

TECHNICAL NOTES AND REFERENCES



Technical Notes

These technical notes provide a very brief outline of some of the key methodology used in PIRLS-05/06. For more detailed information readers are advised to go to the *PIRLS 2006 technical report* edited by Martin, Mullis, and Kennedy (2007) and available on the PIRLS web site (www.pirls.org). The PIRLS 2006 User Guide that is included with the international database also provides a comprehensive overview of the technical aspects of the study. (See Foy & Kennedy, 2008 for details.)

TN1 Weighting

The sampling design required schools to be sampled with probability proportional to size (PPS), and for classrooms to be sampled with equal probabilities. In addition, many countries, including New Zealand, used stratification to improve the precision of their sampling. Weighting was applied to all countries' data to ensure proper survey estimates and to adjust for the fact that the sampling design resulted in differential probabilities of selection for each student within the population. The weighting took into account school-, class-, and student-level information so that the overall sampling weight was a product of the school, class, and student weights.

TN2 Scaling

PIRLS makes use of a multiple-matrix sampling whereby students answer subsets of items from a larger pool of test items. Psychometric scaling techniques based on Item Response Theory (IRT) enable population estimates to be generated even though students do not respond to all the same achievement items.

Three Item Response Theory models are used, corresponding to the three types of assessment questions. For multiple-choice questions a three-parameter logistic model is used, which characterises the item in terms of difficulty, discrimination, and the possibility of guessing. For dichotomous open-response questions, a two-parameter logistic model is used (the possibility of guessing is discounted). For polytomous questions (extended response items with 0, 1, 2, and 3 as possible scores), a generalised partial-credit model was used, which factors in the different scores available to respondents.

The Item Response Theory scaling applied in PIRLS-05/06 uses the plausible value methodology to produce estimates of student proficiency in reading.

TN3 Summary statistics

The IRT scaling procedures generate five imputed scores or plausible values for each student. The differences between the five values, which tend to be very small, reflect the degree of uncertainty in the imputation process. To obtain the best estimate of a statistic, the computation is carried out on each of the five plausible values, and the results are then averaged. The national mean for a country, for example, was calculated as the mean of the weighted means of the plausible values. The international achievement means reported for a background index were calculated by first computing the national mean for each plausible value for each country and then calculating the mean across the countries. The five estimates resulting from this were then averaged to derive the international means presented in this report and in the international PIRLS.

TN4 Standard errors

The standard error is a measure of variability due to sampling when estimating a statistic. It provides a measure for determining the discrepancy between, for example, a sample mean and the true population mean. Ninety-five percent of sample means will lie within approximately plus or minus two (or more accurately 1.96) standard errors of the population mean. The standard error is used for determining confidence intervals.

For example, in 2005/2006 the Year 5 student mean for reading was 532 and the standard error of this statistic was 2.0. Therefore, we can say with 95 percent confidence, that the true mean was between 528 and 536 (i.e., $532 \pm 2 \times 2.0$).

Because of the complexity of the design of PIRLS (a complex survey design for the school sampling and a multiple-matrix design for questionnaire allocation), the calculation of standard errors is not as straightforward as it is for a study that uses simple random sampling and one assessment tool. The standard errors included in this report, which usually appear in brackets after the statistic, incorporate both the sampling variance (the uncertainty due to generalising from the sample to the population) and the imputation variance (the uncertainty due to inferring each student's proficiency from their performance on a subset of the items).

The Jackknife Repeated Replication (JRR) technique is used to estimate the sampling variance. This technique constructs a number of pseudo-replicates of the sample and compares each of the pseudo-replicated samples with that of the original sample. As noted, each student's proficiency is estimated by calculating five plausible values. The variability among these plausible values is used as a measure of the imputation variance. Custom-written SAS programs were used to compute the standard error, incorporating each of the variance components for each statistic.

Significance tests-comparisons of means

In this report all the comparisons that have been made were tested for statistical significance using the t statistic, with the alpha (α) level set at 0.05. The alpha level refers to the probability that a difference exists when in actuality it does not; the probability of making an incorrect inference is 5 percent.

To compare the means of two groups of students that have not been sampled independently of each other (e.g., the means for Year 5 boys and girls), the formula to generate the test statistics computed in this report was:

$$(1) \quad t = \frac{\bar{X}_1 - \bar{X}_2}{se_{diff}}$$

where se_{diff} is computed by combining the JRR and imputation variances. This involves computing the average difference between the two correlated samples (e.g., girls and boys in the same classes/schools) once for each of 75 replicate samples (error due to sampling) and five more times for each of the plausible values (imputation error). Custom-written SAS programs were used to compute the standard error of the mean difference between the two groups. The resulting value, t , is compared to the critical value of 1.96, this being the critical value for a two-tailed test at the alpha 0.05 level of significance (95 percent confidence).

If the means for two groups that were sampled independently are being compared (e.g., boys' achievement across two assessments), then the standard error of the difference is calculated as the square root of the sum of the squared standard errors of each mean:

$$(2) \quad se_{diff} = \sqrt{se_1^2 + se_2^2}$$

Note that in all calculations, unrounded figures are used in these tests, which may account for some results appearing to be inconsistent.

TN5 Multiple comparisons of means

When making a comparison between two means, the value of t must be at least equal to the critical value 1.96 for $\alpha \leq 0.05$ (two-tailed). However, in cases where there are more than two means being compared (e.g., comparisons among the four ethnic groups), there are more sources of measurement error to be considered. The Dunn-Bonferroni procedure has been used in these instances. Essentially, this procedure raises the critical value that t must reach before the (multiple) comparisons can be considered statistically significantly different at the 5 percent level.

Although the Dunn-Bonferroni procedure guards against misinterpreting the outcome of making multiple, simultaneous significance tests, the results can vary depending on the number of groups included in the adjustment. As a rule, the Dunn-Bonferroni procedure has been applied when testing multiple groups *within* a given assessment cycle (e.g., comparisons among the ethnic groupings in 2001). However, when comparing across cycles and for groups separately (e.g., Māori achievement from 2001 to 2005/2006), no adjustments have been made. Nor has this adjustment been made when considering gender comparisons within a group in a given assessment (e.g., comparing Pasifika boys' and girls' mean achievement in 2005/2006).

TN6 Effect sizes

Since statistical significance tests can partly be influenced by the sample sizes, a way of adding meaning to a difference which has been found to be *statistically significant* is to have an understanding of the *magnitude* of the difference. One way to do this is through the use of effect sizes. There are various ways of calculating and using effect sizes (see Rosenthal, 1994). In keeping with the reporting of effect sizes in the New Zealand TIMSS national reports, the following approach was used in PIRLS.

Firstly, the within pooled standard deviation (s_w) of the two groups being compared is calculated for each of the five imputed scale scores using:

$$s_w = \sqrt{\frac{\sum W_1 s_1^2 + \sum W_2 s_2^2}{\sum W_1 + \sum W_2}}$$

where:

W_i is the sample weight of group i

s_i is the standard deviation of the scale score of group i .

Then the effect size between the two groups, Cohen's d , is calculated for each of the five imputed scale scores using:

$$d = \frac{\bar{X}_1 - \bar{X}_2}{s_w}$$

where:

\bar{X}_i is the mean imputed scale score of group i .

The final effect size figure reported in this report is the *mean* effect size of the five imputed scale scores.

Interpreting the effect size

When interpreting an effect size between two groups, technically an effect size of 1.0 indicates a relative advantage of one standard deviation on the utilised measure. In other words, the mean of one group will be a whole standard deviation higher than the mean of the other.

TN7 Missing data and minimum group size for reporting achievement data

For tables in this report, particularly those containing context data, unless a non-response option is reported in the table, data are adjusted so that students with missing values are excluded. In general, around one to three to five percent of students had missing values for each question. For questions where the proportion of students with missing values exceeded three percent the proportion is noted.

Internationally, PIRLS does not report mean achievement scores for groups that represent less than 2.5 percent (rounded) of the population. However, in this report, if the proportion of New Zealand students was estimated to be 2 percent (rounded), as long as there were at least 50 students (from 10 schools) in the 'cell' to estimate the proportion, achievement results are reported.

TN8 Odds and odds ratios

Odds, like probability, are a measurement of chance. The relationship between the two is that the odds of an event occurring is the ratio of the probability of the event occurring to the probability of the event not occurring. That is, if we use o to denote the odds of an event occurring and p the probability, then:

$$o = \frac{p}{1-p}$$

However, odds are better described using a simple example. Suppose a jar contains eight marbles, only six of which are black. The *probability* of selecting a black marble is the ratio of the number of black marbles to the total number of marbles.

$$\text{That is, } \frac{6}{8} = \frac{3}{4}$$

Therefore, the *odds* of selecting a black marble is the ratio of the number of black balls to the number of balls that are *not* black. That is, $\frac{6}{2}$, or commonly notated as 3:1.

The odds ratio (*OR*) is defined as the ratio of the odds of an event for one group (usually the group of interest) occurring to the odds of an event occurring in another group.

In the case of lower achievers discussed in this report, the *odds* of students with particular attributes scoring less than 475 (or lower achievers) were calculated and then compared with the odds of students without the characteristic.

The *OR* was defined as:

[Independent variable] have X times the odds to be in the lower-achievers group as non- [independent variable].
'X' is the odds ratio

$$X = \frac{\text{odds of [independent variable]}}{\text{odds of non-[independent variable]}}$$

TN9 Confidentiality

PIRLS is designed to describe the results or to make inferences about the (estimated) population or sub-groups of Year 5 students, and the types and locations of schools they attended. It is not designed to report on the achievement or attributes of any individuals. Because of the cluster design (selecting a class or classes), this also holds for reporting at the school level. The researchers who are responsible for PIRLS here in New Zealand and internationally treat *all* information collected from students, parents, teachers, and schools during the course of the study confidentially. As a result, no individuals or schools are identified when reporting the results of the study.

This was the statement that was included as a note on the questionnaires respondents answered:

"All information collected in this study will be treated confidentially. At no time will you, other individuals, or your school be identified when reporting the results from this study."

References

- Caygill, R., & Chamberlain, M. (2004). *Progress in International Reading Literacy Study (PIRLS): New Zealand's year 5 student achievement 2001*. Wellington: Research Division, Ministry of Education.
- Caygill, R., Sturrock, F., & Chamberlain, M. (2007). *Mathematics and science achievement in New Zealand: tracking the changes of year 5 students in TIMSS 1994–2002*. Wellington: Research Division, Ministry of Education.
- Chamberlain, G. (2001). *Trends in year 5 students' mathematics and science achievement: results from a New Zealand study based on the Third International Mathematics and Science Study*. Wellington: Research Division, Ministry of Education.
- Chamberlain, M. (2007a). *Mathematics and science achievement in New Zealand: summing up New Zealand's participation in three cycles of TIMSS at year 9*. Wellington: Research Division, Ministry of Education.
- Chamberlain, M. (2007b). *Reading literacy in New Zealand: an overview of New Zealand's results from the Progress in International Reading Literacy Study (PIRLS) 2005/2006*. Wellington: Research Division, Ministry of Education.
- Crooks, T., & Flockton, L. (2005). *Reading and speaking assessment results 2004*. National Education Monitoring Report 34. Dunedin: Educational Assessment Research Unit, University of Otago.
- Foy, P., & Kennedy, A. M., (Eds.). (2008). *PIRLS 2006 user guide for the international database*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Joncas, M. (2007). PIRLS 2006 sampling design. In M. O. Martin, I. V. S. Mullis, & A. M. Kennedy (Eds.). *PIRLS 2006 technical report* (pp. 35–48). Chestnut Hill M.A: TIMSS & PIRLS International Study Center, Boston College.
- Kennedy, A. M., Mullis, I. V. S., Martin, M. O., & Trong, K. (Eds.). (2007). *PIRLS 2006 encyclopedia: a guide to reading education in the forty PIRLS 2006 countries*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Martin, M. O., Mullis, I. V. S., & Gonzalez, E. (2004). Home environments fostering children's reading literacy: results from the PIRLS 2001 study of reading literacy achievement in primary schools in 35 countries. *Proceedings of the IEA International Research Conference 2004: PIRLS volume 3*. Cyprus: University of Cyprus.
- Martin, M. O., Mullis, I. V. S., & Kennedy, A. M. (Eds.). (2007). *PIRLS 2006 technical report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- McDowall, S., Cameron, M., Dingle, R., with Gilmore, A., & MacGibbon, L. (2007). *Evaluation of the Literacy Professional Development Project*. Wellington: Ministry of Education.
- Ministry of Education. (1999). *Report of the Literacy Taskforce*. Wellington: Learning Media.
- Ministry of Education. (2006). *Effective literacy practice in years 5 to 8*. Wellington: Learning Media Ltd.
- Ministry of Education. (2007a). *Literacy learning progressions: draft for consultation*. Wellington: Learning Media Ltd.
- Ministry of Education. (2007b). *Te Reo Matatini: Māori Medium Literacy Strategy*. Wellington: Huia Education.
- Ministry of Education. (2007c). *Pasifika Education Plan monitoring report 2006*. Wellington: Author.
- Ministry of Education. (2008). *Ka Hikitia: Managing for success Māori Education Strategy 2008–2012*. Wellington: Ministry of Education.
- Mullis, I. V. S., Kennedy, A. M., Martin, M. O., & Sainsbury, M. (2006). *PIRLS 2006 assessment framework and specifications*. (2nd ed.). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Mullis, I. V. S., Martin, M. O., & Foy, P. (2005). *TIMSS 2003 international report on achievement in the mathematical cognitive domains: findings from a developmental project*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Mullis, I. V. S., Martin, M. O., Kennedy, A. M., & Flaherty, C. L. (Eds.) (2002). *PIRLS 2001 encyclopedia: IEA's a reference guide to reading education in the countries participating in IEA's Progress in International Reading Literacy Study (PIRLS)*. Chestnut Hill, MA: PIRLS International Study Center, Boston College.
- Mullis, I. V. S., Martin, M. O., Kennedy, A. M., & Foy, P. (2007). *PIRLS 2006 international report: IEA's Progress in International Reading Literacy Study in primary schools in 40 countries*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Nash, R. (2004). The explanation of social differences in reading attainment: an inspection of the PIRLS New Zealand data. *Waikato Journal of Education*, 10, 283-297.

Rosenthal, R. (1994). Parametric measures of effect size. In H Cooper and L V Hedges (Eds.), *The handbook of research synthesis*. New York: Russell Sage Foundation.

Schwippert, K., & Goy, M. (2007). Concluding remarks. In K. Schwippert (Ed.), *Progress in reading literacy: the impact of PIRLS 2001 in 13 countries*. (pp. 265–267). Münster: Waxmann Verlag GmbH.

Tunmer, W. E., Chapman, J. W., & Prochnow, J. E. (2006). Literate cultural capital at school entry predicts later reading achievement. *New Zealand Journal of Educational Studies*, 41, 2, 183-204.

Twist, L. Schagen, I., & Hodgson, C. (2007). *Readers and reading: the national report for England 2006* (PIRLS: Progress in International Reading Literacy Study). Slough: National Foundation for Educational Research.

Whetton, C., & Twist, L. (2003). *What determines the range of reading attainment in a country?* Paper presented at the 29th International Association for Educational Assessment Conference, 6–10 October 2003, Manchester, United Kingdom. Retrieved 15 April 2004, from: <http://www.nfer.ac.uk/>.