Teaching and Teacher Education 132 (2023) 104242

Contents lists available at ScienceDirect

Teaching and Teacher Education

journal homepage: www.elsevier.com/locate/tate

Research paper

How effective is the professional development in which teachers typically participate? Quasi-experimental analyses of effects on student achievement based on TIMSS 2003–2019

Nils Kirsten [a], [b], [*], Jannika Lindvall [a], Andreas Ryve [a], Jan-Eric Gustafsson [c]

[a] School of Education, Culture and Communication, Mälardalen University, Västerås, Sweden
[b] Department of Education, Uppsala University, Uppsala, Sweden
[c] Department of Education and Special Education, University of Gothenburg, Gothenburg, Sweden

ARTICLE INFO

Article history:
Received 9 March 2023
Received in revised form
20 June 2023
Accepted 21 June 2023
Available online xxx

Keywords:
Teachers' professional development
Student achievement
Quasi-experimental analysis
Within-student between-subjects
Structural equation modeling
TIMSS

ABSTRACT

This study examines the effect of teachers' participation in mathematics and science professional development (PD) on student achievement in nationally representative settings. We use data from all OECD countries in the 2003 through 2019 cycles of the Trends in International Mathematics and Science Study (TIMSS) and apply student fixed effects to control for unobserved student characteristics and school quality. We find a small negative average effect of PD participation, with negative effects concentrated among high-performing students. We discuss potential explanations of these results and suggest ways PD studies may inform the PD in which teachers typically participate.

1. Introduction

Annually, about 90% of teachers internationally participate in professional development (PD), with an average participation of 6–9 days (Kirsten, 2020). Studies indicate that such a level of PD participation corresponds to about three percent of the educational budget (Killeen et al., 2002; Miles et al., 2004; Van den Brande & Zuccollo, 2021). These large investments in terms of public spending and teachers' time may be warranted. For example, meta-reviews of quasi-experimental studies and randomized controlled trials (RCTs) of PD programs have shown average effects of 0.42–0.56 standard deviations on teaching (Garrett et al., 2019; Gonzalez et al., 2022; Kraft et al., 2018) and 0.05–0.23 standard deviations on student achievement (Basma & Savage, 2018, 2023; Kraft et al., 2018; Lynch et al., 2019; Sims et al., 2021). However, although quasi-experimental studies and RCTs may provide a robust basis for causal conclusions, the conditions in which the studied PD programs have been conducted differ substantially from those teachers typically face. Most studied programs are conducted in small scale and with high intensity. For example, the median number of teachers participating in the PD programs Garrett et al. (2019) reviewed was 88 and the interventions' typical duration was 20–100 h. Moreover, small-scale programs are often conducted by skilled experts with extensive experience and may invite rather than mandate teachers to participate (Kennedy, 2016). Therefore, the conclusions concerning PD effects in particular PD programs may not be valid for the PD in which teachers typically participate.

There is, in fact, reason to believe that PD quality is difficult to maintain outside of small-scale programs. For example, recent meta-syntheses indicate that when PD programs are scaled up, their effects decrease (Garrett et al., 2019; Kraft et al., 2018). Furthermore, as Hill et al. (2013) and Sims et al. (2021) argue, it is not sufficient that scholars identify which particular PD programs are effective because teachers and schools in general rarely have access to such programs. To increase the likelihood that teachers in general gain access to higher-quality PD, more focus should be devoted to informing PD in typical settings. Studies can contribute by identifying the potential mechanisms through which PD affects

[*] Corresponding author. School of Education, Culture and Communication, Mälardalen University, Västerås, Sweden.
E-mail address: nils.kirsten@mdu.se (N. Kirsten).

https://doi.org/10.1016/j.tate.2023.104242
0742-051X/© 2023 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

outcomes (Sims et al., 2021), identifying how PD can be scaled up (Patfield et al., 2022) and adapted to various contexts (Koellner & Jacobs, 2015), and providing credible estimates of the effects of the PD teachers typically face.

We particularly aim to contribute to the knowledge concerning the last of these areas by exploring the effects of mathematics and science PD in nationally representative settings. We do so by using nationally representative data from all OECD countries participating in the 2003, 2007, 2011, 2015, and 2019 cycles of the *Trends in International Mathematics and Science Study* (TIMSS). This dataset includes information on teachers' participation in mathematics and science PD and their students' mathematics and science achievement, as well as many other variables concerning schools, teachers, and students. Because students are tested in two subject areas (mathematics and science), the dataset enables the use of a quasi-experimental statistical technique – within-student between-subjects analysis – to isolate PD effects from effects due to students' self-selection into schools and schools' general influence on students. The research questions we answer are.

(1) What effect does teachers' participation in mathematics and science PD have on student achievement in nationally representative settings?
(2) How does the effect vary among students at different achievement levels?

2. Theoretical framework

It is well known that teacher quality has a considerable effect on students' learning (e.g., Chetty et al., 2014; Hanushek, 2011; Jackson, 2018). Accordingly, one of the most common motives for teacher PD is that it will improve the quality of teaching and, subsequently, student learning. However, in practice, PD participation may have both negative and positive effects on student achievement. PD may cause negative effects because teachers' time and school resources are taken from other purposes, such as teaching (Harris & Sass, 2011). Teachers may neglect or use new teaching methods superficially (Sykes & Wilson, 2016; Timperley, 2015), and in unfortunate cases, new teaching methods may be detrimental in general or in specific contexts. Mechanisms that may enhance student achievement are, on the other hand, that new teaching methods may improve teaching (i.e., the chief rationale for teachers' PD), positive attention may boost teacher motivation (Hawthorne effects), and PD may improve teachers' self-efficacy and job satisfaction, which in turn may reduce teacher turnover and improve the conditions for long-term development of teaching quality (Allen & Sims, 2017; Coldwell, 2017).

Scholars have proposed several frameworks of PD features that may predict improved student achievement (e.g., Cordingley et al., 2015; Darling-Hammond et al., 2017; Desimone, 2009; Hill & Papay, 2022). For example, Hill and Papay (2022) suggest that there is emerging evidence on some features: teacher collaboration, individual coaching, follow-up meetings concerning implementation difficulties, addressing subject-specific instructional practices and building relationships with students, and providing teachers with concrete instructional materials. However, the mentioned frameworks on effective PD features are not unanimous and there is currently no clear consensus on the PD features that predict improved student achievement (Sims & Fletcher-Wood, 2021). Thus, we suggest that scholars view the PD features outlined by scholars such as Hill and Papay (2022) as well-grounded hypotheses to be tested in further research.

We assume that the positive and negative PD mechanisms listed above exist and that effects are influenced by PD features. The aim of this study is, however, not to tease apart the importance of each mechanism and PD feature but to provide a starting point for such analyses by shedding light on the degree to which the PD in which teachers typically participate affects student results.

Educational interventions are unlikely to affect all students in the same way. For example, studies have shown that effective teachers and schools improve the results most prominently for low-performing students (Burgess et al., 2022; Jackson et al., 2020). However, as Atlay et al. (2019) describe, research on the causes of such heterogenous effects remains scarce although some mechanisms have been suggested. For example, students with difficulties in areas such as attentive behavior, working memory, and phonological processing may benefit more from teaching methods without high demands on those skills (Fuchs et al., 2016). In line with this hypothesis, Atlay et al. (2019) show that students from low socioeconomic backgrounds profit less from teaching that includes challenging tasks and open-ended problem solving, which the authors suggest may occur because challenging tasks may be understandable to and spur already motivated and high-performing students, whereas such tasks cause frustration for low-performing students. Consistent with this finding, a study of a large-scale Swedish PD program showed that PD participation increased the share of instruction time for open-ended problem solving and that positive PD effects on student achievement could only be found for intermediate- and high-performing students (Grönqvist et al., 2021). If PD programs advocate teaching strategies that influence students differently, the effects on student achievement should also vary among student groups. We investigate differential effects because they may be just as important as mean effects.

3. Previous studies of the effects of nationally representative PD

Most existing studies of PD effects in nationally representative settings have presented correlational evidence (e.g., Blömeke et al., 2016; Havard et al., 2018), for example by regressing students' mathematics or science score on teachers' PD participation. The results of such analyses are likely biased. For example, the direction of causality may be reversed if student performance influences teachers' PD participation. Also, correlations may be an artifact of other factors, such as features of school leadership that affect both PD participation and student achievement. Although some confounding factors can be included as control variables (e.g., students' socioeconomic status and teacher characteristics), numerous confounders are unobserved. To circumvent these risks for bias, a small number of PD studies based on nationally representative data have used fixed effects specifications, which control for unobserved factors by analyzing differences within units (countries or students) rather than between units. Two of these studies used a country-level difference-in-differences approach based on the TIMSS 2007/2011 (Gustafsson & Nilsen, 2016) and PIRLS 2006/2016 (Van Damme et al., 2019). Because the analysis is based on within-country differences over time, unobserved differences between countries that remain constant over time do not affect the estimates. Moreover, because reverse causality, such as resource allocation to struggling learners, operate at local rather than country level, this source of bias is also eliminated from the analysis. However, the estimates may still be biased due to factors that vary over time within countries. For example, a curriculum reform may influence both PD participation and student achievement. Furthermore, because the analysis is conducted at the country level, the number of observations is low (=the number of included countries), which leads to large statistical uncertainty.

An alternative approach is to use student-level fixed effects. This has been done in three studies using state-level data from the US to investigate PD's effects on students' change score over time (value-added academic achievement) (Akiba & Liang, 2016; Harris & Sass, 2011; Wallace, 2009). Thereby, pupil, class, and school characteristics are held constant, isolating PD effects from the influence of each student's previous ability and time-invariant school and class factors. These studies showed small and mostly statistically insignificant PD effects on student achievement. Although these studies are methodologically strong, particularly the study by Harris and Sass (2011), which uses data from several years and includes both student and teacher fixed effects, they only address PD effects in four US states.

To provide an estimation of PD effects on student achievement internationally, the present study uses a within-student between-subjects analysis. This is a student fixed effects approach in which PD effects are estimated based on the difference within students between the two subject areas investigated in the TIMSS survey: mathematics and science. Because the analysis is conducted within students, school- and student-level differences that span both subjects, such as school climate, resources, and general student ability, are accounted for and do not bias estimates. However, some risks for bias remain even in this specification. Particularly, it does not account for associations between a student's subject-specific skills and teachers' PD participation, which are caused by either sorting students into classes based on subject-specific ability or by teachers' selection into PD as a result of the class performance in a subject. Therefore, we investigated these risks for bias in several sensitivity analyses, based on the public TIMSS data, a merged dataset for TIMSS and PIRLS 2011, and a Swedish TIMSS 2015 dataset that included administrative data concerning students' grades and standardized test results. In short, the robustness checks supported the main model's results.

The within-student between-subjects technique has been used in several previous studies based on PISA and TIMSS data to investigate instructional time's effect on student achievement (Bietenbeck & Collins, 2023; Lavy, 2015; Rivkin & Schiman, 2015), teaching methods' effect on student achievement (Bietenbeck, 2014; Schwerdt & Wuppermann, 2011), and the effect of teachers' self-efficacy on student achievement (Jerrim et al., 2023). However, no study that we are aware of has used a within-student between-subjects approach to study the effects of teachers' PD on student achievement.

4. Method

4.1. Data

We collected the data used in this study from the 2003, 2007, 2011, 2015, and 2019 cycles of the TIMSS, including grade four and grade eight in mathematics and science. The datasets are available for download from the *TIMSS & PIRLS International Study Center* website.

Various numbers of countries have participated in different TIMSS cycles, ranging from 29 to 66 school systems (grade four, 2003 and 2019, respectively), including benchmarking participants, such as the Quebec province in Canada. The present study is, however, restricted to OECD countries to limit the contextual complexity. Furthermore, we excluded three OECD countries from the sample for particular TIMSS cycles because they failed to meet TIMSS's sample guidelines. We present the final list of included countries in Supplementary material, Section A, available at the journal website.

As outlined in technical reports (e.g., Martin et al., 2020), TIMSS uses a two-stage clustered sampling design. In the first stage, schools are randomly sampled (at least 20 schools in each school system and grade). In the second stage, one or two classes are randomly sampled in each school. Because sampling probability varies between schools, analyses that aim to achieve nationally representative estimates need to use sampling weights provided by the *TIMSS & PIRLS International Study Center*. We used weights which give each country equal weight in the analysis (senwgt) to ensure that differences in sample sizes among countries do not affect results. Furthermore, analyses must adjust for the fact that clustering within schools reduces the variance in the sample. One method to do this is jackknife repeated replication (JRR), which is used in official TIMSS publications. We used this technique to compute means in the descriptive statistics presented in Tables 1 and 2. Another method to adjust for clustering is to use cluster-robust standard errors, which has been shown to produce qualitatively similar results as JRR (Jerrim et al., 2017). We used this method in the more complex analyses of PD effects because it can be implemented by built in commands in the software used (Stata 17 and Mplus 8.8).

The main outcome variables used in the analyses are students' mathematics and science scores. These scores are reported as five plausible values for each student. TIMSS reports five values rather than one because the test booklet for each student only contains a fraction of the total number of test items. Based on item response theory, the test score for each student is estimated using five random draws to reflect the uncertainty of the student's true test score. All analyses concerning test scores must take this uncertainty into account, which in the main model is achieved by the TYPE = IMPUTATION function in the Mplus software. In the estimation of differential effects for students at different achievement levels, each model was estimated five times (once for each plausible value) to enable subsequent computation of sampling and imputation error (cf. Jerrim et al., 2017). Before analysis, we standardized the score variables by subtracting 500 and dividing by 100. Thereby, the estimates can be interpreted as standard deviations of the distribution in the original TIMSS 1995 sample.

An important feature of TIMSS data is that teachers are linked to students because entire classes and their teachers are sampled (this differentiates the TIMSS from the PISA, in which students and teachers within schools are sampled randomly rather than linked to one another). The main independent variable used in the analyses – teachers' PD participation – emanates from the teacher questionnaire. In all TIMSS cycles from 2003, teachers were asked whether they had participated in PD on a given set of topics in the past two years. We considered the five items that were included in the questionnaires for all cycles and both subject areas (mathematics and science): 1) mathematics/science content, 2) mathematics/science pedagogy/instruction, 3) mathematics/science curriculum, 4) integrating information technology into mathematics/science, and 5) mathematics/science assessment.[1] We recoded these variables as dummy variables, with participation = 1. To summarize the information in all PD variables in descriptive statistics, we additionally compiled them into a variable indicating how many of the items were answered “yes.” We performed such compilation for PD items 1–3 (denoted as PD3 in Tables 1 and 2) and all five PD items (denoted as PD5 in Tables 1 and 2).

In the 2015 and 2019 cycles, teachers were additionally asked

[1] Three additional items exist in the TIMSS 2003–2019 data but not in all cycles and grades. Two of these items are, however, not subject specific because teachers were asked whether they had participated in PD in “addressing individual students' needs” and “improving students' critical thinking or problem solving skills” without specifying mathematics or science. This limits the items' usefulness in a between-subject analysis because it is unclear in which subject we should expect effects. The third item only exists for science (“integrating science with other subjects”) and cannot be used in a between-subject analysis.

Table 1
Means, standard errors and missing data of teachers' PD participation, teacher characteristics, and instruction time for grade four students.

	2003	2007	2011	2015	2019	2003–2019
Mathematics						
PD3 participation	1.10	1.00	0.93	0.82	0.91	0.92
SE	0.07	0.06	0.07	0.06	0.06	0.04
Percent missing	5.96	3.73	9.61	5.69	6.61	6.53
PD5 participation	1.67	1.46	1.39	1.25	1.36	1.37
SE	0.10	0.09	0.09	0.08	0.09	0.05
Percent missing	5.96	3.69	9.58	5.69	6.56	6.50
PD days				0.92	1.04	0.97
SE				0.07	0.08	0.06
Percent missing				5.93	6.86	51.74
Teacher education	0.88	0.87	0.90	0.88	0.88	0.88
SE	0.02	0.01	0.02	0.02	0.02	0.01
Percent missing	22.73	8.67	8.58	11.66	11.32	11.59
Teacher age	41.79	43.44	43.02	43.67	43.74	43.08
SE	0.75	0.72	0.72	0.70	0.75	0.42
Percent missing	6.06	2.49	6.61	5.08	5.37	5.21
Teacher experience	17.36	18.58	18.08	17.94	17.71	17.88
SE	0.79	0.78	0.75	0.73	0.78	0.44
Percent missing	7.10	4.09	8.02	6.29	6.03	6.37
Female teacher	0.85	0.85	0.82	0.82	0.82	0.82
SE	0.02	0.02	0.02	0.02	0.03	0.02
Percent missing	5.68	2.36	6.62	4.92	5.36	5.11
Instruction time	3.12	2.99	3.55	3.32	3.33	3.32
SE	0.05	0.05	0.08	0.07	0.07	0.04
Percent missing	15.21	5.88	9.89	8.54	7.72	8.94
Science						
PD3 participation	1.02	0.97	0.90	0.78	0.85	0.88
SE	0.07	0.06	0.07	0.06	0.07	0.04
Percent missing	9.35	5.84	11.08	6.60	8.31	8.22
PD5 participation	1.54	1.43	1.35	1.19	1.29	1.31
SE	0.10	0.09	0.09	0.08	0.09	0.05
Percent missing	9.25	5.81	11.07	6.60	8.31	8.20
PD days				0.88	1.00	0.94
SE				0.07	0.08	0.06
Percent missing				6.97	8.55	52.78
Teacher education	0.88	0.86	0.90	0.88	0.87	0.87
SE	0.02	0.02	0.02	0.02	0.02	0.01
Percent missing	23.86	9.59	8.78	11.77	11.39	11.96
Teacher age	41.75	43.23	42.91	43.61	43.88	43.00
SE	0.76	0.72	0.73	0.72	0.77	0.42
Percent missing	7.13	3.37	6.91	5.28	6.40	5.83
Teacher experience	17.30	18.28	17.93	17.80	17.63	17.69
SE	0.80	0.77	0.76	0.75	0.78	0.45
Percent missing	8.08	5.13	8.25	6.31	7.08	6.95
Female teacher	0.85	0.85	0.82	0.82	0.82	0.82
SE	0.02	0.02	0.02	0.02	0.03	0.01
Percent missing	6.80	3.25	6.89	5.12	6.46	5.74
Instruction time	2.69	2.78	3.35	3.02	2.91	3.05
SE	0.06	0.05	0.08	0.06	0.06	0.04
Percent missing	18.23	12.66	12.45	9.39	10.20	11.74
# of students	49 108	78 549	117 866	136 393	126 381	508 297
# of mathematics classes	2961	5146	6427	7687	7701	29 922
# of science classes	3047	5032	6258	7428	7567	29 332
# of schools	1886	2967	4253	4914	5217	19 237
# of countries	11	17	23	25	25	28

how many hours they had participated in PD in mathematics/science during the past two years. We recoded these variables into days per year.[2] Tables 1 and 2 present the means, standard errors, and missing data of teachers' PD participation in the sample.[3]

In addition to the PD variables, we used several control variables in the analyses, as reported in Tables 1 and 2. Variables concerning teacher characteristics were teacher education, teacher experience, age, and gender. In addition, we used instruction time in mathematics and science as controls because instruction time may moderate PD effects on student achievement.

In analyses without student fixed effects, we used student reported books at home as a proxy for socioeconomic status, which is the socioeconomic status variable that has been shown to have the highest correlation with student achievement as measured in international large-scale assessments (Eriksson et al., 2021).

A comparison of Tables 1 and 2 reveals that grade eight teachers participate significantly more in PD than grade four teachers, with a difference of about one topic during the past two years, based on

[2] Teachers were asked to check one of five response options (none, less than 6 h, 6–15 h, 16–35 h, more than 35 h). To enable computations of mean values, we substituted each of these intervals with its midpoint value (for example, we substituted the interval 6–15 h with the value 10.5 h).

[3] We calculated means and standard errors using teacher-level probability weights and the JRR method, following *TIMSS & PIRLS International Study Center* recommendations. International means are arithmetic averages of country-level means. Instruction time is summed to represent total instruction time per student and subject rather than instruction time per teacher.

Table 2
Means, standard errors and missing data of teachers' PD participation, teacher characteristics, and instruction time for grade eight students.

	2003	2007	2011	2015	2019	2003–2019
Mathematics						
PD3 participation	1.66	1.64	1.49	1.57	1.43	1.60
SE	0.08	0.08	0.08	0.07	0.08	0.06
Percent missing	6.35	4.75	10.56	7.44	7.43	7.41
PD5 participation	2.55	2.47	2.24	2.43	2.25	2.45
SE	0.12	0.11	0.11	0.11	0.11	0.08
Percent missing	6.34	4.72	10.56	7.44	7.43	7.40
PD days				1.92	1.91	1.92
SE				0.11	0.11	0.09
Percent missing				7.57	7.62	57.58
Teacher education	0.70	0.71	0.65	0.69	0.71	0.70
SE	0.03	0.03	0.03	0.03	0.03	0.02
Percent missing	13.89	12.44	15.65	6.94	7.04	10.83
Teacher age	43.62	42.28	43.39	43.45	43.68	43.25
SE	0.78	0.74	0.73	0.71	0.73	0.54
Percent missing	5.49	3.73	9.59	6.33	6.32	6.38
Teacher experience	18.41	17.21	17.23	17.20	16.77	17.56
SE	0.83	0.77	0.75	0.74	0.72	0.56
Percent missing	7.34	6.26	9.93	6.71	6.35	7.30
Female teacher	0.65	0.64	0.62	0.64	0.64	0.65
SE	0.03	0.03	0.03	0.03	0.03	0.02
Percent missing	5.39	3.73	9.41	6.31	6.32	6.32
Instruction time	3.33	3.29	3.63	3.60	3.59	3.49
SE	0.06	0.06	0.07	0.06	0.08	0.05
Percent missing	9.52	6.06	11.39	9.66	9.08	9.25
Science						
PD3 participation	1.59	1.60	1.46	1.36	1.38	1.50
SE	0.07	0.07	0.07	0.07	0.07	0.05
Percent missing	9.01	5.08	9.28	9.73	8.64	8.46
PD5 participation	2.44	2.41	2.21	2.11	2.12	2.31
SE	0.10	0.10	0.10	0.11	0.10	0.07
Percent missing	8.77	5.06	9.28	9.71	8.64	8.41
PD days				1.73	1.86	1.80
SE				0.10	0.10	0.08
Percent missing				8.82	8.80	61.51
Teacher education	0.54	0.60	0.61	0.64	0.65	0.57
SE	0.03	0.03	0.03	0.03	0.03	0.02
Percent missing	13.31	9.78	14.09	8.85	8.47	10.88
Teacher age	42.95	42.68	43.66	43.26	43.52	43.20
SE	0.66	0.70	0.69	0.69	0.70	0.46
Percent missing	6.56	4.18	8.05	7.49	7.28	6.78
Teacher experience	17.48	16.95	16.80	16.39	16.36	17.03
SE	0.70	0.73	0.71	0.70	0.69	0.47
Percent missing	8.72	6.99	8.47	7.89	7.35	7.91
Female teacher	0.63	0.61	0.63	0.63	0.65	0.64
SE	0.03	0.03	0.03	0.03	0.03	0.02
Percent missing	6.56	4.19	8.05	7.43	7.25	6.77
Instruction time	3.78	3.45	4.00	3.59	3.51	3.79
SE	0.09	0.07	0.10	0.09	0.09	0.06
Percent missing	12.11	8.07	13.13	12.05	10.64	11.30
# of students	68 913	63 903	77 265	95 421	80 866	386 368
# of mathematics classes	3 621	4 241	4 958	6 027	5 241	24 088
# of science classes	6 853	6 981	7 792	8 193	8 043	37 862
# of schools	2 552	2 252	2 759	2 941	2 925	13 429
# of countries	16	14	14	16	17	25

the index containing five PD variables. This is also reflected in the fact that grade eight teachers participated in almost one more day of mathematics/science PD than grade four teachers over the last two years (this variable is only available for the 2015 and 2019 cycles). In both grade levels, teachers' PD participation is slightly higher in mathematics than in science.

As Table 1 shows, the grade four means of teacher characteristics variables are almost identical across mathematics and science. This result is not surprising because 73% of the grade four students in the merged TIMSS database (2003–2019) are taught by one and the same teacher in both mathematics and science. Although grade eight students are predominately taught mathematics and science by different teachers, teacher characteristics are still highly similar between mathematics and science, with the exception of teacher education, which was held by 70% of mathematics teachers but only 57% of the science teachers in the merged dataset (2003–2019). Exploration of the data suggests that this is because a subject major (biology, physics, chemistry, or earth science) is more common than an education major for science teachers in comparison to mathematics teachers.

Furthermore, a comparison of Tables 1 and 2 shows that the average instruction time is slightly higher in mathematics than in science in grade four but somewhat higher in science in grade eight. This difference between grades may arise because science is more often taught as several separate subjects than as integrated science in grade eight (the average number of science courses taught in grade eight is 1.8 in the merged dataset for 2003–2019, whereas the corresponding figure for grade four is 1.0).

4.2. Analysis

4.2.1. Within-student between-subjects analysis

The within-student between-subjects approach used in this study is a type of fixed effects analysis. However, in contrast to a traditional fixed effects analysis of student data, we compare values between subject areas rather than over time. The result is highly similar: no factors that are invariant within students over time (in traditional fixed effects analysis) or between subject areas (in within-student between-subjects analysis) affect estimates.

A within-student between-subjects analysis of PD effects on student achievement for two school subjects is most easily conducted by, in a first step, computing the difference in scores between the two subjects ($\Delta\text{score} = \text{score}_{\text{mathematics}} - \text{score}_{\text{science}}$) and the difference in teachers' PD participation across the subjects ($\Delta\text{PD} = \text{PD}_{\text{mathematics}} - \text{PD}_{\text{science}}$). In a second step, Δscore is regressed on ΔPD. A more flexible and versatile method is, however, to subtract each student's mean value from the subject-specific values, which is the most common method used in fixed effects software packages, such as xtreg (Stata). Our implementation of the within-student between-subjects approach is described in further detail in Sections 4.2.2–4.2.4.

Estimations in a within-student between-subjects framework require that one value is reported for each student in each subject. However, some students have more than one teacher in each subject area. This can be handled in two ways: 1) students taught by more than one teacher in each subject area are excluded from the analysis, or 2) values for students taught by more than one teacher in each subject area are averaged (or in some cases summed, e.g., instruction time) into one value per subject area. In the main analysis, we use the second strategy to avoid reducing the sample. However, we also implemented the first strategy in a sensitivity test restricted to students taught by the same teachers in both subject areas, producing substantially similar estimates as the second strategy.

4.2.2. Structural equation modeling

In the present study, we implement the fixed effects approach in a structural equation modeling framework. As several authors (Allison, 2009; Allison et al., 2017; Bollen & Brand, 2010; Newsom, 2015) have discussed, this specification provides several advantages over traditional forms of fixed effects analysis. For example, it provides greater flexibility in model specification and more opportunities to test assumptions concerning model specifications. Furthermore, the structural equation modeling framework makes it possible to use latent variables with several indicator variables, which enables estimation and elimination of the measurement error associated with each indicator variable, thus increasing the statistical power.[4]

Fig. 1 illustrates the main model. The effect of PD participation is constrained to be equal (this coefficient is denoted b in Fig. 1) across the two subject areas. A latent variable captures the within-student fixed effects (such as general cognitive ability and socioeconomic status), which is constrained to influence the scores in each subject area with the same amount (1.0). Students' fixed effects are also assumed to be correlated equally (c) with PD participation in each subject area (this is the fixed effects assumption, which differentiates the model from a random effects model).[5] The factor loadings of each indicator (content, pedagogy, and curriculum) are constrained to be equal between the two subject areas (d-e) to ensure that differences in factor loadings between the subject areas do not affect estimates.

We entered control variables into the model as observed variables affecting the achievement score in each subject area. For readability, though, Fig. 1 does not include them. As described in Section 4.1, the achievement scores are treated as imputed values. Missing data is handled by full information maximum likelihood, which uses every piece of available information, similar to multiple imputation. The measurement errors associated with students' mathematics and science scores are denoted as ε (we also estimated the measurement errors associated with each PD indicator but omitted them from Fig. 1 for greater readability).

4.2.3. Assessing goodness-of-fit and measurement invariance

Specification of latent variables, such as the PD participation variable in the main model, involves investigation of the relationship between indicator variables and latent variables, which is measured as factor loadings. In this case, the factor loadings of two out of five PD items available in the TIMSS data were consistently lower than for the other PD items: around 0.4–0.5 for the PD items concerning IT and assessment, compared to 0.6–0.8 for the items concerning content, pedagogy, and curriculum. The former two variables also had a weaker correlation with the other variables and less frequently received the "yes" response. One possible explanation for this pattern is that typical PD treats content, pedagogy, and curriculum simultaneously whereas IT and assessment are treated more separately and sparsely. Although a factor loading of 0.4 is often considered as acceptable (Wang & Wang, 2019), exclusion of the IT and assessment items improved model fit indices. Therefore, analyses proceeded without the two latter PD items. However, to check whether the effects reported for the main model were also valid with all five PD variables, we estimated this specification as well. The results are highly similar to the main model specification albeit with worse goodness-of-fit values (see Supplementary material, Section B).

We investigated measurement invariance between the subject areas by comparing the model with factor loadings constrained to be equal between the subject areas with a model without this constraint (cf. Newsom, 2015). As Supplementary material, Section C shows, the differences in goodness-of-fit statistics (RMSEA, TLI, CFI, and SRMR) and the regression coefficients between models with/without this constraint are close to zero, with an average difference of 0.003. The χ^2 test indicates that the differences in model fit are statistically significant, but this is to be expected in large samples even when differences are trivial (Wang & Wang, 2019). Because the differences between the models are so small and constrained factor loadings simplify interpretation of the results, we chose the latter model.

4.2.4. Unconditional quantile regression

In addition to average PD effects on student achievement, we

[4] Latent variables are theoretical concepts that are assumed to explain the covariance among observed indicator variables. In the present study, we assumed teachers' responses to a number of PD items to explain teachers' PD participation in each subject area. For example, if a PD item has a factor loading of 0.7, which corresponds to an R^2 value of 0.49 (0.7×0.7), the latent variable explains 49% of the variance in the PD item. The sources of variance in the indicators that the latent variable does not explain are considered as measurement errors. Analyses using latent variables eliminates this measurement error, thereby increasing precision and statistical power.

[5] We also tested a random effects specification because if student fixed effects were uncorrelated with teachers' PD participation, omitting this correlation would produce more precise estimates without causing bias (i.e., the effect estimates would be similar to those in the fixed effects model but the standard errors would be smaller) (Bollen & Brand, 2010). However, a comparison of the models showed that the estimates the random effects model produced differed substantially from the fixed effects model and that the χ^2 test was statistically significant ($p < 0.001$ for the merged 2003–2019 datasets), which supports the fixed effects model.

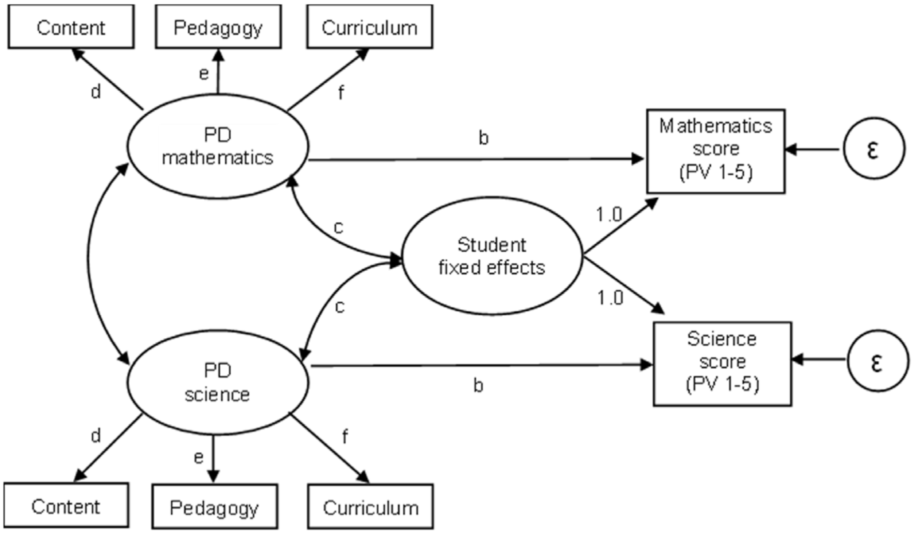

Fig. 1. Within-student between-subjects fixed effects model of PD effects on student achievement.

investigated whether effects varied among students at different achievement levels. One method of doing so is to stratify the sample by achievement level and then estimate the main model on each sample. However, this method severely reduces the sample size and statistical power. A better method is quantile regression because it retains the whole sample. However, in ordinary quantile regression, observations are divided into quantiles based on the dependent variable conditional on the covariates, such as a student's TIMSS score when teacher characteristics and instruction time are held constant. Because this complicates the meaning of a quantile, we used unconditional quantile regression (UQR), which determines the quantiles before regression, so the covariates do not influence them. Therefore, a quantile is defined by the percentage of students above/below a particular TIMSS score. For example, if the 75th quantile corresponds to 579 TIMSS points, 75% of the students scored lower than that. To fit a regression line at a particular quantile, UQR weighs observations above and below the quantile differently.

Because UQR is not available in Mplus, we used the *rifhdreg* Stata command (*recentered influence function regression with high-dimensional fixed effects*, see Rios Avila, 2019). This command enables UQR analysis in combination with fixed effects (within-students between-subjects), probability weights, and clustered standard errors, as in the main model. UQR analyses included all control variables used in the main model.

To mimic the latent PD variable in the main model, we generated the PD variable used in the UQR analysis by confirmatory factor analysis based on the three PD variables in each subject area. Before estimation based on each quantile, we confirmed that the UQR specification produced highly similar results to those from the main model when we estimated the regression for the mean.

5. Results

5.1. The effects of typical PD participation on student achievement

We begin by presenting the results of a model without student fixed effects. In this specification, we pooled mathematics and science data and regressed student achievement on the latent PD participation variable. Tables 3 and 4 report the results with and without control variables. The latent PD variable is scaled by the first indicator variable (PD concerning mathematics/science content), which is also the indicator with the highest factor loading. Therefore, the reported effects should be interpreted as the change in standard deviations of student achievement when PD participation is increased with a unit scaled by the difference between having participated in PD concerning mathematics/science content and not having participated in such PD.

As Tables 3 and 4 show, the estimated effects are all positive and statistically significant ($p < 0.001$). Moreover, the estimated effects are lower when control variables are included, indicating that the correlations between PD and student achievement are not only an effect of PD. We find a similar tendency when the results are separated by subject area (mathematics and science) (Supplementary material, Section D).

However, the effects presented in Tables 3 and 4 (and Supplementary material, Section D) may be biased because omitting student fixed effects means that unobserved student characteristics that influence both teachers' PD participation and student achievement are not taken into account. For example, teachers of high-achieving students may participate more in PD because they have time and resources to do so. The main model uses a within-student between-subjects approach to include student fixed effects, as described in Section 4.2. Tables 5 and 6 present the results with and without control variables.

A comparison of the estimates with and without student fixed effects (Tables 3 and 4 versus Tables 5 and 6) demonstrates considerable differences. The estimates when student fixed effects are included are small and negative in grades four and eight. Furthermore, the estimated effects are unaffected or reinforced by including control variables rather than weakened as in the model without fixed effects. These results indicate that student fixed effects capture most of the variation in teacher characteristics and that differences in instruction time among students explain some of the variation that teachers' PD participation does not explain.

The main model estimates the PD effects on student achievement as −0.017 standard deviations ($p = 0.002$) for grade four and −0.020 standard deviations ($p = 0.010$) for grade eight in the specification including all controls. Therefore, if a teacher has participated in one more unit of PD (e.g., having participated in PD concerning mathematics/science content) during the last two years, the students' achievement is on average 0.017−0.020 standard deviations lower, which is equivalent to a 1.7 to 2.0 TIMSS points decrease (for example, from 500 to ~498 TIMSS points).

Table 3
PD effects on student achievement in a specification without student fixed effects. Data pooled across mathematics and science. Grade four, 2003–2019.

PD participation	No controls	Student controls	+ teacher controls	+ instruction time
1 latent, 3 observed				
Effect	0.090	0.065	0.062	0.065
SE	0.013	0.011	0.011	0.011
p	0.000	0.000	0.000	0.000
RMSEA	0.004	0.004	0.006	0.005
CFI	0.999	0.999	0.993	0.993
TLI	0.997	0.997	0.984	0.984
SRMR	0.003	0.004	0.010	0.009
# of students	508 295	508 295	508 295	508 295
# of schools	19 237	19 237	19 237	19 237
# of countries	28	28	28	28

Table 4
PD effects on student achievement in a specification without student fixed effects. Data pooled across mathematics and science. Grade eight, 2003–2019.

PD participation	No controls	Student controls	+ teacher controls	+ instruction time
1 latent, 3 observed				
Effect	0.132	0.136	0.131	0.120
SE	0.017	0.014	0.014	0.014
p	0.000	0.000	0.000	0.000
RMSEA	0.006	0.006	0.004	0.004
CFI	0.997	0.997	0.995	0.995
TLI	0.992	0.994	0.990	0.990
SRMR	0.006	0.006	0.007	0.007
# of students	385 740	385 740	385 740	385 740
# of schools	13 422	13 422	13 422	13 422
# of countries	25	25	25	25

Table 5
Main model effect estimates of PD on student achievement, grade four, 2003–2019.

PD participation	No controls	Teacher controls	+ instruction time
1 latent, 3 observed			
Effect	−0.010	−0.009	−0.017
SE	0.005	0.005	0.005
p	0.066	0.093	0.002
RMSEA	0.018	0.010	0.010
CFI	0.948	0.957	0.957
TLI	0.923	0.923	0.921
SRMR	0.035	0.021	0.019
# of students	508 295	508 295	508 295
# of schools	19 237	19 237	19 237
# of countries	28	28	28

Table 6
Main model effect estimates of PD on student achievement in grade eight, 2003–2019.

PD participation	No controls	Teacher controls	+ instruction time
1 latent, 3 observed			
Effect	−0.016	−0.016	−0.020
SE	0.008	0.008	0.008
p	0.042	0.044	0.010
RMSEA	0.011	0.007	0.006
CFI	0.978	0.979	0.979
TLI	0.967	0.962	0.962
SRMR	0.026	0.016	0.014
# of students	385 710	385 710	385 710
# of schools	13 421	13 421	13 421
# of countries	25	25	25

Tables 5 and 6 also demonstrate that the goodness-of-fit statistics show good fit compared to the rules of thumb for CFI and TLI ($\geq$0.90), as well as RMSEA and SRMR ($\leq$0.08).[6] In addition to the results for the merged dataset (2003–2019), we performed estimations for each cycle. Fig. 2 presents the mean effects and their confidence intervals, which illustrates that in most cycles, the effect is negative and the confidence intervals are close to or overlap zero. Fig. 2 also illustrates that the confidence intervals of the merged datasets (2003–2019) are narrower because of larger samples. In Supplementary material, Section E, we present tables including the full estimates for each cycle.

5.2. *Heterogenous effects for students at different achievement levels*

As described in Section 4.2, we investigated differences in PD effects for students at different achievement levels using unconditional quantile regression. Figs. 3 and 4 present the results for the merged dataset (2003–2019), with effects and confidence intervals presented for each quantile (see Supplementary material, Section F for the full results).

The results presented in Figs. 3 and 4 indicate that the negative effects of PD estimated in the main model are concentrated among students at higher achievement levels, particularly for grade eight students.

5.3. *Sensitivity analyses*

We conducted several sensitivity analyses to explore the stability of the main model's results and potential sources of bias. This section provides a summary of these analyses. For a full description, see Supplementary material, Sections G-N.

[6] The χ^2 test cannot be conducted for models using imputed data (Mplus treats the plausible values as imputed data).

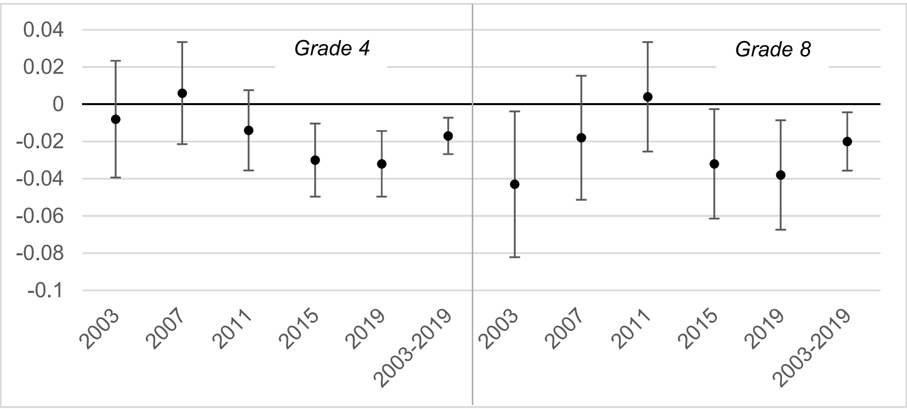

Fig. 2. Effect estimates on student achievement and their 95% confidence intervals (standard deviations). Grade four and eight.

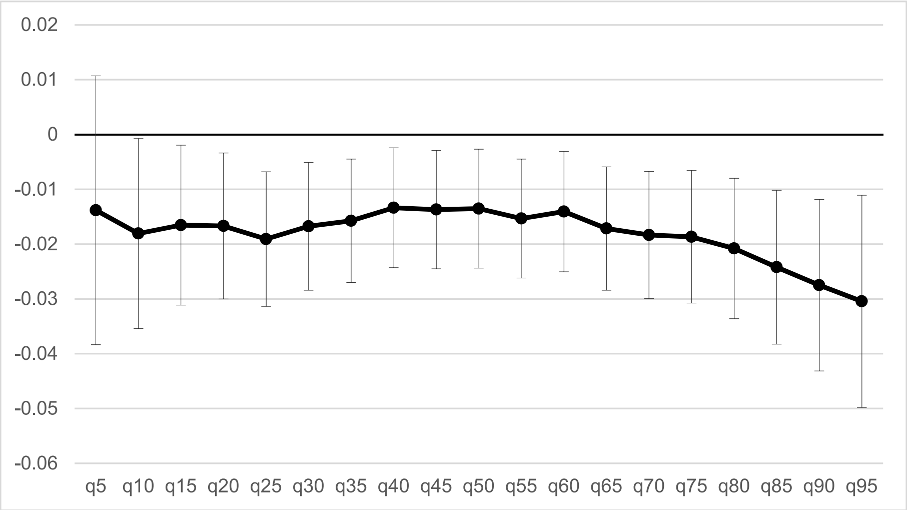

Fig. 3. Effect estimates and their 95% confidence intervals (standard deviations) for students on different achievement levels, grade four, 2003–2019.

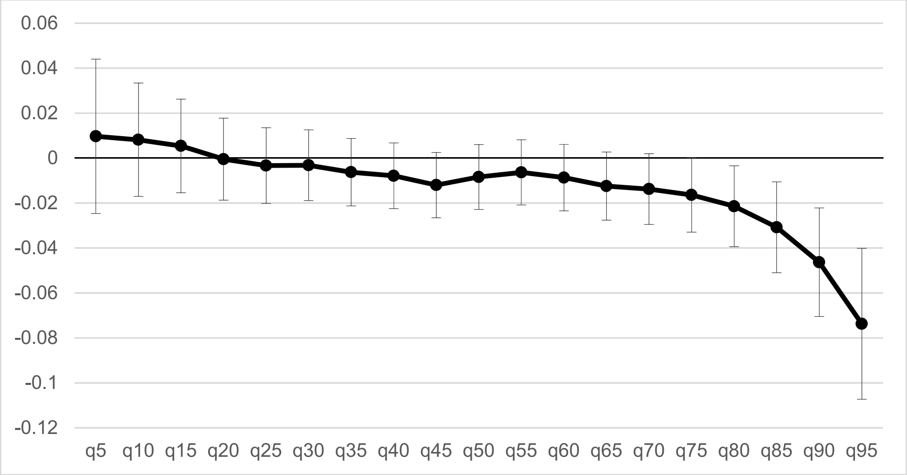

Fig. 4. Effect estimates and their 95% confidence intervals (standard deviations) for students on different achievement levels, grade eight, 2003–2019.

5.3.1. Stability in relation to various PD measures

We explored the stability in relation to various measures of PD participation in model specifications using other PD measures: a) each underlying PD item that was used as an indicator for the latent PD variable in the main model and b) the number of days of PD participation, an item available in the TIMSS 2015 and 2019 cycles. Like in the main model, the estimated PD effects on student achievement in these specifications were small and negative (see

Supplementary material, Section G), indicating that the main model's results remain robust to the different PD measures available in TIMSS.

5.3.2. Adding teacher fixed effects

The teacher characteristics used as control variables in the main model may omit unobserved teacher features which are correlated with both student achievement and PD participation. To explore this potential for bias, we added teacher fixed effects to the model by restricting the analysis to grade four students who were taught by the same teacher in mathematics and science. Therefore, unmeasured teacher characteristics that are fixed across subject areas cannot bias results. The results reported in Supplementary material, Section H show that estimates are comparable to those in the main model albeit closer to zero.

5.3.3. Potential bias caused by subject-specific omitted variables and reverse causality

Not even the teacher fixed effects specification captures subject-specific unobserved characteristics. Therefore, if unobserved teacher characteristics differ between mathematics and science in a way the control variables do not capture, and this difference is systematically associated with both teachers' PD participation and student achievement, this may still bias estimates. One way to check for this bias is to estimate the effect of teachers' subject-specific self-efficacy (a potentially confounding subject-specific variable) on their PD participation in a subject area. The results presented in Supplementary material, Section I indicate that teachers with a higher self-efficacy level participate more in PD. This pattern may hide even more negative PD effects than those estimated in the main model if differences in self-efficacy cause differences in student achievement between subject areas.

The main model's results may also be biased if students are selected into classes based on their subject-specific ability (so that class composition differs between school subjects) and teachers' PD participation correlates with class composition. We first checked this by adding students' previous subject-specific ability as control variables in a TIMSS 2015 dataset including Swedish registry data on previous student performance (standardized tests and grades). The results presented in Supplementary material, Section J show that PD effects differ little between specifications with and without previous achievement as covariates. We conducted a second check by stratifying the sample by schools' tracking status: whether schools used tracking based on subject-specific ability. The results presented in Supplementary material, Section K show great resemblance across all specifications, which indicates that the estimates of the main model are robust to subject-specific tracking policies.

An additional cause for concern is that class-level student achievement may cause teachers to participate more or less in PD. If that is the case, the causality of the model will be reversed. To check the likelihood of such bias, we compared teachers' PD participation between classes in each subject area in which the class average TIMSS score deviated positively and negatively from the school's average score. The analyses presented in Supplementary material, Section L indicate that teachers in classes that deviate positively from the school average participate somewhat more in PD, a tendency that could potentially hide even more negative PD effects than those estimated in the main model.

5.3.4. Heterogenous effects among participating countries

We explored the heterogeneity of effects among countries in the analyses presented in Supplementary material, Section M. These results show that country-level effects are in most cases not statistically significant and show little stability across TIMSS cycles. These fluctuations over time can be interpreted as statistical uncertainty caused by the small country-level sample sizes, in comparison with the full samples used in the main model.

5.3.5. PD effects in subject areas

Finally, a disadvantage of within-student between-subjects analysis is that effects are not differentiated among subject areas. To investigate whether PD effects vary between subject areas, we used a combined TIMSS and PIRLS 2011 dataset to compare PD effects among specifications that excluded one subject area (mathematics, science, or reading) at a time. Supplementary material, Section N shows that the estimated effects were of similar magnitude and sign in all specifications, indicating that the effects of the PD in which teachers typically participate do not vary strongly between subject areas in this dataset.

5.3.6. Within and between variance in PD participation

Fixed effects specifications reduce the variance in the data because differences are typically larger between than within units. This particularly affects the grade four students, as the Supplementary material, Section O shows. Yet, Section O shows that the estimated PD effects on student achievement are very similar to those in the main model (but slightly more negative) when students without variance between subjects in their teachers' PD participation are excluded.

5.3.7. Across subject spill-over effects

When the same teacher teaches students in mathematics and science, PD in one subject could potentially influence the teaching in both subjects. To investigate such spill-over effects, we estimated the main model restricted to students who were taught by different teachers in mathematics and science. The analysis (Supplementary material, Section P) shows that the main model's results (Table 6 as well as E.1 and E.2) are robust to this specification.

6. Discussion

This study's results indicate that the effect of teachers' participation in typical mathematics and science PD on student achievement is close to zero throughout all investigated TIMSS cycles (2003–2019) and both grade levels (four and eight) and that the average effect is slightly negative. Furthermore, the negative effects are particularly clear among high-achieving students. In this section, we discuss this result in relation to findings in previous research, possible mechanisms causing positive and negative PD effects, and methodological challenges in analyses of large-scale assessment data. We also outline study limitations and suggest areas for further research.

6.1. Effect estimates in previous studies of representative PD

The empirical material used in the present study is most similar to that which was used in two studies of nationally representative PD using country-level fixed-effects based on the TIMSS and PIRLS (Gustafsson & Nilsen, 2016; Van Damme et al., 2019). Both studies show positive and statistically significant results (of 0.2 standard deviations and a correlation coefficient of 0.4, respectively). However, these studies' results may be biased because of time-varying factors in each country that are not controlled for, and the estimates may comprise much statistical uncertainty because the number of observations is small. In fact, the effect sizes reported in these studies are surprisingly large in comparison to the effects reported in meta-synthesis of studies of particular PD programs, which range between 0.05 and 0.23 (Basma & Savage, 2018, 2023; Kraft et al., 2018; Lynch et al., 2019; Sims et al., 2021). Because the

PD programs analyzed in these meta-syntheses are likely well-designed and well-funded in comparison to the PD in which teachers typically participate, it is reasonable to expect lower effect sizes in nationally representative PD.

Methodologically, the present study is more similar to analyses of PD effects in US states using fixed effects at the student level albeit in those cases based on achievement growth in mathematics and reading rather than on differences between subjects (Akiba & Liang, 2016; Harris & Sass, 2011; Wallace, 2009). Like the present study, these three studies present small or statistically insignificant effects on student achievement. Therefore, we extend their conclusion that typical PD has very small effects on student achievement from four US states to a database comprising the 25–28 OECD countries participating in the 2003 to 2019 TIMSS cycles.

To our knowledge, no previous study has compared the effect of nationally representative PD on students at different achievement levels. However, studies addressing differential effects based on other types of data indicate that positive teacher and school effects are particularly pronounced for low-performing students (Burgess et al., 2022; Jackson et al., 2020) and that teaching methods focusing on open-ended problem solving are less suitable for students with low socioeconomic status (Atlay et al., 2019). Therefore, our finding that negative effects are concentrated among high-performing students is contrary to those in previous studies. One explanation may be that typical PD encourages teaching methods that are not suitable for high-performing students.

6.2. *What mechanisms explain PD's effects on student achievement?*

Section 2 presents several mechanisms by which teachers' PD participation may improve or impair student achievement. Notably, PD may cause negative effects because it takes time and resources from other purposes but may, on the other hand, enhance student achievement if teaching is improved. In addition to induce loss of instruction time, PD may also disrupt the continuity of teaching and/or increase the use of less-skilled substitute teachers. Therefore, if PD decreases qualitative instruction time, and this negative effect is not balanced by sufficient improvements in teaching quality, the average effect would be negative, in line with the findings in the present study. This mechanism could also be a possible explanation for the finding in several meta-analyses that increasing the duration of PD and coaching interventions is not associated with improved student achievement (Basma & Savage, 2018; Kraft et al., 2018; Lynch et al., 2019). Furthermore, a study of the effects of typical PD indicates that PD may decrease student results during the first year (when PD generates loss of qualitative teaching time) but pay off later albeit with very small effect sizes (Harris & Sass, 2011).

Notably, meta-syntheses indicate that improving teaching one standard deviation translates into improved student achievement of 0.21–0.27 standard deviations (Gonzalez et al., 2022; Kraft et al., 2018). Therefore, even if typical PD indeed improves teaching, the effect may be too small to translate into improved student achievement to the extent that it balances the PD-induced loss of qualitative instruction time.

Because the TIMSS data provides insufficient details on the characteristics of the PD teachers participate in, we were unable to assess the degree to which the teacher-reported PD adheres to frameworks of PD features that may predict effects on student achievement (e.g., Cordingley et al., 2015; Darling-Hammond et al., 2017; Desimone, 2009; Hill & Papay, 2022), such as teacher collaboration, individual coaching, and addressing subject-specific instructional practices. Still, it is a plausible hypothesis that inadequately designed PD causes the lack of positive effects on student achievement. Therefore, one possible policy recommendation could be to enhance the effects of typical PD by adhering to the characteristics of PD programs with demonstrated effects on student achievement. Yet, the fact that effects often decrease when programs are scaled up (Garrett et al., 2019; Kraft et al., 2018) and that promising approaches such as job-embedded or differentiated PD do not always produce measurable effects in typical school districts (Jacob & McGovern, 2015) suggests that improved student achievement should not be assumed even when high-quality characteristics are adhered to. Although we advocate that policy makers take guidance from research on the characteristics of effective PD, we also suggest that PD initiatives be evaluated in small-scale trials before scale-up.

6.3. *Methodological approaches to analyzing data from large-scale assessments*

A comparison of Tables 3 and 4 versus Tables 5 and 6 shows that the inclusion of student fixed effects strongly affects effect estimates. One likely explanation is that the estimates in Tables 3 and 4 are biased because of omitted variables and/or reverse causality. We document mechanisms that may generate these effects: teachers are more likely to participate in PD if their self-efficacy is higher than average (Supplementary material, Section I) and if they teach in classes whose academic performance deviate positively from the school average (Supplementary material, Section L). This underlines the fact that any analysis endeavoring to identify causal effects based on observational data should account for omitted variables and reverse causality. The sensitivity analyses reported in the present study indicates that within-student between-subjects analysis is a promising approach for such studies, as a supplement to fixed effects analysis based on longitudinal data.

We also emphasize that results based on different subsamples – particular TIMSS cycles and particular countries – vary substantially (see Fig. 2 and Supplementary material, Section M), which is at least partially due to larger statistical uncertainty when smaller samples are used. Because analyses based on small samples may be unreliable, we recommend further studies based on large-scale assessment data to include several cycles and countries and to compare estimates among subsamples.

6.4. *Limitations*

The present study can only estimate effects on TIMSS scores that occur within two years after teachers' mathematics and science PD participation. One limitation of this approach is that effects may become apparent later, although most studies have shown PD effects on student achievement already the first year of PD participation (Garrett et al., 2019; Kennedy, 2016; Kraft et al., 2018; Sims et al., 2021). It is, however, possible that indirect PD effects on student achievement occur on an even longer time horizon. For example, if PD improves teachers' self-efficacy and job satisfaction, and thus decreases teacher turnover (Allen & Sims, 2017; Coldwell, 2017), this could lead to positive effects accumulating over time. PD may affect other outcomes as well, such as student motivation or behavior, which we did not explore in this study.

Furthermore, the measures of PD participation that are available in TIMSS lack detail. Because teachers merely state whether they have participated in mathematics/science PD on a given set of topics during the past two years and how many hours they participated in mathematics/science PD, we cannot differentiate effects depending on the quality and form of the PD teachers participated in. Moreover, the items regarding PD participation are based on teachers' self-reported questionnaire responses which could threaten their reliability. However, several studies indicate that when self-report questions concern distinct practices in a limited time (as opposed to questions concerning how well actions are carried out), the consistency between self-reported data and observations is satisfactory (Desimone, 2009; Mayer, 1999; Swan, 2006).

Although a within-student between-subjects approach accounts for bias due to most types of student sorting, some potential causes of bias remain. As outlined in Section 5.3, though, we find no strong indications that the main model's results are biased, particularly not in a way that would change the estimates' sign.

Finally, although a fixed effects approach limits the risks for bias caused by selection, it also limits the variance in the data because the within-student between-subject variance is smaller than the between-student variance. This decreases the analysis' precision so that a larger sample is required to identify statistically significant effects. This is particularly a threat to analyses drawing on small samples, such as specific TIMSS cycles or countries. As Fig. 2 shows, though, datasets including several cycles and countries allow precise estimates.

6.5. Further research on the PD in which teachers typically participate

This and previous studies of typical PD in which student fixed effects were used (Akiba & Liang, 2016; Harris & Sass, 2011; Wallace, 2009) indicate that an expected effect on student achievement of the PD in which teachers typically participate is close to zero. Therefore, it is important to advance the knowledge of whether and how findings of positive PD effects in studies of particular PD programs (Basma & Savage, 2018, 2023; Kraft et al., 2018; Lynch et al., 2019; Sims et al., 2021) are transferrable to the PD in which teachers typically participate. For example, if mechanisms that produce positive PD effects in particular PD programs are identified (e.g., Sims et al., 2021), does implementation of these mechanisms in typical settings (by local PD developers, with realistic levels of resources, etc.) also generate positive PD effects? Although some work has been conducted in this area, such as studies of national or large-scale PD programs (e.g., Jacob & Lefgren, 2004; Lindvall et al., 2021), few studies present credible estimations of how PD policies, such as regulations of PD quality and teachers' PD participation, affect student achievement. Furthermore, adding items to large-scale studies such as the TIMSS, PIRLS, and TALIS concerning features of the PD in which teachers have participated would improve the understanding of typical PD and could enable analyses of how such features moderate PD effects on student achievement in various contexts. Frameworks of effective PD (e.g., Cordingley et al., 2015; Darling-Hammond et al., 2017; Desimone, 2009; Hill & Papay, 2022) and reviews of causal evidence on PD programs' effects (Basma & Savage, 2018, 2023; Kraft et al., 2018; Lynch et al., 2019; Sims et al., 2021) may guide the formulation of such questionnaire items.

Further research may also explore the balance among potential mechanisms through which PD influences student achievement. For example, studies using repeated measures during and after a PD intervention can explore the extent to which PD duration influences teaching and student learning, thus illuminating whether teaching continuously improves during PD programs, whether teaching quality is maintained after PD participation, and whether improved teaching generates larger effects on student achievement when PD-induced time loss is absent. Furthermore, it is also important to study PD effects on outcomes other than academic achievement, such as student absences and effort in class, because such outcomes may be at least as important for students' futures (Jackson, 2018; Kraft, 2019).

7. Funding

This work was supported by the Swedish Research Council [Grant number 2021-03883].

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

A link to the replication package is attached: https://myfiles.uu.se/filr/public-link/file-download/026e0f4c869904620188667a07cf4c18/49020/4406504094379333909/replicationpackage.zip.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.tate.2023.104242.

References

Akiba, M., & Liang, G. (2016). Effects of teacher professional learning activities on student achievement growth. *The Journal of Educational Research, 109*(1), 99–110.

Allen, R., & Sims, S. (2017). Improving science teacher retention: Do national STEM learning network professional development courses keep science teachers in the classroom? Wellcome trust & education datalab. https://www.stem.org.uk/system/files/elibrary-resources/2019/10/science-teacher-retention_0.pdf.

Allison, P. (2009). *Fixed effects regression models*. SAGE Publications.

Allison, P., Williams, R., & Moral-Benito, E. (2017). Maximum likelihood for cross-lagged panel models with fixed effects. *Socius, 3*.

Atlay, C., Tieben, N., Hillmert, S., & Fauth, B. (2019). Instructional quality and achievement inequality: How effective is teaching in closing the social achievement gap? *Learning and Instruction, 63*, Article 101211.

Basma, B., & Savage, R. (2018). Teacher professional development and student literacy growth: A systematic review and meta-analysis. *Educational Psychology Review, 30*(2), 457–481.

Basma, B., & Savage, R. (2023). Teacher professional development and student reading in middle and high school: A systematic review and meta-analysis. *Journal of Teacher Education*. https://doi.org/10.1177/00224871231153084

Bietenbeck, J. (2014). Teaching practices and cognitive skills. *Labour Economics, 30*, 143–153.

Bietenbeck, J., & Collins, M. (2023). New evidence on the importance of instruction time for student achievement on international assessments. *Journal of Applied Econometrics*. https://doi.org/10.1002/jae.2957

Blömeke, S., Olsen, R. V., & Suhl, U. (2016). Relation of student achievement to the quality of their teachers and instructional quality. In T. Nilsen, & J.-E. Gustafsson (Eds.), *Teacher quality, instructional quality and student outcomes: Relationships across countries, cohorts and time* (pp. 21–50). Springer International Publishing.

Bollen, K. A., & Brand, J. E. (2010). A general panel model with random and fixed effects: A structural equations approach. *Social Forces, 89*(1), 1–34.

Burgess, S., Rawal, S., & Taylor, E. S. (2022). *Characterising effective teaching*. University of Bristol; Nuffield Foundation. https://www.nuffieldfoundation.org/wp-content/uploads/2022/05/Burgess-Characterising-Effective-Teaching-Full-Report-April-2022.pdf.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *The American Economic Review, 104*(9), 2633–2679.

Coldwell, M. (2017). Exploring the influence of professional development on teacher careers: A path model approach. *Teaching and Teacher Education, 61*, 189–198.

Cordingley, P., Higgins, S., Greany, T., Buckler, N., Coles-Jordan, D., Crisp, B., Saunders, L., & Coe, R. (2015). Developing great teaching: Lessons from the international reviews into effective professional development. *Teacher Development Trust*. https://tdtrust.org/wp-content/uploads/2015/10/DGT-Full-report.pdf.

Darling-Hammond, L., Hyler, M. E., & Gardner, M. (2017). *Effective teacher professional development*. Learning Policy Institute. https://learningpolicyinstitute.org/sites/default/files/product-files/Effective_Teacher_Professional_Development_REPORT.pdf.

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher, 38*(3), 181–199.

Eriksson, K., Lindvall, J., Helenius, O., & Ryve, A. (2021). Socioeconomic status as a multidimensional predictor of student achievement in 77 societies. *Frontiers in Education, 6*.

Fuchs, L. S., Sterba, S. K., Fuchs, D., & Malone, A. S. (2016). Does evidence-based fractions intervention address the needs of very low-performing students? *Journal of Research on Educational Effectiveness, 9*(4), 662–677.

Garrett, R., Citkowicz, M., & Williams, R. (2019). How responsive is a teacher's classroom practice to intervention? A meta-analysis of randomized field studies. *Review of Research in Education, 43*(1), 106–137.

Gonzalez, K. E., Lynch, K., & Hill, H. C. (2022). *A meta-Analysis of the experimental evidence linking STEM classroom Interventions to teacher knowledge, classroom instruction, and student achievement (EdWorkingPaper No 22-515)*. Annenberg Institute at Brown University. https://doi.org/10.26300/d9kc-4264

Grönqvist, E., Öckert, B., & Rosenqvist, O. (2021). *Does the 'Boost for mathematics' boost mathematics? IFAU.* https://www.ifau.se/Forskning/Publikationer/Working-papers/2021/does-the-boost-for-mathematics-boost-mathematics/.

Gustafsson, J.-E., & Nilsen, T. (2016). The impact of school climate and teacher quality on mathematics achievement: A difference-in-differences approach. In T. Nilsen, & J.-E. Gustafsson (Eds.), *Teacher quality, instructional quality and student outcomes: Relationships across countries, cohorts and time* (pp. 81–95). Springer International Publishing.

Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education Review, 30*(3), 466–479.

Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics, 95*(7), 798–812.

Havard, B., Nguyen, G.-N., & Otto, B. (2018). The impact of technology use and teacher professional development on U.S. National assessment of educational progress (NAEP) mathematics achievement. *Education and Information Technologies, 23*(5), 1897–1918. https://doi.org/10.1007/s10639-018-9696-4. Social Science Premium Collection.

Hill, H. C., Beisiegel, M., & Jacob, R. (2013). Professional development research consensus, crossroads, and challenges. *Educational Researcher, 42*(9), 476–487.

Hill, H. C., & Papay, J. P. (2022). *Building better PL: How to strengthen teacher learning*. RPPL. https://annenberg.brown.edu/sites/default/files/rppl-building-better-pl.pdf.

Jackson, C. K. (2018). What do test scores miss? The importance of teacher effects on non–test score outcomes. *Journal of Political Economy, 126*(5), 2072–2107.

Jackson, C. K., Porter, S. C., Easton, J. Q., & Kiguel, S. (2020). *Who benefits from attending effective high schools?* National Bureau of Economic Research. https://doi.org/10.3386/w28194. Working Paper No. 28194).

Jacob, B. A., & Lefgren, L. (2004). The impact of teacher training on student achievement quasi-experimental evidence from school reform efforts in chicago. *Journal of Human Resources, 39*(1), 50–79.

Jacob, A., & McGovern, K. (2015). *The mirage: Confronting the hard truth about our quest for teacher development*. TNTP. https://eric.ed.gov/?id=ED558206.

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

Jerrim, J., Sims, S., & Oliver, M. (2023). Teacher self-efficacy and pupil achievement: Much ado about nothing? International evidence from TIMSS. *Teachers and Teaching.* https://doi.org/10.1080/13540602.2022.2159365

Kennedy, M. M. (2016). How does professional development improve teaching? *Review of Educational Research, 86*(4), 945–980.

Killeen, K. M., Monk, D. H., & Plecki, M. L. (2002). School district spending on professional development: Insights available from national data (1992-1998). *Journal of Education Finance, 28*(1), 25–49.

Kirsten, N. (2020). Svenska lärares deltagande i kompetensutveckling. En statistisk bearbetning av uppgifter om lärares kompetensutveckling. *PIRLS och PISA 2001-2018 [Swedish teachers' PD participation over time and in international comparison : Statistical analyses of data on teachers' professional development in TALIS, TIMSS, PIRLS, and PISA during the period 2001-2018]. Uppsala universitet*. http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-405426.

Koellner, K., & Jacobs, J. (2015). Distinguishing models of professional development: The case of an adaptive model's impact on teachers' knowledge, instruction, and student achievement. *Journal of Teacher Education, 66*(1), 51–67.

Kraft, M. A. (2019). Teacher effects on complex cognitive skills and social-emotional competencies. *Journal of Human Resources, 54*(1), 1–36.

Kraft, M. A., Blazar, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research, 88*(4), 547–588.

Lavy, V. (2015). Do differences in schools' instruction time explain international achievement gaps? Evidence from developed and developing countries. *The Economic Journal, 125*(588), F397–F424.

Lindvall, J., Helenius, O., Eriksson, K., & Ryve, A. (2021). Impact and design of a national-scale professional development program for mathematics teachers. *Scandinavian Journal of Educational Research, 66*(5), 744–759.

Lynch, K., Hill, H. C., Gonzalez, K. E., & Pollard, C. (2019). Strengthening the research base that informs STEM instructional improvement efforts: A meta-analysis. *Educational Evaluation and Policy Analysis, 41*(3), 260–293.

Martin, M. O., von Davier, M., & Mullis, I. V. (2020). *Methods and procedures: TIMSS 2019 technical report*. International Association for the Evaluation of Educational Achievement (IEA).

Mayer, D. P. (1999). Measuring instructional practice: Can policymakers trust survey data? *Educational Evaluation and Policy Analysis, 21*(1), 29–45.

Miles, K. H., Odden, A., Fermanich, M., & Archibald, S. (2004). Inside the black box of school district spending on professional development: Lessons from five urban districts. *Journal of Education Finance, 30*(1), 1–26.

Newsom, J. T. (2015). *Longitudinal structural equation modeling. A comprehensive introduction*. Routledge, Taylor and Francis Group.

Patfield, S., Gore, J., & Harris, J. (2022). Scaling up effective professional development: Toward successful adaptation through attention to underlying mechanisms. *Teaching and Teacher Education, 116*, Article 103756.

Rios Avila, F. (2019). *Recentered influence functions in Stata: Methods for analyzing the determinants of poverty and inequality* (SSRN Scholarly Paper No. 3378811) https://papers.ssrn.com/abstract=3378811.

Rivkin, S. G., & Schiman, J. C. (2015). Instruction time, classroom quality, and academic achievement. *The Economic Journal, 125*(588), F425–F448.

Schwerdt, G., & Wuppermann, A. C. (2011). Is traditional teaching really all that bad? A within-student between-subject approach. *Economics of Education Review, 30*(2), 365–379.

Sims, S., & Fletcher-Wood, H. (2021). Identifying the characteristics of effective teacher professional development: A critical review. *School Effectiveness and School Improvement, 32*(1), 47–63.

Sims, S., Fletcher-Wood, H., O'Mara-Eves, A., Cottingham, S., Stansfield, C., Van Herwegen, J., & Anders, J. (2021). *What are the characteristics of effective teacher professional development? A systematic review and meta-analysis*. Education Endowment Foundation. https://eric.ed.gov/?id=ED615914.

Swan, M. (2006). Designing and using research instruments to describe the beliefs and practices of mathematics teachers. *Research in Education, 75*(1), 58–70.

Sykes, G., & Wilson, S. M. (2016). Can policy (re)form instruction? In D. H. Gitomer, & C. A. Bell (Eds.), *Handbook of research on teaching* (5th ed., pp. 851–916). American Educational Research Association.

Timperley, H. (2015). Continuing professional development. In J. D. Wright (Ed.), *International encyclopedia of the social & behavioral sciences* (2nd ed., pp. 796–802). Elsevier.

Van Damme, J., Bellens, K., Tielemans, K., & Van Den Noortgate, W. (2019). Do changes in instructional time, professional development of teachers and age of students explain changes in reading comprehension at the country level? An exploration of PIRLS 2006 and 2016. *Education and Self-Development, 14*(2), 10–31.

Van den Brande, J., & Zuccollo, J. (2021). *The cost of high-quality professional development for teachers in England*. Wellcome Trust & Education Policy Institute. https://epi.org.uk/wp-content/uploads/2021/07/2021-Cost-of-quality-teacher-cpd_EPI.pdf.

Wallace, M. R. (2009). Making sense of the links: Professional development, teacher practices, and student achievement. *Teachers College Record, 111*(2), 573–596. Social Science Premium Collection.

Wang, J., & Wang, X. (2019). *Structural equation modeling: Applications using Mplus*. John Wiley & Sons, Incorporated.