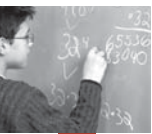**15**

Data Adjudication

INTRODUCTION

This chapter describes the process used to adjudicate the implementation of PISA 2003 in each of the participating countries, and gives the outcomes of the data adjudication. In particular, this chapter reviews the:

- Extent to which each country met PISA sampling standards;

- Outcomes of the national centre and PISA quality monitoring visits;

- Quality and completeness of the submitted data;

- Outcomes of the inter-country reliability study; and

- Outcomes of the translation verification process.

The standards for PISA 2003, which were formally presented to the National Project Managers (NPMs) at the Brussels NPM meeting in February 2001, were used as the basis for the adjudication. The latest version of the standards is available on the PISA Web site (*www.pisa.oecd.org*). The issues covered in those standards are:

- Sampling
- Translation and verification

> - Selection of translators

> - Submission of questionnaire adaptations and modifications for approval

> - Submission of material for translation and verification

- Test administration

> - Selection of test administrators

> - Training of test administrators

> - Security of material

> - Testing session

- Quality monitoring

> - Site visits and training of PISA quality monitors (PQM)

> - Visit by PISA quality monitors

- Coding

> - Single coding

> - Multiple coding

> - PISA international standard indicators [Inter-country-rater-reliability study]

- Data entry and submission

> - Materials submitted

> - Data cleaning

Implementing the standards – quality management

NPMs of countries and adjudicated regions were responsible for implementing the standards based on consortium advice as contained in the study's various operational manuals. During the implementation phase the consortium conducted two quality management activities. The first was quality control performed by consortium staff as they worked with NPMs to implement key parts of the project. As part of the quality control activities, consortium staff checked the work of NPMs and provided advice on rectifying action when required and before critical errors occurred. The second was quality monitoring, which involved the systematic collection of data that monitored the implementation of the standards. For data adjudication it was the information collected during both the quality control and quality monitoring activities that was used to determine the level of compliance to the standards.

Information available for adjudication

The information collected by consortium staff during their quality control activities included communications and documentation exchanged with NPMs. The information available from quality monitoring instruments included:

- PISA quality monitor reports (data collection sheets and general observations);
- Test administrator session reports;
- Main study reviews;
- Sampling forms;
- National centre quality monitor interviews; and
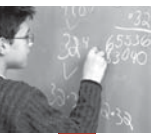- Data cleaning questionnaire.

Each of the quality monitoring instruments addressed different aspects of the standards and were collected at different times during the data collection phase. There are two types of PISA Quality Monitor (PQM) reports, one containing data for each observed session and another detailing the general observations of each quality monitor. The PQM reports contain data related to test administration as well as a record of interview with school co-ordinators. The test administrator session report is completed by each test administrator after each test session and also contains data related to test administration. The data from this report were data-entered by the national centre and submitted as part of the dataset to the consortium. The national centre quality interview schedule contains information on all the standards, as does the main study review. The data submission questionnaire contains information specific to the data and is mainly used for data cleaning purposes.

The national centre quality monitor interview schedule, main study review, and data submission questionnaire are self-declared by the NPM. The PQM data is collected independently of the NPM and can be viewed as being collected by a peer of the test administrator who is nominated by the NPM.

Data adjudication process

The main aim of the adjudication process is to make a single determination on adjudicated data in a manner that is transparent, based on evidence and which is defensible. The data adjudication process achieved this through the following steps:

- *Step 1:* Quality control and quality monitoring data were collected during the data collection phase.

- *Step 2:* Data from quality monitoring instruments were entered into a single quality management database.

- *Step 3:* Experts compiled country-by-country reports that contained quality monitoring data for expert areas.

- *Step 4:* Experts considered the quality monitoring data, along with their quality control information, in order to make a judgement. In this phase the experts collaborated with the project director and data manager to address any identified areas of concern. Where necessary, the relevant NPM was contacted through the project director. At the end of this phase each expert constructed, for each adjudicated-dataset, a summary detailing how the standards had been implemented.

- *Step 5:* The consortium reviewed the reports and made a determination with regard the quality of the data.

It was expected that the data adjudication would result in a range of possible recommendations. Some possible, foreseen recommendations included:

- That some data be removed for a particular country, for example the removal of data for some items, such as open-ended items, or the removal of data for some schools.

- That rectifying action be performed by the NPM, for example providing additional evidence to demonstrate that there is no non-response bias or rescoring open-ended items.

- That the data not be endorsed for use in certain types of analyses.

- That the data not be endorsed for inclusion in the PISA 2003 database.

Throughout the data collection phase, the consortium concentrated its quality control activities to ensure that the highest scientific standards were implemented. However during data adjudication a wider definition of quality was used especially when considering data that was at risk. In particular the underlying criteria used in adjudication was "fitness for use". That is, data was endorsed for use if it was deemed to be fit for meeting the intended purposes of PISA 2003.

### GENERAL OUTCOMES

Overview of response rate issues

The PISA school response rate requirements are discussed in Chapter 4. Figure 15.1 is a scatter plot of the attained PISA school response rates before and after replacements. Those countries that are plotted in the lighter shaded region were regarded as fully satisfying the PISA school response rate criterion.

Canada, United Kingdom, and the United States failed to meet the school response rate requirements. In addition to failing the school response rate requirement, the United Kingdom was the only participant to fail the student response rate requirement (see Table 12.4).

After reviewing the sampling outcomes, the consortium asked Canada, United Kingdom, and The United States, to provide additional data that would assist the consortium in making a balanced judgement about the threat of the non-response to the accuracy of inferences which could be made from the PISA data.
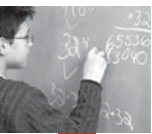
Figure 15.1 ■ **Attained school response rates**
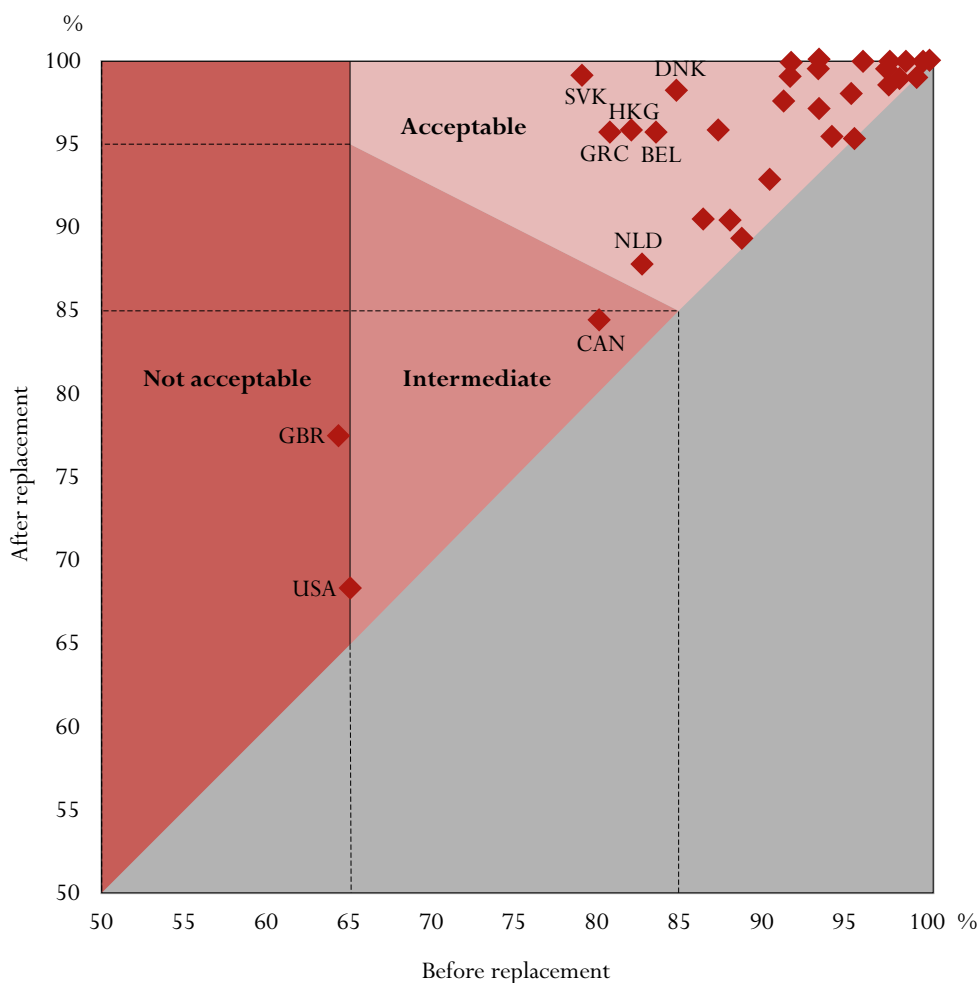
DETAILED COUNTRY COMMENTS

It is important to recognise that the PISA data adjudication is a late but not necessarily final step in a quality assurance process. By the time each country was adjudicated, quality assurance mechanisms (such as the sampling procedures documentation, translation verification, data cleaning and site visits) had identified a range of issues and ensured that they had been rectified at least in the majority of cases. Details on the various quality assurance procedures and their outcomes are documented elsewhere (see Chapter 7 and Appendix 9). Data adjudication focused on residual issues that remained after these quality assurance processes. There were not many such issues and their projected impact on the validity of the PISA results was deemed to be negligible. Unlike sampling issues, which under most circumstances could directly affect all of a country's data, the residual issues identified in other areas have an impact on only a small proportion of the data. For example, coding leniency or severity for a single item in reading has an effect on between just one-third and one half of 1 per cent of the reading data and even for that small fraction, the effect would be minor. Other breaches of standards identified in a small number of countries include

a failure to follow the specified multiple marker design and a failure to involve national committees in instrument development. Where the specified multiple coding design was not implemented, a sufficient level of quality assurance data was usually available to determine the quality of the manual coding.

Australia

Australia fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

Austria

Austria fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

Belgium

Belgium fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

Brazil

Brazil had a very low coverage of the 15-year-old population (54 per cent), due to low rates of enrolment, and that this should be taken into account when interpreting Brazilian data. Further, Brazil did not submit data for the scheduled inter-country coder reliability study and therefore it was not possible to implement the necessary quality assurance procedures for the manually coded items.

The Brazilian data was available for inclusion in the full range of PISA reports.

Canada

There were sampling-related concerns with the Canadian data. The overall exclusion rate of 6.83 per cent exceeded the PISA standard of 5 per cent. The majority of the exclusions (5.26 per cent) were within-school exclusions with large contributions from language-based exclusions and special needs students. The high overall exclusion rate was also contributed to by the exclusion of very small schools, that is, schools having only one or two eligible students. In addition there was also a high ineligible rate of 5.29 per cent, where the ineligibles were about evenly split between drop-outs and transferred students.

The Canadian school response rate, of 79.95 per cent before replacement and 84.38 per cent after all replacements, did not meet PISA standards. Much of Canada's non-responses came from the relatively large province of Ontario. Canada presented evidence to show that the characteristics of non-responding schools in Ontario were not markedly different from those of respondent schools.

It was concluded that the problems observed in the Canadian data had a minimal impact on the data, and inclusion in the full range of PISA 2003 reports was recommended.

Czech Republic

The Czech Republic fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

Denmark

Denmark had an overall exclusion rate of 5.33 per cent, the majority of which were within school exclusions due to language issues. This exceeds the PISA standard of 5 per cent.

Inclusion of Danish data in the full range of PISA 2003 reports was recommended.

Finland

Finland fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

France

The implementation of PISA in France deviated from the internationally recommended procedures in a number of ways. First, France did not implement the school questionnaire. It follows that France cannot be included in those reports and analyses that utilise school questionnaire data. Second, France did not implement the recommended multiple coding design. The alternative design implemented in France, however, was carefully reviewed and it was deemed that the design implemented provided a sufficient level of quality assurance for the coding activities. Third, it was noted that the test administrators were not trained in person as required by the standards. As an alternative, the test administrators were trained through phone calls. Finally, due to local requirements, the PQMs were school inspectors and were not formally independent of the French national centre as was required by the standards.

Given that the PISA quality monitors did not identify problems with the test administration and that the lack of independence of the quality monitors was unlikely to cause problems it was concluded that the identified issues would have no marked effect on the data and it was therefore recommended that all the available French data be included in PISA reports.

Germany

Germany fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

Greece

Problems were identified with the printing and pagination of the instruments administered in Greece. Additional analysis undertaken to examine this issue suggested that at the national level the impact of printing problems on the data were likely to be minimal. It was recommended that the Greek data be included in the full range of PISA 2003 reports.

Hong Kong-China

Hong Kong-China fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

Hungary

Hungary fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.
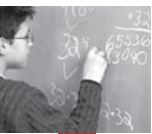
Iceland

Iceland fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

Indonesia

Indonesia did not submit the educational career questionnaire for international verification before the questionnaire was administered in the field. Also, there was evidence of poor translation in some of the administered instruments. The consortium therefore deleted, during the analysis phase, items it identified as poorly translated (see Chapter 5). The quality of the printed instruments was also significantly below that of other PISA countries. While coverage of the PISA population met PISA standards, Indonesia had a low level of 15-year-old enrolment, so coverage of 15-year-olds was just 46 per cent.

It was recommended that all the available Indonesian data be included in PISA 2003 reports.

Ireland

Ireland fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

Italy

Italy fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

Italy, Veneto - Nord Est

The Italian region of Veneto - Nord Est fully met the PISA 2003 standards.

Italy, Trento - Nord Est

The Italian region of Trento - Nord Est fully met the PISA 2003 standards.

Italy, Toscana – Centro

The Italian region of Toscana – Centro fully met the PISA 2003 standards.

Italy, Piemonte - Nord Ovest

The Italian region of Piemonte - Nord Ovest fully met the PISA 2003 standards

Italy, Lombardia - Nord Ovest

The Italian region of Lombardia - Nord Ovest fully met the PISA 2003 standards.

Italy, Bolzano - Nord Est

The Italian region of Bolzano - Nord Est fully met the PISA 2003 standards.

Japan

Japan fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

Latvia

Latvia did not submit the Russian language test instruments to the international verification team for a final optical check. In addition the Russian coding guides were not submitted for verification. In Latvia, 35.4 per cent of the population is assessed in Russian. Analysis of the submitted data suggested that these breaches of PISA 2003 standards had no marked affect on the Latvian data and inclusion in the full range of PISA 2003 reports was recommended.

Liechtenstein

Liechtenstein fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

Luxembourg

In Luxembourg, students were tested in either German or French, depending upon the combination of languages in which they have experienced instruction. The procedures of allocating languages to students were different in PISA 2003 to those applied in PISA 2000. This change in procedures was deemed to prevent the interpretation of trends in Luxembourg between PISA 2000 and PISA 2003.

Luxembourg fully met the PISA 2003 standards, and inclusion in PISA 2003 reports that were not concerned with trends was recommended.
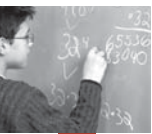
Mexico

For Mexico, it was noted that the quality of the printing and layout of instruments varied in the administered booklets. The originally submitted database included unusually high numbers of inconsistencies between student questionnaire data and tracking forms, which could only be corrected by taking the information provided in the tracking forms as accurate. Some school questionnaire indicators were found to have percentages of missing values around 50 per cent after data cleaning. Consequently, some of these indicators were not included in the final database.

Furthermore, the percentage of ineligible students was very high (8.10 per cent), and this was due mainly to a substantial number of students with invalid or out-of-range incorrect birth dates, and transferred students. The coverage of the national 15-year-old population was low (49 per cent), primarily because of low (58 per cent) enrolment rates of the target population. As the problems encountered with sampling and data collection were not deemed to have marked effects on the results, inclusion in the full range of PISA 2003 reports was recommended.

The Netherlands

The Netherlands fully met the PISA standards, and inclusion in the full range of PISA 2003 reports was recommended.

New Zealand

The within-school samples included a high percentage of ineligible students (5.99 per cent), with these approximately evenly split between drop-outs and transferred students. Additionally, New Zealand had an overall exclusion rate of 5.07 per cent, the majority of which were within-school exclusions due to language issues.

It was recommended that the data for New Zealand be included in the full range of PISA 2003 reports.

Norway

Norway fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

Poland

Norway fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

Portugal

In Portugal, the within-school samples included a high percentage of ineligible students (5.68 per cent), mostly being due to dropouts. It was recommended that the data for Portugal be included in the full range of PISA 2003 reports.

Korea

Korea fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

Russian Federation

The Russian Federation fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

Serbia

In Serbia, a number of sampling-related concerns were noted. First, the overall exclusion rate of 5.66 per cent, does not meet the PISA standard of 5 per cent. Second, the within-school samples included a high percentage of ineligible students (5.74 per cent), with those being mostly drop-outs. Third, while coverage of the PISA population met PISA standards, Serbia had some frame data issues so coverage of 15-year-olds appeared to be just 69 per cent.

In addition, Serbia implemented an unapproved marker design. Follow-up analysis suggested that this had no marked effect on the data.

It was recommended that the Serbian data be included in the full range of PISA 2003 reports.

Spain

It was noted that Spain had high overall exclusion rates (7.29 per cent) that did not meet PISA 2003 standards. This high level of exclusions was largely due to within-school exclusions. Additionally, the third

coverage index was low but has been explained by sources of error in the statistics gathered to obtain the SF2[a] value.

In the Basque country, as noted below, some students were tested in a language spoken at home rather than the official language of instruction. However, the percentage affected in Spain as a whole was very small.

It was recommended that the Spanish data be included in the full range of PISA reports.

Castilla-Leon

The Spanish region of Castilla-Leon had an overall exclusion rate (5.96 per cent, but 4.89 per cent when langue exclusions were removed), and inclusion in the full range of PISA reports was recommended.

Catalonia

The Catalonian multiple-marker data showed a consistent leniency bias across all three domains included in the study. The impact on the overall results, however, was deemed to be small. It was concluded that the Spanish region of Catalonia fully met the PISA standards.

Basque Country

For the Spanish region of the Basque Country, the standard procedure relating to the language of assessment was not followed. All students receiving instruction in bilingual Spanish/Basque settings were tested in Castillian, instead of being given the choice of a Basque or Spanish booklet. Students receiving instruction in Basque immersion schools were only tested in Basque when they had a Basque-speaking mother, a Basque-speaking father and used themselves Basque in their communications at home. All other Basque immersion students were tested in their home language (Castillian) rather than in their language of instruction (Basque). Note that as the Basque Country contains only a small percentage of the Spanish population this deviation does not influence the results for Spain overall.

In all other respects the data for the Basque Country met the PISA standards. The consortium recommended that the Basque Country data be included in the full range of PISA reports and that the data be annotated where it is published to indicate that the PISA results in the Basque Country must be interpreted as the results obtained by the students enrolled in the Basque educational system, but not as the results obtained by the students attending instruction in Basque language.
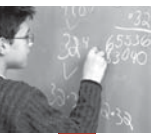
Slovak Republic

The Slovak Republic fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

Sweden

Sweden fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

Switzerland

Switzerland fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

Thailand

Thailand fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

Tunisia

The within-school samples included a high percentage of ineligible students (6.36 per cent). Further, it was noted that the printing quality of the administered instruments varied, and that there were pagination and layout errors in some of the administered booklets. Follow-up analysis did not suggest that the low printing quality had had a material affect on the quality of the data.

Tunisia met the PISA standards, and inclusion in the full range of PISA reports was recommended.

Turkey

While coverage of the PISA population met PISA 2003 standards, Turkey had a low level of 15-year-old enrolment (54 per cent) so coverage of 15-year-olds was just 36 per cent. Turkey has several forms of informal education through which participants receive their training via mail, television, or hands on experience. There are no records of the 15-year-olds that might be in those programmes. This may be one factor explaining the low enrolment in formal education of 15-year-olds.

Turkey fully met the PISA 2003 standards, and inclusion in the full range of PISA 2003 reports was recommended.

United Kingdom

Problems relating to response rate and testing window were identified for the data from the United Kingdom. A poor school response rate resulted in an extension of the three-month testing window, which is required by the PISA technical standards. After the extension of the testing window, the school response rate (64.32 per cent prior to replacements and 77.37 per cent after replacements) and student response rates (77.92 per cent) were still below PISA standards.

The United Kingdom was especially well placed to provide accurate evidence one way or the other as to the existence of non-response bias in the PISA data, because results of national assessment data were available at the school level (for two assessments) and at the individual student level (for one of these assessments) for the entire PISA sample. The United Kingdom national centre prepared a report in February 2004, entitled *PISA 2003 England Sample: Report of an Investigation into Response Bias at the School and Student Level*. While England and Wales were part of the same data collection procedure, data from Scotland, which was adjudicated by the consortium as a separate unit, were fully comparable with results from other OECD countries and with results from PISA 2000.

The conclusion that the PISA sampling referee drew from this analyses was that there was good evidence that the school sample was not substantially biased upwards or downwards, in terms of mean student achievement, as a result of non-response. However, there was evidence that the responding schools were a more homogeneous group in terms of student achievement than the full sample.

For the student sample, the conclusion was that it appeared that student non-response was likely to have induced a bias in achievement. It was not possible to ascertain the exact magnitude of this. However, before

finalising this conclusion, an important additional check was needed. The initial analyses on response rates were carried out before student weights had been calculated by the consortium (see Chapter 8 for a full description of student weights). The PISA sampling referee, therefore, asked the United Kingdom national centre to carry out analyses using the student weights with adjustments for non-response, to see whether these adjustments might have been effective in reducing the non-response bias. These weighted analyses indicated that the weight adjustments did not have an appreciable effect on reducing the non-response bias.

The uncertainties surrounding the sample and its bias are such that PISA 2003 scores for the United Kingdom cannot reliably be compared with those of other countries. They can also not be compared with the performance scores for the United Kingdom from PISA 2000. The regional data from Wales are also not comparable with other countries.

The results are, however, accurate for many within-country comparisons between subgroups (*e.g.* males and females) and for relational analyses. The results for the United Kingdom were included in a separate category below the results for the other participating countries. Other data for the United Kingdom that were not reported in the initial report were made available on the PISA Web site (*www.pisa.oecd.org*).

All international averages and aggregate statistics include the data for the United Kingdom.
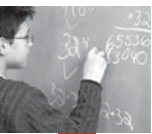
Scotland

Scotland fully met the PISA standards.

United States

Problems relating to response rate and testing window were identified for the data from the United States. As a result of a poor school response rate the consortium approved the use of a second testing window. Both the use of a second testing window and the timing of the window within the school year were breaches of the PISA technical standards. After taking into account the data from the second testing window, the United States data still did not meet the school response standards, the overall school response rate was 64.94 per cent before replacements and 68.12 per cent after replacements. Furthermore, the United States had high overall exclusion rates (7.28 per cent) mostly due to high within school exclusions. These did not seem to be concentrated in any particular category of student (*i.e.* gender, grade, etc.) but were spread over all student types.

Two separate investigations were conducted to validate the United States data. The first investigated the hypothesis that testing students early in the school year would lead to different achievement results than testing students of equivalent age later in the school year, as PISA requires. The hypothesis is that, because of loss of retention of knowledge and skills over the summer period, a student of a particular age (at the time of testing), tested at the beginning of the school year will tend to perform less well on PISA than a student of the same age (relative to the testing date) tested at a later point in the year. Specifically for the United States, this would mean that students born between July 1987 and June 1988, tested in September and October 2003, would not perform as well on average as students born in 1987 and tested in April and May 2003.

Schools were not randomised to testing periods, but rather the time of testing is confounded with the school's willingness to participate in the April-May period. This willingness to participate at this time of

the year might well be associated with student achievement. The United States national centre conducted an analysis that attempted to deal with this issue of confounding. Although no nonrandomised study can ever be entirely conclusive, the evidence was quite strong that use of a later testing time did not impact the average achievement results, either negatively (as hypothesised above) or positively. The multi-level models used showed that while public/private status, school size, percent of minority students, location, and region all had significant relationships with student achievement, time of testing did not. It was also the case that the mean scores of students tested in September and October were almost identical to those tested in April and May, suggesting that this finding is robust (in other words, it is not necessary to rely on model to explain away any differences between the two time periods).

The second study was a non-response bias analysis, conducted on the assumption that the September and October assessments would in fact be included in the data. These analyses were conducted by the United States national centre. The PISA sampling referee reviewed this report and concluded that, there is likely to be relatively little school non-response bias. Region did appear to be significantly related to school response, but it was not a very strong predictor of achievement. It also appeared that the respondent sample was somewhat relatively deficient in Asian and Pacific Islander students. However, the absolute difference in the percentages of these students between the responding sample and full sample is not great (4.4 per cent in the full sample; 3.8 per cent in the responding sample).

The United States data was included in the full range of PISA 2003 reports.

Uruguay

While coverage of the PISA population met PISA standards, Uruguay had a low level of 15-year-old enrolment, so coverage of 15-year-olds was just 63 per cent. It was also noted that the percentage of ineligible students was high (7.78 per cent).

Uruguay met the PISA standards, and inclusion in the full range of PISA reports was recommended.

READER'S GUIDE

Country codes

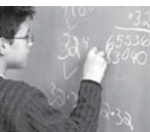The following country codes are used in this report:

OECD countries

AUS	Australia
AUT	Austria
BEL	Belgium
BEF	Belgium (French Community)
BEN	Belgium (Flemish Community)
CAN	Canada
CAE	Canada (English Community)
CAF	Canada (French Community)
CZE	Czech Republic
DNK	Denmark
FIN	Finland
FRA	France
DEU	Germany
GRC	Greece
HUN	Hungary
ISL	Iceland
IRL	Ireland
ITA	Italy
JPN	Japan
KOR	Korea
LUX	Luxembourg
LXF	Luxembourg (French Community)
LXG	Luxembourg (German Community)
MEX	Mexico
NLD	Netherlands
NZL	New Zealand
NOR	Norway
POL	Poland
PRT	Portugal

SVK	Slovak Republic
ESP	Spain
ESB	Spain (Basque Community)
ESC	Spain (Catalonian Community)
ESS	Spain (Castillian Community)
SWE	Sweden
CHE	Switzerland
CHF	Switzerland (French Community)
CHG	Switzerland (German Community)
CHI	Switzerland (Italian Community)
TUR	Turkey
GBR	United Kingdom
IRL	Ireland
SCO	Scotland
USA	United States

Partner countries

BRA	Brazil
HKG	Hong Kong-China
IND	Indonesia
LVA	Latvia
LVL	Latvia (Latvian Community)
LVR	Latvia (Russian Community)
LIE	Liechtenstein
MAC	Macao-China
RUS	Russian Federation
YUG	Serbia and Montenegro (Serbia)
THA	Thailand
TUN	Tunisia
URY	Uruguay

List of abbreviations

The following abbreviations are used in this report:

ACER	Australian Council for Educational Research	**NDP**	National Desired Population
AGFI	Adjusted Goodness-of-Fit Index	**NEP**	National Enrolled Population
BRR	Balanced Repeated Replication	**NFI**	Normed Fit Index
CFA	Confirmatory Factor Analysis	**NIER**	National Institute for Educational Research, Japan
CFI	Comparative Fit Index	**NNFI**	Non-Normed Fit Index
CITO	National Institute for Educational Measurement, The Netherlands	**NPM**	National Project Manager
CIVED	Civic Education Study	**OECD**	Organisation for Economic Cooperation and Development
DIF	Differential Item Functioning	**PISA**	Programme for International Student Assessment
ESCS	Economic, Social and Cultural Status		
ENR	Enrolment of 15-year-olds	**PPS**	Probability Proportional to Size
ETS	Educational Testing Service	**PGB**	PISA Governing Board
IAEP	International Assessment of Educational Progress	**PQM**	PISA Quality Monitor
		PSU	Primary Sampling Units
I	Sampling Interval	**QAS**	Questionnaire Adaptations Spreadsheet
ICR	Inter-Country Coder Reliability Study		
		RMSEA	Root Mean Square Error of Approximation
ICT	Information Communication Technology		
		RN	Random Number
IEA	International Association for the Evaluation of Educational Achievement	**SC**	School Co-ordinator
		SD	Standard Deviation
		SEM	Structural Equation Modelling
INES	OECD Indicators of Education Systems	**SMEG**	Subject Matter Expert Group
		SPT	Study Programme Table
IRT	Item Response Theory	**TA**	Test Administrator
ISCED	International Standard Classification of Education	**TAG**	Technical Advisory Group
		TCS	Target Cluster Size
ISCO	International Standard Classification of Occupations	**TIMSS**	Third International Mathematics and Science Study
ISEI	International Socio-Economic Index	**TIMSS-R**	Third International Mathematics and Science Study – Repeat
MENR	Enrolment for moderately small school		
		VENR	Enrolment for very small schools
MOS	Measure of size	**WLE**	Weighted Likelihood Estimates
NCQM	National Centre Quality Monitor		
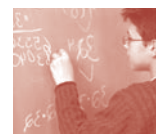
References

Adams, R.J., Wilson, M.R. and **W. Wang** (1997), "The multidimensional random coefficients multinomial logit model", *Applied Psychological Measurement 21,* pp. 1-24.

Aiken, L. R. (1974), "Two scales of attitudes toward mathematics," *Journal for Research in Mathematics Education 5,* National Council of Teachers of Mathematics, Reston, pp. 67-71.

Andersen, Erling B. (1997), "The Rating Scale Model", in van der Linden, W. J. and R.K. Hambleton (eds.), *Handbook of Modern Item Response Theory,* Springer, New York/Berlin/Heidelberg.

Bandura, A. (1986), *Social Foundations of Thought and Action: A Social Cognitive Theory,* Prentice Hall, Englewood Cliffs, N.J.

Baumert, J. and **O. Köller** (1998), "Interest Research in Secondary Level I : An Overview", in L. Hoffmann, A. Krapp, K.A. Renninger & J. Baumert (eds.), *Interest and Learning*, IPN, Kiel.

Beaton, A.E. (1987), *Implementing the New Design: The NAEP 1983-84 Technical Report* (Report No. 15-TR-20), Educational Testing Service, Princeton, N.J.

Bryk, A. S. and **S.W. Raudenbush** (1992), *Hierarchical Linear Models: Applications and Data Analysis Methods,* SAGE Publications, Newbury Park.

Bollen, K.A. and **S.J. Long** (eds.) (1993), *Testing Structural Equation Models*, SAGE publications, Newbury Park.

Branden, N. (1994), *Six Pillars of Self-Esteem*. Bantam, New York.

Brennan, R.L. (1992), *Elements of Generalizability Theory*, American College Testing Program, Iowa City.

Buchmann, C. (2000), *Measuring Family Background in International Studies of Educational Achievement: Conceptual Issues and Methodological Challenges,* paper presented at a symposium convened by the Board on International Comparative Studies in Education of the National Academy of Sciences/National Research Council on 1 November, in Washington, D.C.

Cochran, W.G. (1977), *Sampling Techniques* (3rd edition), Wiley, New York.

Cronbach, L.J., G.C. Gleser, H. Nanda and **N. Rajaratnam** (1972), *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*, Wiley and Sons, New York.

Eccles, J.S. (1994), "Understanding Women's Educational and Occupational choice: Applying the Eccles *et al.* Model of Achievement-Related Choices", *Psychology of Women Quarterly 18*, Society for the Psychology of Women, Washington, D.C., pp. 585-609.

Eccles, J.S. and **A. Wigfield** (1995), "In the mind of the achiever: The structure of adolescents' academic achievement-related beliefs and self-perceptions", *Personality and Social Psychology Bulletin 21*, Sage Publications, Thousand Oaks, pp. 215-225.

Ganzeboom, H.B.G., P.M. de Graaf and **D.J. Treiman** (1992), "A standard international socio-economic index of occupational status", *Social Science Research 21*, Elsevier, pp.1-56.

Gifi, A. (1990), *Nonlinear Multivariate Analysis*, Wiley, New York.

Greenacre, M.J. (1984), *Theory and Applications of Correspondence Analysis*, Academic Press, London.

Grisay, A. (2003), "Translation procedures in OECD/PISA 2000 international assessment", *Language Testing 20,* Holder Arnold Journals, pp.225-240.

Gustafsson, J.E and **P.A. Stahl** (2000), *STREAMS User's Guide, Version 2.5 for Windows,* MultivariateWare, Mölndal, Sweden.

Hacket, G. and **N. Betz.** (1989), "An Exploration of the mathematics Efficacy/mathematics Performance Correspondence", *Journal of Research in Mathematics Education 20,* National Council of Teachers of Mathematics, Reston, pp. 261-273.

Harvey-Beavis, A. (2002), "Student and Questionnaire Development" in OECD, *PISA 2000 Technical Report*, OECD, Paris.

Hatcher, L. (1994), *A Step-by-Step Approach to Using the SAS System for Factor Analysis and Structural Equation Modeling,* SAS Institute Inc., Cary.

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

Jöreskog, K.G. and **Dag Sörbom** (1993), *LISREL 8 User's Reference Guide,* Scientific Software International, Chicago.

Judkins, D.R. (1990), "Fay's Method for Variance Estimation", *Journal of Official Statistics 6,* Statistics Sweden, Stockholm, pp. 223-239.

Kaplan, D. (2000), *Structural Equation Modeling: Foundation and Extensions,* SAGE Publications, Thousand Oaks.

Keyfitz, N. (1951), "Sampling with probabilities proportionate to science: Adjustment for changes in probabilities", *Journal of the American Statistical Association 46,* American Statistical Association, Alexandria, pp.105-109.

Lepper, M. R. (1988), "Motivational considerations in the study of instruction", *Cognition and Instruction 5,* Lawrence Erlbaum Associates, Mahwah, pp. 289-309.

Ma, X. (1999), "A Meta-Analysis of the Relationship Between Anxiety Toward mathematics and Achievement in mathematics", *Journal for Research in Mathematics Education 30,* National Council of Teachers of Mathematics, Reston, pp. 520-540.
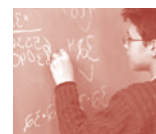
Macaskill, G., R.J. Adams and **M.L. Wu** (1998), "Scaling methodology and procedures for the mathematics and science literacy, advanced mathematics and physics scales", in M. Martin and D.L. Kelly (eds.) *Third International Mathematics and Science Study, Technical Report Volume 3: Implementation and Analysis,* Center for the Study of Testing, Evaluation and Educational Policy, Boston College, Chestnut Hill.

Marsh, H. W. (1990), *Self-Description Questionnaire (SDQ) II: A theoretical and Empirical Basis for the Measurement Of Multiple Dimensions of Adolescent Self-Concept: An Interim Test Manual and a Research Monograph,* The Psychological Corporation, San Antonio.

Marsh, H. W. (1994), "Confirmatory factor analysis models of factorial invariance: A multifaceted approach" *Structural Equation Modeling 1,* Lawrence Erlbaum Associates, Mahwah, pp. 5-34.

Marsh, H. W. (1999), *Evaluation of the Big-Two-Factor Theory of Motivation Orientation: Higher-order Factor Models and Age-related Changes*, paper presented at the 31.62 Symposium, Multiple Dimensions of Academic Self-Concept, Frames of Reference, Transitions, and International Perspectives: Studies From the SELF Research Centre. Sydney: University of Western Sydney.

Masters, G. N. and **B. D. Wright** (1997), "The Partial Credit Model", in W. J. van der Linden and R.K. Hambleton (eds.), *Handbook of Modern Item Response Theory,* Springer, New York/Berlin/Heidelberg.

Meece, J., A. Wigfield and **J. Eccles** (1990), "Predictors of Maths Anxiety and its Influence on Young Adolescents' Course Enrolment and Performance in Mathematics", *Journal of Educational Psychology 82,* American Psychological Association, Washington, D.C., pp. 60-70.

Middleton, J.A. and **P.A. Spanias** (1999), "Findings, Generalizations, and Criticisms of the Research", *Journal for Research in Mathematics Education 30*, National Council of Teachers of Mathematics, Reston, pp. 65-88.

Mislevy, R.J. (1991), "Randomization-based inference about latent variable from complex samples", *Psychometrika 56*, Psychometric Society, Greensboro, pp. 177-196.

Mislevy, R.J. and **K.M. Sheehan** (1987), "Marginal estimation procedures", in A.E. Beaton (ed.), *The NAEP 1983-1984 Technical Report* (Report No. 15-TR-20), Educational Testing Service, Princeton, N.J.

Mislevy, R.J. and **K.M. Sheehan** (1980), "Information matrices in latent-variable models", *Journal of Educational Statistics 14.4,* American Educational Research Association and American Statistical Association, Washington, D.C., and Alexandria, pp. 335-350.

Mislevy, R.J., A.E. Beaton, B. Kaplan and **K.M. Sheehan.** (1992), "Estimating population characteristics form sparse matrix samples of item responses", *Journal of Educational Measurement 29*, National Council on Measurement in Education, Washington, D.C., pp. 133-161.

Multon, K. D., S. D. Brown and **R.W. Lent** (1991), "Relation of self-efficacy beliefs to academic outcomes: A meta-analytic investigation", *Journal of Counselling Psychology 38,* American Psychological Association, Washington, D.C., pp. 30-38.

Muthén, B. O., S. H. C. du Toit and **D. Spisic** (1997), "Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical outcomes", *Psychometrika*, Psychometric Society, Greensboro.

Muthen, L. and **B. Muthen** (2003), *Mplus User's Guide Version 3.1,* Muthen & Muthen, Los Angeles.

Nishisato, S. (1980), *Analysis of Categorical Data: Dual Scaling and its Applications,* University of Toronto Press, Toronto.

OECD (Organisation for Economic Co-Operation and Development) (1999), *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 (2002), *PISA 2000 Technical Report,* OECD, Paris.

OECD (2003), *Student Engagement at School: A Sense of Belonging and Participation: Results from PISA 2000,* OECD, Paris.

OECD (2004a), *The PISA 2003 Assessment Framework: Mathematics, Reading, Science and Problem Solving Knowledge and Skills*, OCED, Paris.

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

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

OECD (2005a), *PISA 2003 Data Analysis Manual: SAS® Users,* OECD, Paris.

OECD (2005b), *PISA 2003 Data Analysis Manual: SPSS® Users,* OECD, Paris.

Owens L. and **J. Barnes** (1992), *Learning Preference Scales*, Australian Council for Educational Research, Hawthorn.

Rasch, G. (1960), *Probabilistic models for some intelligence and attainment tests,* Nielsen and Lydiche, Copenhagen.

Rust, K. (1985), "Variance estimation for complex estimators in sample surveys", *Journal of Official Statistics 1,* Statistics Sweden, Stockholm, pp. 381-397.

Rust, K. and **J.N.K. Rao** (1996), "Variance estimation for complex surveys using replication techniques", *Statistical Methods in Medical Research 5,* Holder Arnold Journals, pp. 283-310.

Sändal, C.E., B. Swensson and **J. Wretman** (1992), *Model Assisted Survey Sampling,* Springer-Verlag, New York.

Schaffer, E. C., P. S. Nesselrodt and **S. Stringfield** (1994), "The Contribution of Classroom Observation to School Effectiveness Research" in Reynolds *et. al.* (eds.), *Advances in School Effectiveness Research and Practice,* Pergamon, Oxford/New York/Tokyo.

Schulz, W. (2003), *Validating Questionnaire Constructs in International Studies. Two Examples from PISA 2000,* paper presented at the Annual Meeting of the American Educational Research Association (AERA) in Chicago, 21-25 April.
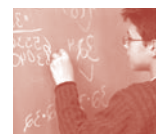
Schulz, W. (2004), "Mapping Student Scores to Item Responses", in W. Schulz and H. Sibberns (eds.), *IEA Civic Education Study. Technical Report*, IEA, Amsterdam.

Sirotnik, K. (1970), "An analysis of variance framework for matrix sampling", *Educational and Psychological Measurement 30,* SAGE Publications, pp. 891-908.

Slavin, R. E. (1983), "When does cooperative learning increase student achievement?" *Psychological Bulletin 94,* American Psychological Association, Washington, D.C., pp. 429-445.

Statistical Solutions (1992), *BMDP Statistical Software*, Statistical Solutions, Los Angeles.

Teddlie, C. and **D. Reynolds** (2000) (eds.), *The International Handbook of School Effectiveness Research,* Falmer Press, London/New York.

Thorndike, R.L. (1973), *Reading Comprehension Education in Fifteen Countries: An Empirical Study*, Almquist & Wiksell, Stockholm.

Travers, K. J., R.A. Garden and **M. Rosier** (1989), "Introduction to the Study", in D.A. Robitaille and R.A. Garden (eds.), *The IEA Study of Mathematics II: Contexts and Outcomes of School Mathematics Curricula,* Pergamon Press, Oxford.

Travers, K. J. and **I. Westbury** (1989), *The IEA Study of Mathematics I: Analysis of Mathematics Curricula,* Pergamon Press, Oxford.

Verhelst, N. (2004), "Generalizability Theory", in Council of Europe, *Reference Supplement to the Preliminary Pilot version of the Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, (Section E), Council of Europe (DGIV/EDU/LANG (2004) 13), Strasbourg.

Warm, T. A. (1989), "Weighted Likelihood Estimation of Ability in Item Response Theory", *Psychometrika* 54, Psychometric Society, Greensboro, pp. 427-45.

Wigfield, A., J. S. Eccles and **D. Rodriguez** (1998), "The development of children's motivation in school contexts", in P. D. Pearson. and A. Iran-Nejad (eds.), *Review of Research in Education 23,* American Educational Research Association, Washington D.C., pp. 73-118.

Wilson, M. (1994), "Comparing Attitude Across Different Cultures: Two Quantitative Approaches to Construct Validity", in M. Wilson (ed.), *Objective Measurement II: Theory into Practice,* Ablex, Norwood, pp. 271-292.

Wolter, K.M. (1985), *Introduction to Variance Estimation*, Springer-Verlag, New York.

Wu, M.L., R.J. Adams and **M.R. Wilson** (1997), *ConQuest: Multi-Aspect Test Software* [computer program], Australian Council for Education Research, Camberwell.

Zimmerman, B.J. and **D.H. Schunk** (eds.) (1989), *Self-Regulated Learning and Academic Achievement. Theory, Research and Practice,* Springer, New York.

Table of Contents