Chapter 5
Teacher Quality and Mean Student Outcomes: A Multi-model Approach

Abstract IEA's Trends in International Mathematics and Science Study (TIMSS) can potentially be used to address important questions about the role of teachers in influencing student outcomes. To establish the potential relationships between student achievement and different types of teacher quality and instructional metrics and how stable these relationships may be across time, a variety of different statistical approaches were implemented and compared. The results from a single-level (unclustered) model, a multilevel model, and a classroom-mean regression model were compared, as was the stability of within-country estimates across time. The potential utility of country-level fixed-effects analysis was also explored. The complex sampling design of TIMSS enables robust standard errors to be generated. The analyses suggested only a weak relationship between teacher quality measures and student outcomes across countries and grade levels, and considerable instability in within-country estimates across time, especially at grade eight.

Keywords International large-scale assessment (ILSA) · Opportunity to learn · Statistical methodology · Teacher education · Teacher experience · Teacher quality · Trends in International Mathematics and Science Study (TIMSS)

5.1 Introduction

To date, research into relationships between teacher characteristics and student outcomes has relied on data from only a few countries, or been restricted to analysis of data collected at single point in time, or both. Such limitations have made it difficult to draw general conclusions about whether particular measures of teacher quality (such as, instructional alignment) are systematically associated with higher learning gains, or whether any such gains are due to other factors or the unique circumstances of a given educational system. In this chapter, we take advantage of the measures of teacher quality within the TIMSS framework, and the extensive international data collected by TIMSS over 20 years, to investigate this problem more thoroughly, using a variety of statistical methods. Our aim was twofold: first, to identify whether given teacher-related metrics were generally related to student

N. Burroughs et al., *Teaching for Excellence and Equity*, IEA Research for Education 6,
https://doi.org/10.1007/978-3-030-16151-4_5

mean outcomes within and across countries; and second, whether these relationships were sensitive to the statistical method employed or a particular sample (namely, data collected from one cycle of TIMSS). This chapter can be considered to be an extended robustness check on the association between teacher quality and student mathematics performance, comparing particular results to different educational contexts (namely differing national education systems), different years, and different methodological choices.

A secondary consideration was the challenge of balancing rigorous quantitative methods with the practical problems of dealing with large-scale datasets. Sophisticated multilevel models with very large sample sizes, such as would be necessary if incorporating all of the TIMSS countries in every cycle into one model, can be computationally quite demanding, even with modern computing power. More sophisticated methods would likely identify more precise and less biased estimates. This chapter presents an exploratory study, aiming to assess whether there were obvious patterns across time and space, preliminary to more detailed research.

We begin by assessing how frequently the relationship between measures of teacher effectiveness and student outcomes were statistically significant across countries. As noted in Chap. 2, the existing research literature has failed to identify any consistent relationships between teacher characteristics (experience, education, teacher content knowledge) and student outcomes, but has indicated that student outcomes may be more strongly associated with teacher behaviors (content coverage and time on teaching mathematics). Thus, we aimed to determine whether this was a consistent finding across different educational systems, and how sensitive these findings were to the statistical method used. To do this, we compared the results of a "full" model, incorporating teacher- and student-level effects, with those from a model that ignored student clustering within classrooms and classroom-level means.

Having considered the effect of the different statistical methods on our findings, we next investigated the stability of these associations across time. Although the country-level averages of teacher quality measures may vary over time, the direction and strength of relationship between teacher and student factors should be more consistent, assuming low measurement error, sufficiently sensitive and reliable instrumentation, and rough institutional stability in a particular educational system. We analyzed the stability of statistical estimates across time, first by comparing multilevel regression coefficients for a given country across multiple years of the TIMSS, and next by conducting a fixed-effect analysis using country-level means. Finally, we assessed the robustness of multilevel model results by comparing a simplified means of calculating standard errors with the more elaborate jackknifing procedure recommended by the various TIMSS user guides.

All of the analyses in this chapter use a basic additive model in which student mean achievement in mathematics is predicted by six teacher variables (experience, preparedness, education, alignment, time spent on mathematics, and teacher gender) and three student control variables (books in the home, language spoken in the home, and student gender). The operationalization of these variables has already been discussed in Chap. 3. We applied this model to each education system participating in TIMSS for every cycle of participation, using standard jackknifed standard errors and five plausible values. Education systems that did not include one of the five main

teacher variables (experience, education, preparedness, alignment, and time spent on mathematics) were excluded from the analysis. Each education system was analyzed separately by year and grade level.

The analysis presented in the main text is summary data combining the results for multiple educational systems. (For detailed country-level results for each statistical model, please consult Appendix B.)

5.2 Consistency Across Pooled, Multilevel, and Classroom-Means Models

We tested the basic linear model using three different statistical models. The first is a simple pooled within-country model using ordinary least squares (OLS) regression. The second is a multilevel model that clusters students within classrooms, with student variables at level one and teacher variables at level two. The third model aggregates student-level variables at the classroom level to create classroom means (in other words, classrooms rather than students are the unit of analysis). These classroom-mean results are analyzed using a single-level model.

We focused on five teacher-level variables: teacher experience (Exp), teacher education to teach mathematics (Mathprep), time spent on teaching mathematics (Mathtime), instructional alignment with national standards (Alignment), and self-reported preparedness to teach mathematics (Prepared). Our analyses included participating TIMSS education systems where all these variables were available and excluded systems where some variables were unavailable. Consequently, we applied our three types of model (pooled, multilevel, and classroom means) to 307 cases at both grades four and eight for the 2003, 2007, 2011, and 2015 cycles of TIMSS. Each of the 307 cases represents an educational system in a single cycle of TIMSS. With three statistical models per case, this comprises 921 separate regressions, with 1535 teacher effectiveness variables compared across models ($307 \times 5 = 1535$) (see Table 5.1).

In terms of the consistency of statistical inference across all three models, our results showed that, among the 1535 comparisons, 1151 (75.0%) produced similar estimates, either significant or non-significant, across the pooled, multilevel, and classroom-means models. Among those comparisons with consistent estimates of significance, 78 (5.1%) comparisons were significant with same direction, and 1073 (69.9%) cases were statistically non-significant ($p > 0.05$). That is, although the estimates of regression coefficients and the standard errors were slightly different across the three models, two-thirds of them were substantively identical. In over two-thirds of comparisons, none of the models identified a statistically significant effect of teacher effectiveness measures on student outcomes.

There was one exceptional case where all three model estimates were statistically significant, but in opposite directions. This unusual case considered alignment of the grade eight curriculum in Malta in the 2007 cycle of TIMSS; coefficient estimates for the single-level and classroom-means models were 160.3 and 103.5,

Table 5.1 Number of model regressions by grade and cycle of TIMSS

Grade	Cycle of TIMSS	Number of education systems with required data	Number of model[a] regressions
Grade four	2003	12	36
Grade four	2007	35	105
Grade four	2011	48	144
Grade four	2015	45	135
Grade eight	2003	31	93
Grade eight	2007	50	150
Grade eight	2011	45	135
Grade eight	2015	41	123
Total		307	921

Notes [a]Models may be pooled single-level models, multilevel models, or classroom-means models

respectively (positive and statistically significant), however, the coefficient estimate for the multilevel model was −33.0 (negative and statistically significant).

We also compared the three models in pairs to assess the statistical significance and directional discrepancy of the regression coefficient estimates. This approach looked at partial consistency among two statistical models, rather than across all three. Sixty-one (4.0%) comparisons were statistically significant with same direction for both the single-level and multilevel model, but not for classroom-means model. For instance, the estimates of grade four teacher experience in Iran for 2003 using the single-level and multilevel models were 1.2 and 1.5, respectively, but the estimate of the same predictor using the classroom-means model was −0.6, which was not statistically significant.

In some cases, even though both the single-level and multilevel models produced statistically significant estimates, the directions opposed each other. For example, when analyzing grade eight data for Jordan in 2007, while the estimate of preparedness in single-level model was 32.1, it was −9.9 using the multilevel model. Similarly, grade eight teacher preparedness for Tunisia in 2007 was 8.7 with single-level model, but −5.5 with the multilevel model.

When making comparisons between the multilevel and classroom-means models, there were 23 (1.5%) comparisons where the estimates of both models were statistically significant with same direction, where this was not true for single-level model. For instance, the estimate of grade four teacher experience for Chinese Taipei in 2003 was identical (0.3) across the three models, but this was not significant for the single-level model (the magnitude of the estimate was small, however). There was also one exceptional case where estimates of time spent on teaching mathematics were significant for both models but in opposite directions. For Iran in 2015 at grade eight, the time spent on teaching mathematics had a positive coefficient of 0.1 in the

multilevel model, but a coefficient of –0.1 for the classroom-means model. However, given the very small effect size, this inconsistency should not be overstated.

There were more consistent results between the single- and classroom-means models. We found that there were 52 (3.4%) comparisons that were statistically significant and in the same direction, where this was not the case for multilevel model. Interestingly, there were no comparisons between the single-level and classroom-mean models that showed opposing directions. However, there were 125 (8.1%) comparisons where the estimates were significant only for the multilevel model.

Focusing on our five teacher-level variables, coefficients differed across the three models with regard to statistical significance for each variable. On examination of the 307 cases (countries per year) for each teacher effectiveness variable, we were able to classify the results into eight categories: (1) significant for only the multilevel model, (2) non-significant for only the multilevel model, (3) significant for both the single-level and multilevel models, (4) significant for both the multilevel and classroom-means models, (5) non-significant for both multilevel and classroom-means models, (6) non-significant for both single-level and multilevel models, (7) significant for all three models, and (8) non-significant for all three models (Fig. 5.1).

Our analysis strongly suggests that, overall, measures of teacher effectiveness have a weak and inconsistent association with student outcomes. The vast majority of the time (66–75% of the time), the three statistical models were in agreement that there was no statistically significant relationship with mean mathematics performance for any measure of teacher effectiveness.

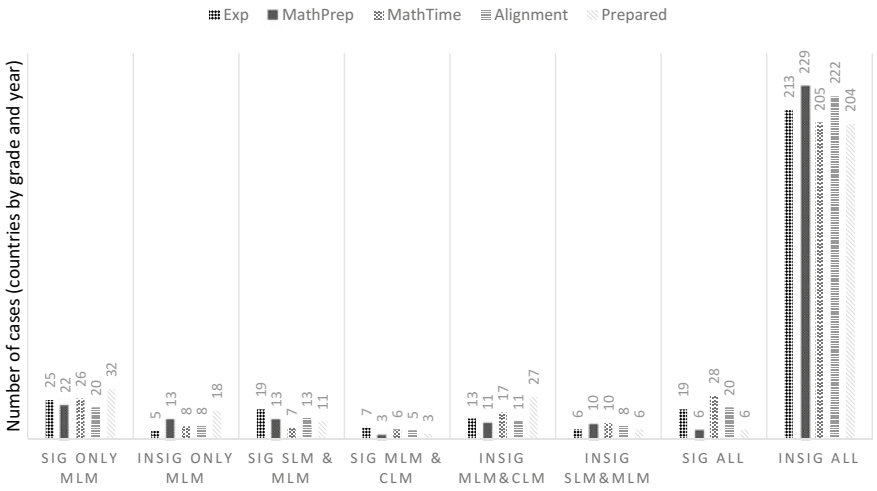

Fig. 5.1 The number of significant (SIG) and non-significant (INSIG) associations between teacher variables and student outcomes by educational system using single-level (SLM), multilevel (MLM), and classroom-means (CLM) models, based on data from TIMSS 2003–2015

When we considered specific teacher-level variables, for teacher experience there were 25 of the 307 cases (8.1%) where the estimates were statistically significant for only the multilevel model. In contrast, five (1.6%) cases of the estimates were statistically non-significant for only the multilevel model. Regarding significance in pairs, there were 19 (6.2%) cases where the estimates were statistically significant for both the single-level and multilevel models. We also found out that seven (2.3%) cases were statistically significant for only the multilevel and classroom-means models. Conversely, in terms of non-significance in pairs, in 13 (4.2%) cases the estimates were non-significant for both the multilevel and classroom-means models, and in six (2.0%) cases the estimates were non-significant for both single-level and multilevel models. Lastly, with respect to consistency across three models, in 19 (6.2%) cases the estimates were statistically significant across all three models. In the remaining 213 (69.4%) cases the estimates were non-significant for all three models.

For teacher education to teach mathematics (Mathprep), the estimates were statistically significant for only the multilevel model in 22 (7.2%) of the 307 cases. In contrast, there were 13 (4.2%) cases where the estimates were statistically non-significant for only the multilevel model. Regarding significance in pairs, the estimates were statistically significant for only the single-level and multilevel models in 13 (4.2%) cases. We also found that only three (1.0%) cases were statistically significant for only the multilevel and classroom-means models. Conversely, in terms of non-significance in pairs, in 11 (3.6%) cases the estimates were non-significant for both the multilevel and classroom-means models, and there were 10 (3.3%) cases where the estimates were non-significant for both the single-level and multilevel models. Lastly, in six (2.0%) cases the estimates were statistically significant across all three models, while in the remaining 229 (74.6%) cases the estimates were non-significant for all three models.

For time spent on teaching mathematics (Mathtime), the estimates were statistically significant for only the multilevel model in 26 (8.5%) of the 307 cases. In contrast, the estimates were statistically non-significant for only the multilevel model in eight (2.6%) cases. Regarding significance in pairs, there were seven (2.3%) cases where the estimates were statistically significant for both the single-level and multilevel models. We also found that only six (2.0%) cases that were statistically significant for both the multilevel and classroom-means models. Conversely, in terms of non-significance in pairs, 17 (5.5%) estimates were non-significant for both the multilevel and classroom-means models, and 10 (3.3%) estimates were non-significant for both the single-level and multilevel models. Lastly, considering consistency across three models, 28 (9.1%) estimates were statistically significant across all three models and the remaining 205 (66.8%) estimates were non-significant for all three models.

For teacher alignment with national standards, estimates were statistically significant for only the multilevel model in 20 (6.5%) of the 307 cases. In contrast, the estimates were statistically non-significant for only the multilevel model in eight (2.6%) cases. Regarding significance in pairs, there were 13 (4.2%) cases where the estimates were statistically significant for both the single-level and multilevel models. We also found that only five (1.6%) cases were statistically significant for both the multilevel and classroom-means models. In terms of non-significance

in pairs, eleven (3.6%) estimates were non-significant for both the multilevel and classroom level model, and in eight (2.6%) cases the estimates were non-significant for both the single-level and multilevel models. Lastly, 20 (6.5%) of the estimates were statistically significant across all three models, although, among these 20 cases, as we mentioned previously, for the grade eight curriculum in Malta in the 2007 cycle of TIMSS, the estimates for the single-level and classroom-means models were positive while the estimate for the multilevel model was negative. The remaining 222 (72.3%) estimates that were non-significant for all three models presented.

For preparedness, there were 32 (10.4%) of 307 cases where the estimates were statistically significant only for the multilevel model. In contrast, 18 (5.9%) estimates were statistically non-significant only for the multilevel model. There were 11 (3.6%) cases in which the estimates were statistically significant for both single-level and multilevel models; among these 11 cases, the directions of the multilevel model estimates for grade eight data for Jordan and Tunisia from 2007 were −9.9 and −5.5, respectively. We also determined that only three (1.0%) cases were statistically significant for both multilevel and classroom-means models. Conversely, in 27 (8.8%) cases the estimates were non-significant for both multilevel and classroom-means models, and in six (2.0%) instances the estimates were non-significant for both the single-level and multilevel models. Lastly, only six (2.0%) educational systems had estimates that were statistically significant across all three models; the remaining 204 (66.4%) systems had estimates that were non-significant for all three models.

When we focused only on the multilevel model, we found that the estimates of 291 (19.0%) cases showed there was a significant relationship between one of the measures of teacher quality (either a characteristic or a behavior) and student performance in mathematics, and there were 1244 (81.0%) non-significant relationships. Among the 291 statistically significant cases, student performance was most commonly related to teacher experience (in 70 cases [22.8%]), followed by teacher education (44 cases [14.3%]), time spent on teaching mathematics (67 cases [21.8%]), alignment (59 cases [18.9%]), and preparedness (52 cases [16.9%]).

Finally, assuming that the results from the multilevel model are unbiased (or the least biased) estimates because of the substantive and empirical importance of student clustering within classrooms, there is a possibility that using the pooled or classroom-mean models leads to incorrect statistical inference. We found that there were 171 (11.1%) cases in which null hypotheses that coefficients were zero in population were incorrectly rejected (type 1 error) by either or both the single-level and classroom-means models. Moreover, there were 125 (8.1%) cases that failed to reject or incorrectly retained the null hypotheses (type 2 error). Although there are circumstances where classroom-means and single-level (non-clustered) analyses may yield substantively similar results, there were enough differences to warrant employing the more complex and computationally burdensome multilevel model.

5.3 Stability of Estimates Across Time

An alternative way of testing the robustness of the associations between teacher quality measures and student outcomes is to examine the stability of estimates for a particular country across time. Although changes in country-level means of (for example) teacher experience could be explained by substantive policy or labor market conditions, the relationship between teacher experience and student outcomes should not vary dramatically over a short span of time within a particular educational system. To test whether this assumption holds, the variables' coefficients estimated for each year were compared to establish whether there were significant differences between them.

We based the theoretical framework used for the comparison procedure on the work of Clogg et al. (1995). They argued that, in large samples, the significance of the difference between the coefficients could be assessed using the following statistics:

$$Z = \frac{\widehat{\beta_1} - \widehat{\beta_2}}{\sqrt{(SE\widehat{\beta_1})^2 + (SE\widehat{\beta_2})^2}}$$

where $\widehat{\beta_1}$ and $\widehat{\beta_2}$ are regression coefficients from the models that are to be compared and SE is the associated standard error of the said estimates. The null hypothesis for such a test would be the equality of the coefficients. When employing a specific statistical test, the underlying assumptions of the test should be satisfied. Clogg et al. (1995) cautioned against the independence of coefficients when using this formula. Since these estimates were taken at different times, there is no reason to suspect dependency of samples and as such, this statistic is valid for the purpose at hand. However, since coefficients from multiple time points are being compared, the issue of multiple comparisons must be addressed (Curran-Everett 2000), in which the error rate increases as increasing numbers of pairs are compared. Several different methods have been suggested to adjust for this problem (Benjamini and Hochberg 1995; Hochberg 1988; Hommel 1988; Weisstein 2004). Here we adopted the correction method suggested by Benjamini and Hochberg (1995), which emphasized control for a false discovery rate while also minimizing the family-wise error rate.

The variables that we assessed were:

- The number of years the teacher has been teaching (Exp)
- A teacher's formal education to teach mathematics (Mathprep)
- Time spent on teaching mathematics (Mathtime)
- Alignment of the topics taught with national standards (Alignment)
- A teacher's self-reported preparation to teach mathematics topics (Prepared).

After sorting the data available from each country and each sample year, 46 countries were found to have participated in TIMSS at least twice across the 2003–2015 sample period at grade four, and 50 countries were found to have participated in TIMSS at least twice across the 2003–2015 sample period at grade

Table 5.2 Within-country variation across time in teacher quality coefficients

Grade level	Variable name	% Countries that show significant difference (proportion in parenthesis)
Grade four	Alignment	13.04 $\left(\frac{6}{46}\right)$
	Mathprep	15.21 $\left(\frac{7}{46}\right)$
	Prepared	15.21 $\left(\frac{7}{46}\right)$
	Mathtime	15.56 $\left(\frac{7}{45}\right)$
	Exp	23.91 $\left(\frac{11}{46}\right)$
Grade 8	Alignment	44.00 $\left(\frac{22}{50}\right)$
	Mathprep	22.00 $\left(\frac{11}{50}\right)$
	Prepared	34.00 $\left(\frac{17}{50}\right)$
	Mathtime	32.00 $\left(\frac{16}{50}\right)$
	Exp	30.00 $\left(\frac{15}{50}\right)$

Notes Significance level is $\alpha = 0.05$. Alignment = proportion of mathematics topics reported as covered by teachers compared with national expectations, Mathprep = index (1–5) of teacher education to teach mathematics, Mathtime = mean number of minutes spent on mathematics teaching per week, Prepared = index (1–4) of self-efficacy to teach mathematics, Exp = experience teaching in years

eight. However, not all of these countries had longitudinal data for every variable, so the exact number of countries differed slightly. For each variable, we determined the percentage and proportion of countries that had significant differences between their estimated coefficients at both grade four and grade eight (Table 5.2).

At grade four, approximately 15% of countries showed a significant difference in estimated coefficients for most of our variables across the sampling period, with the exception of teacher experience. The relationship between teacher experience and student outcomes was found to be inconsistent in around 24% of the countries. Drawing on these results, one plausible hypothesis would be that there is a subset of countries who consistently show significant differences between the variables, hence the similar percentage between variables. However, although some countries did appear more than once, the countries who showed significant changes between the coefficients were randomly distributed. In other words, the temporal instability of relationships did not appear to be a country-specific factor.

At grade eight, the association between teacher instructional alignment and student mean mathematics performance was inconsistent in almost half the sample (22 countries), while the inconsistency for three other variables (preparedness, time on math, and experience) was $\geq$30%. The percentage of countries that exhibited significant differences was higher at grade eight than grade four. There was no clear pattern of countries who showed statistically significant differences in coefficients in all variables at grade eight. However, with such a high number of countries showing

statistically significant variation, countries often demonstrated significant variation in effect for two or more variables.

An important aspect is the existence of countries who not only showed a significant change in the coefficient but also a change in the coefficient sign. A change in the sign of the trend, rather than just the amplitude, implies a higher level of coefficient instability.

Both grade levels showed the existence of statistically significantly differences in the coefficients of all measured variables, and, notably, grade eight data showed much higher percentages of statistically significant differences. High percentages of statistically significant differences can imply an issue with the measurement tools or indicate a highly variant sample across time. While more research and analysis is needed to support these initial findings, the high percentage of statistically unstable coefficients restricts our ability to establish a clear pattern. The problem is further compounded when we consider the existence of a few differences that changed sign. Scale evaluation and adjustment is needed to ensure that observed variability does not stem from the items themselves, rather than from the research population, or as a result of other factors.

This secondary analysis of the multilevel models extracted coefficients and compared them individually. A more holistic approach to the topic of coefficient stability would be to include all of those time points into a single model, thus directly addressing the issue of stability over time; this strategy should be considered in future research.

5.4 Fixed Effect Analysis

To further explore the relationship between teacher quality and student outcomes, we undertook a country-level fixed effects analysis. One of the limitations of the empirical approaches that we used is that they are essentially correlative, making it difficult to attribute causation. A more serious concern is that the models include a limited number of predictors, and hence are subject to unobserved variable bias. The models focus exclusively on measures of teacher characteristics and behavior, and a few student-level indicators that are available in TIMSS. The advantage of fixed-effects models is that analyzing changes within a given unit can provide greater confidence in the association between dependent and independent variables provided the unobserved variables are invariant. Although there is reason to doubt that this assumption holds fully, a fixed-effect analysis should yield somewhat more robust estimates than a cross-sectional regression equation. Our fixed-effect country-level model is restricted to only those countries that participated in the 2007, 2011, and 2015 cycles of TIMSS for a particular grade level, and for educational systems that had country-level estimates available for all variables in the model. It should be noted that education systems are the unit of analysis here; this is not a student- or teacher-level fixed-effect analysis. The “effects” are therefore on the aggregate country level, and may not necessarily apply to particular teachers or students.

Table 5.3 Country-level grade four fixed-effects analysis of the relationship of teacher quality to student outcomes, 2007–2015

Parameter	Estimate	SE	*p*-value
Alignment	−36.89	24.04	0.13
Mathtime	0.14	0.05	0.01
Mathprep	18.55	9.61	0.06
Books	55.50	20.49	0.01
Lang	30.56	7.81	0.00
Prepared	−7.73	10.14	0.45
Exp	−2.01	1.04	0.06
Tmale	−134.89	38.29	0.00

Notes Alignment = proportion of mathematics topics reported as covered by teachers compared with national expectations, Books = index (1–5) of number of books in the home, Lang = index (1–4) of testing language spoken in the home, Mathprep = index (1–5) of teacher education to teach mathematics, Mathtime = mean number of minutes spent on mathematics teaching per week, Prepared = index (1–4) of self-efficacy to teach mathematics, Exp = experience teaching in years, Tmale = index of teacher gender (female = 0, male = 1), Performance = mean student TIMSS score in mathematics, SE = standard error. A *p*-value of 0.05 or lower indicates statistical significance

At grade four, the fixed-effects regression model did uncover a statistically significant association between changes in country-level means of teacher factors and changes in aggregate student mathematics achievement (Table 5.3). Specifically, countries that saw an increase in average time spent on mathematics in grade four saw higher mean student outcomes. Systems whose teachers had increasing levels of education was positive and approached statistical significance ($p = 0.06$), as did a negative relationship between changes in teacher experience and mean mathematics scores. Curricular alignment and teacher self-efficacy had no relationship to student outcomes. Interestingly, teacher gender had a strongly negative association with mathematics outcomes, suggesting that an increasing proportion of male teachers at grade four was associated with weaker mathematics scores. This is a rather curious result and merits more detailed investigation.

At grade eight, time spent on mathematics was again significantly and positively associated with TIMSS mathematics scores, and countries whose teachers reported growing levels of preparedness to teach mathematics also saw higher student outcomes (Table 5.4). None of the other variables approached statistical significance. While these results should not be overstated, they do suggest that, in general, time spent on mathematics may have a positive relationship with student learning and that the failure to uncover a consistent association in other models could be due in part to the exclusion of relevant but unobserved variables.

Two caveats should be kept in mind when interpreting the results of this analysis. As mentioned previously, the virtue of fixed-effects models is that they allow unobserved variables in a given unit (in this case a given educational system) to act as its own control, by identifying change over time. The governing assumption of

Table 5.4 Country-level grade eight fixed effects analysis of the relationship of teacher quality to student outcomes, 2007–2015

Parameter	Estimate	SE	*p*-value
Alignment	5.82	25.02	0.82
Mathtime	0.37	0.12	0.00
Mathprep	16.88	6.35	0.01
Books	23.55	17.21	0.18
Lang	21.96	25.43	0.39
Prepared	−1.50	9.73	0.88
Exp	1.60	1.58	0.32
Tmale	−15.79	54.29	0.77

Notes Alignment = proportion of mathematics topics reported as covered by teachers compared with national expectations, Books = index (1–5) of number of books in the home, Lang = index (1–4) of testing language spoken in the home, Mathprep = index (1–5) of teacher education to teach mathematics, Mathtime = mean number of minutes spent on mathematics teaching per week, Prepared = index (1–4) of self-efficacy to teach mathematics, Exp = experience teaching in years, Tmale = index of teacher gender (female = 0, male = 1), Performance = mean student TIMSS score in mathematics, SE = standard error. A *p*-value = 0.05 or lower indicates statistical significance

such models is that any variables that are excluded are fixed, while all the factors with temporal variation and likely to have a relationship with the outcome of interest are included. This is, of course, an optimistic assumption; there are a number of social, educational, economic, and policy factors that influence student achievement that are likely to have altered during the course of TIMSS (and perhaps even in response to TIMSS testing). Secondly, this analysis aggregates at the education system level, which reduces the number of degrees of freedom (i.e., limits sample size), and is something of a departure for fixed-effects studies, which more typically consider individuals or smaller aggregates (like school districts) rather than entire educational systems. The results of the analysis should therefore be treated with considerable caution.

5.5 An Examination of Standard Errors

One of the distinctive features of large-scale studies like TIMSS is the estimation of standard errors. As described in detail in the TIMSS user guide (Foy 2015), TIMSS uses a complex sampling design; in stage one a random sample of schools is selected, and, in stage two, a randomly-selected intact classroom is selected. If the study were based on a straightforward random sample of all students of a given age or grade level, then the estimation of standard errors could be calculated using conventional means. Instead, TIMSS uses a jackknifing procedure, in which schools are paired and then any calculations (e.g., means and regression coefficients) are

run separately with one of each pair weighted at zero and the weight of the other member of the pair doubled. Standard errors are then estimated by aggregating all of those separate estimates. The IEA has supported calculating means or simple linear regression models through the provision of macros (generated as part of the IEA Database Analyzer, a free specialist analysis package that can be downloaded from www.iea.nl/data). However, when running more complicated models (most especially multilevel models where students are clustered within classrooms), the jackknifing procedure can be computationally quite demanding, taking many hours of computing time to generate models for multiple countries over multiple TIMSS.

The PISA study uses a related, but different approach in the estimation of standard errors. Unlike TIMSS, PISA selects a random population of 15-year-olds and then uses a balanced replicate weight system to calculate standard errors. However, a shortcut that is often used with success when analyzing PISA data is to use adjusted weights, such that the sum of weights equals the number of respondents in the study. An adjustment of this kind is necessary (if for no other reason) to prevent major downward bias in standard errors, since the weights in both TIMSS and PISA are meant to reflect the entire population of students in a country rather than the number of respondents (larger numbers resulting in smaller standard errors). This method has been used to produce results for unpublished studies that are quite similar to the full balanced replicate weight method, and can be convenient when conducting preliminary analysis (given the computational burdens of more formal procedures). However, it should emphasized that this strategy is not technically sound, and any analysis intended for publication or presentation should use the full-scale method. As discussed by Jerrim et al. (2017), there are published studies in reputable journals that have failed to use appropriate statistical procedures, and, although the substantive results have been quite similar, they are not considered reliable.

We conducted a secondary analysis to determine whether the adjusted weight procedure yielded similar standard errors to the jackknifing procedure. All of the multilevel statistical models run using the jackknifing procedure were re-run with student weights reweighted to equal the total number of respondents. We then compared the standard errors of the coefficients between the adjusted weight and jackknifing procedures. The purpose of this analysis was to determine whether the full jackknifing procedure was necessary to avoid downwardly-biased standard errors.

Although there was certainly variation in the magnitude of the differences, on average, we found that standard errors calculated with jackknifing procedure were about twice as large as those calculated using an adjusted base weight (Table 5.5). The difference for grade eight ($\times 2.25$) was larger than for grade four ($\times 1.90$), and was reasonably consistent across years. In nearly every case, the standard errors to account for the complex sampling design were larger with the jackknifing procedure than those using an adjusted base weight; in some cases, four or five times as large. The proportional difference was also greater for teacher-related standard errors ($\times 2.37$) than for student controls ($\times 1.45$). This confirms that it is critical to account for the complex sampling design in TIMSS to avoid the risk of type I errors (false positives).

Table 5.5 Ratio of standard errors using jacknife versus adjusted weight methods

Grade	Year	Intercept	Stmale	Lang	Books	Exp	Tmale	Mathprep	Mathtime	Alignment	Prepared	Mean
4	2003	2.26	0.23	1.53	1.36	2.29	2.36	2.23	2.36	2.45	0.56	1.86
4	2007	1.74	0.20	1.60	1.56	2.35	1.95	2.24	1.72	2.18	2.09	1.86
4	2011	2.00	0.30	1.48	1.66	2.21	2.25	2.39	1.59	2.30	2.27	1.95
4	2015	1.93	0.20	1.44	1.56	2.14	2.19	2.24	2.24	1.98	2.40	1.93
8	2003	2.50	0.25	1.64	1.49	2.86	2.79	2.36	2.42	2.28	2.92	2.25
8	2007	2.27	0.42	1.47	1.47	2.43	2.39	2.06	2.66	2.49	2.22	2.09
8	2011	2.28	0.22	1.70	1.69	2.86	2.88	2.49	2.97	2.58	2.76	2.34
8	2015	2.42	0.48	1.39	1.64	2.71	2.89	2.40	2.91	2.61	2.88	2.33
Mean value		2.18	0.29	1.53	1.55	2.48	2.46	2.30	2.36	2.36	2.26	2.08

Multilevel regression models of teacher quality's relationship to student outcomes

Notes Grade = grade level, Year = year of the TIMSS cycle, Stmale = index of student gender (female = 0, male = 1), Lang = index (1–4) of testing language spoken in the home, Books = index (1–5) of number of books in the home, Exp = experience teaching in years, Tmale = index of teacher gender (female = 0, male = 1), Mathprep = index (1–5) of teacher education to teach mathematics, Mathtime = mean number of minutes spent on mathematics teaching per week, Alignment = proportion of mathematics topics reported as covered by teachers compared with national expectations, Prepared = index (1–4) of self-efficacy to teach mathematics

5.6 Discussion

This chapter has covered a great deal of detailed technical analysis, but the overarching implications are clear. First and most important, using basic statistical analysis based on TIMSS data, the relationship between teacher quality measures and student outcomes appear generally weak and inconsistent. After testing the robustness of the relationship between student outcomes and teacher characteristics in different education systems using alternative statistical models, methods of aggregation, and calculations of standard errors, we found that associations between teacher factors and student outcomes remained modest. When considered over a number of years, the most technically appropriate measure (multilevel estimates with jackknifed standard errors) suggested that there was a negligible relationship between teacher characteristics or teacher behaviors and student outcomes. Although there were education systems and years where teacher effectiveness measures were associated with higher student mathematics performance, this relationship was not robust across time and space.

The weak associations between teacher experience and education mirror much of the research conducted in single-country studies. What is more surprising is that time on mathematics and content coverage has usually uncovered much stronger associations. Most studies that have found a relationship between instructional content and student outcomes have used measures of the volume and intensity of topics covered, as opposed to the alignment with national standards. This finding raises serious questions about the utility of standards-based reform, since there appears to be no strong relationship between teachers' fidelity to mathematics standards and student outcomes. Whether this indicates problems with the measurement of teacher instructional content or the quality of the standards themselves deserves greater attention.

However, our results have a number of limitations, and should not be treated as definitive. The models included were restricted to a fairly limited number of variables and student controls (most seriously, prior student performance, which is not available in the TIMSS dataset), raising the potential for unobserved variable bias. There are also potential problems with the measurements employed, including possible differences in interpretability across varying cultural contexts (within and across educational systems), the indirect measure of teacher professional knowledge through self-reports, and changes in the mathematics topic framework used to measure instructional content coverage.

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57*(1), 289–300.

Clogg, C. C., Petkova, E., & Haritou, A. (1995). Statistical methods for comparing regression coefficients between models. *American Journal of Sociology, 100*(5), 1261–1293.

Curran-Everett, D. (2000). Multiple comparisons: Philosophies and illustrations. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology, 279*(1), R1–R8.

Foy, P. (2015). *TIMSS 2015 user guide for the international database.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from http://timssandpirls.bc.edu/timss2015/international-database/downloads/T15_UserGuide.pdf.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika, 75*(4), 800–802.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika, 75*(2), 383–386.

Jerrim, J., Lopez-Agudo, L., Marcenaro-Gutierrez, O., & Shure, N. (2017). What happens when econometrics and psychometrics collide? An example using the PISA data. *Economics of Education Review, 61,* 51–58.

Weisstein, E. W. (2004). Bonferroni correction. Retrieved from http://mathworld.wolfram.com/BonferroniCorrection.html.