CHAPTER 13

Standard Error Estimation in TIMSS

Bethany Fishbein
Peter Siegel
Umut Atasever
Darryl Cooney
Eugenio Gonzalez

Introduction

To obtain unbiased population estimates of student proficiency, TIMSS uses probability sampling techniques to select students from the national fourth- and eighth-grade student populations and uses matrix-sampling assessment designs to target individual students with a subset of the complete pool of assessment items. This approach keeps the response burden on school systems and students to a minimum, but at the cost of some variance or uncertainty in the reported statistics, such as the means and percentages computed to estimate population parameters.

To quantify this uncertainty, each statistic reported in TIMSS International Results reports is accompanied by an estimate of its standard error. Statistics based on differences between two estimated results also have standard errors, which are used to calculate confidence intervals or to perform tests of statistical significance. For statistics reporting student achievement based on plausible values, standard errors are calculated based on two estimated variance components. The first is referred to as *sampling variance* and reflects the uncertainty due to generalizing from a student sample to the entire student population from which it was drawn. The second is known as *imputation variance* and reflects uncertainty due to inferring students' achievement estimates from their observed performance on a set of achievement items and other achievement-related information. This imputation variance reflects the posterior variance of the achievement estimates given all available information used in the achievement imputation model described in Chapter 11. For reported statistics that are not based on plausible values, the estimates of standard errors are based entirely on sampling variance.

Estimating Sampling Variance

TIMSS uses probability sampling to derive achievement results from national samples of students. Because many such samples are possible but only one sample is drawn, some uncertainty about how well the sample represents the population is expected. The uncertainty caused by sampling students from a target population, known as *sampling variance*, can be estimated from the data of the one sample drawn.

https://doi.org/10.6017/lse.tpisc.timss.rs6950

Whereas estimating the sampling variance from simple random samples is a relatively simple task, estimating the sampling variance from a complex sample design like the one used for TIMSS is a more challenging endeavor. A common way to estimate the sampling variance in these multistage cluster sampling designs is through resampling schemes (Efron, 1982) such as the balanced repeated replication (BRR) and jackknife repeated replication (JRR) techniques (Johnson & Rust, 1992; Quenouille, 1949; Tukey, 1958; Wolter, 1985). TIMSS uses a variation of JRR to estimate sampling variances. JRR was chosen because it is computationally straightforward and provides approximately unbiased estimates of the sampling variance of means, totals, and percentages.

At the core of the JRR technique is the repeated resampling from the observed sample under identical sample design conditions. In the context of TIMSS, this entails grouping primary sampling units into sampling zones based on the TIMSS sample design and conducting repeated draws of subsamples from these zones according to a predetermined scheme. The main features of the TIMSS sample design that JRR incorporates in its repeated replication are the stratification of schools and the clustering of students within schools. This is done by defining jackknife sampling zones as pairs of successive schools according to the sampling frame to model the stratification and clustering from the national samples (see Chapter 3 for information on the TIMSS sample design). The repeated subsampling required by JRR is applied across the sampling zones. The remainder of this section describes the procedure that is followed in TIMSS. The reader is referred to Efron (1982), for example, for explanations of why the procedure outlined below leads to approximately unbiased estimates of the sampling variance.

JRR sampling zones are constructed within explicit strata for each country or benchmarking participant. When schools are sampled, they are ordered within the explicit strata by additional implicit stratification variables and their measure of size. Based on this sorting, successively sampled schools are expected to have similar stratification attributes. When an explicit stratum has an odd number of sampled schools, either by design or because of nonresponding schools, the students in the lone school of the last sampling zone are divided randomly into groups, or according to classroom if more than one class is sampled, to make up two members ("quasi-schools") to calculate JRR standard errors. This results in each sampling zone consisting of two members—either two schools or two "quasi-schools."

In each country, a maximum of 125 zones are created, allowing for as many as 250 participating schools to be assigned to unique JRR zones with two members each. When more than 250 schools are sampled, the additional schools are collapsed into the existing zones in order of selection. The randomization used in the resampling within sampling zones preserves the sampling variance measured in the original sampling zones after collapsing. Note that the JRR sampling zones may be constructed in a different manner under specific national conditions or sampling designs. Country-specific information on these differences for TIMSS 2023 is available in Chapter 9. Appendix 13A shows the school sample size and number of constructed JRR sampling zones, before collapsing, for the participating countries and benchmarking participants in TIMSS 2023.

For estimating the sampling variance, the JRR procedure draws two subsamples based on each sampling zone: one where the first school or quasi-school in the pair is included and the second is removed, and the other where the second school is included and the first is removed. When a school is removed from a sampling zone, the sampling weights of the students in the remaining school are doubled to make up for the omitted school. All students in the other sampling zones are included in both subsamples, and their sampling weights are unchanged. With this process applied in each sampling zone, the JRR procedure yields up to 250 replicate subsamples, each with its own set of replicate sampling weights to account for the successive removal of each school from the pair in any given sampling zone.

The process of creating replicate sampling weights for the replicate subsamples defines replicate factors k_{hi} as follows:

$$k_{hi} = \begin{cases} 2 \text{ for students in school } i \text{ of sampling zone } h \\ 0 \text{ for students in the other school of sampling zone } h \\ 1 \text{ for students in any other sampling zone} \end{cases}$$

These replicate factors are used to compute the replicate sampling weights as follows:

$$W_{hij} = k_{hi} \cdot W_{0j}$$

where W_{0j} is the overall sampling weight of student j, and W_{hij} is the resulting replicate sampling weight of student j when school i from sampling zone h is included, and the other school in the pair is removed.

Exhibit 13.1 illustrates the calculation of the replicate factors necessary to produce the replicate sampling weights. Within each sampling zone, each school or quasi-school is randomly assigned an indicator u_{hi}, coded either 0 or 1, such that one school has a value of 0 and the other a value of 1. This indicator determines how schools within each zone will be successively included and removed. When a school is removed from a zone, the replicate factor is set to 0, and the sampling weights of all students in that school are set to 0. When a school is included, the replicate factor is set to 2, and the sampling weights of all students in that school are doubled. The sampling weights of students in all the other sampling zones remain unchanged.

Exhibit 13.1: Construction of Replicate Factors Across Sampling Zones

Sampling Zone	School Replicate Indicator (u_{hi})	Replicate Factors for Computing JRR Replicate Sampling Weights (k_{hi})											
		Zone 1		Zone 2		Zone 3		...	Zone *h*		...	Zone 125	
		(1)	(2)	(3)	(4)	(5)	(6)		(2*h*-1)	(2*h*)		(249)	(250)
1	0	2	0	1	1	1	1	⋯	1	1	⋯	1	1
	1	0	2										
2	0	1	1	2	0	1	1	⋯	1	1	⋯	1	1
	1			0	2								
3	0	1	1	1	1	2	0	⋯	1	1	⋯	1	1
	1					0	2						
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮	⋮	⋮	⋮	⋮
h	0	1	1	1	1	1	1	⋯	2	0	⋯	1	1
	1								0	2			
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮	⋮
125	0	1	1	1	1	1	1	⋯	1	1	⋯	2	0
	1											0	2

For example, sampling Zone 1 yields two sets of replicate sampling weights, hence the two columns for Zone 1. The first set has doubled sampling weights $(k_{11} = 2)$ for the students in the first school $(u_{11} = 0)$ of Zone 1, zeroed sampling weights $(k_{12} = 0)$ for the students in the second school $(u_{12} = 1)$ of Zone 1, and unchanged sampling weights $(k_{hi} = 1)$ for all students in the other sampling zones, e.g., Zones 2 through 125. This is shown in the first Zone 1 column. The second set of replicate sampling weights (shown in the second Zone 1 column) has zeroed sampling weights $(k_{11} = 0)$ for the students in the first school $(u_{11} = 0)$ of Zone 1, doubled sampling weights $(k_{11} = 2)$ for the students in the second school $(u_{12} = 1)$ of Zone 1, and unchanged sampling weights $(k_{hi} = 1)$ for all students in the other sampling zones.

The process is repeated across all 125 possible sampling zones, generating up to 250 sets of replicate sampling weights. The replicate sampling weights are then used to estimate any statistic of interest up to 250 times. The variation across these JRR estimates is an estimate of the sampling variance.

Given a statistic t to be computed from a national sample, the formula used to estimate the sampling variance of that statistic, based on the JRR algorithm, is given by the following equation:

$$Var_{jrr}(t_0) = \frac{1}{2}\sum_{h=1}^{125}\sum_{i=1}^{2}(t_{hi} - t_0)^2 \qquad (13.1)$$

where the term t_0 denotes the statistic of interest estimated with the unaltered overall student sampling weights W_{0j} and the term t_{hi} denotes the same statistic computed using the set of replicate sampling weights W_{hij} obtained from sampling zone h $(h = 1, \ldots, 125)$, where the i^{th} school (1st or 2nd) in the zone is included and the other removed. Efron (1982) provides a mathematical proof of why the variance can be calculated based on these squared deviations of the t_{hi} from the total sample statistics in JRR-based resampling schemes.

The sampling variance estimated with the JRR method properly accounts for the variation arising from having sampled students using the TIMSS multistage stratified cluster sample design. Its square root estimates the standard error due to sampling for a statistic derived from variables other than those based on plausible values. Examples of such statistics include the mean age of students and the percentage of students with at least one parent with a university degree. When used for statistical inference, the degrees of freedom for this estimate may be determined using the modified Satterthwaite (1941) approach as provided by Johnson and Rust (1992) and discussed by Qian (1998). For consistency with previous cycles, TIMSS 2023 uses an assumption of approximate normality, which was applied in prior cycles to simplify calculations.

Estimating Imputation Variance

Achievement estimates are based on observations of how students perform on a set of achievement items. Any estimate of a student variable, achievement, or self-report constructs based on a finite set of observed response variables is affected by measurement error. Responses to items provide an estimate of students' proficiency or other student characteristics. The responses to these sets of items used to measure student characteristics are not constants but vary over time (students do not always give precisely the same responses) and across different sets of questions.

Uncertainty about students' proficiency is a function of the number of items administered and the interaction of the item characteristics and student proficiency, among other factors. Measurement error is typically larger when fewer items are involved, but some amount of measurement error would always be observed, even if a student would take all the assessment items. However, the entire item pool in any given TIMSS assessment cycle is far too extensive to be administered to any student. Therefore, TIMSS uses a matrix-sampling assessment design whereby each student is given a single test booklet containing only a portion of the entire assessment. The results from all students and booklets are then analyzed using item response theory (IRT) to provide initial estimates of achievement on the TIMSS reporting scale. An imputation model is then applied integrating the results of the IRT analysis with the relationship between contextual variables and achievement. This imputation model is a latent regression model described in Chapter 11 and is used to derive estimates of student performance in the form of plausible values. These plausible values are proficiency estimates that incorporate the portion of measurement uncertainty that can be quantified. Proficiency estimates have an associated

variability due to measurement error. TIMSS follows the customary procedure of imputing multiple plausible values for each student and using the variability among them as a measure of that uncertainty, known as *imputation variance*. Currently, five plausible values are used.

The general procedure for estimating the imputation variance when analyzing student achievement data follows the basic principle developed by Rubin (1987) of performing any statistical analysis once for each imputation and aggregating these multiple sets of results (Mislevy et al., 1992). Thus, in TIMSS for any given achievement-based statistic t, estimating that statistic from each plausible value yields five estimates t_m, $m = 1, \ldots, 5$, all computed using the overall student sampling weights W_{0j}. The final estimate of that statistic, t_0, is the average of these five estimates:

$$t_0 = \frac{1}{5}\sum_{m=1}^{5} t_m.$$

The imputation variance of the statistic t_0 is simply the variance of the five results from the plausible values, computed as follows:

$$Var_{imp}(t_0) = \frac{6}{5}\sum_{m=1}^{5}\frac{(t_m - t_0)^2}{4}$$

where the factor $\frac{6}{5}$ is a correction factor necessary when using the multiple imputation methodology (Rubin, 1987). The total variance of the statistic t_0 is calculated by adding the imputation variance to the sampling variance as follows:

$$Var_{tot}(t_0) = Var_{jrr}(t_0) + Var_{imp}(t_0) \tag{13.2}$$

The sampling variance $Var_{jrr}(t_0)$ for a statistic based on plausible values is the average of the sampling variances calculated with each of the five plausible values $Var_{jrr}(t_m)$, $m = 1, \ldots, 5$, as follows:

$$Var_{jrr}(t_0) = \frac{1}{5}\sum_{m=1}^{5} Var_{jrr}(t_m)$$

where

$$Var_{jrr}(t_m) = \frac{1}{2}\sum_{h=1}^{125}\sum_{i=1}^{2}(t_{mhi} - t_m)^2$$

and t_{mhi} is the appropriate JRR estimate for plausible value m and computed using the set of replicate sampling weights of sampling Zone h where school i is included. The square root of the total variance is the standard error estimate for any statistic based on plausible values, such as the average TIMSS mathematics achievement for girls, or the percentage of students at or above the TIMSS Advanced International Benchmark of mathematics achievement.

Appendices 13B through 13E provide details on the standard errors for the TIMSS 2023 proficiency estimates of participating countries and benchmarking entities in fourth-grade mathematics, fourth-grade science, eighth-grade mathematics, and eighth-grade science, respectively. The exhibits contained in each appendix report the JRR sampling variance, imputation variance, total variance, and the overall standard error for each participant's mean proficiency estimates for the overall subject as well for the content and cognitive domain subscales and the environmental knowledge subscales for science.

Estimating Standard Errors for International Averages

Some exhibits in the TIMSS International Results reports include international averages and their standard errors. For example, TIMSS 2023 Exhibit 1.1.2 reports the international average for the percentages of girls and boys and their fourth-grade mathematics achievement. International averages are computed using the data from participating countries included in the main table of the exhibit. Results from the benchmarking participants are not included in the estimation of international averages.

For any given statistic t_0, its international average is given by

$$t_{intl} = \frac{1}{N}\sum_{k=1}^{N} t_{0k}$$

where N is the number of countries contributing to the international average and t_{0k} is the estimate of our statistic of interest for country k.

The total variance of the international average t_{intl} is given by

$$Var_{tot}(t_{intl}) = \frac{1}{N^2}\sum_{k=1}^{N} Var_{tot}(t_{0k}) \qquad (13.3)$$

where $Var_{tot}(t_{0k})$ is the total variance of our statistic of interest for country k. The standard error of the international average is the square root of the total variance.

For statistics based on plausible values, the total variance includes the sampling and imputation variances, as given in equation (13.2) above. For statistics not based on plausible values, such as percentages, the total variance is based entirely on the sampling variance, as shown in equation (13.1) above.

Estimating Standard Errors for Comparing Results from Independent Samples

Standard errors, along with providing a measure of uncertainty for TIMSS results, are also a necessary part of performing a null hypothesis significance test when comparing two or more estimates of population or subgroup averages. A basic objective of TIMSS is to provide fair and accurate comparisons of student achievement across assessment cycles. For example, TIMSS

2023 Exhibit 1.1.10 is one such example, showing fourth-grade mathematics trend results across the TIMSS assessment cycles. Additionally, the interactive TIMSS 2023 Exhibit 1.1.1 can be used to perform pairwise comparisons across countries for fourth-grade mathematics achievement. All of these comparisons require the computation of a standard error for the difference between two estimates, which has an expected value of zero (indicating no difference).

TIMSS results are reported by way of a statistic such as a mean or percentage, and each statistic is accompanied by its standard error, computed using either equation (13.1) or equation (13.2), as appropriate. Results from different assessment cycles, or from different countries within the same cycle, are treated as independent samples, and computing the standard error of a difference is straightforward.

When computing the difference between two TIMSS results t_A and t_B from independent samples, such as comparing the achievement of countries A and B, or comparing the achievement of a country between assessment cycles A and B, the standard error of that difference is given by

$$SE(t_A - t_B) = \sqrt{Var_{tot}(t_A) + Var_{tot}(t_B)}$$

or, more simply

$$SE(t_A - t_B) = \sqrt{SE(t_A)^2 + SE(t_B)^2}$$

which can be stated as follows: the standard error of the difference between two independent estimates is the square root of the sum of their respective squared standard errors.

It should be noted that this approach to computing standard errors for comparing independent samples assumes the true variance in the two populations are the same, which may not always be the case. Moreover, TIMSS currently does not include estimates for the trend scale linking error in the standard error of the difference for comparing results between assessment cycles. TIMSS 2023 uses an approach consistent with previous cycles to simplify calculations.

Estimating Standard Errors for Comparing Results from Dependent Samples

In the context of TIMSS, results from dependent samples are those statistics derived from the same national or benchmarking sample. The achievement difference between girls and boys, as shown, for example, in TIMSS 2023 Exhibit 1.1.2, is an example of results comparing two dependent samples. This dependence occurs because girls and boys are selected simultaneously from a shared sampling frame of schools and often classrooms. Attributes from girls and boys from the same school tend to be more similar compared to subgroups selected from different schools, thus resulting in a correlation that needs to be accounted for in the

computation of the standard error of their difference. In other words, rather than computing the error as if boys and girls came from two independent samples, the error of the difference is computed using the procedure for a paired-sample *t*-test where the error of the difference is the variance of the differences across all replicate subsamples.

The difference between two statistics is itself a statistic. Therefore, the standard error of any difference between two dependent samples is computed in the same way as any other statistic, as was described earlier. The (up to) 250 sets of replicate weights produce 250 replicate estimates of the difference of interest and equations (13.1) and (13.2) apply, where t_{hi} and t_0 represent the differences between the point estimates for the two groups.

Estimating Standard Errors for Comparing Against International Average

Participating countries may be interested to compare their average achievement results to the international average achievement across TIMSS countries for that cycle. For example, the interactive feature in TIMSS 2023 Exhibit 1.1.1 can identify countries that had average mathematics achievement that was not statistically significantly different from the TIMSS 2023 International Average.

When comparing a country's result to the international average, TIMSS accounts for the fact that the country contributed to the international average and its standard error. To correct for this contribution, the standard error of the difference needs to be adjusted. The total variance of the difference $t_k - t_{intl}$, comparing country k to the international average for a statistic t, is given by

$$Var_{tot}(t_k - t_{intl}) = Var_{tot}(t_{intl}) + \frac{(N-1)^2 - 1}{N^2} Var_{tot}(t_k) \quad (13.4)$$

where N is the number of countries contributing to the international average, $Var_{tot}(t_{intl})$ is the total variance of the international average as computed by equation (13.3), and $Var_{tot}(t_k)$ is the total variance for country k as computed by equation (13.2) for results based on plausible values, or equation (13.1) for results not based on plausible values.

Equation (13.4) can be simplified and expressed in terms of standard errors as follows:

$$SE(t_k - t_{intl}) = \sqrt{SE(t_{intl})^2 + \frac{N-2}{N} SE(t_k)^2}$$

where $SE(t_{intl})$ is the standard error of the international average and $SE(t_k)$ is the standard error for country k.

References

Efron, B. (1982). T*he jackknife, the bootstrap, and other resampling plans*. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM).

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.

Qian, J. (1998). Estimation of the effective degrees of freedom in t-type tests for complex data. In *Proceedings of the Section on Survey Research Methods, American Statistical Association* (pp. 704–708). http://www.asasrms.org/Proceedings/papers/1998_119.pdf

Quenouille, M. H. (1949). Problems in plane sampling. *The Annals of Mathematical Statistics, 20*(3), 355–375. https://doi.org/10.1214/aoms/1177729989

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons.

Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika, 6*, 309–316. https://doi.org/10.1007/BF02288586

Tukey, J. W. (1958). Bias and confidence in not quite large samples. *The Annals of Mathematical Statistics, 29*(2), 614. https://doi.org/10.1214/aoms/1177706647

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

↓ Appendix 13A: Number of Schools and JRR Sampling Zones in the TIMSS 2023 Samples

↓ Appendix 13B: Summary Statistics and Standard Errors for Proficiency – Grade 4 Mathematics

↓ Appendix 13C: Summary Statistics and Standard Errors for Proficiency – Grade 4 Science

↓ Appendix 13D: Summary Statistics and Standard Errors for Proficiency – Grade 8 Mathematics

↓ Appendix 13E: Summary Statistics and Standard Errors for Proficiency – Grade 8 Science