Social Sciences & Humanities Open 8 (2023) 100605

Contents lists available at ScienceDirect

Social Sciences & Humanities Open

journal homepage: www.sciencedirect.com/journal/social-sciences-and-humanities-open

Review Article

Invalid estimates and biased means. A replication of a recent meta-analysis investigating the effect of teacher professional development on pupil outcomes

Joshua Fullard [a,b,*]

[a] Warwick Business School, University of Warwick, UK
[b] Institute for Social and Economics Research, University of Essex, UK

ARTICLE INFO

Keywords:
Meta-analysis
Replication
Teacher professional development

ABSTRACT

Meta-analyses can play an important role in educational research. Aggregating results from different, but comparable, studies to demonstrate the range of plausible effect sizes can be highly informative. However, the validity of a meta-analysis mean depends on the quality of the studies included. If a meta-analysis includes studies that are not valid tests of the meta-analysis hypothesis this can introduce bias into the estimate. In this review we replicate a recent meta-analysis investigating the effect of teacher professional development on pupil outcomes to illustrate the importance of maintaining data quality. We demonstrate that when we exclude the studies that are invalid tests of the meta-analysis hypothesis, adjust the authors biased selection criteria and include a valid study that the authors inappropriately exclude the effect size falls from 0.09 to −0.008.

1. Introduction

Teachers are, perhaps, the most important profession in society. After all the role teachers play in educating and inspiring young people has large societal benefits (Aaronson et al., 2007; Hanushek et al., 2016). These benefits include an increase in democratic participation, a reduction in crime and an improvement in health outcomes (Cutler & Lleras-Muney, 2006; Groot & van den Brink, 2010; Marshall, 2016). Consequently, improving teacher effectiveness is high on the policy agenda around the world (Caena, 2011; Desimone, 2009).

One approach to improve teacher effectiveness is to recruit more effective teachers.[1] While this is theoretically possible, it is unlike to occur for three reasons. First, it is challenging to identify effective teachers, so it is not clear if new entrants are any more effective than those they are replacing (Hanushek & Rivkin, 2006). Second, more than half of the countries in Europe as well as Australia and USA are already facing acute teacher recruitment challenges so there is unlikely to be any scope to focus on teacher effectiveness (Birch et al., 2018; DfE, 2022; Nguyen et al., 2022). In the UK, for instance, policymakers have persistently failed to recruit enough new teachers over the last decade (roughly 33,000 entrants per year) to replace those who leave (36,000–41,000 leavers per year) causing pupil to teacher ratios to increase from 17.3 in 2010 to 19.1 in 2019 (Fullard, J., 2022). Third, even if policymakers could identify and recruit more effective teachers it would take time for them to train and become integrated into the education system. Furthermore, this will be mitigated by the high rate of attrition, in England roughly 1 in 3 teachers leave within the first five years, and the fact that evidence from the USA suggests that more effective teachers are more likely to leave (Long & Danechi, 2021; Wiswall, 2013). Due to the significant challenges associated with improving teacher effectiveness through recruitment, policies aimed at improving the effectiveness of existing teachers seem like the next best approach.

The empirical literature shows that teacher effectiveness generally improves with experience, particularly novice teachers – although the magnitude and persistence of the effect is not well established (Chingos & Peterson, 2011; Podolsky et al., 2019; Rockoff, 2004). For instance Wiswall (2013) finds large improvements in teacher effectiveness among established teachers while Chingos and Peterson (2011) finds that teachers might actually become less effective over time. With this literature in mind, if a relatively inexpensive professional development (PD) programme could improve teacher effectiveness more rapidly, it could be a cost-effective approach to improving pupil attainment.

In recent years several studies have attempted to understand the effect of teacher PD on pupil attainment through meta-analysis. These

* Corresponding author. University of Warwick, Warwick Business School, Scarman Road, Coventry, United Kingdom.
E-mail address: Joshua.Fullard@wbs.ac.uk.

[1] Teacher effectiveness (or quality) in this paper refers to teachers' ability to improve student outcomes measured by test scores. This is discussed in Section 4.

https://doi.org/10.1016/j.ssaho.2023.100605
Received 24 March 2023; Received in revised form 5 June 2023; Accepted 18 June 2023
Available online 21 June 2023

include Lynch et al. (2019) and Sims et al. (2021) where they have found large, positive, statistically significant effect sizes and concluded that PD has a positive effect on pupil attainment – with results like this is unsurprising that teacher PD has become imbedded in both UK and US official government guidance. Schools in England, for instance, currently spend around £1.5bn per year on teacher PD (roughly £2950 per teacher).

Although there are exceptions, such as Angrist and Lavy (2001), there are concerns that these meta-analyses' are overstating the evidence base as the literature generally finds no effect (Jackson et al., 2014). Even multiyear teacher training and certification programs are generally shown to have no effect on pupil outcomes (Buddin & Zamarro, 2009; Harris & Sass, 2011; Koedel et al., 2015). Chingos and Peterson (2011), for instance, concludes that, despite the challenges identifying and recruiting effective teachers, it remains easier to recruit an effective teacher than to train one.

Meta-analysis has long been a popular tool used in educational research (Kulik & Kulik, 1989; Slavin, 1984).[2] However, researchers have long held concerns over the impact that the quality of studies included have on the meta-analysis estimates, and the corresponding conclusions (Almaatouq et al., 2022; Sharpe, 1997; Simonsohn et al., 2022). One of the main issues here is that not all empirical research is equally valid. Some studies offer valid tests of the meta-analysis hypothesis, but many do not. For instance, in an early evidence review of the effect of teacher PD on student achievement Yoon et al. (2007) found that, of over 1300 studies, only 9 were valid tests of the meta-analysis hypothesis.

From a poor identification strategy and/or experimental design to reporting errors and data fraud there are a range of factors that can invalidate a study. The problem with including invalid studies in a meta-analysis is that the different sources of invalidity are likely to go in the same direction – and will bias the meta-analysis estimate (Simonsohn et al., 2022).

In this paper we review Fletcher-Wood and Zuccollo (2020), a recent meta-analysis investigating the effect of teacher PD on pupil outcomes. Specifically, we investigate the extent to which their findings are driven by the inclusion of invalid studies. While there are several recent meta-analyses investigating similar research questions such as Kraft et al. (2018), Lynch et al. (2019) and Sims et al. (2021), Fletcher-Wood and Zuccollo (2020) was selected due to its: narrow focus, policy relevance and lack of substantive peer review.

Firstly, Fletcher-Wood and Zuccollo (2020) is relatively narrow in focus. The paper only considers studies that use Randomised Control Trials (RCT's) investigating the effect of teacher PD on pupil attainment (measured by test scores). As there are objective criteria RCT's must meet to be considered successful, and it is clear if the RCT is measuring PD, identifying which studies are (not) valid estimates of the meta-analysis hypothesis in this setting should be straightforward. Second, the results from this meta-analysis are directly used in a cost-benefit analysis of teacher PD to inform the policy debate (Van den Brande & Zuccollo, 2021). Third, this meta-analysis was performed by one of the UK's most influential policy organisation, the Education Policy Institute. The meta-analysis we review is cited in both the Governments Teaching Development Framework and Ofsted's review of teachers' professional development (DfE, 2020; Ofsted, 2023).[3] As it is not clear that this research has undergone the same scientific scrunchy as research published in academic journals – and they report an effect size almost twice as large as comparable studies such as Sims et al. (2021) who use a similar approach - investigating if these results and the corresponding conclusions are valid seem like an important contribution. These results will also inform a larger debate about the role that policy organisations, whose research does not undergo significant peer review, should have on informing public debate and shaping the policy agenda.

In this paper we demonstrate that the positive effect found by Fletcher-Wood and Zuccollo (2020) is entirely driven by the inclusion of invalid studies and the authors selection criteria. First, of the 49 estimates (from 42 studies) included only 14 (from 10 studies) are valid estimates of the meta-analysis hypothesis. Second, the authors selection criteria "when multiple outcome measure were supplied and all seemed to reflect similar dimensions, we chose the first listed" introduces upward bias into the meta-analysis as the positive, statistically significant effects, are more likely to be reported first. Third, the authors include a valid estimate of the meta-analysis hypothesis in the evidence review but, without justification, exclude it from the meta-analysis. When the invalid studies are removed, the bias in the selection criteria adjusted, and valid study included the effect size falls from 0.09 to −0.008. This provides reason to doubt the claim that PD improves student learning, based on the evidence provided in this meta-analysis. In addition, the claim that an increase in PD will increase each student's lifetime earnings by over £6000 and a net societal benefit of over £61bn over ten years – calculated using this meta-analysis mean – is unlikely to be true.

In section 2 we discuss the empirical challenges estimating the effect of PD and the criteria a RCT must meet to be considered successful and identify which studies are invalided estimates of the meta-analysis hypothesis and should be excluded. In Section 3 we replicate the meta-analysis removing the invalid studies,using an unbiased selection criteria and preform robustness checks. In Section 4 we discuss the results.

2. Empirical challenges estimating the effect of professional development

Investigating the effect of professional development (PD) on pupil outcomes can be empirically challenging because the decision to engage in PD is likely to be correlated with other factors that influence pupil outcomes, many of which are unlikely to be observed by the researcher. Consequently, regressing pupil outcome on teacher PD would not identify the causal effect of PD due to omitted variable bias.

To give an example, there are two forms of selection that would bias the estimates. The first is within school selection. Within a school more effective teachers might be more likely to engage in PD which would produce an upwards bias. Conversely, we might also have a situation where less effective teachers might perceive greater returns to PD and are more likely to engage. This would produce a downward bias.

The second form of bias is between school selection. Schools that offer more PD opportunities might also offer teachers additional complementary resources/support that improves pupil outcomes. This would produce an upward bias. Conversely, schools that offer more PD, due to financial constraints, might provide teachers with less resources/support in other areas. This would produce a downward bias.

As the direction of the bias is unclear researchers investigating the effect of PD have often turned to experimental methods such as Randomised Control Trials (RCT). In the simplest case a RCT in this setting would be teachers getting randomly sorted into two groups where one group gets some additional PD (Treatment Group), and the other group does not (control group). Due to randomisation the treatment and control group should be statistically similar on observable (and unobservable) characteristics therefore any difference in pupil attainment between the two groups post hoc is the causal effect of PD.

There are three essential principles in experimental design that are necessary conditions for causal inference. These are randomisation, sample size and design. A fundamental issue in causal inference is that we cannot observe the counterfactual – we cannot observe how pupils would have performed had their teacher not received the PD. The

[2] By Meta-analysis we specifically refer to as the statistically combining of findings from different research studies to produce an overall effect size, or meta-analysis mean.

[3] Ofsted is the body responsible for inspecting and regulating schools in England.

second-best approach is to observe two statistically identical groups. It is not enough to control for observable characteristics as there could be differences that are not observed by the researcher. Randomisation is the only way to balance unobservable characteristics. Random assignment does not mean that there is a statistically identical participant in the control group for every treatment participant. But it does mean that the participants in the treatment group are identical to those in the control group, on average, if the groups are sufficiently large. This brings us to the second essential principle, sample size. Experiments require enough participants to make sure that the results are not due to poor identification or chance.

To identify the effect of an intervention RCT's need to be carefully designed so that the only differences between the treatment and control group is that intervention. Even if the randomisation is successful and the sample size is sufficiently large the RCT might not causally identify the effect of PD if the RCT is not designed to be able to identify the effect of PD. For example, if the treatment group is exposed to an intervention of PD and something else such as a curriculum change, then the effect of PD is not causally identified. This is referred to as overidentification – there is not enough identifying variation in the data (treatment vs control group) to isolate the effect that PD has on pupil attainment. This is why RCT's often have multiple treatment arms to causally identify the role that different factors have.

2.1. Invalid estimates of professional development

With the three essential principals of experimental design in mind, we are going to discuss why 32 of the 42 original studies in Fletcher--Wood and Zuccollo (2020) are invalid tests of the meta-analysis hypothesis and should be excluded from the analysis. To clearly identify which studies were included in the original meta-analysis these are numbered (e.g., Study 1, Study 2 …. Study 42). Table 1 shows which studies correspondents to each number.[4]

2.2. A control group is essential

To evaluate the effect of a PD intervention on pupil outcomes it is necessary to have a control group to identify the counterfactual – what would have happened to pupil attainment had their teachers not been exposed to PD. Study 3 contains two different treatment groups (two groups are exposed to different interventions) and no control group. This design is not appropriate for causal interpretation and should not be included in the meta-analysis.

2.3. Comparing similar groups

In a RCT the authors need to demonstrate that the randomisation was successful and that they are comparing two statistically similar groups. It is the responsibility of the authors to demonstrate that the groups are balanced. If they do not demonstrate this the study should be excluded from the meta-analysis as the reader cannot be confidence that any difference in outcomes is not driven by group differences on observable or unobservable characteristics. In this case Study 1 provides no evidence that the treatment and control groups are balanced and therefore should be excluded.

The other studies all include balance tables which demonstrate if the randomisation was successful, or not. In this setting, if the randomisation was successful, we would expect to see balance on key school level characteristics (the randomisation is performed at the school level), teacher level characteristics (teachers are the ones who complete the PD) and pupil level characteristics (our outcome of interest is pupil attainment).

Table 1
List of studies included in the Meta Analysis with the numbers that they are referred to in the text and their corresponding citation. Column 3 indicates if the study is a valid estimate of the Meta Analysis hypothesis and how many valid estimates the study contains.

Study number	Study name	Valid Studies (Number of estimates)
Study 1	Jacobs et al. (2007)	
Study 2	Fisher et al. (2011)	
Study 3	Sailors and Price (2010)	
Study 4	Allen et al. (2015)	
Study 5	Heller et al. (2007)	
Study 6	Motteram et al. (2016)	
Study 7	Penuel et al. (2011)	Yes (1)
Study 8	Greenleaf et al. (2011)	
Study 9	Campbell and Malkus (2011)	
Study 10	Cotabish et al. (2013)	
Study 11	Allen et al. (2011)	Yes (1)
Study 12	Hanley et al. (2016)	
Study 13	Glazerman et al. (2010)	
Study 14	Matsumura et al. (2013)	
Study 15	Jay et al. (2017)	
Study 16	Parkinson et al. (2015)	
Study 17	Papay et al. (2020)	
Study 18	Speckesser et al. (2018)	Yes (2)
Study 19	Vignoles et al. (2015)	
Study 20	Anders et al. (2018)	
Study 21	Miller et al. (2017)	
Study 22	Jerrim et al. (2015)	
Study 23	Wiggins et al. (2019)	
Study 24	Humphrey et al. (2018)	
Study 25	Murphy et al. (2017)	Yes (1)
Study 26	Gersten et al. (2010)	
Study 27	Rose et al. (2017)	
Study 28	Kitmitto et al. (2018)	
Study 29	Boylan et al. (2018)	
Study 30	Wiggins et al. (2017)	
Study 31	Garet et al. (2011)	Yes (1)
Study 32	Hanley et al. (2015)	
Study 33	Worth et al. (2017)	
Study 34	McNally (2014)	Yes (1)
Study 35	Education (2016)	Yes (2)
Study 36	Sloan et al. (2018)	
Study 37	Tracey et al. (2019)	Yes (1)
Study 38	Santagata et al. (2010)	
Study 39	Thurston et al. (2016)	
Study 40	Biggart (2015)	
Study 41	Borman et al. (2008)	Yes (3)
Study 42	Garet et al. (2008)	Yes (1)

Of the studies included in the meta-analysis 12 have statistically significant differences on key covariates. Some of the differences are very large. In Study 23 there are differences in school characteristics (60% of the schools in the control group are large compared to 35% of the treatment group) and pupil characteristics (15% of the pupils in the control group have English as an additional language compared to 5% of the treatment). In Study 13, Study 26 and Study 38 there are statistically significant differences in teachers' levels of educational attainment (69.8% vs 59.5% in teaching certification, 6% vs 14% in post-masters' qualifications and 96% vs 73% in university qualification respectively). In Study 15 there are differences on the proportion of students with English as an additional language (53% vs 47%), Study 24 contains differences on pupils' gender (50.4% vs 54.9%), Free School Meal (FSM) eligibility (27% vs 23%) and special educational needs (23.1% vs 18%), Study 40 has differences on pupils gender (56% vs 44%) and FSM eligibility (36% vs 43%), Study 5 contains differences by ethnicity (35% vs 25% African America) and English as an addition language (14% vs 22%) and Study 32 has differences on pupil gender (47.1% vs 41.5%).[5]

[4] The numbering is determined by the order they were presented in the original meta-analysis.

[5] Study 32 also has differential attrition rates (15% in treatment group compared to 19% in the control).

Study 2 has a very small sample (16 teachers) and they are not balanced on teacher characteristics (sex or subjects taught). Finally, Study 27 and Study 8 contain statistically significant differences in pupils' baseline levels of educational attainment. In these studies, it is not clear that the randomisation was successful and that they are comparing two statistically similar groups.[6] Therefore, these are invalid estimates of the causal effect of PD and should be excluded from the meta-analysis.

2.4. Attrition and data quality

While the randomisation might have been successful non-random attrition (or missing data) could bias the estimates - other differences between the groups might be driving the results.

There are three studies included in the meta-analysis that have issues with attrition and data quality. Study 14 has an extremally high attrition rate (almost 50%) and Study 22 has statistically significant differences in attrition between groups. Finally, in Study 12, 25% of the pupil outcomes are missing and it is not clear what, if any, of the intervention has been delivered. These studies should be excluded due to concerns over attrition and data quality.

2.5. Studies not investigating professional development

There are four studies that are included in the meta-analysis that are not investigating the effect of teacher PD. Study 9 investigate the benefit of having an additional experienced member of leadership staff in a school while Study 33 and Study 17 investigate the role that teacher peer observations have on pupil outcomes. Study 17 explicitly states that their intervention is an alternative to PD as "the empirical evidence suggest little effect [of professional development] on teacher performance". Finally, Study 6 investigates the impact that students participating in biweekly reflective activities (both individual and in class) has on outcomes. These studies should not be included in the meta-analysis because they are not investigating the effect of PD.

2.6. Overidentification

Many of the studies that are included in the meta-analysis are investigating the effect of multiple interventions on pupil outcomes, not just PD. This is not necessarily a problem as long as the RCT is designed to identify the effect of PD individually. Study 42, for example, uses two different treatment arms to investigate the role of PD (treatment A) and the role of PD and school coaching (treatment B) compared to a control group.

However, eleven studies in the meta-analysis include multiple interventions in one treatment group so the effect of PD is not cleanly identified. Study 10, Study 29, Study 30 and Study 36 all investigate the effect of a curriculum change, the provision of additional resources and PD. Study 16, Study 21 and Study 48 investigate the impact of PD and additional resources as well as school leaders receiving additional support. Study 4, Study 19 and Study 20 investigates the roll of PD, additional peer support and additional leadership support (Study 20 only). Study 28 investigates the effect of PD plus additional teaching materials. Finally, Study 39 investigates the effect of PD, additional resources (for both teachers and support staff), additional peer support and additional collaboration with parents.

In all these studies the impact of PD is not clearly identified because they are combined with other intervention(s) that could plausibly impact pupil attainment. Therefore, these studies are not valid estimates of the meta-analysis hypothesis and should be excluded from the analysis.

2.7. Publication bias and selection criteria

It is not just the inclusion of invalid estimates that could bias the meta-analysis mean. Publication bias compounded by the authors section criteria also introduces additional upward bias in this setting.

A general problem for any meta-analyses is publication bias - studies that show statistically significant results are more likely to be published that those that do not show an effect (Begg, 1994). In this setting studies that find PD has a positive effect on pupil outcomes are more likely to be published. While there are empirical strategies to deal with publication bias the meta-analysis does not make any adjustments or caveat their results by stating that the meta-analysis mean is likely to be positively biased (Stanley, 2005). This issue is further compounded by the authors selection criteria.

In the meta-analysis the authors state that "when multiple outcome measure were supplied and all seemed to reflect similar dimensions, we chose the first listed." As studies generally report the largest (positive) effect first, this selection criteria amounts to only including the positive/statistically significant effects from studies that contain multiple measures of pupil outcomes. For example, Study 18 finds no effect on Math or English GCSE attainment, but these results are not included in the meta-analysis because they are not presented first - even though the results for Math and English attainment are more clearly aligned with the measure of attainment from other studies in the meta-analysis. This selection criteria introduces additional upward bias into the meta-analysis mean. To demonstrate this, we use a new selection rule. Instead of selecting the measures of attainment that the authors report first as a default, we select the measures of attainment that are most comparable to those used in other studies. Of the valid studies three contain multiple estimates (Study 18, Study 35 and Study 41). In Study 18 we select the Math and English GCSE measures (instead of Attainment 8 scores). In Study 35 in select the same measures as those that were included in the original meta-analysis (Math and English GCSE measures) and in Study 41 we select all three science measures not just the measure that is presented first.

3. Replication

Table 2 shows how the mean effect size changes when we remove the invalid studies that were originally included (columns 2–3), adjust the authors bias selection criteria when there are multiple estimates in a valid study (column 4) and include a valid study that was inappropriately excluded (column 5).

The original study calculates a mean effect size of 0.09 using 49 estimates obtained from 42 studies (Table 2 column 1). As discussed, we remove 32 of these studies from the meta-analysis because they do not estimate the causal effect of PD on pupil attainment (Table 3).

The remaining 10 studies contain 14 estimates of the causal effect of teacher PD on pupil attainment.[7] Of these 14 estimates 2 are statistically different from 0 (Study 7 finds a positive effect and Study 41 finds a negative effect). The remaining 12 estimates are statistically indistinguishable from 0. The mean effect size for these 14 estimates is 0.004 (Table 2 Column 4).

3.1. Selection criteria

In the meta-analysis replication above, we used a different approach for deciding which measures of attainment should be included when there are multiple measures in a valid study. We believe that the authors approach is flawed (selecting the measure presented first) and

[6] There are different weighting strategies to try to deal with statistically significant differences between treatment and group groups (i.e., inverse probability weighting, propensity score matching and regression adjustment). None of these studies make these adjustments.

[7] The remaining 10 studies are: Study 7, Study 11, Study 18, Study 25, Study 31, Study 34, Study 35, Study 37, Study 41 and Study 42.

Table 2
Shows how the Meta Analysis mean (standard error) changes when we sequentially exclude invalid studies (column 2–3), adjust the selection criteria (column 4), include a valid study that was incorrectly excluded (column 5) and use a less restrictive definition of Professional Development (columns 6–8).

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	0.092 (0.017)	0.054 (0.034)	0.016 (0.049)	0.004 (0.038)	−0.008 (0.033)	0.016 (0.032)	0.004 (0.029)	0.013 (0.036)
Invalid Studies I		X	X	X	X	X	X	X
Invalid Studies II			X	X	X	X	X	X
Selection Criteria				X	X	X	X	
Include Valid Study Excluded					X		X	X
Less Restrictive Definition of PD						X	X	X
Number of studies (estimates)	42 (49)	21 (25)	10 (12)	10 (14)	11 (16)	13 (17)	14 (19)	14 (16)

Note: invalid studies I (row 1) exclude all the studies that are invalid estimates of the meta-analysis hypothesis for the following reasons: they do not have a control group, do not provide evidence of balance between control group or treatment group, contain statistically significant differences on key covariates, do not investigate the effect of PD or have issues around data quality. Invalid studies II (row 2) exclude studies that are not designed to estimate the causal effect of PD due to over-identification. Selection criteria (row 3) refers to adjusting the authors original selection criteria (the measure presented first) to a less biases selection criteria. Include valid study excluded (row 4) includes a valid study of the meta-analysis hypothesis that is incorrectly excluded. Less restrictive definition of PD (row 5) includes studies whose PD intervention also included the provision of additional resources and teaching materials.

Table 3
Summary of studies excluded from the meta-analysis.

Number of studies excluded	Studies	Reason for Exclusion
1	Study 3	Do not have a control group
1	Study 1	Do not provide evidence of balanced treatment and control groups
12	Study 2, Study 5, Study 8, Study 13, Study 15, Study 23, Study 24, Study 26, Study 27, Study 38, Study 40, Study 32	Statistically significant differences on key covariates
4	Study 6, Study 9, Study 17, Study 33	Studies do not investigate the effect of PD
11	Study 4, Study 10, Study 16, Study 19, Study 20, Study 21, Study 28, Study 29, Study 30, Study 36, Study 39	Not designed to estimate the causal effect of PD (overidentification)
3	Study 12, Study 14, Study 22	Non-random attrition and data quality

introduces upward bias into the meta-analysis. To demonstrate this, if we use authors original selection criteria the mean effect size from the 10 valid studies increases from 0.004 (Table 2, Column 4) to 0.016 (Column 3).

So far, we have only considered the papers that the authors used in their meta-analysis. However, there are four studies that Fletcher-Wood and Zuccollo (2020) include in the evidence review but are not used to calculate the meta-analysis mean. We agree with the decision to remove three of the studies – for instance, Sailors and Price (2015) is rightly excluded because they do not have a control group. However, they also exclude Rimm-Kaufman et al. (2014) from calculating the meta-analysis mean. It is not clear why this paper is excluded as it is a valid study of the meta-analysis hypothesis. When the two estimates from this study are included the mean effect size falls from 0.004 (Table 2 column 4) to −0.008 (Column 5). This demonstrates that the authors decisions about which studies to include, or not, and which estimates to use in these studies introduces significant upward bias into the meta-analysis estimate. This will be discussed further in section 5.

3.2. Definition of professional development

In the replication we remove 11 studies because they are not designed to estimate the causal effect of PD – the treatment group(s) are exposed to multiple interventions, so they are unable to causally identify the effect of PD. We believe this is appropriate because a study that investigates the effect of PD and a curriculum change, for instance, isn't informative about the effect of PD as it is not clear how much of the effect, if any, is driven by PD compared to the curriculum change.

Nevertheless, it is plausible that PD often works in tandem with other light-touch interventions, such as the provision of additional resources and teaching materials. If we relax our definition of PD and include studies where the intervention also provides teachers with additional resources and teaching materials (Study 16, Study 21 and Study 28) we find that the mean effect size increases to 0.016 (Table 2 column 6).[8] This estimate does need to be treated with caution. While it is possible that the combination of PD and additional teaching resources might make PD more productive it is also possible that the effect is entirely driven by the provision of additional resources.

4. Discussion

In this paper we demonstrate that the positive effect found by Fletcher-Wood and Zuccollo (2020) is entirely driven by the inclusion of invalid studies and the authors selection criteria. First, of the 49 estimates (from 42 studies) included only 14 (from 10 studies) are valid estimates of the meta-analysis hypothesis. Second, the authors selection criteria "when multiple outcome measure were supplied and all seemed to reflect similar dimensions, we chose the first listed" introduces upward bias into the meta-analysis as the positive, statistically significant effects, are more likely to be reported first. When the invalid studies are removed, and the bias in the selection criteria adjusted, the effect size falls from 0.09 to 0.004. Moreover, when we include valid estimates of the meta-analysis hypothesis that the authors excluded without justification the effect size falls to −0.008. This provides reason to doubt the claim that PD improves student learning, based on the evidence provided in this meta-analysis. In addition, the claim that an increase in PD will increase each student's lifetime earnings by over £6000 and a net societal benefit of over £61bn over ten years – calculated using this meta-analysis mean – is unlikely to be true.

Reviewing the studies included in this meta-analysis suggest that a general improvement in experimental and empirical methods would help researchers design more robust experiments. First, almost 1 in 4 of the studies reviewed have statistically significant differences on key covariates. This suggests poor experimental design. In a large experiment, where data on schools, teachers and students is collected before the intervention(s) takes place, researchers should be able to check if there are any differences between groups and rerandomize if necessary (Bruhn & McKenzie, 2009). Second, other forms of bias could influence the estimates if the experiments are not carefully designed. For instance, experimenter demand effects (EDE) and/or efficiency wage effects might bias the estimates in this context and, without a carefully designed experiment (such as including a placebo treatment arm), it is

[8] The mean effect size of these three new studies included are 0.07.

challenging to identify them. For instance, the teachers know they are taking part in a study, and this might change their behaviour. This is a problem because EDE are likely to be stronger in the treatment group (those who are more actively involved in the study) than the control group (business as usual). In this setting EDE are likely to exist and produce upward bias. It is also possible that teachers who are offered more PD might feel more valued and therefore work harder or more productively in some way that improves pupil outcomes. As Britton and Propper (2016) and Fullard (2021) provide evidence of efficiency wage effects in teaching, and the effect is likely to be positively correlated with teachers' enjoyment of the intervention(s), this is likely to also produce upward bias. Without a carefully designed experiment it is challenging to identify these sources of bias. Third, a researcher completing a meta-analysis of RCT's should have the necessary training to be able to identify if the RCT was carried out successfully or not. For instance, studies that do not have a control group should easily be identified as invalid. An improvement in experimental and empirical methods in educational research would help researchers design more robust experiments in the future and more effectively evaluate the quality of existing studies.

So far, we have assumed that the choices that the authors made were well intended but misguided. However, the structure of their research might have affected their choices, consciously or otherwise. Recall that the results from the meta-analysis were used in a larger project evaluating the costs and benefits of increasing teacher PD. It is plausible that if the researchers did not find a positive effect in the meta-analysis, then the funders might not have proceeded to the second (much larger) stage of the project. Afterall, there is not much point doing a cost benefit analysis when the meta-analysis finds no meaningful benefit. In addition, a larger effect is more likely to engage audiences and gain policy traction than a smaller effect – or no effect at all. Combine these incentives with a lack of substantive peer review, or research transparency, and this might explain why studies that found large positive effects, such as Study 1 and Study 3, were included in the meta-analysis despite clear flaws (no control group and no evidence of balance respectively) while empirically strong papers which found negative effects were excluded without justification (such as Rimm-Kaufman et al. (2014)).

In our paper we replicate Fletcher-Wood and Zuccollo (2020) and evaluate the impact of PD on teacher effectiveness by measuring the impact on student outcomes. One limitation of this paper is that we only assess the validity of the study designs (are they investigating the causal effect of PD on pupil outcomes) not the comparability of the studies. The studies in the meta-analysis use a range of different interventions (with different intensities) where the effect on pupil outcomes is measured on different subject areas (i.e., Math, Science and English) in different ways (i.e., standardised tests, internal schools' assessments and research developed tests) - see Fletcher-Wood and Zuccollo (2020) for a detailed discussion. This is potentially problematic. For instance, it is not immediately obvious that a study measuring a weekly intervention on students Maths scores on a standardized test is directly comparable to a termly intervention on students English scores on a school quiz. This is an important distinction in this setting because teacher effectiveness can be influenced by a range of different factors and can be measured in several different ways – some more reliable than others (Brooks & Springer, 2022; Dolton & Marcenaro-Gutierrez, 2011; Hanushek & Rivkin, 2006). Evaluating the comparability of the PD interventions, how the effect on pupil attainment is measured, and how these influence the meta-analysis conclusions seem like a promising area of future research.

A second limitation of this paper is that it is unclear if pupil outcomes should be the primary outcome of interest – it might be more appropriate to first look at the impact of PD on teacher behaviour, and then investigate if these changes transfer to student attainment. This is especially important in this setting as the effect sizes are generally small and the measures of student attainment can be very noisy. Evaluating how PD impacts other outcomes such as teacher/student wellbeing seem like promising areas of future research.

In this paper we found evidence to doubt the doubt the claim that PD improves student learning, based on the evidence provided in Fletcher-Wood and Zuccollo (2020) and the subsequent policy conclusions. Given the large role that policy organisations have on informing the public debate and shaping the policy agenda the research community needs to have a serious debate about what measures can be put in place to maintain research standards in these organisations. A requirement for policy organisations to make their data and code publicly available for replication purposes as well as having a substantive independent peer review system in place to check research, before it influences the policy debate, seem like a good place to start.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: I would like to disclose a potential competing interest. Between September 2019 and June 2021, I worked for the Education Policy Institute, one of the two organisations who performed the meta-analysis I am reviewing. I was not involved in the meta-analysis that I am reviewing in any way.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics, 25*(1), 95–135.

Allen, J. P., Hafen, C. A., Gregory, A. C., Mikami, A. Y., & Pianta, R. (2015). Enhancing secondary school instruction and student achievement: Replication and extension of the My Teaching Partner-Secondary intervention. *Journal of Research on Educational Effectiveness, 8*(4), 475–489.

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science, 333*(6045), 1034–1037.

Almaatouq, A., Griffiths, T. L., Suchow, J. W., Whiting, M. E., Evans, J., & Watts, D. J. (2022). *Beyond playing 20 questions with nature: Integrative experiment design in the social and behavioral sciences* (pp. 1–55). Behavioral and Brain Sciences.

Anders, J., Stokes, L., Hudson-Sharp, N., Dorsett, R., Rolfe, H., George, A., Buzzeo, J., & Munro-Lott, N. (2018). *Mathematical Reasoning: Evaluation report and executive summary*.

Angrist, J. D., & Lavy, V. (2001). Does teacher training affect pupil learning? Evidence from matched comparisons in jerusalem public schools. *Journal of Labor Economics, 19*(2), 343–369.

Begg, C. B. (1994). Publication bias. *The handbook of research synthesis, 25*, 299–409.

Biggart, A. (2015). *Quest: Evaluation report and executive summary*.

Birch, P., Balcon, M.-P., Bourgeois, A., Davydovskaia, O., & Tremosa, S. P. (2018). *Teaching careers in Europe: Access, progression and support. Eurydice report*. European Commission: Education, Audiovisual and Culture Executive Agency.

Borman, G. D., Gamoran, A., & Bowdon, J. (2008). A randomized trial of teacher development in elementary science: First-year achievement effects. *Journal of Research on Educational Effectiveness, 1*(4), 237–264.

Boylan, M., Demack, S., Wolstenholme, C., Reidy, J., & Reaney, S. (2018). *ScratchMaths: Evaluation report and executive summary*.

Britton, J., & Propper, C. (2016). Teacher pay and school productivity: Exploiting wage regulation. *Journal of Public Economics, 133*, 75–89.

Brooks, C. D., & Springer, M. G. (2022). *Evaluating teacher effectiveness: A review of historical developments and current trends* (pp. 127–149). The Routledge handbook of the economics of education.

Bruhn, M., & McKenzie, D. (2009). In pursuit of balance: Randomization in practice in development field experiments. *American Economic Journal: Applied Economics, 1*(4), 200–232.

Buddin, R., & Zamarro, G. (2009). Teacher qualifications and student achievement in urban elementary schools. *Journal of Urban Economics, 66*(2), 103–115.

Caena, F. (2011). *Literature review teachers' core competences: Requirements and development*. European Commission Thematic Working Group 'Professional Development of Teachers.

Campbell, P. F., & Malkus, N. N. (2011). The impact of elementary mathematics coaches on student achievement. *The Elementary School Journal, 111*(3), 430–454.

Chingos, M. M., & Peterson, P. E. (2011). It's easier to pick a good teacher than to train one: Familiar and new results on the correlates of teacher effectiveness. *Economics of Education Review, 30*(3), 449–465.

Cotabish, A., Dailey, D., Robinson, A., & Hughes, G. (2013). The effects of a STEM intervention on elementary students' science knowledge and skills. *School Science & Mathematics, 113*(5), 215–226.

Cutler, D. M., & Lleras-Muney, A. (2006). *Education and health: Evaluating theories and evidence*. USA: National bureau of economic research Cambridge, Mass.

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher, 38*(3), 181–199.
DfE. (2020). *National professional qualification (NPQ): Leading teacher development Framework*.
DfE, A. (2022). *Issues paper: Teacher workforce shortages*.
Dolton, P., & Marcenaro-Gutierrez, O. D. (2011). If you pay peanuts do you get monkeys? A cross-country analysis of teacher pay and pupil performance. *Economic Policy, 26*(65), 5–55.
Education, I.f. E. (2016). *Teacher effectiveness enhancement programme evaluation report and executive summary*.
Fisher, D., Frey, N., & Lapp, D. (2011). Coaching middle-level teachers to think aloud improves comprehension instruction and student reading achievement. *The Teacher Educator, 46*(3), 231–243.
Fletcher-Wood, H., & Zuccollo, J. (2020). *The effects of high-quality professional development on teachers and students: A rapid review and meta-analysis*. Education Policy Institute.
Fullard, J. (2021). Relative wages and pupil performance, evidence from TIMSS. In *ISER Working Paper Series, 2021*.
Fullard, J. (2022). *Teacher diversity in England 2010-2021*.
Garet, M. S., Cronen, S., Eaton, M., Kurki, A., Ludwig, M., Jones, W., Uekawa, K., Falk, A., Bloom, H. S., & Doolittle, F. (2008). *The impact of two professional development interventions on early reading instruction and achievement. NCEE 2008-4030*. National Center for Education Evaluation and Regional Assistance.
Garet, M. S., Wayne, A. J., Stancavage, F., Taylor, J., Eaton, M., Walters, K., Song, M., Brown, S., Hurlburt, S., & Zhu, P. (2011). *Middle school mathematics professional development impact study: Findings after the second year of implementation. NCEE 2011-4024*. National Center for Education Evaluation and Regional Assistance.
Gersten, R., Dimino, J., Jayanthi, M., Kim, J. S., & Santoro, L. E. (2010). Teacher study group: Impact of the professional development model on reading instruction and student outcomes in first grade classrooms. *American Educational Research Journal, 47*(3), 694–739.
Glazerman, S., Isenberg, E., Dolfin, S., Bleeker, M., Johnson, A., Grider, M., & Jacobus, M. (2010). *Impacts of comprehensive teacher induction: Final results from a randomized controlled study. NCEE 2010-4027*. National Center for Education Evaluation and Regional Assistance.
Greenleaf, C. L., Litman, C., Hanson, T. L., Rosen, R., Boscardin, C. K., Herman, J., Schneider, S. A., Madden, S., & Jones, B. (2011). Integrating literacy and science in biology: Teaching and learning impacts of reading apprenticeship professional development. *American Educational Research Journal, 48*(3), 647–717.
Groot, W., & van den Brink, H. M. (2010). The effects of education on crime. *Applied Economics, 42*(3), 279–289.
Hanley, P., Bohnke, J., Slavin, B., Elliott, L., & Croudace, T. (2016). *Let's think secondary science: Evaluation report and executive summary*.
Hanley, P., Slavin, R., & Elliott, L. (2015). *Thinking, doing, talking science: Evaluation report and executive summary*. Education Endowment Foundation.
Hanushek, E. A., & Rivkin, S. G. (2006). Teacher quality. *Handbook of the Economics of Education, 2*, 1051–1078.
Hanushek, E. A., Ruhose, J., & Woessmann, L. (2016). It pays to improve school quality: States that boost student achievement could reap large economic gains. *Education, 16*(3), 52–61.
Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics, 95*(7–8), 798–812.
Heller, J. I., Curtis, D. A., Rabe-Hesketh, S., & Verboncoeur, C. J. (2007). *The effects of" Math pathways and pitfalls" on students' mathematics achievement*. Online Submission: National Science Foundation Final Report.
Humphrey, N., Hennessey, A., Ashworth, E., Frearson, K., Black, L., Petersen, K., Wo, L., Panayiotou, M., Lendrum, A., & Wigelsworth, M. (2018). *Good behaviour game: Evaluation report and executive summary*. Education Endowment Foundation.
Jackson, C. K., Rockoff, J. E., & Staiger, D. O. (2014). Teacher effects and teacher-related policies. *Annual Review of Economics, 6*(1), 801–825.
Jacobs, V. R., Franke, M. L., Carpenter, T. P., Levi, L., & Battey, D. (2007). Professional development focused on children's algebraic reasoning in elementary school. *Journal for Research in Mathematics Education, 38*(3), 258–288.
Jay, T., Willis, B., Thomas, P., Taylor, R., Moore, N., Burnett, C., Merchant, G., & Stevens, A. (2017). *Dialogic teaching: Evaluation report and executive summary*.
Jerrim, J., Austerberry, H., Crisan, C., Ingold, A., Morgan, C., Pratt, D., Smith, C., & Wiggins, M. (2015). *Mathematics mastery: Secondary evaluation report*. Education Endowment Foundation.
Kitmitto, S., González, R., Mezzanote, J., & Chen, Y. (2018). Thinking, doing. *Talking Science*. https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/thinking-doing-talking-science-effectiveness-trial.
Koedel, C., Parsons, E., Podgursky, M., & Ehlert, M. (2015). Teacher preparation programs and teacher quality: Are there real differences across programs? *Education Finance and Policy, 10*(4), 508–534.
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.
Kulik, J. A., & Kulik, C.-L. C. (1989). Meta-analysis in education. *International Journal of Educational Research, 13*(3), 221–340.
Long, R., & Danechi, S. (2021). *Teacher recruitment and retention in England* (p. 7222). House of Commons Briefing Paper Number.
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.
Marshall, J. (2016). Education and voting conservative: Evidence from a major schooling reform in great britain. *The Journal of Politics, 78*(2), 382–395.
Matsumura, L. C., Garnier, H. E., & Spybrook, J. (2013). Literacy coaching to improve student reading achievement: A multi-level mediation model. *Learning and Instruction, 25*, 35–48.
McNally, S. (2014). *Hampshire hundreds: Evaluation report and executive summary*. Education Endowment Foundation.
Miller, S., Biggart, A., Sloan, S., & O'Hare, L. (2017). *Success for all: Evaluation report and executive summary*. Education Endowment Foundation.
Motteram, G., Choudry, S., Kalambouka, A., Hutcheson, G., & Barton, H. (2016). *ReflectED: Evaluation report and executive summary*. Education Endowment Foundation.
Murphy, R., Weinhardt, F., Wyness, G., & Rolfe, H. (2017). *Lesson study: Evaluation report and executive summary*. Education Endowment Foundation.
Nguyen, T. D., Lam, C. B., & Bruno, P. (2022). Is there a national teacher shortage? A systematic examination of reports of teacher shortages in the United States. In *EdWorkingPaper*.
Ofsted. (2023). *Independent review of teachers' professional development in schools: Phase 1 findings*.
Papay, J. P., Taylor, E. S., Tyler, J. H., & Laski, M. E. (2020). Learning job skills from colleagues at work: Evidence from a field experiment using teacher performance data. *American Economic Journal: Economic Policy, 12*(1), 359–388.
Parkinson, J., Salinger, T., Meakin, J., & Smith, D. (2015). *Results from a three-year i3 impact evaluation of the Children's Literacy Initiative (CLI): Implementation and impact findings of an intensive professional development and coaching program*. Washington, DC: American Institutes for Research.
Penuel, W. R., Gallagher, L. P., & Moorthy, S. (2011). Preparing teachers to design sequences of instruction in earth systems science: A comparison of three professional development programs. *American Educational Research Journal, 48*(4), 996–1025.
Podolsky, A., Kini, T., & Darling-Hammond, L. (2019). *Does teaching experience increase teacher effectiveness? A review of us research*. Journal of Professional Capital and Community.
Rimm-Kaufman, S. E., Larsen, R. A., Baroody, A. E., Curby, T. W., Ko, M., Thomas, J. B., Merritt, E. G., Abry, T., & DeCoster, J. (2014). Efficacy of the responsive classroom approach: Results from a 3-year, longitudinal randomized controlled trial. *American Educational Research Journal, 51*(3), 567–603.
Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *The American Economic Review, 94*(2), 247–252.
Rose, J., Thomas, S., Zhang, L., Edwards, A., Augero, A., & Roney, P. (2017). *Research learning communities: Evaluation report and executive summary*. Education Endowment Foundation.
Sailors, M., & Price, L. R. (2010). Professional development that supports the teaching of cognitive reading strategy instruction. *The Elementary School Journal, 110*(3), 301–322.
Sailors, M., & Price, L. (2015). Support for the improvement of practices through intensive coaching (SIPIC): A model of coaching for improving reading instruction and reading achievement. *Teaching and Teacher Education, 45*, 115–127.
Santagata, R., Kersting, N., Givvin, K. B., & Stigler, J. W. (2010). Problem implementation as a lever for change: An experimental study of the effects of a professional development program on students' mathematics learning. *Journal of Research on Educational Effectiveness, 4*(1), 1–24.
Sharpe, D. (1997). Of apples and oranges, file drawers and garbage: Why validity issues in meta-analysis will not go away. *Clinical Psychology Review, 17*(8), 881–901.
Simonsohn, U., Simmons, J., & Nelson, L. D. (2022). Above averaging in literature reviews. *Nature Reviews Psychology, 1*(10), 551–552.
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.
Slavin, R. E. (1984). Meta-analysis in education: How has it been used? *Educational Researcher, 13*(8), 6–15.
Sloan, S., Gildea, A., Miller, S., & Thurston, A. (2018). *Zippy's Friends Evaluation report and executive summary*.
Speckesser, S., Runge, J., Foliano, F., Bursnall, M., Hudson-Sharp, N., Rolfe, H., & Anders, J. (2018). *Embedding formative assessment: Evaluation report and executive summary*.
Stanley, T. D. (2005). Beyond publication bias. *Journal of Economic Surveys, 19*(3), 309–345.
Thurston, A., Roseth, C., O'Hare, L., Davison, J., & Stark, P. (2016). *Talk of the Town. Evaluation report and executive summary*. Education Endowment Foundation.
Tracey, L., Boehnke, J., Elliott, L., Thorley, K., Bowyer-Crane, C., & Ellison, S. (2019). *Grammar for Writing: Evaluation report and executive summary*.
Van den Brande, J., & Zuccollo, J. (2021). *The effects of high-quality professional development on teachers and students: A cost-benefit analysis*.
Vignoles, A., Jerrim, J., & Cowan, R. (2015). *Mathematics mastery: Primary evaluation report*. February 2015.
Wiggins, M., Jerrim, J., Tripney, J., Khatwa, M., & Gough, D. (2019). *The RISE project: Evidence-informed school improvement*. Education Endowment Foundation.
Wiggins, M., Parrao, C. G., Austerberry, H., & Ingold, A. (2017). *Foreign Language learning in primary school: Evaluation report and executive summary*. Education Endowment Foundation.

Wiswall, M. (2013). The dynamics of teacher quality. *Journal of Public Economics, 100*, 61–78.
Worth, J., Sizmur, J., Walker, M., Bradshaw, S., & Styles, B. (2017). *Teacher observation: Evaluation report and executive summary*. Education Endowment Foundation.
Yoon, K. S., Duncan, T., Lee, S. W.-Y., Scarloss, B., & Shapley, K. L. (2007). *Reviewing the evidence on how teacher professional development affects student achievement. issues & answers. rel 2007-no. 033*. Regional Educational Laboratory Southwest (NJ1).