
Box 9.2	SAS® syntax for computing the percentages of students by proficiency level in science and its standard errors by using the PROC_FREQ_NO_PV macro (e.g. PISA 2006).....	138
Box 9.3	SAS® syntax for computing the percentage of students by proficiency level in science and its standard errors by using the PROC_FREQ_PV macro (e.g. PISA 2006).....	140
Box 9.4	SAS® syntax for computing the percentage of students by proficiency level and its standard errors by gender (e.g. PISA 2006).....	140
Box 9.5	SAS® syntax for generating the proficiency levels in mathematics (e.g. PISA 2003).....	141
Box 9.6	SAS® syntax for computing the mean of self-efficacy in mathematics and its standard errors by proficiency level (e.g. PISA 2003).....	142
<hr/>		
Box 10.1	SAS® syntax for merging the student and school data files (e.g. PISA 2006).....	148
Box 10.2	Question on school location in PISA 2006.....	149
Box 10.3	SAS® syntax for computing the percentage of students and the average performance in science, by school location (e.g. PISA 2006).....	149
<hr/>		
Box 11.1	SAS® syntax for computing the mean of job expectations by gender (e.g. PISA 2003).....	154
Box 11.2	SAS® macro for computing standard errors on differences (e.g. PISA 2003).....	157



Box 11.3	Alternative SAS® macro for computing the standard error on a difference for a dichotomous variable (e.g. PISA 2003).....	158
Box 11.4	SAS® syntax for computing standard errors on differences which involve PVs (e.g. PISA 2003).....	160
Box 11.5	SAS® syntax for computing standard errors on differences that involve PVs (e.g. PISA 2006).....	162
<hr/>		
Box 12.1	SAS® syntax for computing the pooled OECD total for the mathematics performance by gender (e.g. PISA 2003).....	170
Box 12.2	SAS® syntax for the pooled OECD average for the mathematics performance by gender (e.g. PISA 2003).....	171
Box 12.3	SAS® syntax for the creation of a larger dataset that will allow the computation of the pooled OECD total and the pooled OECD average in one run (e.g. PISA 2003).....	172
<hr/>		
Box 14.1	SAS® syntax for the quarter analysis (e.g. PISA 2006).....	189
Box 14.2	SAS® syntax for computing the relative risk with five antecedent variables and five outcome variables (e.g. PISA 2006).....	193
Box 14.3	SAS® syntax for computing the relative risk with one antecedent variable and one outcome variable (e.g. PISA 2006).....	194
Box 14.4	SAS® syntax for computing the relative risk with one antecedent variable and five outcome variables (e.g. PISA 2006).....	194
Box 14.5	SAS® syntax for computing effect size (e.g. PISA 2006).....	196
Box 14.6	SAS® syntax for residual analyses (e.g. PISA 2003).....	200
<hr/>		
Box 15.1	Normalisation of the final student weights (e.g. PISA 2006).....	207
Box 15.2	SAS® syntax for the decomposition of the variance in student performance in science (e.g. PISA 2006).....	208
Box 15.3	SAS® syntax for normalising PISA 2006 final student weights with deletion of cases with missing values and syntax for variance decomposition (e.g. PISA 2006).....	211
Box 15.4	SAS® syntax for a multilevel regression model with random intercepts and fixed slopes (e.g. PISA 2006).....	214
Box 15.5	SAS® output for the multilevel model in Box 15.4.....	214
Box 15.6	SAS® syntax for a multilevel regression model (e.g. PISA 2006).....	216
Box 15.7	SAS® output for the multilevel model in Box 15.6.....	217
Box 15.8	SAS® output for the multilevel model with covariance between random parameters.....	218
Box 15.9	Interpretation of the within-school regression coefficient.....	220
Box 15.10	SAS® syntax for a multilevel regression model with a school-level variable (e.g. PISA 2006).....	221
Box 15.11	SAS® syntax for a multilevel regression model with interaction (e.g. PISA 2006).....	222
Box 15.12	SAS® output for the multilevel model in Box 15.11.....	222
Box 15.13	SAS® syntax for using the multilevel regression macro (e.g. PISA 2006).....	224
Box 15.14	SAS® syntax for normalising the weights for a three-level model (e.g. PISA 2006).....	226
<hr/>		
Box 16.1	SAS® syntax for testing the gender difference in standard deviations of reading performance (e.g. PISA 2000).....	233
Box 16.2	SAS® syntax for testing the gender difference in the 5th percentile of the reading performance (e.g. PISA 2006).....	235
Box 16.3	SAS® syntax for preparing a data file for the multilevel analysis.....	238



Box 16.4	SAS® syntax for running a preliminary multilevel analysis with one PV	239
Box 16.5	SAS® output for fixed parameters in the multilevel model.....	239
Box 16.6	SAS® syntax for running multilevel models with the PROC_MIXED_PV macro	242
<hr/>		
Box 17.1	SAS® macro of PROC_MEANS_NO_PV.sas.....	250
Box 17.2	SAS® macro of PROC_MEANS_PV.sas.....	253
Box 17.3	SAS® macro of PROC_FREQ_NO_PV.sas.....	256
Box 17.4	SAS® macro of PROC_FREQ_PV.sas.....	259
Box 17.5	SAS® macro of PROC_REG_NO_PV.sas.....	263
Box 17.6	SAS® macro of PROC_REG_PV.sas.....	266
Box 17.7	SAS® macro of PROC_CORR_NO_PV.sas.....	270
Box 17.8	SAS® macro of PROC_CORR_PV.sas.....	273
Box 17.9	SAS® macro of PROC_DIF_NO_PV.sas	276
Box 17.10	SAS® macro of PROC_DIF_PV.sas	279
Box 17.11	SAS® macro of QUARTILE_PV.sas	282
Box 17.12	SAS® macro of RELATIVE_RISK_NO_PV.sas.....	288
Box 17.13	SAS® macro of RELATIVE_RISK_PV.sas.....	291
Box 17.14	SAS® macro of EFFECT_SIZE_NO_PV.sas	296
Box 17.15	SAS® macro of EFFECT_SIZE_PV.sas.....	298
Box 17.16	SAS® macro of PROC_MIXED_NO_PV.sas.....	301
Box 17.17	SAS® macro of PROC_MIXED_PV.sas	306
<hr/>		
Box A1.1	Descriptive statistics of background and explanatory variables.....	318
Box A1.2	Background model for student performance.....	319
Box A1.3	Final net combined model for student performance.....	320
Box A1.4	Background model for the impact of socio-economic background.....	321
Box A1.5	Model of the impact of socio-economic background: “school resources” module	322
Box A1.6	Model of the impact of socio-economic background: “accountability practices” module	323
Box A1.7	Final combined model for the impact of socio-economic background.....	323
LIST OF FIGURES		
Figure 1.1	Relationship between social and academic segregations.....	27
Figure 1.2	Relationship between social segregation and the correlation between science performance and student HISEI	27
Figure 1.3	Conceptual grid of variable types.....	29
Figure 1.4	Two-dimensional matrix with examples of variables collected or available from other sources	30
<hr/>		
Figure 2.1	Science mean performance in OECD countries (PISA 2006).....	38
Figure 2.2	Gender differences in reading in OECD countries (PISA 2000).....	38
Figure 2.3	Regression coefficient of ESCS on mathematic performance in OECD countries (PISA 2003).....	39
Figure 2.4	Design effect on the country mean estimates for science performance and for ESCS in OECD countries (PISA 2006)	42
Figure 2.5	Simple random sample and unbiased standard errors of ESCS on science performance in OECD countries (PISA 2006)	43



Figure 4.1	Distribution of the results of 36 students.....	60
Figure 4.2	Sampling variance distribution of the mean.....	62
Figure 5.1	Probability of success for two high jumpers by height (dichotomous).....	82
Figure 5.2	Probability of success for two high jumpers by height (continuous).....	83
Figure 5.3	Probability of success to an item of difficulty zero as a function of student ability.....	83
Figure 5.4	Student score and item difficulty distributions on a Rasch continuum.....	86
Figure 5.5	Response pattern probabilities for the response pattern (1, 1, 0, 0).....	88
Figure 5.6	Response pattern probabilities for a raw score of 1.....	89
Figure 5.7	Response pattern probabilities for a raw score of 2.....	90
Figure 5.8	Response pattern probabilities for a raw score of 3.....	90
Figure 5.9	Response pattern likelihood for an easy test and a difficult test.....	91
Figure 5.10	Rasch item anchoring.....	92
Figure 6.1	Living room length expressed in integers.....	96
Figure 6.2	Real length per reported length.....	97
Figure 6.3	A posterior distribution on a test of six items.....	98
Figure 6.4	EAP estimators.....	99
Figure 8.1	A two-dimensional distribution.....	127
Figure 8.2	Axes for two-dimensional normal distributions.....	127
Figure 13.1	Trend indicators in PISA 2000, PISA 2003 and PISA 2006.....	179
Figure 14.1	Percentage of schools by three school groups (PISA 2003).....	198
Figure 15.1	Simple linear regression analysis versus multilevel regression analysis.....	205
Figure 15.2	Graphical representation of the between-school variance reduction.....	215
Figure 15.3	A random multilevel model.....	216
Figure 15.4	Change in the between-school residual variance for a fixed and a random model.....	218
Figure 16.1	Relationship between the segregation index of students' expected occupational status and the segregation index of student performance in reading (PISA 2000).....	244
Figure 16.2	Relationship between the segregation index of students' expected occupational status and the correlation between HISEI and students' expected occupational status.....	245

LIST OF TABLES

Table 1.1	Participating countries/economies in PISA 2000, PISA 2003, PISA 2006 and PISA 2009.....	21
Table 1.2	Assessment domains covered by PISA 2000, PISA 2003 and PISA 2006.....	22
Table 1.3	Correlation between social inequities and segregations at schools for OECD countries.....	28
Table 1.4	Distribution of students per grade and per ISCED level in OECD countries (PISA 2006).....	31
Table 2.1	Design effect and type I errors.....	41
Table 2.2	Mean estimates and standard errors.....	45



Table 2.3	Standard deviation estimates and standard errors.....	45
Table 2.4	Correlation estimates and standard errors.....	45
Table 2.5	ESCS regression coefficient estimates and standard errors.....	46
<hr/>		
Table 3.1	Height and weight of ten persons	52
Table 3.2	Weighted and unweighted standard deviation estimate	52
Table 3.3	School, within-school, and final probability of selection and corresponding weights for a two-stage, simple random sample with the first-stage units being schools of equal size.....	54
Table 3.4	School, within-school, and final probability of selection and corresponding weights for a two-stage, simple random sample with the first-stage units being schools of unequal size	54
Table 3.5	School, within-school, and final probability of selection and corresponding weights for a simple and random sample of schools of unequal size (smaller schools)	55
Table 3.6	School, within-school, and final probability of selection and corresponding weights for a simple and random sample of schools of unequal size (larger schools)	55
Table 3.7	School, within-school, and final probability of selection and corresponding weights for PPS sample of schools of unequal size	56
Table 3.8	Selection of schools according to a PPS and systematic procedure.....	57
<hr/>		
Table 4.1	Description of the 630 possible samples of 2 students selected from 36 students, according to their mean.....	61
Table 4.2	Distribution of all possible samples with a mean between 8.32 and 11.68.....	63
Table 4.3	Distribution of the mean of all possible samples of 4 students out of a population of 36 students.....	64
Table 4.4	Between-school and within-school variances on the mathematics scale in PISA 2003.....	67
Table 4.5	Current status of sampling errors.....	67
Table 4.6	Between-school and within-school variances, number of participating schools and students in Denmark and Germany in PISA 2003	68
Table 4.7	The Jackknives replicates and sample means.....	70
Table 4.8	Values on variables X and Y for a sample of ten students.....	71
Table 4.9	Regression coefficients for each replicate sample.....	71
Table 4.10	The Jackknife replicates for unstratified two-stage sample designs.....	72
Table 4.11	The Jackknife replicates for stratified two-stage sample designs.....	73
Table 4.12	Replicates with the Balanced Repeated Replication method.....	74
Table 4.13	The Fay replicates	75
<hr/>		
Table 5.1	Probability of success when student ability equals item difficulty.....	84
Table 5.2	Probability of success when student ability is less than the item difficulty by 1 unit.....	84
Table 5.3	Probability of success when student ability is greater than the item difficulty by 1 unit	84
Table 5.4	Probability of success when student ability is less than the item difficulty by 2 units	85
Table 5.5	Probability of success when student ability is greater than the item difficulty by 2 units.....	85
Table 5.6	Possible response pattern for a test of four items.....	87
Table 5.7	Probability for the response pattern (1, 1, 0, 0) for three student abilities.....	87
Table 5.8	Probability for the response pattern (1, 0) for two students of different ability in an incomplete test design.....	91
Table 5.9	PISA 2003 test design	93

Table 6.1	Structure of the simulated data.....	100
Table 6.2	Means and variances for the latent variables and the different student ability estimators.....	100
Table 6.3	Percentiles for the latent variables and the different student ability estimators.....	101
Table 6.4	Correlation between HISEI, gender and the latent variable, the different student ability estimators.....	101
Table 6.5	Between- and within-school variances.....	102
<hr/>		
Table 7.1	HISEI mean estimates.....	107
Table 7.2	Squared differences between replicate estimates and the final estimate.....	108
Table 7.3	Output data file exercise1 from Box 7.2.....	111
Table 7.4	Available statistics with the PROC_MEANS_NO_PV macro.....	111
Table 7.5	Output data file exercise2 from Box 7.3.....	112
Table 7.6	Output data file exercise3 from Box 7.4.....	112
Table 7.7	Percentage of girls for the final and replicate weights and squared differences.....	113
Table 7.8	Output data file exercise4 from Box 7.5.....	114
Table 7.9	Output data file exercise5 from Box 7.6.....	115
Table 7.10	Output data file exercise6 from Box 7.7.....	116
Table 7.11	Output data file exercise6_criteria from Box 7.7.....	117
Table 7.12	Output data file exercise7 from Box 7.8.....	117
<hr/>		
Table 8.1	The 405 mean estimates.....	120
Table 8.2	Mean estimates and their respective sampling variances on the science scale for Belgium (PISA 2006).....	121
Table 8.3	Output data file exercise6 from Box 8.2.....	123
Table 8.4	Output data file exercise7 from Box 8.3.....	123
Table 8.5	The 450 regression coefficient estimates.....	125
Table 8.6	HISEI regression coefficient estimates and their respective sampling variance on the science scale in Belgium after accounting for gender (PISA 2006).....	125
Table 8.7	Output data file exercise8 from Box 8.5.....	125
Table 8.8	Output data file exercise9 from Box 8.6.....	126
Table 8.9	Correlation between the five plausible values for each domain, mathematics/quantity and mathematics/space and shape.....	128
Table 8.10	The five correlation estimates between mathematics/quantity and mathematics/space and shape and their respective sampling variance.....	129
Table 8.11	Standard deviations for mathematics scale using the correct method (plausible values) and by averaging the plausible values at the student level (pseudo-EAP) (PISA 2003).....	131
Table 8.12	Unbiased shortcut for a population estimate and its standard error.....	132
Table 8.13	Standard errors from the full and shortcut computation (PISA 2006).....	132
<hr/>		
Table 9.1	The 405 percentage estimates for a particular proficiency level.....	138
Table 9.2	Estimates and sampling variances per proficiency level in science for Germany (PISA 2006).....	139
Table 9.3	Final estimates of the percentage of students, per proficiency level, in science and its standard errors for Germany (PISA 2006).....	139
Table 9.4	Output data file exercise6 from Box 9.3.....	140
Table 9.5	Output data file exercise7 from Box 9.4.....	140
Table 9.6	Mean estimates and standard errors for self-efficacy in mathematics per proficiency level (PISA 2003).....	143
Table 9.7	Output data file exercise8 from Box 9.6.....	143



Table 10.1	Percentage of students per grade and ISCED level, by country (PISA 2006)	146
Table 10.2	Output data file exercise1 from Box 10.3	150
Table 10.3	Output data file exercise2 from Box 10.3	150
Table 11.1	Output data file exercise1 from Box 11.1	155
Table 11.2	Mean estimates for the final and 80 replicate weights by gender (PISA 2003)	155
Table 11.3	Difference in estimates for the final weight and 80 replicate weights between females and males (PISA 2003)	157
Table 11.4	Output data file exercise2 from Box 11.2	158
Table 11.5	Output data file exercise3 from Box 11.3	159
Table 11.6	Gender difference estimates and their respective sampling variances on the mathematics scale (PISA 2003)	159
Table 11.7	Output data file exercise4 from Box 11.4	160
Table 11.8	Gender differences on the mathematics scale, unbiased standard errors and biased standard errors (PISA 2003)	161
Table 11.9	Gender differences in mean science performance and in standard deviation for science performance (PISA 2006)	161
Table 11.10	Regression coefficient of HISEI on the science performance for different models (PISA 2006)	163
Table 11.11	Cross tabulation of the different probabilities	163
Table 12.1	Regression coefficients of the index of instrumental motivation in mathematics on mathematic performance in OECD countries (PISA 2003)	169
Table 12.2	Output data file exercise1 from Box 12.1	170
Table 12.3	Output data file exercise2 from Box 12.2	171
Table 12.4	Difference between the country mean scores in mathematics and the OECD total and average (PISA 2003)	174
Table 13.1	Trend indicators between PISA 2000 and PISA 2003 for HISEI, by country	180
Table 13.2	Linking error estimates	182
Table 13.3	Mean performance in reading by gender in Germany	184
Table 14.1	Distribution of the questionnaire index of cultural possession at home in Luxembourg (PISA 2006)	188
Table 14.2	Output data file exercise1 from Box 14.1	190
Table 14.3	Labels used in a two-way table	190
Table 14.4	Distribution of 100 students by parents' marital status and grade repetition	191
Table 14.5	Probabilities by parents' marital status and grade repetition	191
Table 14.6	Relative risk for different cutpoints	191
Table 14.7	Output data file exercise2 from Box 14.2	193
Table 14.8	Mean and standard deviation for the student performance in reading by gender, gender difference and effect size (PISA 2006)	195
Table 14.9	Output data file exercise4 from Box 14.5	197
Table 14.10	Output data file exercise5 from Box 14.5	197
Table 14.11	Mean of the residuals in mathematics performance for the bottom and top quarters of the PISA index of economic, social and cultural status, by school group (PISA 2003)	199

Table 15.1	Between- and within-school variance estimates and intraclass correlation (PISA 2006).....	209
Table 15.2	Output data file “ranparm1” from Box 15.3.....	212
Table 15.3	Output data file “fixparm3” from Box 15.6.....	217
Table 15.4	Output data file “ranparm3” from Box 15.6.....	217
Table 15.5	Variance/covariance estimates before and after centering.....	219
Table 15.6	Output data file of the fixed parameters file.....	221
Table 15.7	Average performance and percentage of students by student immigrant status and by type of school.....	223
Table 15.8	Variables for the four groups of students.....	223
Table 15.9	Comparison of the regression coefficient estimates and their standard errors in Belgium (PISA 2006).....	224
Table 15.10	Comparison of the variance estimates and their respective standard errors in Belgium (PISA 2006).....	225
Table 15.11	Three-level regression analyses.....	226
Table 16.1	Differences between males and females in the standard deviation of student performance (PISA 2000).....	234
Table 16.2	Distribution of the gender differences (males – females) in the standard deviation of the student performance.....	234
Table 16.3	Gender difference on the PISA combined reading scale for the 5 th , 10 th , 90 th and 95 th percentiles (PISA 2000).....	235
Table 16.4	Gender difference in the standard deviation for the two different item format scales in reading (PISA 2000).....	236
Table 16.5	Random and fixed parameters in the multilevel model with student and school socio-economic background.....	237
Table 16.6	Random and fixed parameters in the multilevel model with socio-economic background and grade retention at the student and school levels.....	241
Table 16.7	Segregation indices and correlation coefficients by country (PISA 2000).....	243
Table 16.8	Segregation indices and correlation coefficients by country (PISA 2006).....	244
Table 16.9	Country correlations (PISA 2000).....	245
Table 16.10	Country correlations (PISA 2006).....	246
Table 17.1	Synthesis of the 17 SAS® macros.....	249
Table A2.1	Cluster rotation design used to form test booklets for PISA 2006.....	324
Table A12.1	Mapping of ISCED to accumulated years of education.....	449
Table A12.2	ISCO major group white-collar/blue-collar classification.....	451
Table A12.3	ISCO occupation categories classified as science-related occupations.....	451
Table A12.4	Household possessions and home background indices.....	455
Table A12.5	Factor loadings and internal consistency of ESCS 2006 in OECD countries.....	465
Table A12.6	Factor loadings and internal consistency of ESCS 2006 in partner countries/economies.....	466



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

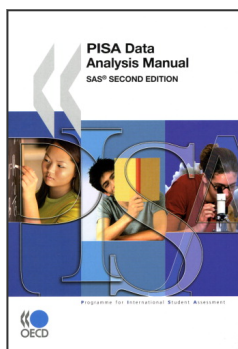
All syntaxes and macros in this manual can be copied from the PISA website (www.pisa.oecd.org). The 17 SAS® macros presented in Chapter 17 need to be saved under "c:\pisa\macro\", before starting analysis. Each chapter of the manual contains a complete set of syntaxes, which must be done sequentially,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

This work is published under the responsibility of the Secretary-General of the OECD. The opinions expressed and arguments employed herein do not necessarily reflect the official views of OECD member countries.

This document and any map included herein are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

You can copy, download or print OECD content for your own use, and you can include excerpts from OECD publications, databases and multimedia products in your own documents, presentations, blogs, websites and teaching materials, provided that suitable acknowledgment of OECD as source and copyright owner is given. All requests for public or commercial use and translation rights should be submitted to rights@oecd.org. Requests for permission to photocopy portions of this material for public or commercial use shall be addressed directly to the Copyright Clearance Center (CCC) at info@copyright.com or the Centre français d'exploitation du droit de copie (CFC) at contact@cfcopies.com.