Haberman *et al.*
*Large-scale Assessments in Education* (2024) 12:29
https://doi.org/10.1186/s40536-024-00214-x

Large-scale Assessments in Education

RESEARCH Open Access

# Teacher-centered analysis with TIMSS and PIRLS data: weighting approaches, accuracy, and precision

Shelby J. Haberman[1], Sabine Meinck[2,3*] and Ann-Kristin Koop[2]

*Correspondence:
ann-kristin.koop@iea-hamburg.de

[1] Haberman Statistics, Brookline, MA, USA
[2] Research and Analysis Unit , International Association for the Evaluation of Educational Achievement (IEA), Hamburg, Germany
[3] Sampling Unit, International Association for the Evaluation of Educational Achievement (IEA), Hamburg, Germany

## Abstract

This paper extends existing work on teacher weighting in student-centered surveys by looking into aspects of practical implementation of deriving and using weights for teacher-centered analysis in the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS). The formal conditions to compute teacher-centered weights are detailed, including mathematical equations. We provide a proposal on how to define the targeted populations as well as how to collect data that is needed to derive teacher-centered weights, yet currently unavailable. We also tackle the issue of teacher nonresponse by proposing a respective adjustment factor, as well as mentioning the challenge of multiple selection probabilities when teachers teach in multiple schools. The core part of the paper focuses on studying the level of accuracy that can be expected when estimating teacher population characteristics. We use TIMSS 2019 data and simulate likely scenarios regarding the variance in weights. The results show that (i) the different weighting scenarios lead to relatively similar estimates; however, the differences between the scenarios are sufficient to justify the recommendation to use correctly derived teacher weights; (ii) differences between estimated standard errors based on complex sampling and corresponding estimates based on simple random sampling are sufficiently consistent to support use of a procedure to estimate standard errors that accounts for both sample weights and the complex sampling design; (iii) sample sizes and variance in weights significantly limit estimate precision, so that total population estimates with sufficient precision are available in the majority of countries but subpopulation features are generally not sufficiently precise. To provide a critical evaluation of our results, we recommend implementation of the proposed method in one or more countries. This recommended study will permit examination of logistical considerations in implementation of required changes in data acquisition and will provide data to replicate the analysis with teacher-centered weights.

**Keywords:** Complex sampling, Weights, Teachers, TIMSS, PIRLS

## Introduction

Many contemporary international large-scale assessments (ILSA), for example the Trends in International Mathematics and Science Study (TIMSS, Martin et al., 2020), the Progress in International Reading Literacy Study (PIRLS, Martin et al., 2017), and

the Programme for International Student Assessment (PISA, OECD, 2019a), investigate student populations. Others cover teachers, the most prominent one is the Teaching and Learning International Survey (TALIS, OECD, 2019b). There is a third type of ILSA that attempts to cover both teacher and student populations within one study, requiring compromises regarding the optimization of the sampling designs. Examples for such studies are the International Civic and Citizen Study (ICCS, Schulz et al., 2018) and the International Computer and Information Literacy Study (ICILS, Fraillon et al., 2020), which both target eighth grade students and their teachers and aim for fully representative samples for both groups. While this solution sounds intriguing and cost-efficient, it comes with a severe disadvantage, that is, there is no direct linkage between teachers and students, hence, for example, teachers' attitudes and teaching styles cannot be related directly to their students' characteristics and outcomes.

TIMSS and PIRLS are among the most well-known ILSAs in the world, with more than 50 participating countries and educational systems. Since 1995, TIMSS every four years has investigated attainment in mathematics and science of students in fourth and eighth grades. Since 2001, PIRLS every five years has studied reading literacy of students in fourth grade. A rich array of contextual information is gathered in both studies from both the students themselves and individuals involved in students' learning: school principals, parents, and teachers of the sampled students. Even though TIMSS and PIRLS are designed to provide information on student learning, and analyzing teacher-level characteristics is not part of the studies' analytical objectives, scholars are interested to use the information that is collected from teachers. However, analyzing teacher data from these studies is not straightforward. In this paper, we consider TIMSS 2019. This survey provides summary results for teachers on variables ranging from years of experience to job satisfaction for different educational systems, subjects taught, and grade taught. For example, the average years of experience in Albania of a student's mathematics teacher in Grade 4 is estimated to be 22 (Mullis et al., 2020, page 390). This average does not necessarily estimate in Albania the mean years of experience of a mathematics teacher for Grade 4. Instead the reported average estimates a weighted mean of years of experience. For a given teacher, the weight is a sum over students taught in a given grade and subject of the fraction of instruction provided. In the Albanian example, each student has only one mathematics instructor, so that the weight is proportional to the number of students taught. The TIMSS 2019 User Guide (Fishbein et al., 2021) warns users of the TIMSS 2019 database of this difference between these two averages:

> *The teachers in the TIMSS 2019 International Database do not constitute representative samples of teachers in the participating countries. Rather, they are the teachers of nationally representative samples of students. Therefore, analyses with teacher data should be made with students as the units of analysis and reported in terms of students who are taught by teachers with a particular attribute. (Fishbein et al., 2021, p. 13)*

This warning reflects two distinct issues. The sampling design does not ensure that sampled teachers are a representative sample of all teachers in an educational system, and the data collection does not permit a weighting adjustment to allow use of the sampled teachers to estimate mean characteristics of the population of teachers.

Although TIMSS emphasizes assessment of achievement of students, in line with Hooper et al. (2022), we argue that simple modifications of forms provided by participating schools permit development of teacher-centered sampling weights that allow use of the sample of teachers in TIMSS and PIRLS for estimation of means of characteristics of the teacher populations of participating educational systems.

In this paper, we will start by proposing a teacher population definition for the surveyed grades and subjects in TIMSS. Next, we will briefly review weighting in TIMSS and current inferences that implicitly use sample student weights to provide sample teacher weights. We refer to these weights hereafter as student-centered teacher weights (s-tchwgt). They are useful for research questions dealing with the relationship between teachers and students. By revisiting the results from Hooper et al. (2022), we will then introduce sample teacher-centered teacher weights (t-tchwgt) that can be used if the interest is on teachers themselves rather than on their students.

Thereafter, we will apply the findings of Hooper et al. (2022) to determine how to obtain the information needed to derive teacher-centered weights and how to examine accuracy of estimates based on t-tchwgt. Because the current data from TIMSS and PIRLS do not now permit application of the approaches proposed by Hooper et al. (2022) from a theoretical perspective, results of a simulation study will be presented examining the expected precision of the proposed teacher-centered estimates. To inform this simulation, we considered existing data from TIMSS 2019. Complications such as weight adjustments for non-response and multiple chances of selections when teachers teach in multiple schools will be considered. The paper will close with conclusions concerning the feasibility in practice of teacher-centered estimates and with recommendations concerning implementation of such estimates.

Because TIMSS and PIRLS use the same sampling design (Joncas and Foy, 2012), the findings of this research are fully applicable to other iterations of TIMSS, and to PIRLS. The notation we use for our paper can be found in Table 14.

### Defining international target populations of teachers for TIMSS and PIRLS

The introduction of revised teacher weights in TIMSS will facilitate analyses on the teacher level without the need to use students as units of analysis and reporting. To draw direct conclusions about a teacher population with equally weighted teachers, it is important to agree on an unambiguous definition of this population. This section attempts a proposal for such definition in line with the assumptions in the remainder of this paper. According to the authors' knowledge, there is no explicit definition of the population of teachers in either TIMSS or PIRLS. However, as specified in the TIMSS technical documentation (LaRoche et al., 2020), TIMSS invites all mathematics and science teachers of the selected classes to participate. The same applies for reading/language teachers of the participating PIRLS classes (Martin et al., 2017). To allow the current selection mechanism to align with the procedures proposed in this paper, we suggest to include all mathematics and science teachers who instruct students in the target grade, i.e., fourth and/or eighth grade for TIMSS, and all reading/language

teachers of fourth-graders for PIRLS. The proposed definition corresponds to the following TIMSS and PIRLS international target population definition of students:[1]

> *Fourth grade (TIMSS and PIRLS)*
>
> *All students enrolled in the grade that represents four years of schooling counting from the first year of ISCED Level 1, providing the mean age at the time of testing is at least 9.5 years (LaRoche et al., 2020, sect. 3.4)*
>
> *Eighth grade (TIMSS only)*
>
> *All students enrolled in the grade that represents eight years of schooling counting from the first year of ISCED Level 1, providing the mean age at the time of testing is at least 13.5 years (LaRoche et al., 2020, sect. 3.4)*

To these student target populations correspond four distinct teacher target populations in TIMSS: mathematics teachers of fourth-grade classes, science teachers of fourth-grade classes, mathematics teachers of eighth-grade classes, and science teachers of eighth-grade classes; and one teacher target population in PIRLS: reading/language teachers of fourth-grade classes, as follows:

> *Fourth grade (TIMSS and PIRLS; mathematics, science, and reading/language teachers)*
>
> *All teachers teaching mathematics [science, reading/language] to students enrolled in the grade that represents four years of schooling counting from the first year of ISCED Level 1, providing the student mean age at the time of testing is at least 9.5 years (LaRoche et al., 2020, sect. 3.4)*
>
> *Eighth grade (TIMSS only; mathematics and science teachers)*
>
> *All teachers teaching mathematics [science] to students enrolled in the grade that represents eight years of schooling counting from the first year of ISCED Level 1, providing the student mean age at the time of testing is at least 13.5 years (LaRoche et al., 2020, sect. 3.4)*

It is important to note that the teacher target populations are not mutually exclusive; e.g., a mathematics teacher of fourth-grade students can also be a science teacher of eighth-grade students, or a teacher might teach multiple subjects to the same class. Moreover, teachers can teach at different schools. All teachers are considered equally, regardless of the hours taught. We further suggest to define the subjects *science* and *mathematics* based on the content domains of the assessment. Thus, subjects related to mathematics must cover at least one of the following content domains: number, measurement, geometry, algebra, data, or probability (Lindquist et al., 2017). Subjects related to science must cover at least one of the following content domains: life science, biology, chemistry, physical science, physics, or earth science (Centurino and Jones, 2017). Even though we have tried to give as accurate a definition as possible, there may still be contested cases. For example, if several teachers teach the same subject to the same class, the general rule

[1] International Standard Classification of Education (ISCED).

is that all teachers are part of the target population. We propose that a teacher associated with a class is not considered part of the target population only if one of the following conditions applies: the teacher is not at all involved in instructing the students, the teacher clearly only has a supporting role, the teacher is in training, or the teacher's role in delivering instruction is otherwise very limited. Furthermore, in accordance with the proposed definition, teachers who do not teach the respective target grade and/or subject during the TIMSS testing period are not considered part of the target population.

Due to the multistage sampling procedure of TIMSS and PIRLS, the listing of teachers is inter-related with the sampling of schools and classes. In order not to jeopardize the core objectives of the studies and to keep procedures simple and cost-efficient, exclusion criteria for teachers must align with the exclusion criteria for schools and classes. Thus, teachers are excluded if they only instruct students in excluded schools or excluded classes. For instance, to a limited extent, TIMSS and PIRLS permit countries to exclude very small schools. At the class level, participating countries are allowed to exclude classes in which all students are either non-native speakers or have functional or intellectual disabilities.

#### Weighting in TIMSS

In this section, we will summarize the usual sampling procedures applied in TIMSS (Joncas and Foy, 2012), as this knowledge is built upon in the following sections.

In TIMSS, multistage sampling is used to obtain student samples for assessment of achievement in mathematics and science in the fourth and eighth grade (LaRoche et al., 2020). This procedure is not designed to facilitate sampling of teachers. To consider procedural changes to facilitate inferences on teachers, we examine the sampling procedure used in TIMSS for an educational system with $N$ schools, $H$ strata, $C$ classes, and $S$ students in the target grade. At the initial stage, within stratum $h$, schools are sampled with probability proportional to size (PPS), where ideally the size measure for a school $i$ is defined as the number of students $S_{hi}$ in the target grade. A school and two replacement schools are selected simultaneously from the $N_h$ schools in the stratum. The original school is used if it participates. The first replacement school is used if the original school does not participate but the first replacement school does. The second replacement school is used if neither the original school nor the first replacement school participates but the second replacement school does. After adjustments for nonresponse, participating sampled school $i$ from explicit stratum $h$ has a sampling weight $F_{hi1} = A_{h1}M_h/(n_h m_i)$. This weight involves the size measure $m_i$ for sampled school $i$, the sum $M_h$ of size measures for all schools in stratum $h$, and the school non-participation adjustment $A_{h1}$ for stratum $h$. For stratum $h$, the adjustment $A_{h1}$ depends on the number $n_h$ of participating sampled schools and the number $n_{hnr}$ of cases in which neither the originally sampled school nor its two replacement schools participated. The adjustment $A_{h1} = (n_h + n_{hnr})/n_h$. If schools in stratum $h$ are certain to participate, then $A_{h1}$ is always 1 and the inverse of $F_{hi1}$ is the exact probability that school $i$ participates. The mechanisms used in TIMSS for adjusting nonresponse are based on the assumption that observations are missing at random within the adjustment cells. However, since this assumption cannot be definitively proven, strict requirements on participation rates are enforced Meinck (2015a).

Within a school, classes are usually randomly drawn with equal probability of selection, and the class then has a weight inversely proportional to its probability of selection. As in the case of sampled schools, adjustment is made for non-participation. Let $\delta_i$ be the number of participating classes in school $i$ out of the number $c_i$ of sampled classes. Let $C_i$ be the total number of eligible classes in school $i$. Let $A_{h2}$, the class non-participation adjustment for stratum $h$ be $n_h$ divided by the sum over participating schools $i$ in the stratum of the class participation fractions $\delta_i/c_i$. The class weight component for sampled class $j$ of sampled school $i$ is then $F_{hij2} = A_{h2}C_i/c_i$. The overall weighting of class $j$ of school $i$ is $G_{hij2} = F_{hi1}F_{hij2}$. The inverse of $G_{hij2}$ estimates the joint probability that school $i$ and class $j$ are both sampled and participate.

In some cases, classes within schools are divided into strata, and classes are randomly selected within strata. This approach could be used, for example, if schools have classes with different language of instruction, and they aim for a specific sample size for both languages. Such stratification of classes within schools is used by some countries in recent TIMSS and PIRLS studies. Simple changes in arguments must then be made.

Within classes, let $n_{ij}$ be the number of students in the class, let $n_{ij1}$ be the number of selected students in the class, let $n_{ij3}$ be the number of selected students in the class who participate, and let $n_{ij2}$ be the number of students sampled who might have participated. (It is possible due to class changes that $n_{ij2}$ and $n_{ij1}$ differ.) Students who are selected and participate receive weight component $F_{ij3} = (n_{ij}/n_{ij1})(n_{ij2}/n_{ij3})$. The final weight for a participating student is $G_{hij3} = G_{hij2}F_{ij3}$. The inverse of $G_{hij3}$ is the estimated joint probability that student $k$ is a sampled and participating member of sampled and participating class $j$ from sampled and participating school $i$. If non-participation does not exist for schools and classes and all students in a class are sampled, then the student weight $G_{hij3}$ reduces to $M_hC_i/(n_hm_ic_i)$. TIMSS also allows subsampling of students within classes. In this case, classes are sampled with PPS and students within classes are sampled with systematic simple random sampling (systematic SRS). This procedure was however used exclusively for Singapore during the last cycles of the studies. For simplicity we do not extend the paper for this special case; however, such an extension is straightforward. Let $Y$ be a real student measurement variable with value $Y_{ijk}$ for student $k$ from class $j$ of school $i$, and let $\bar{Y}$ be the mean of the $S$ values of $Y$. The estimated mean $\bar{Y}_s$ is then the ratio estimate with numerator equal to the sum of $G_{hij3}Y_{ijk}$ over observed students $k$, classes $j$, and school $i$ for which $Y_{ijk}$ is available and denominator equal to the corresponding sum of $G_{hij3}$ over observed students $k$, classes $j$, and school $i$ for which $Y_{ijk}$ is available (Hájek, 1971).

#### Two types of teacher weights: student- and teacher-centered weights

Scholars familiar with the TIMSS data will be aware that teacher weights are already provided in publicly-available data files. In this research paper, however, we distinguish two types of teacher weights. The teacher weights that are already available are linked to the students of the responding teachers. These weights are labeled teacher weights (*TCHWGT*) in the TIMSS 2019 data base. To emphasize their relation with the student population, we call these weights student-centered teacher weights (s-tchwgt). If s-tchwgt is used, students are the units of analysis. These weights are derived by dividing the final student estimation weight by the number of teachers related to

an individual student. For example, suppose a student has a final weight of 10 and two science teachers. In this case, the student dataset is duplicated and merged to the data of both teachers, and s-tchwgt for each case in the resulting file has a value of 10/2=5. As pointed out in the introduction, this weight is useful to describe average features of target grade students. It allows statements such as: "50 percent of students in country X have science teachers with a postgraduate degree."

The second type of teacher weights, which are the subject of this research paper, provide an approach for teacher-centered analysis and will be named teacher-centered teacher weights (t-tchtwgt). With reference to the above example, t-tchwgt could be used to estimate the number of science teachers in the targeted teacher population who completed a postgraduate degree. In the following section, we will present the issue in a more formal way.

To describe the current student-centered teacher weights in TIMSS, consider a teacher variable $U$ with value $U_{it}$ for teacher $t$ in target school $i$ for a specific subject (mathematics or science). We begin with the student-centered case. For each student $k$ in class $j$ of school $i$, let $K_{ijk}$ be the number of teachers the student has for the subject under consideration. Let the student-centered population weight $W_{it}$ of teacher $t$ in school $i$ be the sum of the fractions $1/K_{ijk}$ for all students $k$ in a class $j$ who are taught by teacher $t$. The student-centered population mean $\bar{U}_W$ of the teacher variable $U$ is the ratio with numerator equal to the sum of the products $W_{it}U_{it}$ for teachers $t$ in target schools $i$ and denominator equal to the corresponding sum $S$ of the weights $W_{it}$. Recall that the target population has $S$ students. The population mean $\bar{U}_W$ is also the population mean over all students $k$ in classes $j$ in schools $i$ of the average of the $U_{it}$ for the $K_{ijk}$ teachers $t$ who instruct the student. For sampled teacher $t$ of sampled and participating school $i$, let the student-centered sampling weight $W_{its}$ be the sum of $G_{hij3}/K_{ijk}$ over sampled and participating students $k$ from sampled and participating classes $j$ of school $i$ who have teacher $t$. Then the student-centered estimated mean $\bar{U}_{Ws}$ is the ratio with numerator equal to the sum of the products $W_{its}U_{it}$ over sampled teachers $t$ from sampled and participating schools $i$ for whom $U_{it}$ is observed and denominator equal to the sum of the $W_{its}$ over sampled teachers $t$ from sampled and participating schools $i$ for whom $U_{it}$ is observed. The estimates $\bar{U}_{Ws}$ are used in TIMSS.

In the case of teacher-centered weights, let $D_i$ be the number of teachers in school $i$ for a targeted subject, let $D_+$ be the sum of the $D_i$ over all target schools $i$, and let $\Sigma(U)$ be the total of the $U_{it}$ for the $D_+$ teachers $t$ in target schools $i$. The teacher-based mean $\bar{U}$ of the teacher variable $U$ for teachers $t$ in target schools $i$ is just the sample mean of the $U_{it}$ over teachers $t$ in schools $i$. With current data, $\bar{U}$ cannot be estimated. Nonetheless, it is possible to consider how $\bar{U}$ and $\bar{U}_W$ compare. To aid in comparison, let $V_{it} = D_+W_{it}/S$ be the adjusted student-centered population weight, so that the average $\bar{V}$ of the $V_{it}$ is 1. Then $\bar{U}$ is the average of the products $\bar{V}U_{it}$, while $\bar{U}_W$ is the average of the products $V_{it}U_{it}$. If either the student-centered population weights $W_{it}$ are constant, so that each $W_{it}$ is the average number $S/D_+$ of students per teacher, or the variables $U_{it}$ are constant, so that each $U_{it}$ is $\bar{U}$, then $\bar{U}_W$ and $\bar{U}$ are equal. Arguments here are most appropriate if no teachers teach the same target subject in the same grade at more than one school. Otherwise, some modifications are required.

To establish an upper bound on the difference $|\bar{U}_W - \bar{U}|$ for the case in which neither the teacher variables $U_{it}$ nor the student-centered population weights $W_{it}$ are constant, let $\sigma(U)$ be the population standard deviation of the teacher variables $U_{it}$ for teachers $t$ in target schools $i$, so that $\sigma(U)$ is the square root of the mean of the squared deviations $(U_{it} - \bar{U})^2$, and let $\sigma(W)$ be the corresponding population standard deviation of the student-centered weights $W_{it}$ for teachers $t$ in schools $i$. By assumption, both $\sigma(U)$ and $\sigma(W)$ are positive. Let the population correlation coefficient of the $U_{it}$ and $W_{it}$ be $\rho(U, W)$. The difference between $\bar{U}_W$ and $\bar{U}$ is the average of the products $(V_{it} - 1)U_{it}$. Because the average of the differences $(V_{it} - 1)$ is 0, the average of the products $(V_{it} - 1)\bar{U}$ is also 0. Thus the difference $\bar{U}_W - \bar{U}$ is the average of the products $(V_{it} - 1)(U_{it} - \bar{U})$. This average is the population covariance $\gamma(V, U)$ of the $V_{it}$ and the $U_{it}$. If $\rho(V, U)$ denotes the population correlation $\gamma(V, U)/[\sigma(V)\sigma(U)]$, then it follow that

$$\bar{U}_W - \bar{U} = \sigma(V)\sigma(U)\rho(V, U)/S. \tag{1}$$

Thus a small absolute relative difference $|\bar{U}_W - \bar{U}|/\sigma(U)$ results if either the standard deviation of the adjusted weight variables $V_{it}$ is small or the absolute value of the correlation coefficient of the $V_{it}$ and $U_{it}$ is small. If all classes in the target population have only one teacher for the subject of interest and all teachers teach the same number of students, then this standard deviation is 0.

#### Teacher-centered inference: methods

A simple change in data collection permits direct study of teachers of students in the target population (Hooper et al., 2022). The key is to record, for each sampled teacher in a particular grade and subject in a participating school, the total number of classes taught by that teacher in the same school, subject, and grade. In this way, two approaches described herein have been proposed to estimate the distribution of teacher variables in the target population of teachers (Hooper et al., 2022). Horwitz-Thompson estimation (Horvitz and Thompson, 1952), which is abbreviated as HT, is a traditional method to obtain unbiased estimates of sums of population variables under sampling without replacement. The other approach, multiplicity-adjusted indirect sampling (MAIS), provides simplified analysis that involves possible multiple-counting of the same teacher. Both approaches lead to unbiased estimation of sums of teacher variables in the target population if non-participation adjustments are not required. HT has the advantage of fixed weights but requires simple random sampling of classes within schools. MAIS has the advantage of applicability to sampling of classes by methods not equivalent to simple random sampling. In addition, MAIS is much easier to describe, so that it will be emphasized in applications. Theoretical results are derived for variances and their estimates for both the HT and MAIS approaches, however, due to its wider applicability, the MAIS approach will be used to obtain indications of the potential accuracy of estimated means of teacher variables for individual educational systems.

Because the information required for the analysis is not currently obtained in TIMSS, analysis considers plausible scenarios for teacher weights rather than direct use of teacher weights. In addition to consideration of variances, this paper also treats

problems of teacher non-response via approaches similar to those used in TIMSS for student non-response, class non-response, and school non-response.

In both approaches under consideration, the procedure for sampling classes is the standard one in TIMSS. The two approaches HT and MAIS diverge once classes are sampled. Let $d_{it}$ of the $C_i$ classes be taught for a given subject, mathematics or science, at least in part by teacher $t$, and let $d_{its}$ of the $c_i$ sampled classes be taught by that teacher. Let $\delta_{it}$ be the number of sampled teachers who participate in school $i$. Let the teacher non-participation adjustment $A_{ht}$ in stratum $h$ be $n_h$ divided by the sum over participating schools $i$ in the stratum of the fractions $\delta_{it}/d_{it}$.

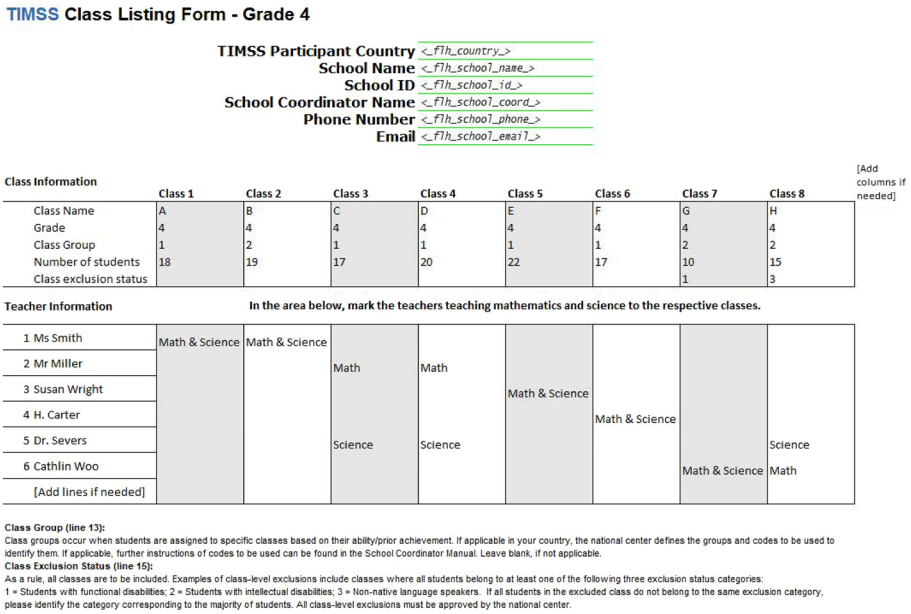
As in the development of student-centered weights, let $D_i$ be the number of teachers $t$ in the school, and let $D_+$ be the sum of the $D_i$ over schools in the target population. The challenge is estimating $\bar{U}$ by use of the participating teachers $t$ associated with the $c_i$ classes sampled from each sampled school $i$.

To describe the HT approach to teacher-centered weights, consider computing the probability that a teacher $t$ from school $i$ is in a sampled class given that school $i$ has been sampled. If $c_i$ classes are sampled randomly, so that $C_i - c_i$ classes are not sampled, then the probability $T_{it}$ that teacher $t$ is sampled is 1 if $C_i - c_i < d_{it}$. Otherwise,

$$T_{it} = 1 - \prod_{a=0}^{c_i-1} \frac{C_i - d_{it} - a}{C_i - a}. \tag{2}$$

The formula for $C_i - c_i < d_{it}$ applies because it is impossible in this case for teacher $t$ not to be sampled. The alternative case holds since the product of $C_i - d_{it} - a$ over non-negative integers $a < c_i$ is the number of ordered samples of classes of size $c_i$ that do not include teacher $t$ and the product of $C_i - a$ over non-negative integers $a < c_i$ is the total number of ordered samples of classes of size $c_i$. In the simplest case, $c_i = 1$ and $T_{it} = d_{it}/C_i$. Then the sampling weight $W_{itH} = F_{hi1}A_{ht}/T_{it}$ for participating sampled teacher $t$ from school $i$. The teacher-centered sample mean $\bar{U}_H$ based on the HT approach is then the ratio estimate with numerator equal to the sum of the products $W_{itH}U_{it}$ over participating sampled teachers $t$ in participating and sampled schools $i$ for which $U_{it}$ is observed and denominator equal to the sum of the $W_{itH}$ over participating sampled teachers $t$ in participating and sampled schools $i$ for which $U_{it}$ is observed. As expected from Horwitz-Thompson estimation, for a school $i$ with no non-participation of teachers and all $U_{it}$ observed for sampled teachers, the sum of $U_{it}/T_{it}$ over sampled teachers $t$ estimates the sum $U_{i+}$ of $U_{it}$ over all targeted teachers $t$ in the school. The sum of the products $W_{itH}U_{it}$ over sampled and participating teachers $t$ in sampled and participating schools $i$ then estimates the sum of the $U_{it}$ over all teachers $t$ in schools $i$ from the target population.

In the MAIS approach, the sample weight $W_{itM} = G_{ih2}A_{ht}d_{its}/d_{it}$ if teacher $t$ is sampled and participates in sampled and participating school $i$. The teacher-centered sample mean $\bar{U}_M$ based on the MAIS approach is then the ratio with numerator equal to the sum of the products $W_{itM}U_{it}$ over participating sampled teachers $t$ in participating and sampled schools $i$ for which $U_{it}$ is observed and denominator equal to the sum of the $W_{itM}$ over participating sampled teachers $t$ in participating and sampled schools $i$ for which $U_{it}$ is observed. If $d_{it} > 1$ for a sampled teacher $t$ in school $i$, then



**TIMSS Class Listing Form - Grade 4**

| | |
|---|---|
| **TIMSS Participant Country** | <_flh_country_> |
| **School Name** | <_flh_school_name_> |
| **School ID** | <_flh_school_id_> |
| **School Coordinator Name** | <_flh_school_coord_> |
| **Phone Number** | <_flh_school_phone_> |
| **Email** | <_flh_school_email_> |

**Class Information** [Add columns if needed]

| | Class 1 | Class 2 | Class 3 | Class 4 | Class 5 | Class 6 | Class 7 | Class 8 |
|---|---|---|---|---|---|---|---|---|
| Class Name | A | B | C | D | E | F | G | H |
| Grade | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
| Class Group | 1 | 2 | 1 | 1 | 1 | 1 | 2 | 2 |
| Number of students | 18 | 19 | 17 | 20 | 22 | 17 | 10 | 15 |
| Class exclusion status | | | | | | | 1 | 3 |

**Teacher Information** **In the area below, mark the teachers teaching mathematics and science to the respective classes.**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1 Ms Smith | Math & Science | Math & Science | | | | | | |
| 2 Mr Miller | | | Math | Math | | | | |
| 3 Susan Wright | | | | | Math & Science | | | |
| 4 H. Carter | | | | | | Math & Science | | |
| 5 Dr. Severs | | | Science | Science | | | | Science |
| 6 Cathlin Woo | | | | | | | Math & Science | Math |
| [Add lines if needed] | | | | | | | | |

**Class Group (line 13):**
Class groups occur when students are assigned to specific classes based on their ability/prior achievement. If applicable in your country, the national center defines the groups and codes to be used to identify them. If applicable, further instructions of codes to be used can be found in the School Coordinator Manual. Leave blank, if not applicable.
**Class Exclusion Status (line 15):**
As a rule, all classes are to be included. Examples of class-level exclusions include classes where all students belong to at least one of the following three exclusion status categories:
1 = Students with functional disabilities; 2 = Students with intellectual disabilities; 3 = Non-native language speakers. If all students in the excluded class do not belong to the same exclusion category, please identify the category corresponding to the majority of students. All class-level exclusions must be approved by the national center.

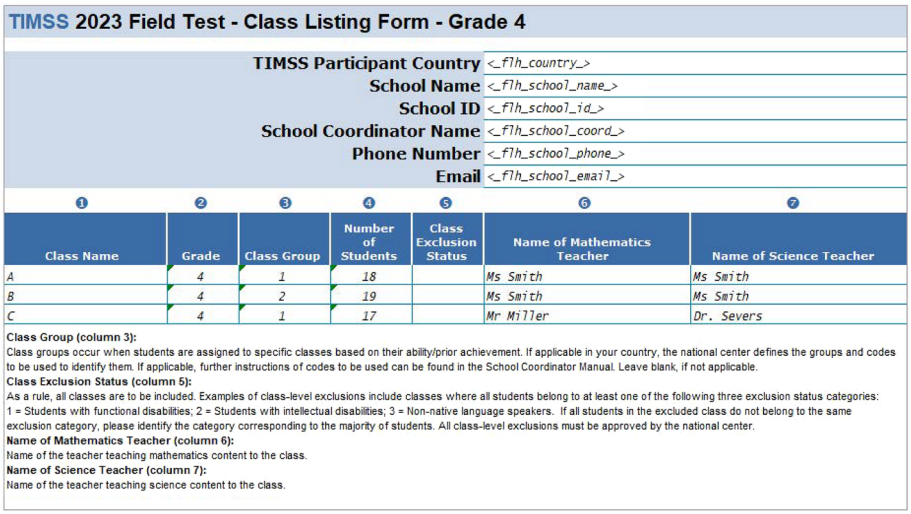
**Fig. 1** TIMSS fourth grade adjusted class listing form

the count $d_{its}$ and the sample weight $W_{itM}$ are not constant. Nonetheless, $d_{it}c_i/C_i$ is the expected value of the number $d_{its}$ of times teacher $t$ teaches a sampled class. This expected value is also the product of the probability $T_{it}$ that $d_{its} > 0$ and the expected value of $d_{its}$ given that $d_{its} > 0$. It follows that $d_{its}$ given that $d_{its}$ is positive has expected value $d_{it}c_i/C_i$, so that $W_{itM}$ and $W_{itT}$ have the same expected value given selection of teacher $t$. As a consequence, both the MAIS and HT approaches provide comparable estimates of the teacher-centered mean $\bar{U}$. Although the simpler form of the MAIS estimate is an attraction in a comparison with the HT estimate, a more important consideration is that MAIS can be employed when simple random sampling of classes is not present as long as the expected value of $d_{its}$ is $d_{it}c_i/C_i$. The HT approach must be modified if simple random sampling of classes is not employed within schools.

In a number of cases, the HT and MAIS approaches coincide. If, for all schools $i$, either the number of sampled classes $c_i$ is 1, $c_i = C_i$, or the number $d_{it}$ of classes each teacher $t$ instructs is always 1, then $W_{itH} = W_{itM}$ for all sampled and participating teachers $t$ and $\bar{U}_M = \bar{U}_H$.

### Teacher-centered inferences in TIMSS: changes needed in data collection

Although the current sampling procedure and data collection in TIMSS do not permit simple inferences about the distribution of characteristics for teachers who participate in instruction of mathematics or science in the fourth or eighth grade, it is possible to add a new school-level form to permit such inferences without changing other aspects of sample design and data collection described in Johansone (2020). For each grade examined (4 or 8), the required new form for a participating school $i$ includes a list of the $C_i$ classes eligible for sampling. The list specifies for each eligible class all teachers of mathematics or science who instruct at least some class students.



**TIMSS 2023 Field Test - Class Listing Form - Grade 4**

| | |
|---|---|
| **TIMSS Participant Country** | <_flh_country_> |
| **School Name** | <_flh_school_name_> |
| **School ID** | <_flh_school_id_> |
| **School Coordinator Name** | <_flh_school_coord_> |
| **Phone Number** | <_flh_school_phone_> |
| **Email** | <_flh_school_email_> |

| ❶ Class Name | ❷ Grade | ❸ Class Group | ❹ Number of Students | ❺ Class Exclusion Status | ❻ Name of Mathematics Teacher | ❼ Name of Science Teacher |
|---|---|---|---|---|---|---|
| A | 4 | 1 | 18 | | Ms Smith | Ms Smith |
| B | 4 | 2 | 19 | | Ms Smith | Ms Smith |
| C | 4 | 1 | 17 | | Mr Miller | Dr. Severs |

**Class Group (column 3):**
Class groups occur when students are assigned to specific classes based on their ability/prior achievement. If applicable in your country, the national center defines the groups and codes to be used to identify them. If applicable, further instructions of codes to be used can be found in the School Coordinator Manual. Leave blank, if not applicable.
**Class Exclusion Status (column 5):**
As a rule, all classes are to be included. Examples of class-level exclusions include classes where all students belong to at least one of the following three exclusion status categories: 1 = Students with functional disabilities; 2 = Students with intellectual disabilities; 3 = Non-native language speakers. If all students in the excluded class do not belong to the same exclusion category, please identify the category corresponding to the majority of students. All class-level exclusions must be approved by the national center.
**Name of Mathematics Teacher (column 6):**
Name of the teacher teaching mathematics content to the class.
**Name of Science Teacher (column 7):**
Name of the teacher teaching science content to the class.

**Fig. 2** TIMSS 2023 fourth grade class listing form

Figure 1 presents an example of such a listing form. It would replace the currently used class listing form presented in Fig. 2). We acknowledge that this list is more complex than the current class listing form and requires some additional work by the school coordinators. We therefore recommend a field trial to provide a thorough usability test. With the new listing form, it is straightforward to determine the number $d_{it}$ of classes taught, at least in part, by a teacher $t$ in school $i$. It is quite common in the fourth grade to have a single teacher who provides all mathematics and science instruction for a class. In this case, values of $d_{it}$ will typically be small. On the other hand, it is much less common in the eighth grade for only a single teacher to provide all mathematics and science instruction for a class. Thus larger values of $d_{it}$ may be encountered. Given the new form, no other procedures in TIMSS need be changed in order to replace student-centered weights by teacher-centered weights.

### Adjustment for teachers in multiple schools

If a teacher works in the target grade and subject in more than one school in the target population, then the selection probability is affected. We propose to handle this situation as done in other studies like ICCS (Zuehlke and Vandenplas, 2011), ICILS (Meinck and Cortes, 2015), and TALIS (OECD, 2014). This is, we propose to add in the teacher questionnaire the question: "At the moment, in how many other schools do you teach mathematics [/science] to target grade students?". Based on the response, another weight adjustment factor would be included into the computation of the teacher weights, calculated as the inverse of the total number of schools a teacher teaches target grade students in the respective subject. E.g., the total weight of a science teacher teaching this subject to target grade students in two schools will be halved. Note that this weight adjustment factor is called the "teacher multiplicity factor" or "teacher multiplicity adjustment" in the studies cited above, but should not be confused with the multiplicity adjustment of the MAIS approach. Both address the issue of multiple selection probabilities of teachers, the difference however is that one handles multiple selection probabilities within the

**Table 1** Number of TALIS 2018 participating education systems having a specified weighted percentage of teachers working at multiple schools

| % | Mathematics teachers | | Science teachers | |
|---|---|---|---|---|
| | ISCED 1 | ISCED 2 | ISCED 1 | ISCED 2 |
| < 5 | 13 | 34 | 13 | 31 |
| 5–10 | 0 | 9 | 0 | 9 |
| >10 | 1 | 4 | 1 | 7 |
| *N* | 14 | 47 | 14 | 47 |

sampled school, and the other one in different schools (whether sampled or not). For a more formal description of the computation see, e.g., Meinck and Cortes (2015).

To gain insights if weight adjustments for teachers working at more than one schools would be needed in practice, we analyzed the TALIS 2018 database[2]. TALIS is a teacher and school leader survey with 48 participating education systems in the 2018 cycle. The core target population is lower secondary school teachers (ISCED level 2), but countries can also survey lower and upper secondary schools (ISCED 1 and 3). For each education system a sample of about 200 schools and 20 teachers per school was drawn (OECD, 2019b). Table 1 shows the number of TALIS 2018 participating education systems that have a specified weighted percentage of teachers who indicated working at more than one school. The weighted percentage of teachers reporting working at more than one school is less than five for most of the education systems. But there are also education systems in all four groups for which the estimated percentage of such teachers exceeds 10. Note that it is likely to happen even more rarely that teachers teach the TIMSS and PIRLS target grades in multiple schools, as ISCED levels cover multiple grades while TIMSS and PIRLS cover just one grade. This finding implies that weight adjustments might be necessary for only a limited number of educational systems, and it supports our decision to ignore this issue for the study following later.

### Sample sizes

To explore the use of TIMSS and PIRLS data for the practical implementation of teacher-centered weights, the teacher sample sizes of both studies were investigated by using the TIMSS 2019[3] and PIRLS 2016[4] databases. The TIMSS 2019 sample sizes for teachers and schools, were calculated separately for each participating country or benchmarking system[5] and for each of the four defined populations (see Table 13 in the Appendix). Within each population, only unique teacher identifiers (IDs, variable *IDTEACH* in the TIMSS and PIRLS databases) and unique school IDs (variable *IDSCHOOL* in the TIMSS and PIRLS databases) were considered. One result of this approach is that a teacher of two sampled classes is only considered as one teacher in

[2] OECD, TALIS 2018 Database, https://www.oecd.org/education/talis/talis-2018-data.htm (assessed on July 21st, 2022).

[3] TIMSS 2019 International Database, https://www.iea.nl/data-tools/repository/timss, (assessed on January 21st, 2022).

[4] PIRLS 2016 International Database, https://www.iea.nl/data-tools/repository/pirls, (assessed on January 21st, 2022).

[5] Since TIMSS 2003, TIMSS introduced a so-called Benchmarking Program, which also allows sub-entities of countries to participate in the survey (Martin and Mullis, 2004). We will use the term educational system for a participating country or benchmarking system in the following.

**Table 2** Number of TIMSS 2019 educational systems by teacher sample size (categorized)

| Sample size | <150 | 150–199 | 200–249 | 250–299 | 300–349 | ≥ 350 | N |
|---|---|---|---|---|---|---|---|
| *Grade 4* | | | | | | | |
| Mathematics | 0 | 14 | 27 | 8 | 6 | 9 | 64 |
| Science | 1 | 18 | 22 | 8 | 6 | 9 | 64 |
| *Grade 8* | | | | | | | |
| Mathematics | 2 | 13 | 15 | 5 | 2 | 9 | 46 |
| Science | 2 | 7 | 11 | 5 | 2 | 19 | 46 |

**Table 3** TIMSS 2019: summary of teacher sample sizes

| | |
|---|---|
| Grade 4: Mathematics teachers | |
| Min | 155 (Pakistan) |
| Max | 1073 (United Arab Emirates) |
| International mean | 273 |
| Range | 41 of 64 education systems have a sample size that lies between 150 and 249 teachers. |
| Grade 4: Science teachers | |
| Min | 145 (Hong Kong) |
| Max | 1036 (United Arab Emirates) |
| International mean | 266 |
| Range | 40 of 64 education systems have a sample size that lies between 150 and 249 teachers. |
| Grade 8: Mathematics teachers | |
| Min | 142 (England) |
| Max | 1036 (United Arab Emirates) |
| International mean | 271 |
| Range | 28 of 46 education systems have a sample size that lies between 150 and 249 teachers. |
| Grade 8: Science teachers | |
| Min | 141 (England) |
| Max | 1180 (United Arab Emirates) |
| International mean | 382 teacher |
| Range | 26 of 46 education systems have a sample size that exceeds 250 teachers. 19 of these education systems have a sample size that exceeds 350 teachers. |

the calculation of the respective sample size. The same approach was taken for PIRLS, where only one teacher population would be considered, that is reading/language teachers of fourth-grade students.

In TIMSS the sample sizes of teachers vary substantially among participating education systems (summarizing statistics for the four teacher populations can be found in the Tables 2 and 3). For example, the teacher samples of fourth-grade mathematics teachers in Pakistan, Northern Ireland, and Hong Kong SAR are rather small (below 160) whereas the United Arab Emirates' sample size is 1073. Overall, the sample size exceeds, with few exceptions, 150 in all teacher populations and the minimum sample size of schools over all populations is at least 98 (Malta). This seems to be a promising finding in regard to future teacher-centered analyses. On average sample sizes vary between 266 (fourth-grade mathematics teachers) and 382 (eighth-grade science

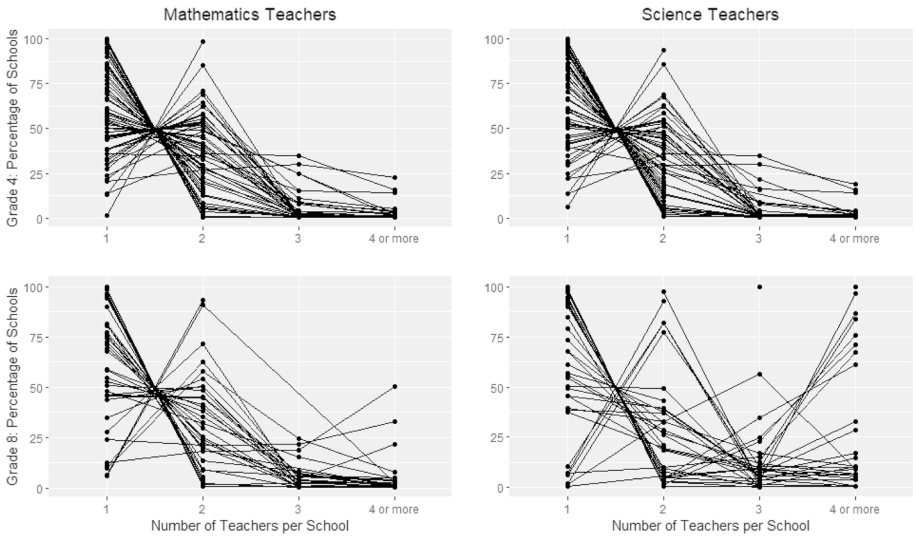



**Fig. 3** Number of participating teachers per school in TIMSS 2019 by education system

teachers). Differences in sample sizes can be explained by several factors such as the school and class sample sizes, the number of teachers associated with a class and the non-response rate.

Due to the sampling procedures in TIMSS, student sample sizes (which ultimately determine school and class sample sizes) significantly affect the size of the teacher samples, being generally positively correlated. For example, England with 3365 sampled students has the lowest student sample size in the eighth grade (Martin et al., 2020, Exhibit 9.6) and accordingly a below-average teacher sample size. The opposite is the case for the United Arab Emirates, where the 22,334 participating students is by far the highest student sample size in the eighth grade (Martin et al., 2020, Exhibit 9.6) and with 1036 mathematics and 1180 science teachers the largest teacher sample sizes.

A comparison of sample sizes of mathematics versus science teachers in the fourth grade shows that the two sample sizes do not differ much in most of the educational systems. This result is partly due to an overlap of science and mathematics teachers in the fourth grade. In 43 educational systems more than 50% of the mathematics teachers teach science in addition; and in 18 education systems even more than 90% of the mathematics teachers teach science in addition. Exceptions are educationalsystems like Bahrain, Kuwait and South Africa. These educational systems have as many mathematics as science teachers and no overlap between these groups. When comparing educational systems that participated in both surveys, TIMSS for the fourth grade and TIMSS for eighth grade, most of them (27 out of 38) have a larger science teacher sample in the eighth grade compared to the fourth grade.

The sample sizes of teachers were also analyzed on school level. Figure 3 displays the percentages of schools with a given number of participating teachers per school in TIMSS 2019, lines combine the values for a given education system. As can be seen from the figure, there is substantial variation in between countries regarding the obtained number of teachers per school, affecting the total sample size of teachers. In the majority of sampled schools in all countries, only one or two teachers are obtained. This result

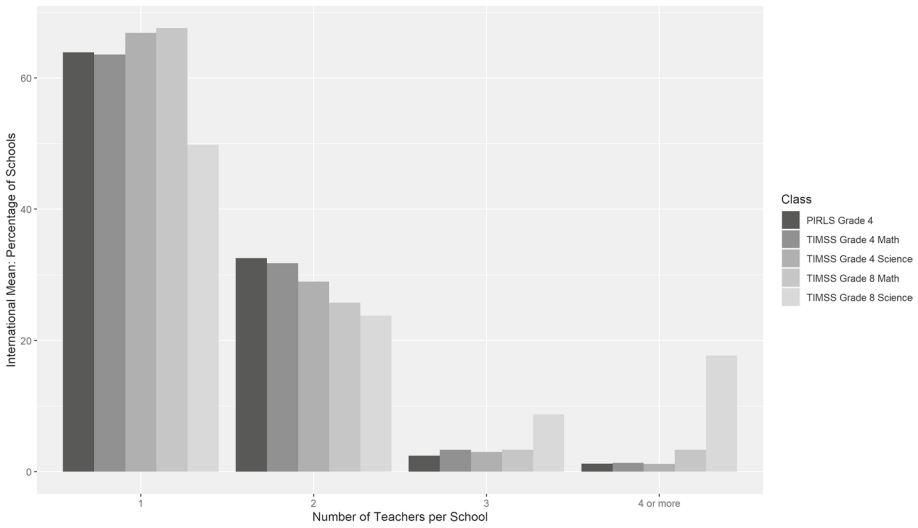



**Fig. 4** Number of participating teachers per school in TIMSS 2019 (international average)

can also be concluded from Fig. 4, which shows the international mean percentage of schools that have 1, 2, 3 or more than 3 teachers per school. The situation is slightly different when looking at eighth-grade science teachers, where data of four or more teachers is collected from each school in a significant number of countries, related to the fact that specialist teachers of the different sciences (physics, chemistry, earth science, biology etc.) exist and respond to the questionnaires. Consequently, given the current TIMSS sampling design, the sample size for the four teacher populations of interest can vary in between a minimum determined by the minimum school and class sample size (150 schools with one class in TIMSS), multiplied by the school, class and teacher participation rate, and a relatively large number in countries with large school samples, multiple selected classes within schools, or where structural conditions require multiple teachers teaching a class. Very small countries with school censuses (e.g., Malta) may have even smaller samples.

The sample sizes of fourth-grade teachers in PIRLS show similar pattern as the ones in TIMSS. Sample sizes of teachers vary between 122 (Macao SAR) and 1119 (Canada). On average educational systems have a sample size of 271 teachers. In most of the participating schools, one or two teachers participated in the survey. More information about the sample sizes in PIRLS can be found in Figs. 4, 5 and Table 13.

#### Sample variances for estimates of teacher variables

Efforts described above to achieve teacher-centered teacher weights are only reasonable if the results have an acceptable level of precision. In the following, we investigate what would be likely levels of sampling variance when estimating teacher population characteristics. Large sampling variance could be due, among other factors, to relatively small samples or relatively large variance of weights. An acceptable level of sampling variance could be determined in various ways. One standard involves the accuracy of student-centered teacher summaries that TIMSS currently reports. Another standard is based on the regular TIMSS requirements for measurement of student achievement

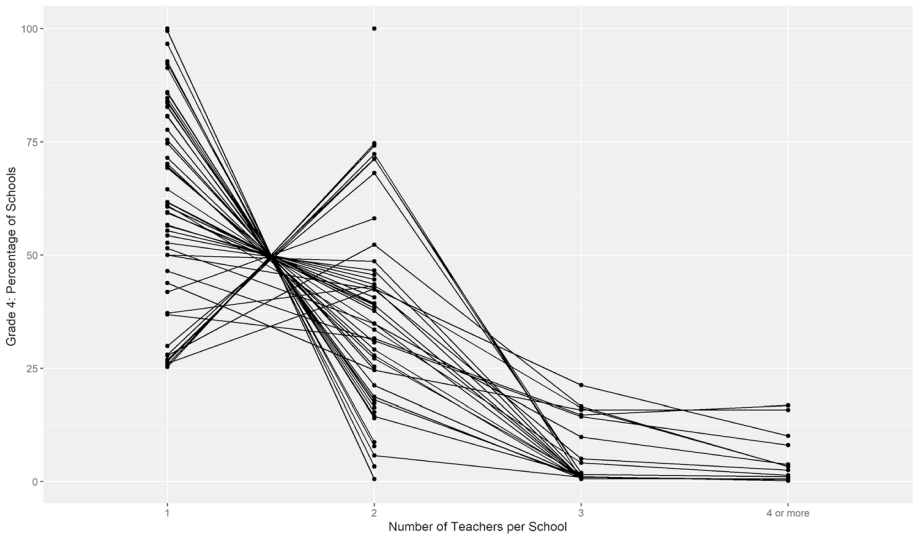



**Fig. 5** Number of participating teachers per school in PIRLS 2016 by education system

that national student samples should provide for a standard error no greater than .035 standard deviation units for the country's mean achievement. Sample estimates of any student-level percentage estimate (e.g., a student background characteristic) should have a confidence interval of ±3.5% (LaRoche et al., 2020). Given the relatively small teacher samples, this precision cannot be reached, even if the design effect of estimates associated with the teacher samples would be close to 1 due to clustering effects expected to be negligible (very small cluster sizes; teacher variables have lower intra-class correlation coefficients than student variables (Meinck, 2015b)). However, given the sample sizes presented in the Table 13 (see Appendix), many but not all precision levels can be expected to correspond to an effective sample size of at least 150, a value that translates to a standard error of .08 standard deviation units. We claim that teacher population estimates reaching these respective minimum levels of precision can be deemed satisfactory. Moreover, it might be informative to compare the sampling variance of an estimator based on teacher-centered versus student-centered teacher weights (Dumais and Morin, 2019; Schulz, 2020). We use TIMSS 2019 data for the analysis. However, because we are missing one important piece of information to compute the teacher-centered teacher weights, namely how many classes a participating teacher teaches, we consider some plausible scenarios to suggest possible results of teacher-centered weights. These scenarios clearly do not obviate the importance of a pilot study to examine teacher-centered weights, but they do provide some indication of how results for teacher-centered weights might differ from those from student-centered weights.

In this discussion, student-centered weights for teacher characteristics are computed according to current reporting practice in TIMSS 2019. For teacher-centered weights, results are obtained for approximations of the MAIS approach. We consider the following two scenarios.

**Scenario 1:** Class-centered weights. The teacher-centered MAIS weight $W_{itM}$ for teacher $t$ in school $i$ of stratum $h$ is certainly no greater than the sum $W_{itC}$ of the class weights $G_{hij2}$ for all the sampled classes $j$ associated with teacher $t$. This sum is used for

**Table 4** Items used in comparisons of teacher weights

| Item/Scale | Grade 4 | Grade 8 | Scale level | Min | Max |
|---|---|---|---|---|---|
| By the end of this school year, how many years will you have been teaching altogether? | ATBG01 | BTBG01 | Scale | 0 | 60 |
| Are you female or male? | ATBG02 | BTBG02 | D | 1 | 2 |
| How old are you? | ATBG03 | BTBG03 | C (6) | 1 | 6[a] |
| I have too much material to cover in class | ATBG09B | BTBG09B | C (4) | 1 | 4 |
| I need more time to assist individual students | ATBG09E | BTBG09E | C (4) | 1 | 4 |
| Scale: job satisfaction | ATBGTJS | See below. | S | 4.8 | 11.7 |
| | See above. | BTBGTJS | S | 5.3 | 11.7 |
| *Exclusive mathematics teacher variable* | | | | | |
| In the past two years, how many hours in total have you spent in formal (in-service/professional development, e.g., workshops, seminars, etc.) for mathematics? | ATBM10 | BTBM23 | C (5) | 1 | 5 |
| *Exclusive science teacher variable* | | | | | |
| In the past two years, how many hours in total have you spent in formal (in-service/professional development, e.g., workshops, seminars, etc.) for science? | ATBS09 | BTBS22 | C (5) | 1 | 5 |

[a]Categories have different size

S: Scale D: Dichotomous C: Categorical

In case of a categorical variable, the number of categories is given in brackets

class-centered weights. The class-centered weight $W_{itC}$ is $W_{itM}$ if teacher $t$ teaches all classes, so that $d_{its} = d_{it} = C_i$, or if teacher $t$ only teaches a single class, so that $d_{its} = d_{it}$ if $t$ is sampled.

**Scenario 2:** School-centered weights. Because the class factor $F_{hij2}$ is always at least 1, the expected value $W_{itH}$ of $W_{itC}$ for a sampled teacher $t$ is always at least as large as the school weight $F_{hi1}$. In a few educational systems participating in TIMSS 2019, $F_{hi1} = W_{itM} = W_{itH}$. This situation only applies to Malta and Pakistan for the fourth grade for mathematics and science because all classes and teachers are sampled.

To assess the accuracy of the weighted means under study, jackknife repeated replication (JRR) for schools was employed as in TIMSS 2019 and a parallel analysis (SRS) was employed based on the classical formula for estimation of the variance of a weighted mean under simple random sampling (Cochran, 1977, Chapter 6). As in the JRR results, a finite sampling correction is not used. JRR has the advantage of consistency with current practice, but it should be emphasized that the resulting estimated standard errors need not be accurate. The use of JRR and the use of SRS are both based on assumptions of random sampling with replacement that clearly do not apply given that populations of schools are finite, sampling of schools is without replacement, and sampling of schools within strata is systematic with a random start (Kish and Frankel, 1974). The issue of appropriateness of use of JRR in TIMSS also applies to existing student-centered weights. Nonetheless, the estimates may provide some guidance concerning reasonable expectations.

As an added check, unweighted results assuming simple random sampling with replacement of teachers in an educational system were obtained and both JRR and SRS were applied.

The full table of results is very large. For each of the two grades and two subjects, seven items were considered for this analysis (see Table 4; for further details on variables and

**Table 5** Unweighted and weighted means for grade 4 (average across countries)

| Subject | Item | Unweighted | Student centered | Class centered | School centered |
|---|---|---|---|---|---|
| Math | ATBG01 | 17.093 | 17.154 | 16.921 | 16.800 |
| Math | ATBG02 | 1.184 | 1.180 | 1.187 | 1.195 |
| Math | ATBG03 | 3.766 | 3.769 | 3.751 | 3.739 |
| Math | ATBG09B | 1.951 | 1.948 | 1.983 | 2.008 |
| Math | ATBG09E | 1.554 | 1.549 | 1.573 | 1.586 |
| Math | ATBGTJS | 10.080 | 10.108 | 10.100 | 10.107 |
| Math | ATBM10 | 2.729 | 2.748 | 2.726 | 2.713 |
| Science | ATBG01 | 16.696 | 16.717 | 16.521 | 16.355 |
| Science | ATBG02 | 1.187 | 1.179 | 1.186 | 1.192 |
| Science | ATBG03 | 3.750 | 3.746 | 3.732 | 3.723 |
| Science | ATBG09B | 1.985 | 1.983 | 2.020 | 2.049 |
| Science | ATBG09E | 1.582 | 1.581 | 1.602 | 1.615 |
| Science | ATBGTJS | 10.070 | 10.102 | 10.090 | 10.096 |
| Science | ATBS09 | 2.343 | 2.363 | 2.347 | 2.344 |

scales see Martin et al. (2020)). We considered exclusively items that would provide interesting information on characteristics of the teacher population such as gender, age, teaching experience, job satisfaction etc. We did not consider variables that are related to a specific class and would hence not be suitable for teacher-centered analysis. Occasionally teacher responses were missing or inconsistent. The teacher's responses for science or mathematics were defined as the average of the responses not missing if more than one teacher questionnaire was available.

For simplicity, this study primarily involves the study of weighted means; however, other summary statistics could easily be examined with the same methodology. For example, cumulative distribution functions can certainly be examined.

For the fourth grade, TIMSS 2019 provides data for 64 educational systems, while in the eighth-grade, data for 46 educational systems are available. Thus in all, our analysis results in a table with 1540 rows. Table columns include the code and name of the educational system, the grade, the subject, the number of observations with item responses, the number of observations with omitted responses, the four estimated means, and the four estimated standard errors. Hence the full table is too large for presentation in this paper; however, it is available in supplementary materials as an R data frame and as an Excel spreadsheet.

A simple summary of results for the raw means and three weighted means is provided in Tables 5 and 6. Because variables vary considerably in their ranges, corresponding summaries of weighted standard deviations are provided in Tables 7 and 8. These summaries are averages across participating educational systems for each grade, subject, and item. Thus by themselves they only provide a rough notion of results. Nonetheless it is worth noting that different weighting approaches do yield relatively similar average results across countries.

In terms of effect sizes in which the difference of means for an item, country, subject, and grade is divided by the square root of the average of the corresponding variances, the average absolute value of the effect size for student-weighted versus class-weighted means is 0.036, while the corresponding average for student-weighted

**Table 6** Unweighted and weighted means for grade 8 (average across countries)

| Subject | Item | Unweighted | Student centered | Class centered | School centered |
|---|---|---|---|---|---|
| Math | BTBG01 | 15.891 | 16.034 | 15.971 | 15.810 |
| Math | BTBG02 | 1.395 | 1.398 | 1.397 | 1.397 |
| Math | BTBG03 | 3.709 | 3.729 | 3.720 | 3.702 |
| Math | BTBG09B | 2.050 | 2.029 | 2.069 | 2.114 |
| Math | BTBG09E | 1.616 | 1.608 | 1.631 | 1.660 |
| Math | BTBGTJS | 9.958 | 9.965 | 9.968 | 9.990 |
| Math | BTBM23 | 3.300 | 3.318 | 3.284 | 3.240 |
| Science | BTBG01 | 15.464 | 15.566 | 15.491 | 15.377 |
| Science | BTBG02 | 1.364 | 1.361 | 1.362 | 1.373 |
| Science | BTBG03 | 3.709 | 3.715 | 3.713 | 3.704 |
| Science | BTBG09B | 2.048 | 2.031 | 2.075 | 2.120 |
| Science | BTBG09E | 1.641 | 1.631 | 1.661 | 1.693 |
| Science | BTBGTJS | 9.893 | 9.900 | 9.910 | 9.911 |
| Science | BTBS22 | 3.282 | 3.293 | 3.269 | 3.239 |

**Table 7** Unweighted and weighted standard deviations for grade 4 (average across countries)

| Subject | Item | Unweighted | Student centered | Class centered | School centered |
|---|---|---|---|---|---|
| Math | ATBG01 | 9.897 | 9.828 | 9.877 | 9.908 |
| Math | ATBG02 | 0.347 | 0.343 | 0.352 | 0.360 |
| Math | ATBG03 | 1.070 | 1.063 | 1.072 | 1.071 |
| Math | ATBG09B | 0.794 | 0.796 | 0.806 | 0.814 |
| Math | ATBG09E | 0.665 | 0.657 | 0.675 | 0.685 |
| Math | ATBGTJS | 1.643 | 1.626 | 1.633 | 1.619 |
| Math | ATBM10 | 1.221 | 1.210 | 1.216 | 1.222 |
| Science | ATBG01 | 9.862 | 9.811 | 9.851 | 9.848 |
| Science | ATBG02 | 0.344 | 0.336 | 0.344 | 0.352 |
| Science | ATBG03 | 1.069 | 1.062 | 1.070 | 1.072 |
| Science | ATBG09B | 0.811 | 0.812 | 0.822 | 0.827 |
| Science | ATBG09E | 0.685 | 0.685 | 0.699 | 0.704 |
| Science | ATBGTJS | 1.658 | 1.645 | 1.652 | 1.645 |
| Science | ATBS09 | 1.164 | 1.158 | 1.159 | 1.160 |

versus school-weighted means is 0.069. These average effect sizes are relatively small. Averages within grades and subjects vary little. Figure 6 provides an illustration of the similarity of student-centered (x-axis for each panel) and class-centered means (y-axis for each panel) in the case of science in the eighth grade (complementary figures for all other scenarios—school-centered means, mathematics and science both grades—can be found in the Appendix, see Figs. 8, 9, 10, 11, 12, 13, 14). To place all items on the same scale, the minimum value of the item score is subtracted from the mean and the result is divided by the range of the item score. Thus all values are between 0 and 1. For reference, the diagonal line has intercept 0 and slope 1. Clearly all points are very close to the line.

**Table 8** Unweighted and weighted standard deviations for grade 8 (average across countries)

| Subject | Item | Unweighted | Student centered | Class centered | School centered |
|---|---|---|---|---|---|
| Math | BTBG01 | 9.489 | 9.451 | 9.526 | 9.589 |
| Math | BTBG02 | 0.458 | 0.457 | 0.459 | 0.460 |
| Math | BTBG03 | 1.067 | 1.059 | 1.068 | 1.067 |
| Math | BTBG09B | 0.820 | 0.814 | 0.824 | 0.830 |
| Math | BTBG09E | 0.685 | 0.679 | 0.692 | 0.702 |
| Math | BTBGTJS | 1.681 | 1.664 | 1.666 | 1.660 |
| Math | BTBM23 | 1.215 | 1.201 | 1.212 | 1.216 |
| Science | BTBG01 | 9.385 | 9.314 | 9.325 | 9.345 |
| Science | BTBG02 | 0.454 | 0.450 | 0.453 | 0.457 |
| Science | BTBG03 | 1.054 | 1.043 | 1.043 | 1.039 |
| Science | BTBG09B | 0.829 | 0.822 | 0.832 | 0.843 |
| Science | BTBG09E | 0.693 | 0.684 | 0.702 | 0.724 |
| Science | BTBGTJS | 1.722 | 1.713 | 1.709 | 1.709 |
| Science | BTBS22 | 1.217 | 1.205 | 1.210 | 1.220 |

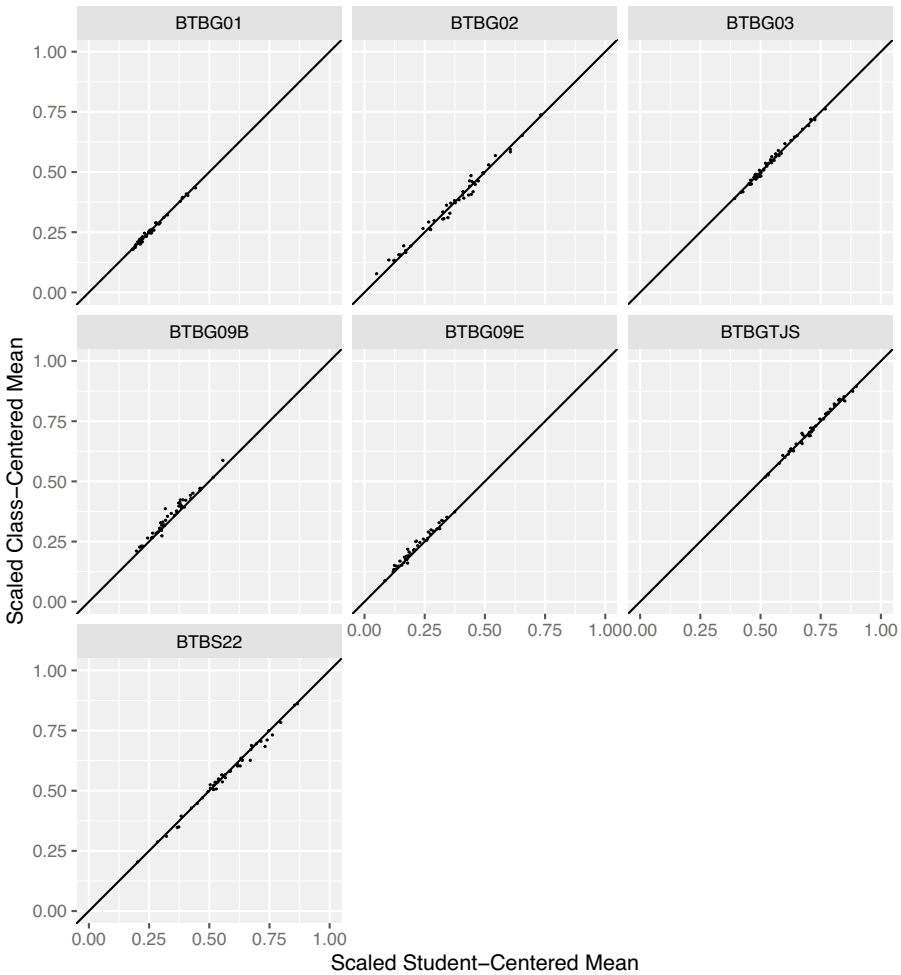



**Fig. 6** Scaled student-centered means versus class-centered means: eighth grade science

**Table 9** Design effects for unweighted and weighted means: grade 4 (average across countries)

| Subject | Item | Unweighted | Student centered | Class centered | School centered |
|---|---|---|---|---|---|
| Math | ATBG01 | 1.018 | 1.022 | 1.026 | 1.004 |
| Math | ATBG02 | 0.972 | 1.359 | 1.453 | 1.177 |
| Math | ATBG03 | 1.017 | 1.331 | 1.366 | 1.144 |
| Math | ATBG09B | 1.028 | 1.145 | 1.163 | 1.077 |
| Math | ATBG09E | 1.012 | 1.124 | 1.137 | 1.057 |
| Math | ATBGTJS | 1.084 | 2.421 | 2.485 | 1.502 |
| Math | ATBM10 | 1.054 | 1.156 | 1.167 | 1.064 |
| Science | ATBG01 | 1.038 | 1.014 | 1.030 | 1.006 |
| Science | ATBG02 | 0.958 | 1.453 | 1.510 | 1.205 |
| Science | ATBG03 | 1.053 | 1.381 | 1.421 | 1.179 |
| Science | ATBG09B | 1.016 | 1.127 | 1.136 | 1.074 |
| Science | ATBG09E | 1.027 | 1.089 | 1.098 | 1.041 |
| Science | ATBGTJS | 1.042 | 2.347 | 2.385 | 1.479 |
| Science | ATBS09 | 1.067 | 1.133 | 1.121 | 1.033 |

Nonetheless, despite the reported averages, it should be noted that effect sizes can sometimes be large. The most extreme case for comparison of student-centered and class-centered weights occurs in Pakistan for mathematics in the fourth grade for item *ATBM10*. In this case, the student-centered weighted mean is 2.683 and the class-centered weighted mean is 2.170. The respective weighted standard deviations are 1.671 and 1.479, so the effect size is 0.325. For comparison of student-centered and school-centered weighted means, the most extreme case is in the United States for mathematics in the eighth grade for item *BTBM23*. The student-centered weighted mean is 3.528, and the school-centered weighted mean is 2.910. The respective weighted standard deviations are 1.132 and 1.157. The corresponding effect size is 0.540. In these two instances, the difference in weighted means can have substantial effect on interpretations of results.

Standard errors are usually a significant concern in large-scale assessments because these studies rely on complex samples. These samples are characterized by various features such as stratification and clustering which prevent using standard formula (assuming SRS) to estimate standard errors (Lohr, 1999). Looking at standard error estimates using both the SRS and the JRR approach we investigate whether this may also be a concern for teacher-centered analysis. A summary of design effects is provided in Tables 9 and 10.

These design effects are averages over countries of squares of the ratios of standard errors from JRR and SRS. Average design effects are often close to 1, especially in the unweighted case, but average ratios are much higher in weighted cases for the scales *ATBGTJS* and *BTBGTJS*. Thus the design effects indicate a small but non-negligible effect of the complex design on standard errors, likely clustering and unequal weights being the driving forces (see Meinck and Vandenplas (2021) for more details). The most extreme design effects are quite large. In the case of school-centered means for item *ATBG02* in Latvia in the fourth-grade mathematics, the design effect is about 28.9, however, there is a fundamental difficulty in this case because only one of 200 sampled teachers of mathematics in the fourth-grade reports being male. In this case,

**Table 10** Design effects for unweighted and weighted means: grade 8 (average across countries)

| Subject | Item | Unweighted | Student centered | Class centered | School centered |
|---|---|---|---|---|---|
| Math | BTBG01 | 1.022 | 1.034 | 1.033 | 0.995 |
| Math | BTBG02 | 0.960 | 1.163 | 1.287 | 1.105 |
| Math | BTBG03 | 1.036 | 1.383 | 1.499 | 1.203 |
| Math | BTBG09B | 1.034 | 1.151 | 1.218 | 1.072 |
| Math | BTBG09E | 1.041 | 1.153 | 1.221 | 1.094 |
| Math | BTBGTJS | 1.038 | 2.385 | 2.897 | 1.552 |
| Math | BTBM23 | 1.068 | 1.251 | 1.418 | 1.118 |
| Science | BTBG01 | 1.089 | 1.084 | 1.145 | 1.046 |
| Science | BTBG02 | 0.956 | 1.256 | 1.345 | 1.135 |
| Science | BTBG03 | 1.071 | 1.610 | 1.780 | 1.294 |
| Science | BTBG09B | 1.071 | 1.174 | 1.224 | 1.104 |
| Science | BTBG09E | 1.067 | 1.210 | 1.276 | 1.064 |
| Science | BTBGTJS | 1.069 | 2.616 | 3.116 | 1.577 |
| Science | BTBS22 | 1.159 | 1.367 | 1.577 | 1.161 |

**Table 11** Ratio of SRS standard errors to standard deviation for grade 4 (average across countries)

| Subject | Item | Unweighted | Student centered | Class centered | School centered |
|---|---|---|---|---|---|
| Math | ATBG01 | 0.067 | 0.077 | 0.078 | 0.089 |
| Math | ATBG02 | 0.066 | 0.076 | 0.078 | 0.091 |
| Math | ATBG03 | 0.066 | 0.077 | 0.078 | 0.089 |
| Math | ATBG09B | 0.067 | 0.078 | 0.079 | 0.092 |
| Math | ATBG09E | 0.067 | 0.078 | 0.080 | 0.093 |
| Math | ATBGTJS | 0.066 | 0.075 | 0.076 | 0.086 |
| Math | ATBM10 | 0.067 | 0.077 | 0.078 | 0.090 |
| Science | ATBG01 | 0.068 | 0.078 | 0.079 | 0.091 |
| Science | ATBG02 | 0.067 | 0.075 | 0.078 | 0.090 |
| Science | ATBG03 | 0.067 | 0.077 | 0.079 | 0.091 |
| Science | ATBG09B | 0.068 | 0.080 | 0.081 | 0.094 |
| Science | ATBG09E | 0.068 | 0.080 | 0.082 | 0.095 |
| Science | ATBGTJS | 0.067 | 0.077 | 0.078 | 0.089 |
| Science | ATBS09 | 0.069 | 0.080 | 0.081 | 0.092 |

instability of estimates of standard errors (and design effects) is not surprising. On the other hand, for class-centered means, 14.4, the largest design effect, arises in Australia for mathematics in the eighth grade for item *BTBGTJS*, pointing to a substantial clustering effect regarding job satisfaction of eighth-grade mathematics teachers in this country (i.e., teachers within the same school tend to have similar job satisfaction levels), inflated by the high variance in weights. For student-centered means, the most extreme ratio, 11.8, arises in Dubai for item *ATBGTJS* for mathematics in the fourth grade. Given these results, further analysis will be based on JRR, and a clear recommendation for using standard error estimation methods accounting for the complex designs is warranted.

**Table 12** Ratio of JRR standard errors to standard deviation for grade 8 (average across countries)

| Subject | Item | Unweighted | Student centered | Class centered | School centered |
|---|---|---|---|---|---|
| Math | BTBG01 | 0.068 | 0.078 | 0.077 | 0.091 |
| Math | BTBG02 | 0.065 | 0.083 | 0.085 | 0.103 |
| Math | BTBG03 | 0.068 | 0.090 | 0.091 | 0.108 |
| Math | BTBG09B | 0.068 | 0.085 | 0.085 | 0.100 |
| Math | BTBG09E | 0.068 | 0.084 | 0.086 | 0.103 |
| Math | BTBGTJS | 0.068 | 0.114 | 0.119 | 0.138 |
| Math | BTBM23 | 0.070 | 0.086 | 0.089 | 0.102 |
| Science | BTBG01 | 0.065 | 0.073 | 0.073 | 0.087 |
| Science | BTBG02 | 0.061 | 0.077 | 0.079 | 0.095 |
| Science | BTBG03 | 0.065 | 0.088 | 0.089 | 0.106 |
| Science | BTBG09B | 0.065 | 0.078 | 0.078 | 0.095 |
| Science | BTBG09E | 0.065 | 0.078 | 0.080 | 0.096 |
| Science | BTBGTJS | 0.065 | 0.111 | 0.115 | 0.131 |
| Science | BTBS22 | 0.068 | 0.082 | 0.085 | 0.097 |

As evident from Tables 11 and 12, standard errors are a major concern in any of the weighted means under study. We noted above that a standard error of .08 standard deviation units might be deemed acceptable, however, even the average ratio between standard errors and standard deviations[6] is higher for most variables and weighting scenarios, meaning more than half of the ratios for specific countries are higher than this value. Student-centered and class-centered estimates have similar ratios of standard errors to standard deviations, and results for school-centered weights are a bit worse. The least satisfactory results are associated with the job satisfaction scales *ATBGTJS* and *BTBGTJS*.

To check more thoroughly on the issue of standard errors, it is helpful to examine cumulative distribution functions of JRR ratios of standard errors to weighted standard deviation (scaled JRR standard errors). Figure 7 provides an example for school-centered weighted means (complementary figures for other grades, subjects and weighting scenarios can be found in the Appendix, see Figs. 15, 16, 17, 18, 19, 20, 21), with the ratio of standard error to standard deviation on the x-axis for each panel and the cumulative distribution function on the y-axis for each panel. Clearly results are rather variable for different educational systems. As evident from the vertical line at 0.08, it is certainly not uncommon for ratios to be less than 0.08; however, occasionally ratios are about 0.3, pointing to very imprecise estimates. A basic issue is the existence of enough responses, depending not only on sample size but also on participation. For example, the value of 0.332 for England involves only 86 responses, due to low participation rates at both school and teacher level. On the other hand, the issue is a bit more complicated. For example, for item *BTBGTJS* in the United States, 426 responses are present but the ratio is 0.240. Some explanation is provided in terms of the effective sample size measure equal to the ratio of the square of the

[6] Contrasting standard errors against standard deviations allows direct comparisons of sampling precision between populations or variables with different scales. In other contexts, the coefficient of variation is often used instead, but it has some drawbacks, for example it does not work well for scales with a mean of zero.

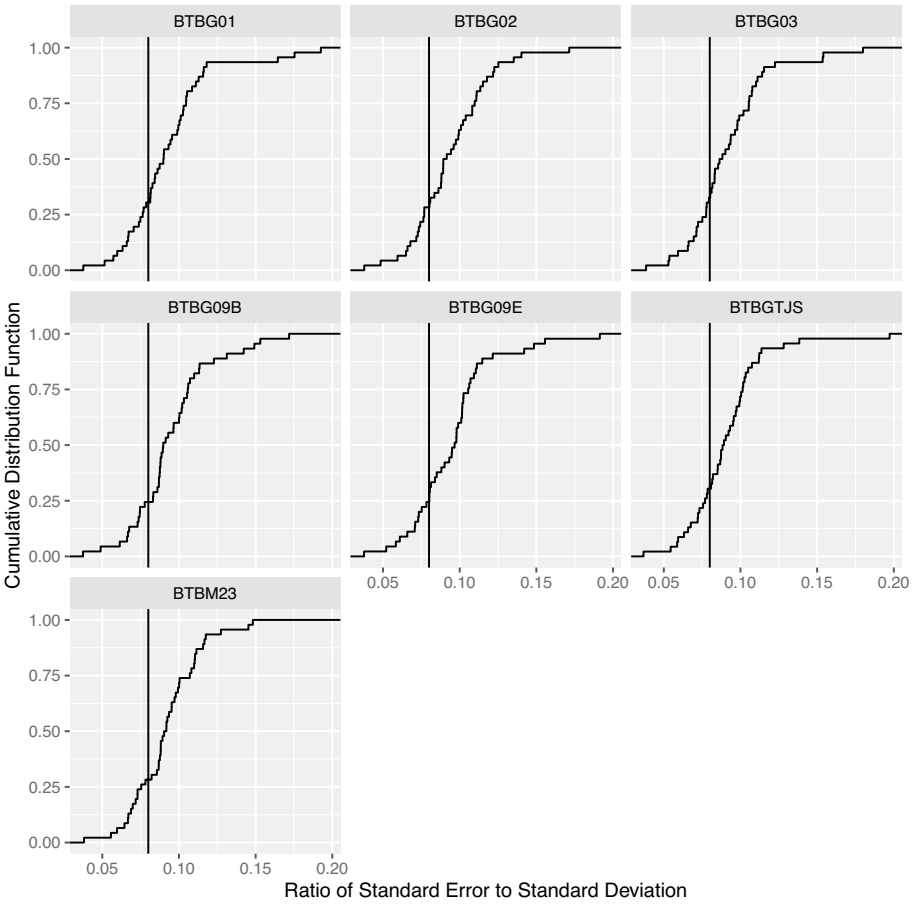



**Fig. 7** Scaled JRR standard errors for School-centered means: eighth grade science

sum of the weights to the sum of the squares of the weights (Kish, 1965, p. 259). In the case of the United States, the sample size for science teachers in eighth grade is 468, but the effective sample size for school-centered weights is only 32.7, pointing to a very large design effect of almost 15. The effective sample size for the United States is so low because some sampled schools have very low probabilities of being sampled and hence very high weights. These very low probabilities reflect very small school sizes. The effect on the weights could even not be compensated by a method applied in TIMSS and PIRLS to minimize fluctuations in sampling weights, that is, set uniform selection probabilities when sampling small schools. For example, for eighth grade one sampled school had only one sampled student and another had only two sampled students. This result reflects a decision not to exclude very small schools from the American sample and a decision in TIMSS not to apply methods to reduce unusually high weights, which may be reconsidered in future cycles of TIMSS. The exclusion for small schools is not unusual in other educational systems participating in TIMSS, and standardizing this approach may be an effective measure to avoid large variance in weights also for the student sample. At the moment, TIMSS allows exclusion of small schools covering up to 2% of the student population. For example, in Gauteng and Western Cape, schools in the sample for eighth grade must have at least

10 students. Another reasonable approach to consider is the application of exclusion rules for teacher analysis not applied for student analysis due to the much smaller number of teachers in an educational system.

Overall, according to the considered scenarios, teacher-centered analysis seems to be possible with fairly reasonable precision using the MAIS approach, although some limits exist for specific variables and educational systems. In any case, the results suggest that analysis of teachers in any educational system participating in TIMSS generally cannot effectively examine subgroups given the number of teachers sampled.

## Summary, conclusions and recommendations

TIMSS and PIRLS expend significant effort and cost to collect and analyze data for an elaborate explanatory model covering student achievement in the areas of mathematics, science, and reading, and the contexts of learning these subjects. The ability to analyze teacher-level characteristics from proper samples drawn from teacher populations is not included in their study designs, as choices had to be made to keep the costs and complexity levels of these studies manageable. Still, a rich array of data related to teacher characteristics is collected, and scholars wish using this data to investigate characteristics of teachers. This paper builds on the work by Hooper et al. (2022), extending their introduction of two approaches to derive weights for teacher-centered analysis using TIMSS and PIRLS data by looking into aspects of practical implementation of these approaches.

We began with proposing a definition for teacher target populations, tied to the grades and subjects they teach, in line with the focus of the two large-scale assessments. This definition should help to correctly and comprehensively identify all in-scope teachers within schools sampled for TIMSS and PIRLS, being a requisite for accurate estimation of population characteristics. We then formalized the computation of teacher-centered weights and using them to derive teacher-centered population estimates, and discuss some issues and limitations related with this. We highlighted the utility of both, student-centered and teacher-centered analysis, depending on the research question to be answered, and disentangled the differences between the two types of weights. Next we suggest a procedure and form on how to collect data about teachers that is needed to derive teacher-centered weights, yet currently unavailable. This step is key if in future cycles teacher-centered weights should be derived in TIMSS and PIRLS. Alternative forms or procedures may work, and optimal solutions may depend on the particular situation in participating countries. We however recommend here a standardized procedure that can be applied in all countries, a feature that is important in ILSA to support their dense timelines, high quality standards, and production modes. Collecting this additional information demands slightly more work by school coordinators, and a small adjustment in operations, that may be well justifiable given the possible gain in knowledge.

We also tackle the issue of non-response by proposing a non-response adjustment factor in line with existing approaches in ILSA, as well as mentioning the challenge of multiple selection probabilities when teachers teach in multiple schools, where we refer to solutions applied in other ILSA.

The core part of the paper focuses on studying the level of accuracy that can be expected when estimating teacher population characteristics. We look into sample sizes as they are a fundamental factor related with precision. Then we use TIMSS 2019 data and simulate likely scenarios regarding the variance in weights. Identifying the MAIS method as the method most effective for TIMSS and PIRLS as it can handle within-school stratification, we continue only with this method. The results show that the different weighting scenarios (including using no weights) lead to relatively similar estimates, at least on average, however with large enough differences for specific variables and countries to warrant the recommendation to use teacher-centered weights for analysis of teacher populations rather than student-centered weights. Second, results provide evidence to use weights and an algorithm to estimate standard errors that accounts for the complex sampling design, as standard error estimates would otherwise be systematically biased. We find further that sample sizes and variance in weights are significantly limiting estimate precision. Especially the large variation in weights induces particularly large design effects. Hence, while characteristics of whole teacher populations can be estimated with sufficient precision in the majority of countries, we discourage estimating subpopulation features (such as, for example, job satisfaction of male teachers), and we strongly recommend that, to avoid unreasonable interpretations, analysts with research questions should thoroughly check sample sizes and variances in weights of the populations of interest. However, if such research questions are deemed of high interest, national research coordinators should discuss options to adjust the sampling design for their countries. Options that would not jeopardize the core objective of TIMSS and PIRLS (that is, studying students) include increasing the number of schools or classes (and thereby teachers) selected and extending the teacher survey to teachers not sampled by way of student sampling.

The results presented here are of limited reliability as they are based on plausible scenarios rather than real data that permit computation of teacher-centered weights. Therefore, the next step is actual implementation in one or more countries, followed by replicating the analysis presented here with real data, which would allow a critical evaluation of our results.

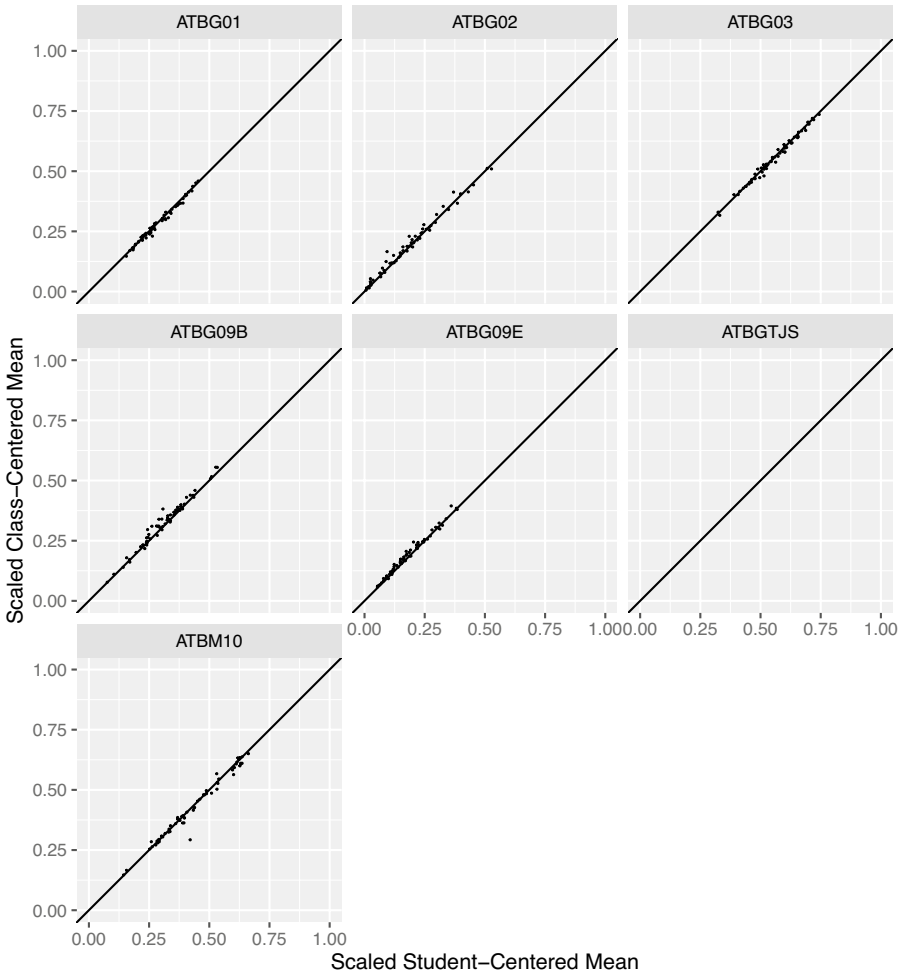



**Fig. 8** Scaled student-centered means versus class-centered means: grade 4 mathematics

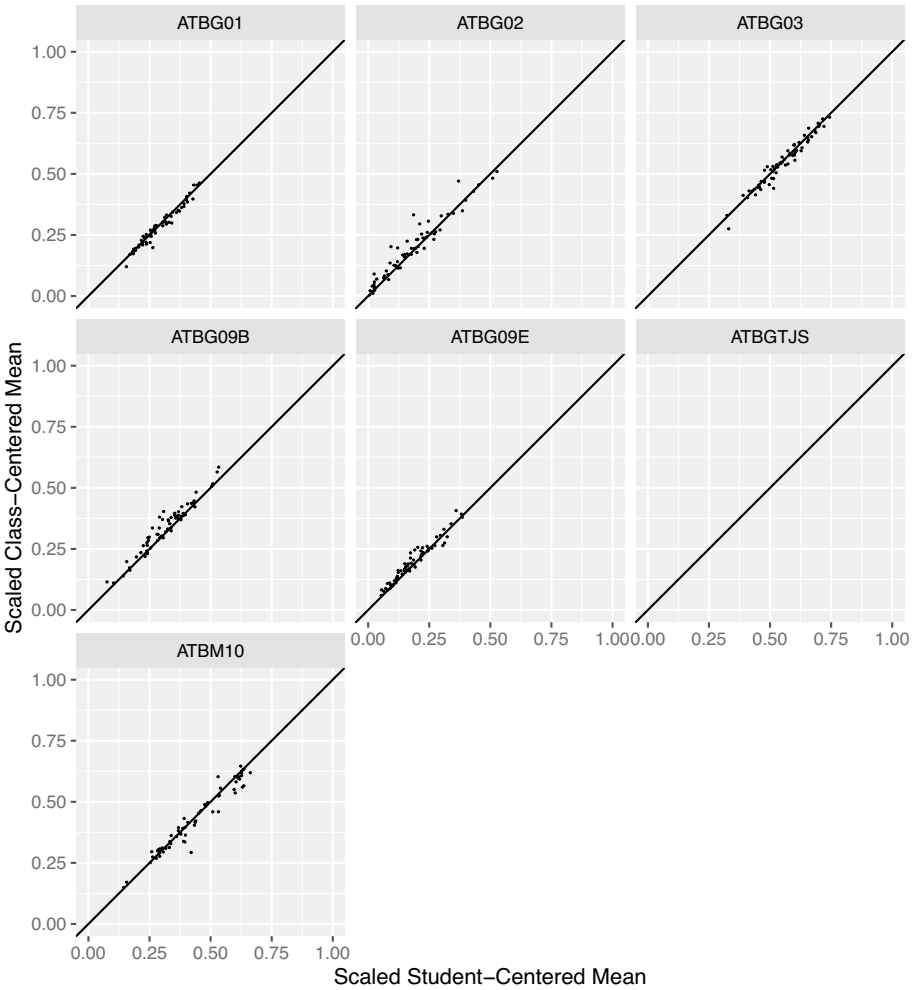



**Fig. 9** Scaled student-centered means versus school-centered means: grade 4 mathematics

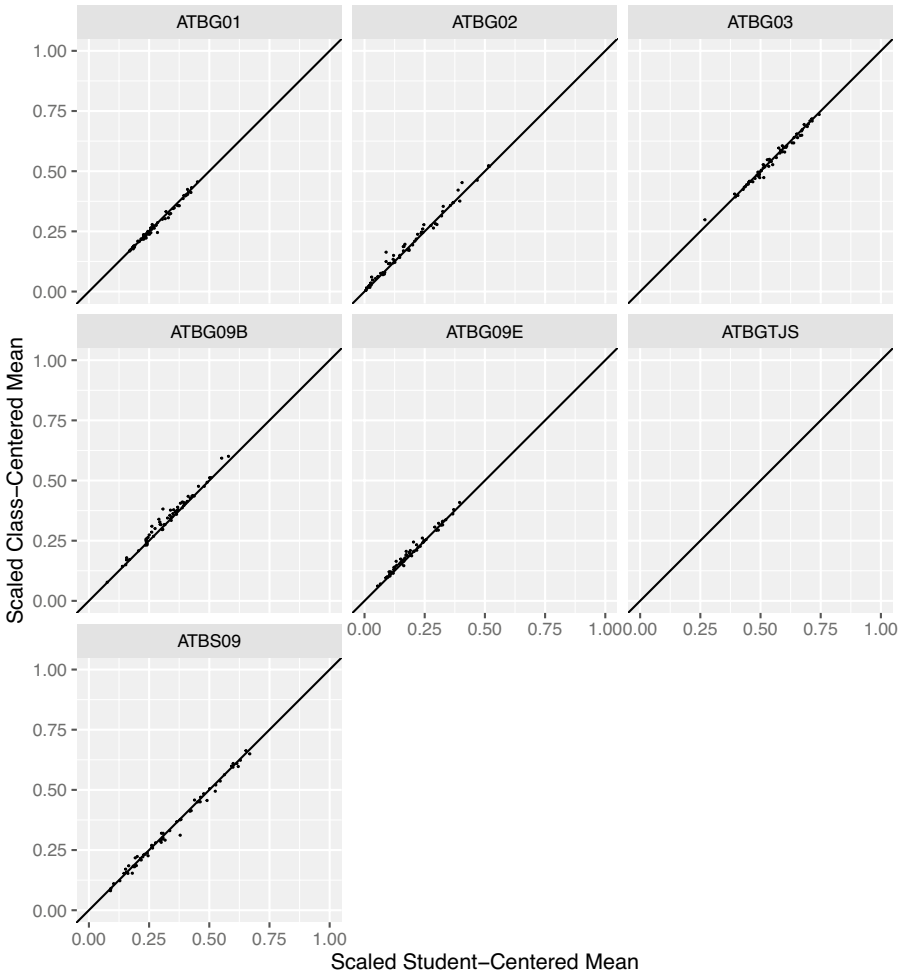



**Fig. 10** Scaled student-centered means versus class-centered means: grade 4 science

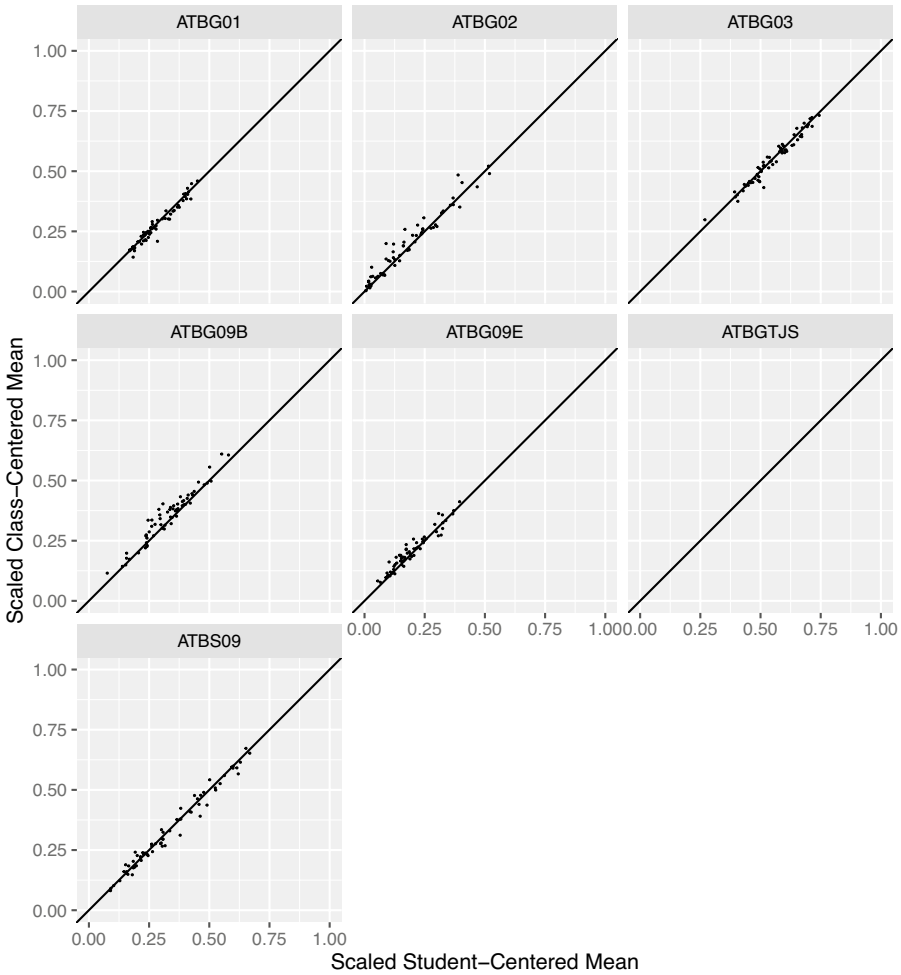



**Fig. 11** Scaled student-centered means versus school-centered means: grade 4 science

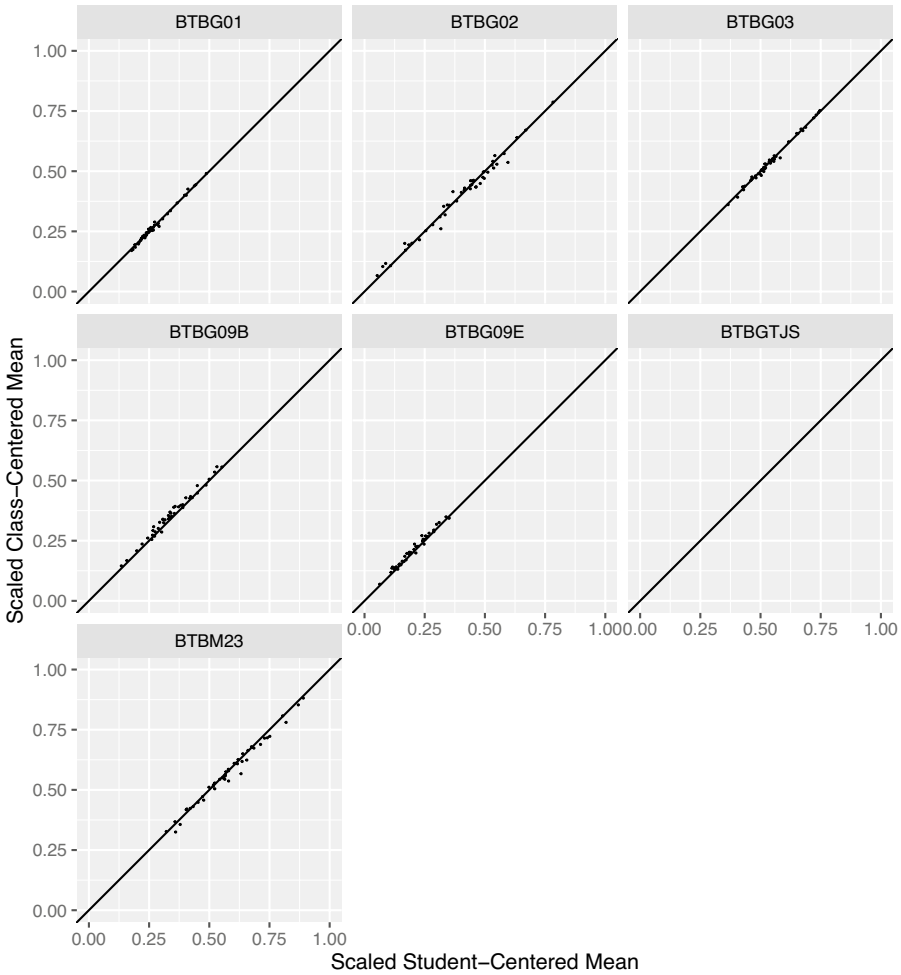



**Fig. 12** Scaled student-centered means versus class-centered means: grade 8 mathematics

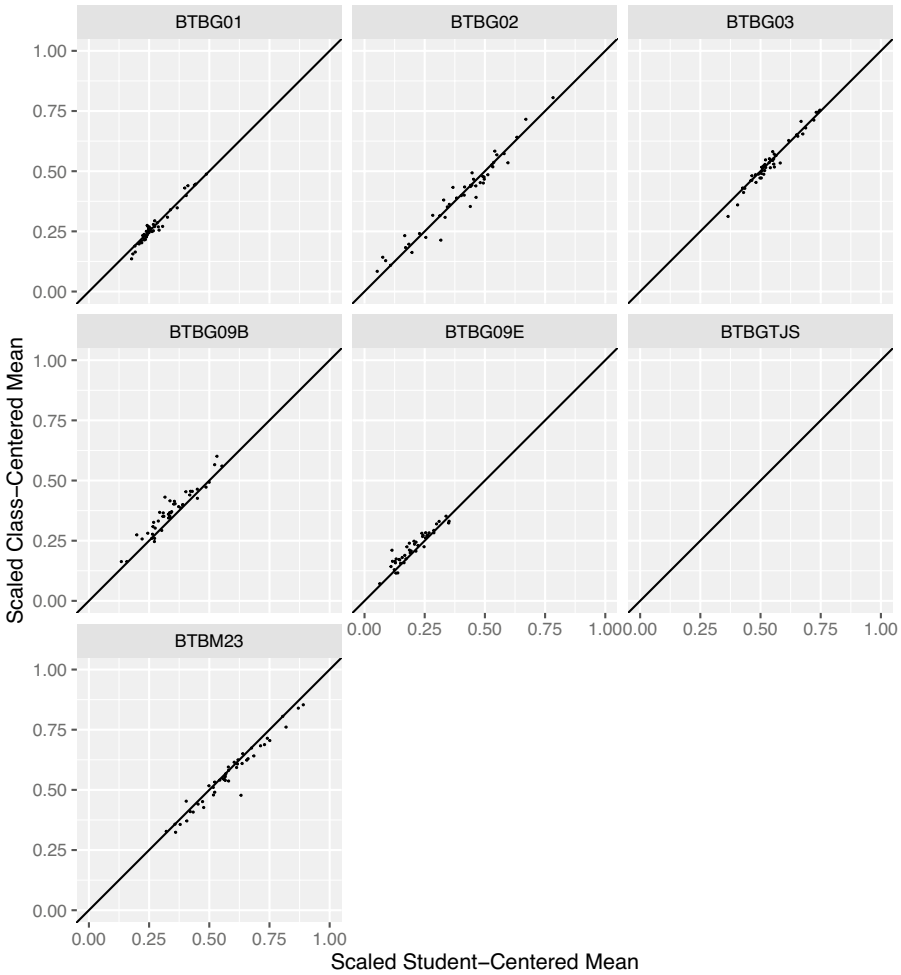



**Fig. 13** Scaled student-centered means versus school-centered means: grade 8 mathematics

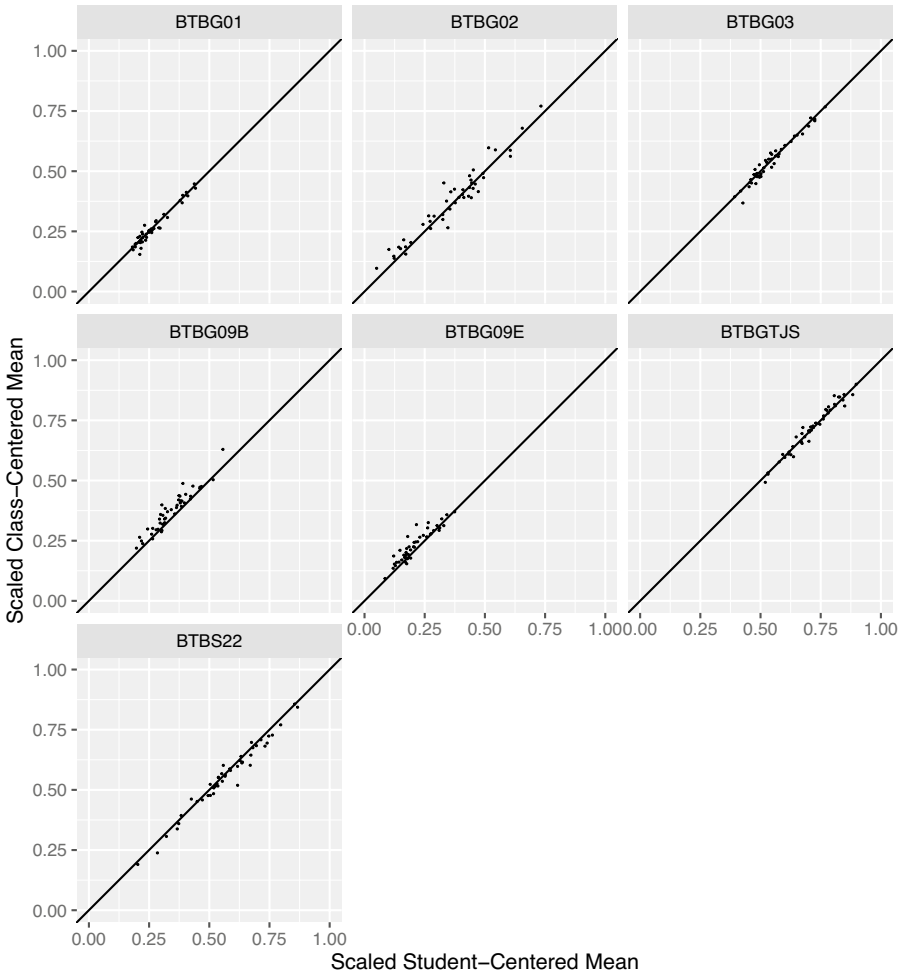



**Fig. 14** Scaled student-centered means versus school-centered means: grade 8 science

## Appendix

See Tables 13, 14 and Figs. 15, 16, 17, 18, 19, 20, 21.

**Table 13** Achieved sample sizes in TIMSS and PIRLS

| Achieved sample sizes teachers (T) and schools (S) | TIMSS | | | | | | | | PIRLS | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Grade 4 | | | | Grade 8 | | | | Grade 4 | |
| | Math | | Science | | Math | | Science | | Rd./Lng. | |
| Educational system | T | S | T | S | T | S | T | S | T | S |
| Albania | 204 | 167 | 203 | 167 | | | | | | |
| Armenia | 212 | 150 | 201 | 150 | | | | | | |
| Australia | 402 | 287 | 369 | 287 | 444 | 284 | 739 | 284 | 531 | 286 |
| Austria | 303 | 193 | 303 | 193 | | | | | 259 | 150 |
| Azerbaijan | 243 | 194 | 243 | 194 | | | | | 298 | 170 |
| Bahrain | 217 | 185 | 217 | 185 | 233 | 112 | 273 | 112 | 208 | 182 |
| Belgium (Flemish) | 283 | 147 | 276 | 147 | | | | | 277 | 148 |
| Belgium (French) | | | | | | | | | 254 | 158 |
| Bosnia and Herzegovina | 334 | 178 | 334 | 178 | | | | | | |
| Bulgaria | 209 | 151 | 210 | 151 | | | | | 213 | 153 |
| Canada | 913 | 704 | 906 | 704 | | | | | 1119 | 926 |
| Chile | 179 | 169 | 172 | 169 | 173 | 164 | 207 | 164 | 154 | 154 |
| Chinese Taipei | 216 | 162 | 177 | 162 | 311 | 203 | 222 | 203 | 176 | 150 |
| Croatia | 263 | 153 | 263 | 153 | | | | | | |
| Cyprus | 229 | 151 | 168 | 151 | 170 | 98 | 436 | 98 | | |
| Czech Republic | 264 | 152 | 257 | 152 | | | | | 270 | 157 |
| Denmark | 190 | 166 | 190 | 166 | | | | | 186 | 185 |
| Egypt | | | | | 169 | 169 | 169 | 169 | | |
| England | 159 | 139 | 159 | 139 | 142 | 136 | 141 | 136 | 210 | 170 |
| Finland | 326 | 158 | 317 | 158 | 358 | 154 | 786 | 154 | 295 | 151 |
| France | 300 | 155 | 300 | 155 | 188 | 150 | 343 | 150 | 284 | 163 |
| Georgia | 220 | 154 | 215 | 154 | 175 | 145 | 629 | 145 | 285 | 200 |
| Germany | 216 | 203 | 218 | 203 | | | | | 227 | 208 |
| Hong Kong SAR | 157 | 139 | 145 | 139 | 184 | 136 | 146 | 136 | 150 | 138 |
| Hungary | 252 | 149 | 249 | 149 | 272 | 154 | 603 | 154 | 206 | 149 |
| Iran, Islamic Rep. of | 224 | 224 | 224 | 224 | 220 | 220 | 220 | 220 | 309 | 271 |
| Israel | 231 | 150 | | | 444 | 157 | 261 | 157 | 159 | 159 |
| Ireland | | | 231 | 150 | 560 | 149 | 395 | 149 | 219 | 148 |
| Italy | 229 | 162 | 229 | 162 | 209 | 158 | 209 | 158 | 217 | 149 |
| Japan | 230 | 147 | 154 | 147 | 210 | 142 | 155 | 142 | | |
| Jordan | | | 235 | 235 | | | 235 | 235 | | |
| Kazakhstan | 224 | 168 | 224 | 168 | 223 | 168 | 843 | 168 | 234 | 172 |
| Korea, Rep. of | 187 | 151 | 195 | 151 | 228 | 168 | 235 | 168 | | |
| Kosovo | 219 | 145 | 219 | 145 | | | | | | |
| Kuwait | 168 | 164 | 168 | 164 | 173 | 171 | 173 | 171 | | |
| Lebanon | | | | | 204 | 204 | 612 | 204 | | |
| Latvia | 203 | 154 | 189 | 154 | | | | | 216 | 150 |
| Lithuania | 249 | 207 | 250 | 207 | 247 | 194 | 761 | 194 | 243 | 195 |
| Macao SAR | | | | | | | | | 122 | 57 |
| Malta | 210 | 98 | 209 | 98 | | | | | 206 | 95 |

**Table 13** (continued)

| Achieved sample sizes teachers (T) and schools (S) | TIMSS | | | | | | | | PIRLS | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Grade 4 | | | | Grade 8 | | | | Grade 4 | |
| | Math | | Science | | Math | | Science | | Rd./Lng. | |
| **Educational system** | **T** | **S** | **T** | **S** | **T** | **S** | **T** | **S** | **T** | **S** |
| Malaysia | | | | | 265 | 177 | 266 | 177 | | |
| Montenegro | 361 | 140 | 361 | 140 | | | | | | |
| Morocco | 282 | 264 | 282 | 264 | 260 | 251 | 510 | 251 | 372 | 360 |
| Netherlands | 182 | 112 | 182 | 112 | | | | | 226 | 132 |
| New Zealand | 426 | 160 | 404 | 160 | 372 | 134 | 278 | 133 | 411 | 188 |
| North Macedonia | 239 | 150 | 239 | 150 | | | | | | |
| Northern Ireland | | | 156 | 134 | | | | | 161 | 134 |
| Norway (5) | 237 | 150 | 235 | 150 | 263 | 157 | 225 | 157 | 211 | 150 |
| Oman | 246 | 228 | 246 | 228 | 241 | 228 | 242 | 228 | 356 | 306 |
| Pakistan | 155 | 139 | 155 | 139 | | | | | | |
| Philippines | | | 180 | 180 | | | | | | |
| Poland | 225 | 149 | 189 | 149 | | | | | 214 | 148 |
| Portugal | 314 | 181 | 314 | 181 | 185 | 156 | 363 | 156 | 318 | 218 |
| Qatar | | | 251 | 242 | 198 | 152 | 234 | 152 | 378 | 216 |
| Romania | | | | | 219 | 198 | 552 | 198 | | |
| Russian Federation | 200 | 200 | 200 | 200 | 207 | 204 | 749 | 204 | 213 | 206 |
| Saudi Arabia | 222 | 220 | 220 | 220 | 218 | 209 | 231 | 209 | 202 | 202 |
| Serbia | 214 | 165 | 214 | 165 | | | | | | |
| Singapore | 371 | 187 | 362 | 187 | 296 | 153 | 295 | 153 | 354 | 177 |
| Slovak Republic | 268 | 157 | 251 | 157 | | | | | 333 | 220 |
| Slovenia | | | | | | | | | 253 | 160 |
| South Africa (5) | 297 | 297 | 297 | 297 | 542 | 519 | 536 | 519 | | |
| Spain | 509 | 501 | 514 | 501 | | | | | 678 | 629 |
| Sweden | 194 | 145 | 178 | 145 | 214 | 150 | 314 | 150 | 214 | 154 |
| Trinidad And Tobago | | | | | | | | | 195 | 151 |
| Turkey (5) | 180 | 180 | 180 | 180 | 181 | 181 | 181 | 181 | | |
| United Arab Emirates | 1073 | 688 | 1036 | 688 | 1036 | 623 | 1180 | 623 | 652 | 468 |
| United States | 480 | 287 | 469 | 287 | 445 | 273 | 468 | 273 | 208 | 158 |
| *Benchmark participants* | | | | | | | | | | |
| Abu Dhabi, UAE | 386 | 247 | 368 | 247 | 374 | 230 | 405 | 230 | 177 | 151 |
| Andalusia, Spain | | | | | | | | | 188 | 150 |
| Buenos Aires, Argentina | | | | | | | | | 188 | 150 |
| Dubai, UAE | 328 | 199 | 326 | 199 | 301 | 163 | 359 | 163 | 304 | 174 |
| Eng/Afr/Zulu - RSA (5) | | | | | | | | | 147 | 125 |
| Gauteng, RSA (9) | | | | | 150 | 150 | 150 | 150 | | |
| Madrid, Spain | 168 | 167 | 167 | 167 | | | | | 168 | 168 |
| Moscow City, Russian Fed. | 174 | 150 | 174 | 150 | 205 | 150 | 710 | 150 | 173 | 150 |
| Norway (4) | | | | | | | | | 221 | 154 |
| Ontario, Canada | 240 | 163 | 241 | 163 | 198 | 158 | 201 | 158 | 251 | 188 |
| Quebec, Canada | 228 | 148 | 222 | 148 | 148 | 124 | 150 | 124 | 166 | 127 |
| Western Cape, RSA (9) | | | | | 171 | 149 | 165 | 149 | | |

**Table 14** Notations

| Symbol | Notations |
|---|---|
| *Educational system* | |
| $N$ | Schools |
| $H$ | Strata |
| $C$ | Classes |
| $S$ | Students |
| *School weight* | |
| $F_{hi1}$ | Sampling weight of school $i$, if school $i$ participates |
| $A_{h1}$ | School nonparticipation adjustment for stratum $h$ |
| $h$ | Explicit stratum |
| $i$ | Sampled school |
| $m_i$ | Size measure for sampled school $i$ |
| $M_h$ | Sum of size measures for all schools in stratum $h$ |
| $M$ | Total number of students in the target population |
| $n_h$ | Total number of sampled schools in the stratum |
| $S_{hi}$ | Number of students in the target grade |
| $N_h$ | Number of schools attended by the studentsin the target population |
| *Class weight* | |
| $G_{hij2}$ | Overall class weight |
| $F_{hij2}$ | Class weight component for sampled class $j$ of sampled school $i$ |
| $A_{h2}$ | Class nonparticipation adjustment for stratum $h$ |
| $C_i$ | Number of eligible classes in school $i$ |
| $c_i$ | Number of sampled classes in school $i$ |
| $\delta_i$ | Number of participating classes among sampled classesin school $i$ |
| *Student weight* | |
| $G_{hij3}$ | Final weight for a participating student |
| $F_{ij3}$ | Weight multiplier for a selected and participating student |
| $n_{ij}$ | Number of students in the class |
| $n_{ij1}$ | Number of selected students in the class |
| $n_{ij2}$ | Number of students sampled who mighthave participated. (It is possible due to class changes that $n_{ij2}$ and $n_{ij1}$ differ.) |
| $n_{ij3}$ | Number of selected students in the class who participated |
| *Student-centered weight* | |
| $Y_{ijk}$ | Measurement of student $k$ from class $j$ of school $i$ |
| $K_{ijk}$ | Number of teachers for a subject (mathematics or science) of student $k$ in a participating selected class $j$ from participating selected school $i$ |
| *Teacher-centered weight* | |
| $T_{it}$ | Probability that teacher $t$ is sampled |
| $U_{it}$ | Teacher variable for teacher $t$ in school $i$ |
| $d_{it}$ | Number of classes a teacher teaches out of all classes in the school |
| $C_i$ | Number of classes in the school |
| $c_i$ | Number of sampled classes |
| *Properties of teacher weights* | |
| $W = S/M, W_{it}$ | Student-centered teacher weight for teacher $t$ in target schools $i$ |
| $S_{it}$ | The sum of the fractions $1/K_{ijk}$ for all students $k$ in a class $j$ of school $i$ who are taught by teacher $t$ |
| $\bar{U}_S$ | Student-based average |
| *HT approach* | |
| $W_{itT}$ | Final teacher sampling weight according to the HT approach |
| $F_{itT} = 1/T_{it}$ | Teacher component of the final sampling weight |
| $E\,U$ | Teacher-based average, of $U_{it}$ for teachers $t$ in target schools $i$ |

**Table 14** (continued)

| Symbol | Notations |
|---|---|
| $D_i$ | Number of teachers in school $i$ |
| $D_+$ | Sum of the $D_i$ over all target schools $i$ |
| $\Sigma(U)$ | Total of the $U_{it}$ for the $D_+$ teachers $t$ in target schools $i$ |
| $U$ | Real variable defined for all combinations of schools and teachers of a subject who teach students in the target population |
| $U_{i+}$ | Sum of $U_{it}$ for all teachers in school $i$ |
| $U_{++}$ | Population sum |
| $S(W_T U)$ | Sum of the $W_{itT} U_{it}$ for sampled teachers $t$ in sampled schools $i$ |
| *MAIS approach* | |
| | *-notation refers to Mais approach |
| $W_{itT*}$ | Final teacher sampling weight according to the MAIS approach |
| $F_{itT*}$ | Teacher component of the final teacher sampling weight according to the HT approach according to the MAIS approach |
| *Measures of dispersion* | |
| $\sigma$ | Population standard deviation |
| $\rho$ | Population correlation coefficient |
| $p_{hi}$ | Probability: Sampling of Schools |
| $p_{hij}$ | Probability: Sampling both of the distinct schools $i$ and $j$ in stratum $h$ |
| $\zeta_{hij}$ | $p_{hi}p_{hj} - p_{hij}$ |
| $v_{hi} = p_{hi}(1 - p_{hi})$ | Variance associated with the probability $p_{hi}$ |
| $\gamma(A, B)$ | Covariance of $A$ and $B$ |
| *Adjustment for nonresponse* | |
| $A_{hT}$ | Adjustment factor for nonparticipating teachers |
| $\delta_{Ti}$ | Number of participating teachers in school $i$ |
| $G_{hitT*}$ | Teacher sampling weight if nonresponse is ignored (MAIS) |
| $d_{ijt}$ | Number of classes teacher $t$ teaches in the class stratum that includes class $j$ |
| $G_{hitTc}$ | Sum of the class weights $G_{hij2}$ for all sampled classes $j$ in school $i$ associated with teacher $t$ |
| $G_{hitT}$ | Weight of sampled teacher $t$ under HT |

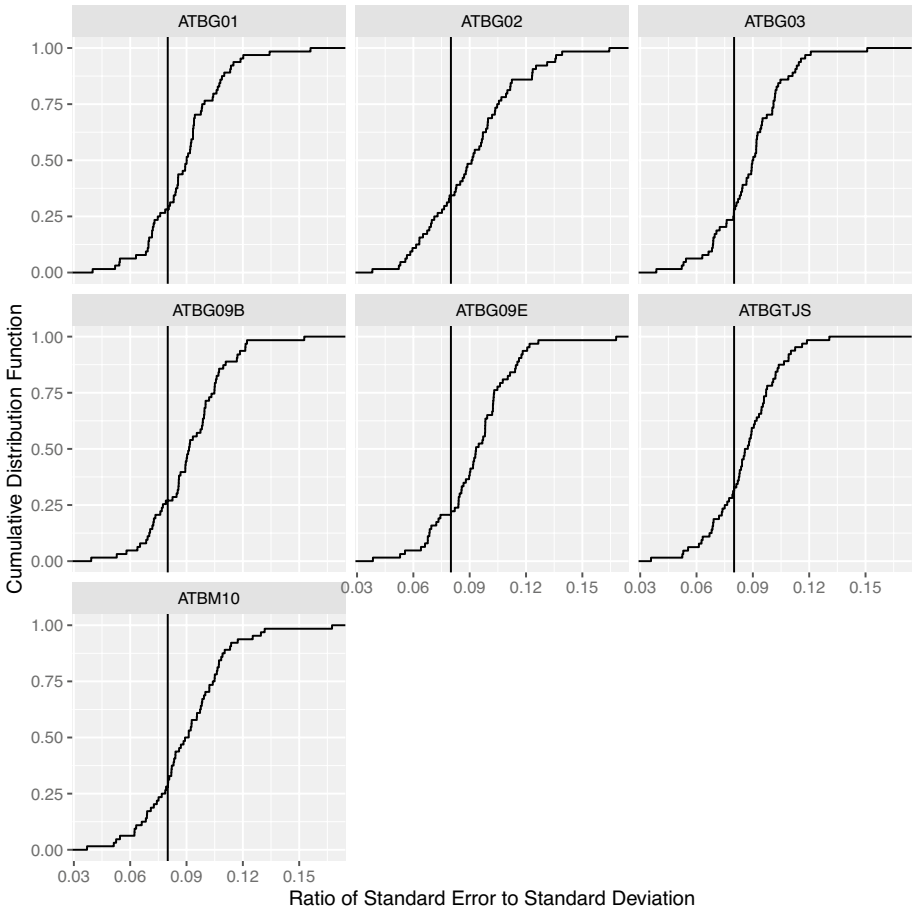



**Fig. 15** Cumulative distribution function of scaled JRR standard errors for school-centered means: grade 4 mathematics

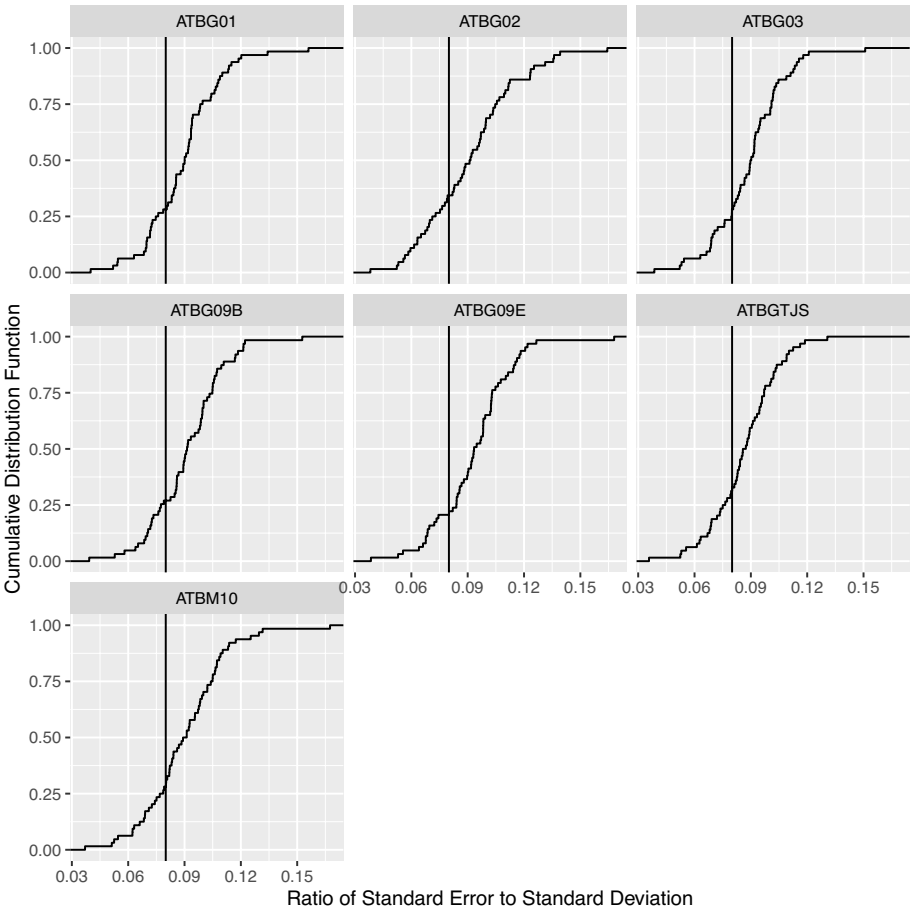



**Fig. 16** Cumulative distribution function of scaled JRR standard errors for school-centered means: grade 4 science

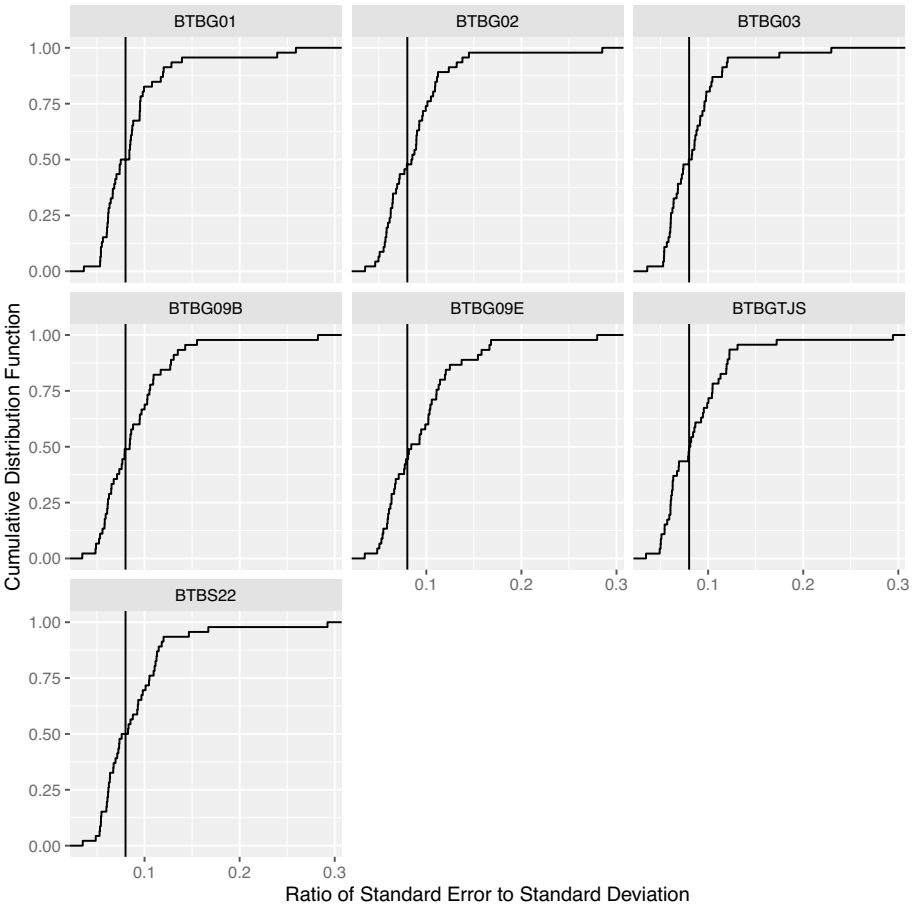



**Fig. 17** Cumulative distribution function of scaled JRR Standard errors for school-centered means: grade 8 mathematics

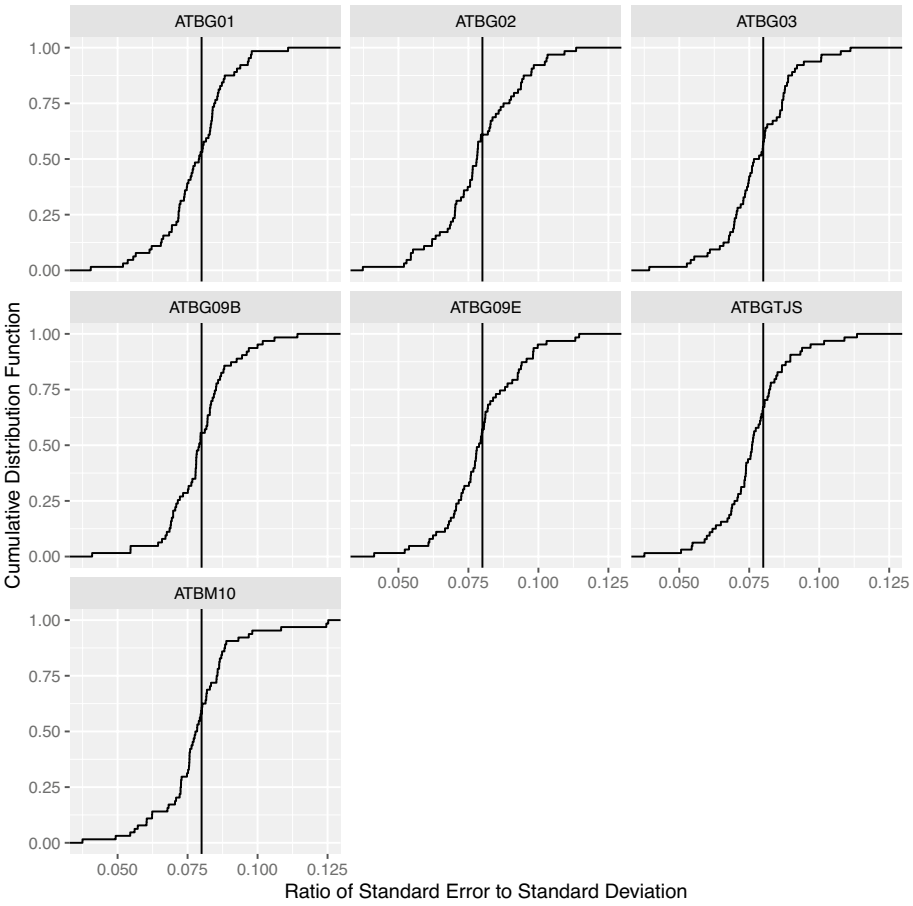



**Fig. 18** Cumulative distribution function of scaled JRR standard errors for class-centered means: grade 4 mathematics

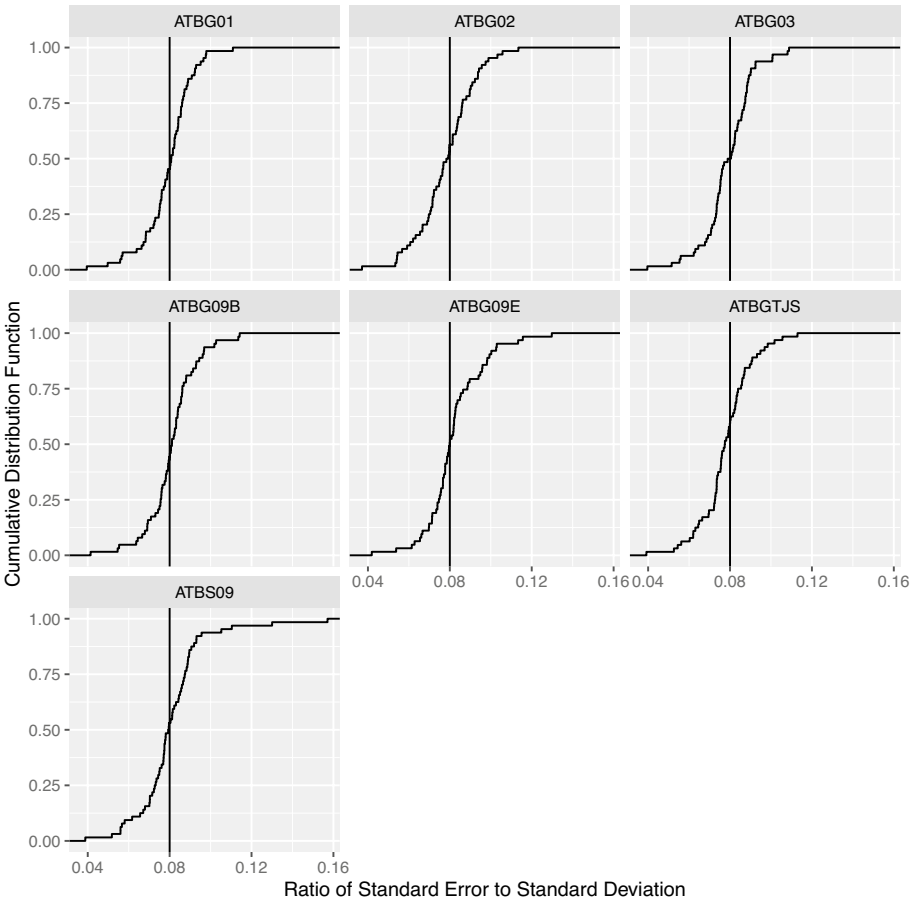



**Fig. 19** Cumulative distribution function of scaled JRR standard errors for class-centered means: grade 4 science

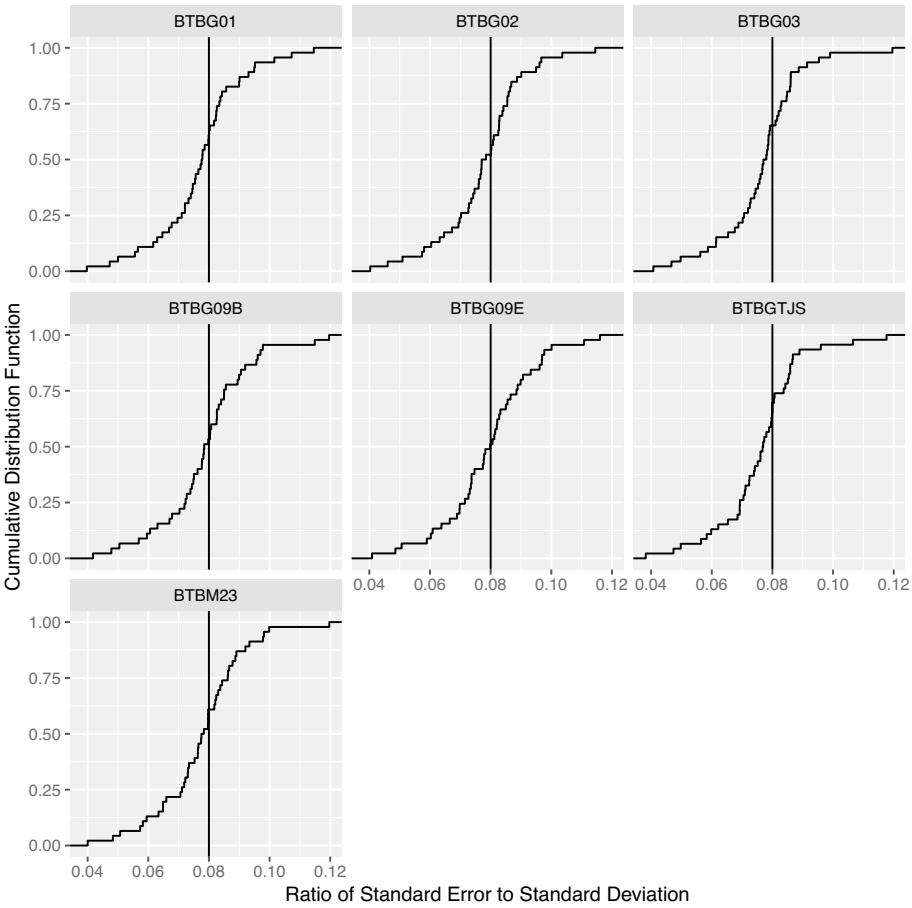



**Fig. 20** Cumulative distribution function of scaled JRR standard errors for class-centered means: grade 8 mathematics

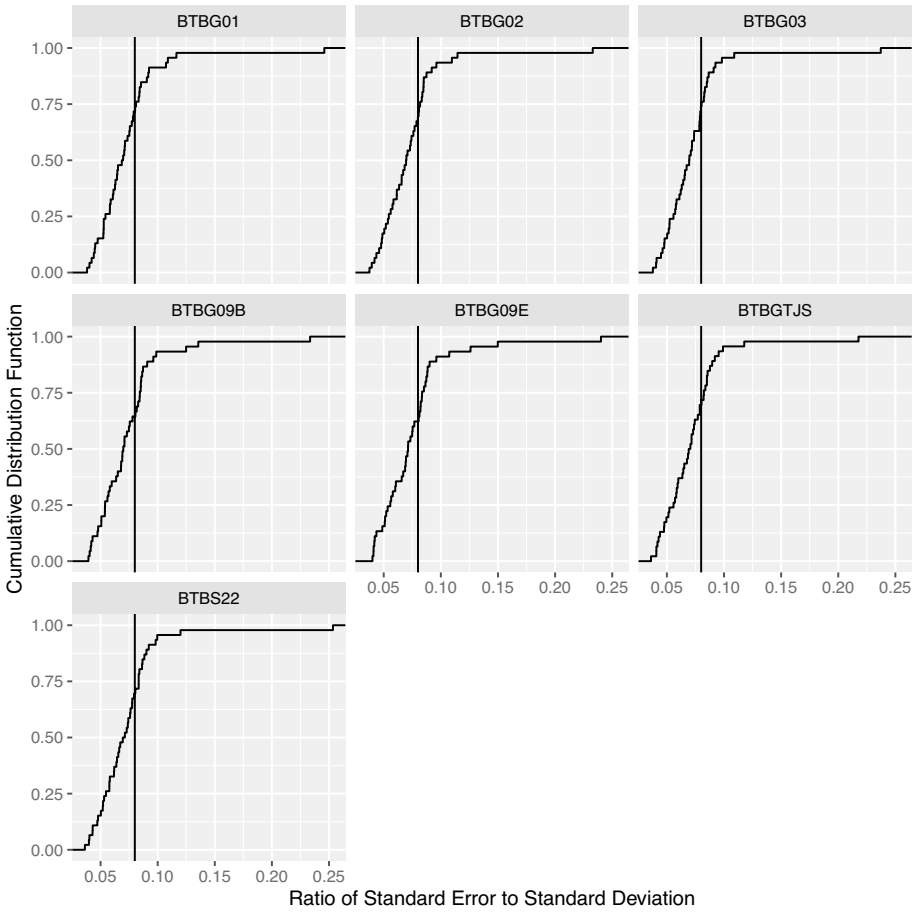



**Fig. 21** Cumulative distribution function of scaled JRR standard errors for class-centered means: grade 8 science

**Abbreviations**

| | |
|---|---|
| ICCS | International Civic and Citizenship Study |
| ICILS | International Computer and Information Literacy Study |
| IEA | International Association for the Evaluation of Educational Achievement |
| ILSA | International large-scale assessments |
| HT | Horwitz-Thompson |
| ISCED | International Standard Classification of Education |
| JRR | Jackknife repeated replication |
| MAIS | Multiplicity-adjusted indirect sampling |
| OECD | Organization for Economic Co-operation and Development |
| PPS | Probability proportional to size |
| PIRLS | Progress in International Reading Literacy Study |
| SRS | Simple random sampling |
| s-tchwgt | Student-centered teacher weights |
| t-tchwgt | Teacher-centered teacher weights |
| TALIS | Teaching and Learning International Survey |
| TIMSS | Trends in International Mathematics and Science Study |

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40536-024-00214-x.

Supplementary Material 1. Analysis Results.

**Acknowledgements**
We would like to extend our sincere thanks to Diego Cortes and Umut Atasever for their most helpful review of this paper.

**Author contributions**
Shelby Haberman conducted the key analyses for this paper and was a major contributor in writing the manuscript. Sabine Meinck steered the project, contributed in major ways to the conceptional design of the research and structure of the paper, and wrote parts of the manuscript. Ann-Kristin Koop conducted supplementary analyses for the manuscript. She was responsible for definitions and explanatory parts in the manuscript and wrote parts of the paper. All authors read and approved the final manuscript.

**Funding**
The research is funded by the International Association for the Evaluation of Educational Achievement (IEA).

**Availability of data and materials**
The datasets generated and/or analyzed during the current are available in the following repositories: IEA, TIMSS: https://www.iea.nl/data-tools/repository/timss IEA, PIRLS: https://www.iea.nl/data-tools/repository/pirls OECD, TALIS: https://www.oecd.org/education/talis/talis-2018-data.htm.

## Declarations

**Ethics approval and consent to participate**
Not applicable.

**Consent for publication**
The authors give consent for the publication which can include data, graphics, and tables.

**Competing interests**
The authors declare that they have no competing interests.

Received: 22 November 2022 Accepted: 24 July 2024
Published online: 20 August 2024

**References**

Centurino, V.A.S., & Jones, L.R. (2017). TIMSS 2019 science framework. In: Mullis, I.V.S., Martin, M.O. (eds.) TIMSS 2019 Assessment Frameworks, Chestnut Hill, pp. 27–56 Chap. 2. http://timssandpirls.bc.edu/timss2019/frameworks/

Cochran, W.G. (1977). *Sampling techniques*, 3rd ed. Wiley.

Dumais, J., & Morin, Y. (2019). Sample design. In: Publishing, O. (ed.) TALIS 2018 Technical Report, pp. 96–108. OECD Publishing, Paris Chap. 5. http://www.oecd.org/education/talis/

Fishbein, B., Foy, P., & Liqun Yin, L. (2021). TIMSS 2019 User Guide for the International Database (2nd Ed.). TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College and International Association for the Evaluation of Educational Achievement (IEA), Chestnut Hill, MA. https://timss2019.org/international-database/downloads/TIMSS-2019-User-Guide-for-the-International-Database-2nd-Ed.pdf

Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Duckworth, D. (eds.). (2020). IEA International Computer and Information Literacy Study 2018 Technical Report. International Association for the Evaluation of Educational Achievement (IEA), Amsterdam . https://www.iea.nl/studies/iea/iccs/2016

Hájek, J. (1971). Discussion of "An essay on the logical foundation of survey sampling, part one" by D. Basu. In: Godambe, V.P., Sprott, D.A. (eds.) Foundations of Statistical Inference, p. 236. Holt, Rinehart, and Winston

Hooper, M., Broer, M., Yarnell, L. M., & Holmes, J. (2022). Talking about teachers: Would sampling weight adjustments allow for teacher-centric inferences in future timss assessments? *Studies in Educational Evaluation, 73*, 101148. https://doi.org/10.1016/j.stueduc.2022.101148

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association, 47*, 663–685. https://doi.org/10.1080/01621459.1952.10483446

Johansone, I. (2020). Survey operations procedures for TIMSS 2019. In: Martin, M.O., von Davier, M., Mullis, I.V.S. (eds.) Methods and Procedures: TIMSS 2019 Technical Report, Chestnut Hill, MA, pp. 6–1628 . Chap. 6. https://timssandpirls.bc.edu/timss2019/methods

Joncas, M., & Foy, P. (2012). Sample design in TIMSS and PIRLS. In: Martin, M.O., von Davier, M., Mullis, I.V.S. (eds.) Methods and procedures: TIMSS and PIRLS 2011, Chestnut Hill, MA, pp. 1–21 . Chap. 3. https://timssandpirls.bc.edu/methods/

Kish, L. (1965). *Survey sampling*. Wiley.

Kish, L., & Frankel, M.R. (1974). Inference from complex samples. Journal of the Royal Statistical Society. Series B (Methodological) 36, 1–37 . https://doi.org/10.2307/2984767

LaRoche, S., Joncas, M., & Foy, P. (2020). Sample design in TIMSS 2019. In: Martin, M.O., von Davier, M., Mullis, I.V.S. (eds.) Methods and Procedures: TIMSS 2019 Technical Report, Chestnut Hill, MA . Chap. 3. https://timssandpirls.bc.edu/timss2019/methods

Lindquist, M., Philpot, R., Mullis, I.V.S., & Cotter, K.E. (2017). TIMSS 2019 mathematics framework. In: Mullis, I.V.S., Martin, M.O. (eds.) TIMSS 2019 Assessment Frameworks, Chestnut Hill, pp. 11–26 . Chap. 1. http://timssandpirls.bc.edu/timss2019/frameworks/

Lohr, S. L. (1999). *Sampling: Design and analysis*. Duxbury Press.

Martin, M.O., & Mullis, I.V.S. (2004). Overview of TIMSS 2003. In: Martin, M.O., Mullis, I.V.S., Foy, P., Chrostowski, S.J. (eds.) TIMSS 2003 Technical Report, pp. 2–21. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College, Chestnut Hill, MA . Chap. 1. https://timssandpirls.bc.edu/timss2003i/technicalD.html

Martin, M.O., Mullis, I.V.S., Hooper, M. (eds.) (2017). Methods and Procedures in PIRLS 2016. TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College and International Association for the Evaluation of Educational Achievement (IEA), Chestnut Hill, MA . https://timssandpirls.bc.edu/publications/pirls/2016-methods.html

Martin, M.O., von Davier, M., Mullis, I.V.S. (eds.) (2020). Methods and Procedures: TIMSS 2019 Technical Report. TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College and International Association for the Evaluation of Educational Achievement (IEA), Chestnut Hill, MA . https://timssandpirls.bc.edu/timss2019/methods/

Meinck, S. (2015a). Computing sampling weights in large-scale assessments in education. survey insights: Methods from the field. Survey Insights: Methods from the Field, Weighting: Practical Issues and 'How to' Approach . https://doi.org/10.13094/SMIF-2015-00004

Meinck, S. (2015b). Sampling design and implementation. In: Fraillon, J., Schulz, W., Friedman, T., Ainley, J., Gebhardt, E. (eds.) International Computer and Information Literacy Study 2013 Technical Report, Amsterdam, pp. 67–86 . Chap. 6. https://www.iea.nl/publications/technical-reports/icils-2013-technical-report

Meinck, S., & Cortes, D. (2015). Sampling weights, nonresponse adjustments and participation rates. In: Fraillon, J., Schulz, W., Friedman, T., Ainley, J., Gebhardt, E. (eds.) International Computer and Information Literacy Study 2013 Technical Report, Chestnut Hill, MA, pp. 87–112 . Chap. 7. https://www.iea.nl/publications/technical-reports/icils-2013-technical-report

Meinck, S., & Vandenplas, C. (2021). In: Nilsen, T., Stancel-Piątak, A., Gustafsson, J.-E. (eds.) Sampling Design in ILSA, pp. 1–25. Springer, Cham . https://doi.org/10.1007/978-3-030-38298-8_25-1.

Mullis, I.V.S., Martin, M.O., Foy, P., Kelly, D.L., Fishbein, B. (eds.) (2020). TIMSS 2019 International Results in Mathematics and Science. TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College and International Association for the Evaluation of Educational Achievement (IEA), Chestnut Hill, MA . https://timssandpirls.bc.edu/timss2019/

OECD. (2014). TALIS 2013 Technical Report. OECD Publishing . https://www.oecd.org/education/school/TALIS-technical-report-2013.pdf

OECD. (2019a). PISA 2018 Assessment and Analytical Framework. OECD Publishing . https://doi.org/10.1787/b25efab8-en

OECD. (2019b). TALIS 2018 Technical Report. OECD Publishing . http://www.oecd.org/education/talis/TALIS_2018_Technical_Report.pdf

Schulz, W. (2020). The reporting of ICILS 2018 results. In: Fraillon, J., Ainley, J., Schulz, W., Friedman, T., Duckworth, D. (eds.) IEA International Computer and Information Literacy Study 2018: Technical Report, pp. 221–234. International Association for the Evaluation of Educational Achievement (IEA) 2020, Amsterdam. Chap. 13. https://www.iea.nl/publications/technical-reports/icils-2018-technical-report

Schulz, W., Carstens, R., Losito, B., Fraillon, J. (eds.). (2018). International Civic and Citizenship Education Study 2016: Technical Report. International Association for the Evaluation of Educational Achievement (IEA), Amsterdam. https://www.iea.nl/studies/iea/iccs/2016

Zuehlke, O., & Vandenplas, C. (2011). Sampling weights and participation rates. In: Schulz, W., Ainley, J., Fraillon, J. (eds.) ICCS 2009 Technical Report, Amsterdam, pp. 69–88. Chap. 7. https://www.iea.nl/studies/iea/iccs/

## Publisher's Note