Estimating Standard Errors for the TIMSS and PIRLS 2011 Achievement Scales

Pierre Foy

To obtain estimates of students' proficiency in mathematics, science and reading that were both accurate and cost-effective, TIMSS and PIRLS 2011 made extensive use of probability sampling techniques to sample students from national fourth and eighth grade student populations, and applied matrix-sampling assessment designs to target individual students with a subset of the complete pool of assessment items. Statistics such as means and percentages computed from these student samples were used to estimate corresponding population parameters. This approach made efficient use of resources, in particular keeping student response burden to a minimum, but at a cost of some variance or uncertainty in the statistics. To quantify this uncertainty, each statistic in the **TIMSS and PIRLS 2011 International Reports** (Mullis, Martin, Foy, & Arora, 2012; Martin, Mullis, Foy, & Stanco, 2012; Mullis, Martin, Foy, & Drucker, 2012) is accompanied by an estimate of its standard error. For statistics based on plausible values, these standard errors incorporate components reflecting the uncertainty due to generalizing from student samples to the entire fourth or eighth grade student populations (sampling variance), and to inferring students' performance on the entire assessment from their performance on the subset of items that they took (imputation variance). For variables other than plausible values, standard errors are due entirely to sampling variance.

Estimating Sampling Variance

The TIMSS and PIRLS 2011 sample design applied a stratified multistage cluster-sampling technique to the problem of selecting efficient and accurate samples of students while working with schools and classes. This design capitalized on the structure of the student population (i.e., students grouped in classrooms within schools) to derive student samples that permitted efficient and economical data collection. However, such a complex sampling design complicates the task of computing standard errors to quantify sampling variability.

When, as in TIMSS and PIRLS, the sample design involves multistage cluster sampling, there are several options for estimating sampling errors that

avoid the assumption of simple random sampling (Wolter, 1985). The jackknife repeated replication technique (JRR) was chosen because it is computationally straightforward and provides approximately unbiased estimates of the sampling errors of means, totals, and percentages.

The variation on the JRR technique used in TIMSS and PIRLS 2011 is described in Johnson and Rust (1992). It assumes that the primary sampling units (PSUs), which are almost always schools in TIMSS and PIRLS, can be paired in a manner consistent with the sample design, with each pair regarded as members of a pseudo-stratum for variance estimation purposes. When used in this way, the JRR technique appropriately accounts for the combined effect of the between- and within-PSU contributions to the sampling variance. The general use of JRR entails systematically assigning pairs of schools to sampling zones, and randomly selecting one of these schools to have its contribution doubled and the other to have its contribution zeroed, so as to construct a number of "pseudo-replicates" of the original sample. The statistic of interest is computed once for the entire original sample, and once again for each jackknife pseudo-replicate sample. The variation between the estimates for each of the jackknife replicate samples and the original sample estimate is the jackknife estimate of the sampling error of the statistic.

Constructing Sampling Zones for Estimating Sampling Variance

To apply the JRR technique used in TIMSS and PIRLS 2011, the sampled schools were paired and assigned to a series of groups known as sampling zones. This was done at Statistics Canada by working through the list of sampled schools in the order in which they were selected and assigning the first and second participating schools to the first sampling zone, the third and fourth participating schools to the second zone, and so on. In total, 75 zones were used, allowing for 150 schools per country. When more than 75 zones were constructed, they were collapsed to keep the total number to 75.

Sampling zones were constructed within design domains, or explicit stratum. When an explicit stratum had an odd number of schools, either by design or because of school non-response, the students in the remaining school were randomly divided to make up two "quasi" schools for the purposes of

calculating the jackknife standard error.[1] Each sampling zone then consisted of a pair of schools or "quasi" schools.

Within each sampling zone, each school was assigned an indicator (u_j), coded randomly to 0 or 1, such that one school had a value of 0, and the other a value of 1. This indicator determined whether the weights for the sampled students in the school in this zone were to be doubled ($u_j = 1$) or zeroed ($u_j = 0$) for the purposes of creating the pseudo-replicate samples.

Applying the Jackknife Repeated Replication Method

To compute a statistic t from the sample of a country, the formula for the sampling variance estimate of the statistic t, based on the JRR algorithm used in TIMSS and PIRLS 2011, is given by the following equation:

$$Var_{jrr}(t) = \sum_{h=1}^{H} \left[t(J_h) - t(S) \right]^2$$

where H is the total number of sampling zones in the sample of the country under consideration. The term $t(S)$ corresponds to the statistic of interest for the whole sample computed with the overall sampling weights (as described in **Sample Design in TIMSS and PIRLS**). The term $t(J_h)$ denotes the same statistic using the h^{th} jackknife replicate sample J_h and its set of replicate sampling weights, which are identical to the overall sampling weights, except for the students in the h^{th} sampling zone. For the students in the h^{th} zone, all students belonging to one of the randomly selected schools of the pair were removed, and the students belonging to the other school in the zone were included twice. In practice, this was accomplished by recoding to zero the weights for the students in the school to be excluded from the replication, and multiplying by two the weights of the remaining students within the h^{th} pair. Each sampled student was assigned a vector of 75 replicate sampling weights , where h took values from 1 to 75. If W_{0i} was the overall sampling weight of student i, the h replicate weights for that student were computed as

$$W_{hi} = W_{0i} \cdot k_{hi}$$

where

$$k_{hi} = \begin{cases} 2 \cdot u_j & \text{if student } i \text{ is in school } j \text{ of sampling zone } h \\ 1 & \text{otherwise} \end{cases}$$

1 If the remaining school consisted of 2 sampled classrooms, each classroom became a "quasi" school.

The school-level indicators u_j determined which students in a sampling zone would get zero weights and which ones would get double weights, based on the school from which the students were sampled. The process of setting the k_{hi} values for all sampled students and across all sampling zones is illustrated in Exhibit 1. Thus, the computation of the JRR variance estimate for any statistic in TIMSS and PIRLS 2011 required the computation of the statistic up to 76 times for any given country: once to obtain the statistic for the full sample based on the overall weights W_{0i}, and up to 75 times to obtain the statistics for each of the jackknife replicate samples J_h using a set of replicate weights W_{hi}.

Exhibit 1: Construction of Replicate Weights Across Sampling Zones in TIMSS and PIRLS 2011

Sampling Zone	School Replicate Indicator (u_i)	Replicate Factors for Computing JRR Replicate Sampling Weights (k_{hi})						
		1	2	3	⋯	h	⋯	75
1	0	0	1	1	⋯	1	⋯	1
	1	2						
2	0	1	0	1	⋯	1	⋯	1
	1		2					
3	0	1	1	0	⋯	1	⋯	1
	1			2				
⋮	⋮	⋮	⋮	⋮	⋱	⋮	⋯	⋮
h	0	1	1	1	⋯	0	⋯	1
	1					2		
⋮	⋮	⋮	⋮	⋮	⋯	⋮	⋱	⋮
75	0	1	1	1	⋯	1	⋯	0
	1							2

In the TIMSS and PIRLS 2011 analyses, 75 replicate weights were computed for each country regardless of the number of actual zones within the country. If a country had fewer than 75 zones, then the additional replicate weights where h was greater than the number of zones within the country were all made equal to the overall sampling weight. Although this involved some redundant computations, having 75 replicate weights for each country had no effect on the magnitude of the error variance computed using the jackknife formula and

it simplified the computation of standard errors for numerous countries at a time. All standard errors presented in the TIMSS and PIRLS 2011 international reports were computed using SAS programs developed at the TIMSS & PIRLS International Study Center.

Estimating Imputation Variance

For variables other than plausible values, standard errors were the result solely of sampling variation, and were computed using the JRR technique. However, the situation for plausible values was more complicated. As described elsewhere,[2] the TIMSS and PIRLS 2011 item pool was far too extensive to be administered in its entirety to any one student, and so a matrix-sampling assessment design was adopted whereby each student was given a single test booklet containing only a part of the entire assessment. The results for all of the booklets were then aggregated using item response theory to provide results for the entire assessment. Multiple imputation was used to derive reliable estimates of student performance (plausible values) on the assessment as a whole, even though each student responded to just a subset of the assessment items. Because every student proficiency estimate incorporates a random element, TIMSS and PIRLS 2011 followed the customary procedure of generating five estimates for each student and using the variability among them as a measure of the imputation uncertainty, or error.

The general procedure for estimating imputation variance using plausible values is described in Mislevy, Beaton, Kaplan, and Sheenan (1992). First, compute the statistic t for each set of M plausible values. The statistics t_m, where m = 1, 2, …, 5, can be anything estimable from the data, such as a mean, the difference between means, percentiles, and so forth.

Once the statistics t_m are computed, the imputation variance is then calculated as:

$$Var_{imp} = \left(1 + \frac{1}{M}\right) Var\left(t_1, \cdots, t_m\right)$$

where M is the number of plausible values used in the calculation, and $Var(t_1,...,t_M)$ is the usual variance of the M estimates computed using each plausible value.

2 The TIMSS and PIRLS 2011 assessment design is described in **Assessment Framework and Instrument Development**.

Combining Sampling and Imputation Variance

The standard errors of all proficiency statistics (i.e., statistics based on plausible values) reported by TIMSS and PIRLS include both sampling and imputation variance components. These standard errors were computed using the following formula:

$$Var\left(t_{pv}\right) = Var_{jrr}\left(t_1\right) + Var_{imp}$$

where $Var_{jrr}(t_1)$ is the sampling variance computed for the first plausible value[3] and Varimp is the imputation variance. The IDB Analyzer – analytic software provided with the international database – automatically computes standard errors as described in this section. More information on how to use the IDB Analyzer can be found in The ***TIMSS 2011 International Database and User Guide*** (Foy, Arora, & Stanco, 2013) and the ***PIRLS 2011 International Database and User Guide*** (Foy & Drucker, 2013).

Estimating Standard Errors Details contains basic summary statistics for overall mathematics, science, and reading achievement in the TIMSS and PIRLS 2011 assessments. Each exhibit presents the student sample size, the mean and standard deviation averaged across the five plausible values, the jackknife sampling error for the mean, and the overall standard error for the mean, which includes the imputation error. **Estimating Standard Errors Details** also contains tables showing the same summary statistics for the mathematics and science content and cognitive domains at the fourth and eighth grades.

References

Foy, P., Arora, A., & Stanco, G.M. (Eds.). (2013). TIMSS 2011 international database and user guide. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Foy, P., & Drucker, K.T. (Eds.). (2013). PIRLS 2011 international database and user guide. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Johnson, E.G., & Rust, K.F. (1992). Population inferences and variance estimation for NAEP data. *Journal of Educational Statistics,* 17(2), 175-190.

Mislevy, R.J., Beaton, A., Kaplan, B.A., & Sheehan, K. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, 29(2), 133-161.

3 Under ideal circumstances and with unlimited computing resources, the JRR sampling variance would be computed for each of the plausible values and the imputation variance as described here. This would require computing the same statistic up to 380 times (once overall for each of the five plausible values using the overall sampling weights, and then 75 times more for each plausible value using the complete set of replicate weights). An acceptable shortcut, however, is to compute the JRR sampling variance component using only one plausible value (the first one), and then the imputation variance using the five plausible values. Using this approach, a statistic needs to be computed only 80 times.

Mullis, I.V.S., Martin, M.O., Foy, P. & Arora, A. (2012). *TIMSS 2011 international results in Mathematics.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Martin, M.O., Mullis, I.V.S., Foy, P. & Stanco, G.M. (2012). *TIMSS 2011 international results in Science.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I.V.S., Martin, M.O., Foy, P. & Drucker, K. T. (2012). *PIRLS 2011 international results in Reading.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Wolter, K. M. (1985). *Introduction to variance estimation.* New York: Springer-Verlag.