

Appendix E: Technical Notes

TN.1: SUMMARY OF THE SCALING PROCESS

As outlined in the introduction, TIMSS-98 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 enable population estimates to be generated even though students do not respond to all of the same achievement items. The Item Response Theory scaling used in TIMSS-98 uses the plausible value methodology to produce estimates of student proficiency in mathematics and science.

Essentially, the following steps were taken to generate the overall mathematics and science scores as well as the scale scores for the subtopics for each subject:

1. *Scaling* — each assessment item is examined against a model which expresses the probability that respondents with different abilities will achieve a certain score on the item. Three Item Response Theory models are used corresponding to the three types of test items.

For multiple-choice items a three parameter logistic model was used, which characterises the item in terms of difficulty, discrimination and the possibility of guessing. For dichotomous open-response items a two parameter logistic was used (the possibility of guessing is discounted). For polychotomous items, a generalised partial credit model was used, which factors in the different scores available to respondents (eg 0, 1, 2 and 3 rather than just right or wrong).

2. *Generating plausible values* — to acknowledge the measurement error surrounding the assessment of an individual's abilities, five separate ability estimates are made for each individual. These estimates are referred to as 'plausible values' and are based on information from student achievement, background data and the item parameters. Together, these plausible values are used to determine unbiased population estimates.
3. *Transforming the scaled scores on to a reporting scale*. For reporting purposes the scaled scores are transformed onto a more accessible reporting scale with a mean of 500 and standard deviation of 100.

The information in this note has largely been adapted from Johnson (1998). The interested reader is referred to Martin et al (2000) for a comprehensive explanation of the scaling process.

TN.2: STANDARD ERROR AND THE JACKKNIFE REPEATED REPLICATION METHOD

Due to the complexity of the design of TIMSS-98 the calculation of standard errors is not as straight forward as it is for a study which uses simple random sampling. Not only is there the issue that each student receives five imputed values (see Technical Note TN.1) for each of the scales used to estimate their underlying ability, there is also a sampling 'cluster effect' to consider. This effect arises because children are selected from schools in clusters (intact classrooms) and individuals in a cluster tend to have more similar characteristics than individuals chosen randomly from a population. The cluster effect effectively reduces the efficiency of the sample.

Because of these complexities standard errors must be estimated using numerical algorithms, which amount roughly to 'simulating' variations in the sampling of class groups, for example by successively leaving whole classes out of the sample, and calculating how the mean is affected. There are a number of different numerical methods or algorithms available for estimating standard errors. The TIMSS studies use the Jackknife Repeated Replication method (Gonzalez & Foy, 2000).

TN.3: SIGNIFICANCE TESTING

Throughout this report, differences have been described as statistically significant where $\alpha = 0.05$ (95% confidence). Although some results were highly statistically significant (eg $p < 0.01$) these have not been explicitly identified in this report. Conservative methods have been used to calculate the significance of differences between mean scores achieved by different population groups. The use of conservative methods increases confidence that where differences are reported as significant they have not occurred by chance. That is, there is less likelihood of making a Type I error of rejecting a null-hypothesis of 'no differences between the groups' when it should be accepted. Conversely, however, there is an increased likelihood of making a Type II error — that is accepting a null-hypothesis of 'no differences between the groups' when it should be rejected. This needs to be considered when interpreting the statistics presented in this report.

The t statistic was calculated using:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{se_1^2 + se_2^2}}$$

Note that t may be slightly under estimated for some groups because of the assumption that population groups have been independently sampled. For example, the sample of boys and the sample of girls are not completely independent and it would be more accurate to jackknife the differences between boys and girls. However, due to time constraints on reporting these data, the above formula, assuming sample independence, was applied uniformly. The results reported here are thus consistent with the use of conservative methods when examining differences between groups (there is less likelihood of making a Type I error).

TN.4: 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 $\alpha \leq 0.05$ (2-tailed). However, in cases where there are more than two means being compared, there are more sources of measurement error to be considered. There are a number of ways to correct for this; in this report the Dunn-Bonferroni procedure has been used. Essentially, this procedure raises the critical value that t must reach before the (multiple) comparisons can be considered statistically significantly different at the five percent level.

It was sometimes the case, however, that when the achievement means of two groups were compared, after adjusting the significance test for multiple comparisons the differences were found to be not statistically significant; yet if the same groups' means had been compared in isolation the difference *would* have been considered statistically significant. In other words a Type II error may have occurred whereby the null-hypothesis of 'no difference' is accepted when in fact it is possible that there is sufficient evidence to confidently reject the null-hypothesis.

Take, for example, the science achievement means of Pakeha (541) and Asian (517). Analysis of this difference yields a t of 2.16. If these two group-means were compared in isolation it could be said they are significantly different because t exceeds the critical value of 1.96 ($\alpha = 0.05$, 2-tailed). However, in this report, these groups' means are compared alongside two other groups' means and are not found to be significantly different from each other, as the critical value has been raised to 2.64 to compensate for the measurement errors inherent in the six comparisons.

TN.5 EFFECT SIZES

As well as determining whether differences in mean achievement are statistically significant it is useful to give an impression of the *magnitude* of the difference. One way of doing this is through the use of effect sizes. There are various ways of calculating and using effect sizes (Rosenthal, 1994). For the purpose of this report we have used the following method.

Calculating the effect size

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 (see Wilkinson et al, 2000 for a useful outline of effect sizes and their uses).

TN.6: MISSING STUDENTS

Analyses in this report are based on all students with achievement data. However, there is invariably data missing from students with regard to *context* variables for two reasons:

1. the student was absent from the session in which the context questionnaire was administered; or
2. the student participated in the session but did not answer the specific question(s) under examination.

Typically, between two and four percent of students are missing from contextual analyses reported in this document.