Chapter 3
Methodology: Constructing a Socioeconomic Index for TIMSS Trend Analyses

Abstract To assess education system trends in the inequality of educational outcomes, a robust socioeconomic index for TIMSS trend analysis is needed. This chapter begins by outlining the TIMSS data and sample design, as well as changes in the sample design over time, with special emphasis on the aspects that are specifically related to family socioeconomic status and student achievement. To analyze trends over the entire 20 years of TIMSS data, the analysis was limited to education systems that participated in the first cycle in 1995, the most recent cycle in 2015, and at least one other administration in between. After assessing the completeness of the data, 13 educational systems were included in the study: Australia, Hong Kong, Hungary, Islamic Republic of Iran, Lithuania, New Zealand, Norway, Republic of Korea, Russian Federation, Singapore, Slovenia, Sweden, and the United States. Items used for constructing the SES index were the number of books at home, home possessions, and the highest level of education of either parent. Students in each educational system were grouped into high- and low-SES groups based on the SES distribution for a given year. Constructing a consistent measure of SES across all TIMSS cycles is an important contribution to research that uses TIMSS for trend analyses. In addition to analyzing the achievement gaps over time, examining trends in performance among low-SES students in each education system provides additional information on how education systems are addressing the issues facing disadvantaged students.

3.1   TIMSS Data and Sample Characteristics

We used the TIMSS grade eight public-use data from 1995 through 2015 to establish how inequalities of education outcomes have changed between 1995 and 2015, and to assess whether education systems have managed to increase the performance of disadvantaged students. TIMSS has been conducted every four years since 1995 to monitor trends in mathematics and science performance of students across education systems. Every participating education system provides a representative sample of students by adopting a two-stage random sample design. Typically, in each participating education system, a sample of schools is drawn at the first stage, and one or more intact classes of students from each of the sampled schools are selected at the second stage (LaRoche et al. 2016). Although most features have remained constant over the different TIMSS cycles, there have also been several significant changes in sample design, country participation, and questionnaire administration.

First, the target population has changed slightly. The first cycle of TIMSS in 1995 identified three target populations; one of them was students enrolled in the two adjacent grades, which maximized coverage of 13-year-olds (Foy et al. 1996). At the time of testing, most students were either in the grade seven or grade eight. This practice was refined for the 1999 cycle of TIMSS, and resulted in only grade eight students being assessed. To maintain comparability, for our study, we therefore only included grade eight students for most education systems in the 1995 assessment in our trend analyses, which is in alignment with the practice outlined in the *TIMSS 1999 international mathematics report* (Mullis et al. 2000) and *TIMSS 2015 international results in mathematics* (Mullis et al. 2016, appendix A.1, at http://timssandpirls.bc.edu/timss2015/international-results/timss-2015/mathematics/appendices/).[1] Norway was the only exception, because Norway only included grade six and seven students in its 1995 sample. However, according to the TIMSS 2015 report, the sample of upper-grade students (grade seven) in Norway in 1995 was comparable to that in 2015 (see Mullis et al. 2016, appendix A.1). Therefore, in the case of Norway, we kept the sample of grade seven students in 1995 for trend comparison (Gonzalez and Miles 2001).[2]

Second, although many education systems have participated in TIMSS over the last 20 years, not every education system participated in each cycle. To analyze trends over the entire 20 years of TIMSS data, we therefore limited our analysis to those education systems that participated in the first cycle in 1995, the most recent

---

[1]In the TIMSS 1999 international mathematics report Mullis et al. (2000) examined trends in mathematics achievement between 1995 and 1999. The 1995 average scale score was calculated for grade eight students only in Exhibit 1.3 (see pp. 34–36).

[2]According to the *TIMSS 1999 user guide for the international database*, the TIMSS 1999 target grade was the upper grade of the TIMSS 1995 population 2 and was expected to be the grade eight in most countries. However, for Norway, it was the seventh grade (see Exhibit 5.2 Grades tested in TIMSS 1995-Population 2). Please refer to https://timss.bc.edu/timss1999i/data/bm2_userguide.pdf

cycle in 2015, and at least one other intermediate administration cycle. This produced a potential sample of 18 education systems.[3]

However, according to the 2015 TIMSS international results in mathematics (see Mullis et al. 2016, appendix A.1), many education systems' previous data cannot be used for trend analysis to 2015. This is primarily due to improved translations or increased population coverage. For example, the data for Australia in 1999, Kuwait in 1995 and 2007, Canada in 1995 and 1999, Israel in 1995, 1999, 2003, and 2007, Slovenia in 1999, and Thailand in 1995 were not considered comparable to 2015 data. Therefore, four education systems (Canada, Israel, Kuwait, and Thailand) had to be excluded from the analyses because 1995 data cannot be used for trend analyses.

In addition, given that our primary focus is SES-related information, we excluded England from our study since it did not have data for parental education in 1995, 1999, and 2007. In total, our analytical sample is limited to the following 13 education systems (Table 3.1):

- Australia, Hong Kong, Hungary, Islamic Republic of Iran, Lithuania, Republic of Korea, Russian Federation, Singapore, Slovenia, and the United States (education systems that participated in all six cycles); and
- New Zealand (which participated in 1995, 1999, 2003, 2011, and 2015), Norway (which participated in 1995, 2003, 2007, 2011, and 2015), and Sweden (which participated in 1995, 2003, 2007, 2011, and 2015).

Finally, for trend analysis, several adjustments were made to follow the approach used by Mullis et al. (2016). First, IEA has a policy that students should not fall under the minimum average age of 13.5 years (for grade eight) at the time of testing (see Mullis et al. 2016, appendix C.10). Therefore, New Zealand assessed students in grade nine across multiple cycles. The results for grade nine students in 1995, 1999, 2011, and 2015 are deemed comparable to those for grade eight students who participated in 2003 in New Zealand. Second, although Slovenia assessed grade eight students in 1995, the results for grade eight students in 1995 are not deemed comparable to those in other cycles. Therefore, data for grade seven students in 1995 is used for trend analysis. Third, in Lithuania, the results for students assessed in Polish or Russian in 2015 are deemed not comparable to previous cycles. Therefore, trend results only include students assessed in Lithuanian and do not include students assessed in Polish or Russian in 2015.

---

[3]The potential sample included (1) 11 education systems that participated in all six cycles: Australia, Hong Kong, Hungary, Islamic Republic of Iran, Israel, Republic of Korea, Lithuania, Russian Federation, Singapore, Slovenia, the United States, and (2) seven education systems that participated in both 1995 and 2015 and in at least one other administration: England, New Zealand, Norway, Sweden, Thailand, Canada, and Kuwait.

Table 3.1 Samples for each education system in each TIMSS assessment year

Education system	Sample characteristics	TIMSS cycle					
		1995	1999	2003	2007	2011	2015
Australia	Overall student sample	12,852	4032	4791	4069	7556	10,338
	Grade level(s) used for trend analysis	G8	n/c	G8	G8	G8	G8
	Number of students in trend sample	7392	n/c	4791	4069	7556	10,338
Hong Kong	Overall student sample	6752	5179	4972	3470	4015	4155
	Grade level(s) used for trend analysis	G8	G8	G8	G8	G8	G8
	Number of students in trend sample	3339	5179	4972	3470	4015	4155
Hungary	Overall student sample	5978	3183	3302	4111	5178	4893
	Grade level(s) used for trend analysis	G8	G8	G8	G8	G8	G8
	Number of students in trend sample	2912	3183	3302	4111	5178	4893
Islamic Republic of Iran	Overall student sample	7429	5301	4942	3981	6029	6130
	Grade level(s) used for trend analysis	G8	G8	G8	G8	G8	G8
	Number of students in trend sample	3694	5301	4942	3981	6029	6130
Lithuania	Overall student sample	5056	2361	4964	3991	4747	4347
	Grade level(s) used for trend analysis	G8	G9	G8	G8	G8	G8
	Number of students in trend sample	2525	2361	4964	3991	4747	2933
New Zealand	Overall student sample	6867	3613	3801	n/a	5336	8142
	Grade level(s) used for trend analysis	G9	G9	G8	n/a	G9	G9
	Number of students in trend sample	3683	3613	3801	n/a	5336	8142
Norway	Overall student sample	5736	n/a	4133	4627	3862	4795
	Grade level(s) used for trend analysis	G7	n/a	G8	G8	G8	G8
	Number of students in trend sample	3267	n/a	4133	4627	3862	4795
Republic of Korea	Overall student sample	5827	6114	5309	4240	5166	5309
	Grade level(s) used for trend analysis	G8	G8	G8	G8	G8	G8
	Number of students in trend sample	2920	6114	5309	4240	5166	5309
Russian Federation	Overall student sample	8160	4332	4667	4472	4893	4780
	Grade level(s) used for trend analysis	G8	G8	G8	G8	G8	G8
	Number of students in trend sample	4022	4332	4667	4472	4893	4780

Table 3.1 (continued)

Education system	Sample characteristics	TIMSS cycle					
		1995	1999	2003	2007	2011	2015
Singapore	Overall student sample	8285	4966	6018	4599	5927	6116
	Grade level(s) used for trend analysis	G8	G8	G8	G8	G8	G8
	Number of students in trend sample	4644	4966	6018	4599	5927	6116
Slovenia	Overall student sample	5606	3109	3578	4043	4415	4257
	Grade level(s) used for trend analysis	G7	n/c	G7 and G8	G8	G8	G8
	Number of students in trend sample	2898	n/c	3578	4043	4415	4257
Sweden	Overall student sample	8855	n/a	4256	5215	5573	4090
	Grade level(s) used for trend analysis	G8	n/a	G8	G8	G8	G8
	Number of students in trend sample	1949	n/a	4256	5215	5573	4090
United States	Overall student sample	10,973	9072	8912	7377	10,477	10,221
	Grade level(s) used for trend analysis	G8	G8	G8	G8	G8	G8
	Number of students in trend sample	7087	9072	8912	7377	10,477	10,221

Source International Association for the Evaluation of Educational Achievement (IEA) Trends in International Mathematics and Science Study (TIMSS) 1995, 1999, 2003, 2007, 2011, and 2015 Mathematics and Science Assessments (see www.iea.nl/data)
Notes G7 grade seven, *G8* grade eight, *n/a* not applicable because the education system did not participate in this cycle, *n/c* not comparable to 2015 data

3.2 Construction of a Proxy Measure for Socioeconomic Status

To address the research questions, we first needed to construct a comparable proxy measure for socioeconomic status across the different TIMSS administration cycles. The TIMSS home educational resources (HER) index measures important aspects of SES, but it is not applicable for trend comparisons across all cycles for several reasons.

First and foremost, the HER index was constructed by different measurement methods in different cycles. In 1995 and 1999, the HER index was a simple combination of several background variables, including the number of books at home, number of home possessions, and parents' education, which were combined into three levels: high, medium, and low. For example, students at the high level were those with more than 100 books in the home, all three educational possessions (computer, study desk, and dictionary), and at least one college-educated parent. This index made interpretation easy since each category had its own corresponding characteristics. However, since 2011, the HER index has been constructed using IRT

Table 3.2 Home possession items by TIMSS cycle

Item	TIMSS cycle					
	1995	1999	2003	2007	2011	2015
Common items	Computer	Computer	Computer	Computer	Computer	Computer/ tablet
	Study desk	Study desk	Study desk	Study desk	Study desk	Study desk
Year-specific items	Dictionary	Dictionary	Dictionary	Dictionary	n/a	n/a
	Calculator	Calculator	Calculator	Calculator	n/a	n/a
	n/a	n/a	n/a	Internet connection	Internet connection	Internet connection
	n/a	n/a	n/a	n/a	Own room	Own room
	n/a	n/a	n/a	n/a	Books of your own	n/a
	n/a	n/a	n/a	n/a	n/a	Own mobile phone
	n/a	n/a	n/a	n/a	n/a	Gaming system

Notes n/a item was not present in this cycle of assessment

scaling methodology (Martin et al. 2011), which allows for the analysis of more fine-grained differences in home educational resources between students, and enables forward comparability for future administrations even if the components of the index should change in the future. The current form of the HER is, however, not comparable to the earlier index. In addition to that, in 2003 and 2007, no HER index was constructed for TIMSS.

Second, the components of the HER index changed because the available home possession items that students are asked about in the student questionnaire have changed over time (Table 3.2). For example, an internet connection at home was not part of the questionnaire before 2007, but is now an important component of the current HER scale. The only common items across all cycles are a computer and study desk. However, in 2015, the question regarding having a computer at home was also changed, resulting in two variables: one asking if a student owns a computer or tablet at home and a second one asking if a student shares a computer or tablet with others in the home.

It was clear that constructing a consistent measure of SES that can be applied across all TIMSS cycles would be of immense value to researchers who wished to use TIMSS for trend analyses. We therefore developed a modified version of the HER index to address this issue. Our SES measure, which we here term SES*, does not represent the full SES construct as usually defined by parental education, family income, and parental occupation.[4] While the construction of such an index serves a specific purpose in this study, we believe that the SES* index that is proposed here is

---

[4]An asterisk is added to denote the conceptual difference (Please refer to Chap. 1 for more details).

sufficiently closely related to the later IRT-scaled HER versions to yield highly relevant and valid results. This index can thus also be beneficially applied to other future studies that intend to use the SES* variable for analysis over multiple administrations.[5]

3.2.1 Components of the SES* Measure

Our SES* measure, which as mentioned in the introduction is a modified version of the HER index, is constructed using three common components across the six cycles of TIMSS. These components include (1) number of books at home, (2) number of home possessions, and (3) the highest level of education of either parent.

Number of Books at Home The information is derived from the student questionnaire asking how many books students have at home. There are five categories, coded (0) to (4): (0) 0 to 10 books; (1) 11 to 25 books; (2) 26 to 100 books; (3) 101 to 200 books; and (4) more than 200 books.

Number of Home Possessions This information comes from questions asking students whether they have each of a list of items at home. Since there are only two common items (computer and study desk) across all cycles, the total number of home possessions ranges from 0 to 2. One caveat needs to be mentioned for 2015. The question regarding having a computer at home was changed to two variables: one asking if a student owns a computer or tablet at home and the other one asking if a student shares a computer or tablet with others at home. We coded a positive response to either of these questions as a "1". Despite the addition of tablet in 2015, the correlations of the other SES* components with the computer/tablet variable were comparable with those found in 2011 (computer alone), with the scoring of either response as a "1". We therefore believe that the addition of tablet in 2015 did not substantially change the construct being measured, and that the SES* index remains consistent over time.

Highest Level of Education of Either Parent This is a derived variable constructed from both the father's and mother's highest educational levels. The categories of the source variables were grouped into five levels in line with the 1995 survey, coded as follows: (0) less than lower secondary; (1) completed lower secondary; (2) completed upper secondary; (3) postsecondary nontertiary education; and (4) completed university or higher. "I don't know" responses were treated as missing.

---

[5]Ideally, a scaled HER index could be constructed for prior years so that analysis with this index would be possible across all TIMSS administrations. However, this exceeds the scope of this research project.

3.2.2   Multiple Imputation of Missing Values

The main components of the SES* index have different degrees of missingness. Of specific concern is parental education, which on average has missing values of around 20%, depending on administration year and education system. Since dropping such a large part of the sample would undermine the generalizability of the findings, especially when the students with missing values tended to come from lower ability levels, multiple imputation was used for all missing values of the SES* index components. Instead of imputing the "highest level of parental education" variable directly, we imputed father's and mother's education separately, compared them after imputation, and then generated the highest level of parental education for the SES* index. We imputed the missing values of SES* index variables five times using multiple imputation chained equations before constructing the SES* index. Imputation using chained equations is known for its flexibility in handling different types of variables (for example binary, categorical, and continuous; Hughes et al. 2014), with our variables of interests being mostly categorical. The imputation is achieved by using the observed values for a given individual and the observed relations in the data for other participants (Schafer and Graham 2002).

In addition, since TIMSS data include multiple education systems across multiple years, we decided to impute the missing data for each year first and only then create a database of all years. The advantage of this approach was that we maximally used available information for a given year since the questionnaires have been modified over time and thus available relevant variables differ by year. In the imputation model, we included all analytic variables that were included in our final analysis, other common home possession items available for all education systems in each year, plausible values of achievement score, and other related variables (such as language spoken at home). After imputation, the correlation between these variables in each year was compared between the original dataset and the imputed dataset, and the results suggested the imputation preserved the overall relationship among variables very well. The student sampling weight was taken into account in the imputation model, as shown in a case study of conducting multiple imputation for missing data in TIMSS (Bouhlila and Sellaouti 2013).

3.2.3   The SES* Index

After imputation, we constructed the SES* index, ranging from 0 to 10 points, by assigning numerical values to each category of each of the three components (Table 3.3). We applied this to the 13 education systems' data for the 2011 and 2015 cycles, and found that this index has a relatively high correlation with the HER scale (2011: $r = 0.87$; 2015: $r = 0.84$). We also compared the variance in mathematics performance explained by the SES* index and by the HER scale in 2011 and 2015 for these 13 education systems. In 2011, the SES* index explained 23.7% of the variance in mathematics, while the HER index explained 23.6% of the

Table 3.3   SES* index construction

SES* component	Categories	Score
Highest level of parental education	Less than lower secondary education	0
	Completed lower secondary education	1
	Completed upper secondary education	2
	Post-secondary, non-tertiary education	3
	Completed university or higher	4
Home possessions	None	0
Computer/tablet	1 home possession	1
Study desk	2 home possessions	2
Number of books at home	0–10 books	0
	11–25 books	1
	26–100 books	2
	101–200 books	3
	More than 200 books	4

variance. In 2015, the SES* index explained 17.8% of the variance in mathematics, while the HER index explained 19.1% of the variance. This suggests that the proposed SES* index is highly correlated with the current HER scale and explains a similar amount of the variance in students' achievement.

The overall weighted distribution and corresponding average mathematics score for all participating education systems in 1995 and 2015 suggests that the distribution of this index is somewhat left skewed (Figs. 3.1 and 3.2). One possible explanation might be that many education systems in our analytic sample have an overall high level of SES*. More importantly, the results clearly suggest that each additional point in the SES* index is associated with higher average mathematics scores. In 1995, the TIMSS achievement score was scaled to have an international average value of 500 and a standard deviation of 100 points for participating countries. On average, the difference in mathematics scores between students with the lowest SES* (0 points) and the highest SES* (10 points) is around 150 points, which is 1.5 times the standard deviation of TIMSS scores. Furthermore, the positive correlation between the SES* index and mathematics scores is not only true overall but also holds across all education systems individually.

3.2.4   Defining High- and Low-SES* Groups

To calculate the achievement gap between students with high- and low-SES* backgrounds over time, we first needed to define the criterion or cut-off points corresponding to high- and low-SES* backgrounds. Among the different approaches for establishing cut-offs, the main choices are either (*a*) using common cut-offs across educational systems and years, or (*b*) defining education system-specific low-SES* versus high-SES* groups based on the distribution of the SES* index for a given year.
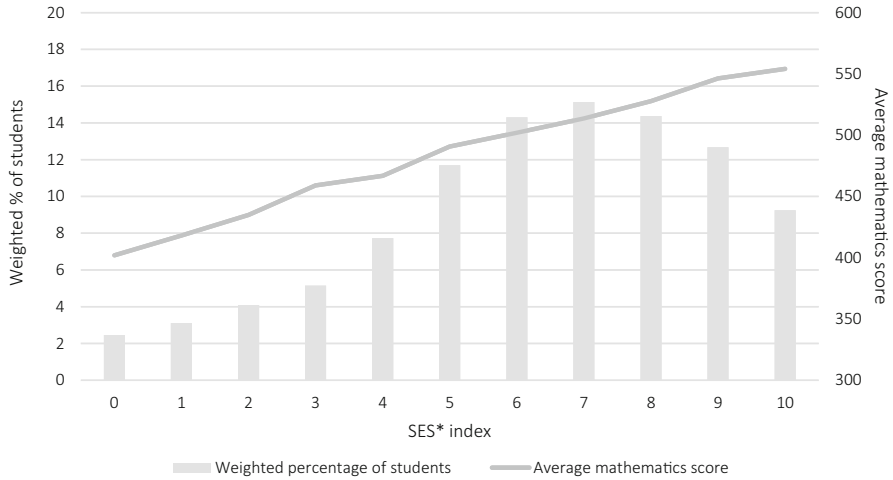
Fig. 3.1 Weighted percentage of students and average mathematics score by SES* index, 1995. (*Note* In 1995, 50,332 students in the 13 selected education systems were included in the analysis)
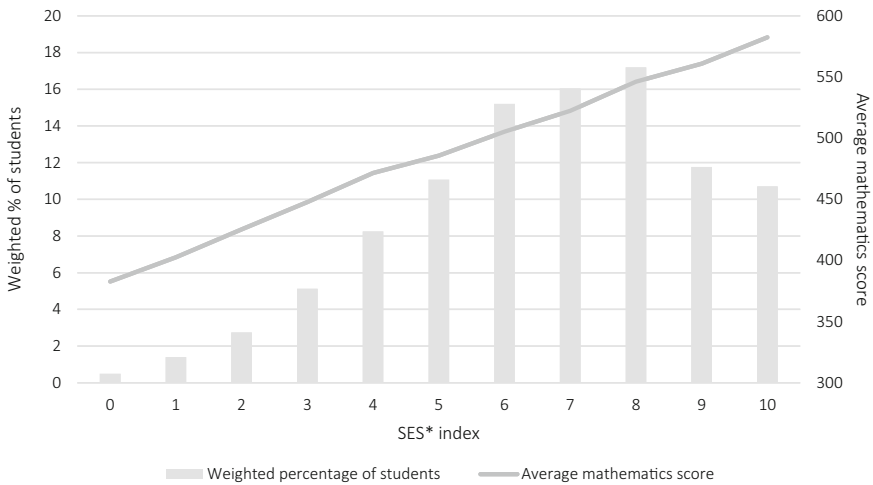

Fig. 3.2 Weighted percentage of students and average mathematics score by SES* index, 2015. (*Note* In 2015, 76,159 students in the 13 selected education systems were included in the analysis)

Common Cut-Offs   Given the weighted distribution of the sum-score SES* index for all students in all participating education systems across all 20 years, we found that an index value of three corresponded to about the 21st percentile of all students, whereas a value of eight points corresponded to the 81st percentile (see Table 3.4).

Table 3.4 Common cut-offs by overall distribution of SES* index (cumulative proportion)

SES* index	Average percentage cut-off	Australia (1995) percentage cut-off	Islamic Republic of Iran (1995) percentage cut-off
0	3	3	22
1	7	4	44
2	13	6	63
3	21	10	76
4	31	18	84
5	42	30	91
6	56	46	95
7	68	60	97
8	81	74	99
9	91	87	100
10	100	100	100

As a first test, we applied these cut-off points to all students. Students with three or fewer points were defined as the low-SES* group and those with eight or more points were defined as the high-SES* group. As can be expected, this approach led to very unbalanced groups when the results were examined by education system. For example, in Australia in 1995, only 10% of students would have been placed into the low-SES* group, while 26% would have been in the high-SES* group. By contrast, in Iran, about 76% of students would have been placed in the low-SES* group, with only 1% in the high-SES* group (Table 3.4).

Thus, common cut-offs tend to generate unbalanced groups in certain education systems since individual education systems' specific situations are not taken into account. While these may be the actual percentages for high- and low-SES* students across educational systems, SES* is a relative concept when viewed within an educational system. That is, what is perceived as high or low SES* is society dependent. And it is the perception which is important, because what is perceived to be real is real in its consequences. Therefore, we decided to establish education system specific cut-offs for each year. Given each education system's distribution of SES* in each year, we used quartiles as cut-offs; students in the bottom quartile were considered low SES*, while students in the top quartile were considered high SES* (see the Appendix for a sensitivity analysis using quintiles versus quartiles and additional information). This approach generated better grouping results because it takes local context into consideration.

Another challenge was how to establish exact 25th or 75th percentile cut-offs using an index with a range of only 11 points in total. Considering the cumulative proportions of students at each SES* point in Australia in 1995 (Table 3.5), we found that students with eight points on the index corresponded to the 73th percentile, while students with nine points corresponded to the 86th percentile. Establishing the bottom quartile was also difficult, since four points corresponded to the 15th percentile, while five points corresponded to the 27th percentile.

To address this issue, we decided to randomly split the sample of students at the cut-off point below or above the 25th and 75th percentiles and then combine it with a

Table 3.5 Weighted distribution of SES* index for Australia, 1995

SES*	Proportion (%)	Cumulative proportion (%)
0	0	0
1	1	1
2	2	3
3	4	6
4	8	15
5	13	27
6	16	43
7	15	59
8	14	73
9	13	86
10	14	100

Note The results are based on one of the imputed datasets

random subsample from the adjacent group, resulting in top and bottom categories containing 25% of students. Again, using Australia in 1995 as our example, to obtain the bottom quartile, we needed another 10% of students in addition to those having 0 to 4 points on the SES* index. Therefore, we randomly selected a subsample of the Australian students who participated in 1995 and who scored five SES* index points to create a sample comprising 25% of students as the bottom SES* category (another way to consider this is that if 27% of students are at index point five, then the sample contains 2 % more students than needed for the bottom quartile, so 2% of students, in absolute terms, have to be randomly excluded from the five-point subsample). Applying the same strategy to every individual education system and year guaranteed that the bottom- and top-quartile SES* groups always represented exactly 25% of students from a given education system in any given year.

3.3 Analytic Approach

3.3.1 Plausible Values and Imputed Datasets

One significant analytic challenge underlying this work was how to simultaneously use the existing five plausible values of achievement scores while incorporating results from the multiple imputation procedure for the missing values of SES* background variables. One approach might be to conduct nested multiple imputation, in which the plausible values imputation is nested within the background variable imputation (Weirich et al. 2014). However, that would have required an extra step back to item responses, and the imputation model would highly depend on the final analytic model, meaning that other studies using this SES* index would have to create their own models. More importantly, the TIMSS & PIRLS International Study Center had clearly stated that principal components for a large number of student background variables were included as conditioning variables to improve the reliability of the estimated student proficiency scores (Foy and Yin 2016). It is reasonable to

believe that the components in our SES* index, which are very important student background variables, were included in the TIMSS conditioning models for proficiency estimation. Therefore, we used the existing plausible values of achievement scores in TIMSS to impute missing values in the SES* component variables, together with other relevant variables, resulting in five imputed datasets.

After imputation, one possibility for using the imputed SES* variable was to average the SES* values among the five imputed datasets and thus generate a single SES* index score for each student. To validate this approach, we randomly selected 10% of cases in each country, replaced the existing value of parental education with "missing", imputed the pseudo-missing values using the same imputation model, and then compared the imputed values with actual values. However, the validation results were not satisfactory, since a simple average of the five imputed values presented a quite different distribution from the actual values because it overlooked the variance between the imputed values. Therefore, we decided not to average the imputed values for SES* but to treat the five imputed SES* values as plausible values (Kevin Macdonald, personal communication, 10 March 2018) and conduct analyses with the PV module in Stata 14 software (Macdonald 2008). This approach allowed us to simultaneously use the five plausible values of the TIMSS achievement scores and the five imputed values for the SES* index in the analyses for this report.

3.3.2   Measuring Educational Inequality

Ferreira and Gignoux (2011) described methods for measuring both inequality in achievement (which they saw as being expressed simply by the degree of variability in the outcome measure), and inequality in opportunity (they proposed a meaningful summary statistic for this would be the amount of variance explained obtained from an OLS regression of students' test scores on a vector C of individual circumstances). Another approach was used by Green et al. (2015) in an application using international adult skills surveys. Their measure was a "social origins gradient" representing the point difference in scores that can be predicted for an individual when the education level of his or her parent(s) is increased from the bottom unit to the top unit (for example from "less than high school" to "college education").

We opted for yet another different approach, one that we believe is better suited for trend analysis of educational inequality. To answer the first research question, *"How has the inequality of education outcomes due to family socioeconomic status changed for different education systems between 1995 and 2015"*, we calculated the achievement gap over time between students in low- and high-SES* groups in terms of the average TIMSS achievement score. The larger the gap, the larger the role of SES* in determining educational outcomes.

In addition, we examined whether the changes in achievement gap between high- and low-SES* students across years were statistically significant. Since these calculations are computationally quite demanding, we focused on providing significance testing for changes in achievement gaps only between the following

years: (1) 1995 versus 2003, (2) 2003 versus 2015, and (3) 1995 versus 2015. For example, to investigate if the change in the gap between 1995 and 2003 was statistically significant, the following regression model was conducted:

$$\widehat{Y}_i = \beta_0 + \beta_1(SES^*) + \beta_2(Year_j) + \beta_3(SES^* * Year_j) + \varepsilon_i$$

Where $\widehat{Y}_i$ is the predicted achievement score (that is, either mathematics or science) for student i in a given education system after controlling for other predictors; β_0 is the mean achievement score for low-SES* students in a given education system in 1995; β_1 is the mean score difference between low- and high-SES* students in a given education system in 1995; and β_2 is the coefficient for a categorical variable indicating the year of assessment. The reference group is 1995, therefore, the coefficient is the mean score difference between students who participated in 2003 and those in 1995, after controlling for other predictors. Meanwhile β_3 is the coefficient for an interaction term between SES* and the assessment year. This reflects how much the achievement gap between low- and high-SES* students in 2003 differs from the achievement gap in 1995, and, therefore, the p-value for β_3 indicates whether the achievement gap in 2003 is statistically different from the achievement gap in 1995. Following the same logic, we conducted similar comparisons of the achievement gaps between 2003 and 2015, and between 1995 and 2015.

While seeing trends in the SES achievement gaps is important, they can hide important changes over time. For example, there might be no change in the size of the SES* gap over time because neither group has changed over time, and, in another case, the SES* gap may not change because both the lower and upper groups have changed in the same direction over time. Because gaps can close or widen for different reasons, it is also important to examine how the most disadvantaged students are doing over time, as proposed by our second research question, *"To what extent have education systems managed to increase the academic performance of disadvantaged students between 1995 and 2015?"* To address this, we analyzed the trend in performance among low-SES* students in each education system from 1995 to 2015. Specifically, we tracked the percentage of low-SES* students who performed at or above the TIMSS international intermediate benchmark (that is, 475 points) for each education system over time.

3.3.3 Country-Level Indicators in the Educational Systems and the Macroeconomic Context

To better understand our findings in the larger context in which education systems operate, we obtained macroeconomic and other indicators from the TIMSS encyclopedias as well as data from external sources from 1995 to 2015. The external sources we consulted include the World Bank, the UNESCO Institute for Statistics, the CIA's World Factbook, the OECD Income Distribution Database, the

Table 3.6 Sources for country-level economic indicators

Indicator	Source	Link
GDP per person (current US$)	The World Bank Open Data	https://data.worldbank.org/indicator/NY.GDP.PCAP.CD
	The World Factbook 2018. Washington, DC: Central Intelligence Agency	https://www.cia.gov/library/publications/the-world-factbook/index.html
Total percent of government expenditure on education	The World Bank Open Data	https://data.worldbank.org/indicator/se.xpd.totl.gb.zs
	The UNESCO Institute for Statistics	http://data.uis.unesco.org/Index.aspx?queryid=183
Total percent of GDP spent on education	The World Bank Open Data	https://data.worldbank.org/indicator/se.xpd.totl.gd.zs
	The UNESCO Institute for Statistics	http://data.uis.unesco.org/?queryid=181
Gini index	The World Bank Open Data	https://data.worldbank.org/indicator/SI.POV.GINI
	The World Factbook 2018. Washington, DC: Central Intelligence Agency	https://www.cia.gov/library/publications/the-world-factbook/rankorder/2172rank.html
	The OECD Income Distribution Database	https://stats.oecd.org/index.aspx?queryid=66670
Top 10% share pre-tax national income	The World Inequality Database	http://wid.world/data/

World Inequality Database, and local education agencies (see Table 3.6). We used this to interpret our findings against changes in the social context of each education system over the 20 years of TIMSS.

References

Bouhlila, D. S., & Sellaouti, F. (2013). Multiple imputation using chained equations for missing data in TIMSS: A case study. *Large-scale Assessments in Education, 1*(1), 4.

Ferreira, F. H. G. & Gignoux, J. (2011). *The measurement of educational inequality: Achievement and opportunity*. Policy Research working paper series no. 5873. Washington, DC: The World Bank.

Foy, P., & Yin, L. (2016). Scaling the TIMSS 2015 achievement data. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in TIMSS 2015* (pp. 13.1–13.62). Chestnut Hill: TIMSS & PIRLS International Study Center, Boston College. Retrieved from http://timss.bc.edu/publications/timss/2015-methods/chapter-13.html

Foy, P., Rust, K., & Schleicher, A. (1996). Sample design. In M. O. Martin & D. L. Kelly (Eds.), *Third International Mathematics and Science Study (TIMSS) technical report, Volume I: Design and development* (pp. 87–91). Chestnut Hill: TIMSS & PIRLS International Study Center, Boston College.

Gonzalez, E. J., & Miles, J. A. (Eds.). (2001). *TIMSS 1999: User guide for the international database*. Chestnut Hill: TIMSS & PIRLS International Study Center, Boston College. Retrieved from https://timss.bc.edu/timss1999i/data/bm2_userguide.pdf

Green, A., Green, F., & Pensiero, N. (2015). Cross-country variation in adult skills inequality: Why are skill levels and opportunities so unequal in Anglophone countries? *Comparative Education Review, 59*(4), 595–618.

Hughes, R. A., White, I. R., Seaman, S. R., Carpenter, J. R., Tilling, K., & Sterne, J. A. (2014). Joint modelling rationale for chained equations. *BMC Medical Research Methodology, 14*(1), 1–17.

LaRoche, S., Joncas, M., & Foy, P. (2016). Sample design in TIMSS 2015. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in TIMSS 2015* (pp. 3.1–3.37). Chestnut Hill: TIMSS & PIRLS International Study Center, Boston College. Retrieved from http://timss.bc.edu/publications/timss/2015-methods/chapter-3.html

Macdonald, K. (2008). *PV: Stata module to perform estimation with plausible values*. Statistical Software Components S456951. Boston College Department of Economics, revised 12 Feb 2014. Retrieved from https://ideas.repec.org/c/boc/bocode/s456951.html

Martin, M. O., Mullis, I. V., Foy, P., & Arora, A. (2011). Creating and interpreting the TIMSS and PIRLS 2011 context questionnaire scales. In M. O. Martin & I. V. S. Mullis (Eds.), *Methods and procedures in TIMSS and PIRLS* (pp. 1–11). Chestnut Hill: TIMSS & PIRLS International Study Center, Boston College. Retrieved from https://timssandpirls.bc.edu/methods/pdf/TP11_Context_Q_Scales.pdf

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Gregory, K. D., Garden, R. A., O'Connor, K. M., Chrostowski, S. J., & Smith, T. A. (2000) *TIMSS 1999 international mathematics report. Findings from IEA's repeat of the Third International Mathematics and Science Study at the eighth grade.* Chestnut Hill: TIMSS & PIRLS International Study Center, Boston College. Retrieved from https://timss.bc.edu/timss1999i/pdf/T99i_Math_All.pdf

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS 2015 international results in mathematics*. Chestnut Hill: TIMSS & PIRLS International Study Center, Boston College. Retrieved from http://timssandpirls.bc.edu/timss2015/international-results

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147.

Weirich, S., Haag, N., Hecht, M., Böhme, K., Siegle, T., & Lüdtke, O. (2014). Nested multiple imputation in large-scale assessments. *Large-scale Assessments in Education, 2*(1), 9.