University of Zurich

Department of Economics

Working Paper Series

ISSN 1664-7041 (print)
ISSN 1664-705X (online)

Working Paper No. 438

Same-Sex Teacher Effects

Alexandra de Gendre, Jan Feld, Nicolás Salamanca and Ulf Zölitz

Revised version, May 2024

Same-Sex Teacher Effects*

Alexandra de Gendre
(University of Melbourne)

Jan Feld
(Victoria University of Wellington)

Nicolás Salamanca
(University of Melbourne)

Ulf Zölitz
(University of Zürich)

May 2024

Abstract

The idea that students benefit from same-sex teachers has motivated policies around the world aimed at reducing gender inequalities. However, we do not know the size or generalizability of such same-sex teacher effects. We fill this gap by conducting a meta-analysis and our own study using data from 90 countries. Our meta-analysis summarizes the literature, highlights that estimates are difficult to compare because of differences in methods, and shows evidence of publication bias. Our multi-country study overcomes these shortcomings by providing many comparable estimates which are free of publication bias. Those estimates are ideal for learning about the generalizability of an effect. Our results reveal an interesting pattern. In primary education, effects vary substantially by country and outcome. In secondary education, same-sex teacher effects are near-universally positive for all countries and outcomes. Our paper showcases how we can use multi-country studies to learn about the generalizability of an effect.

Keywords: Same-sex teachers, role models, STEM, teachers, external validity, multi-country study, gender role models, standardized test scores, grades, job preferences, meta-analysis, meta-science.

JEL classification: I21, I24, J24

* de Gendre: Department of Economics, The University of Melbourne, LCC and IZA, a.degendre@unimelb.edu.au; Feld: School of Economics and Finance, Victoria University of Wellington and IZA, jan.feld@vuw.ac.nz; Salamanca: Melbourne Institute: Applied Economics & Social Research, The University of Melbourne, LCC and IZA, n.salamanca@unimelb.edu.au; Zölitz: University of Zürich, Department of Economics and Jacobs Center for Productive Youth and Child Development, CESifo, CEPR, IZA, ulf.zoelitz@econ.uzh.ch. We gratefully acknowledge financial support from the University of Zurich URPP Equality of Opportunity. This research was supported partially by the Australian government through the Australian Research Council's Centre of Excellence for Children and Families over the Life Course (Project ID CE200100025). Elisa Alonso, Ana Bras, Matthew Bonci, Timo Haller, Andrea Hofer, Francesco Serra, Madeleine Smith, Albert Thieme and Anna Valyogos provided outstanding research assistance. We received valuable comments from Luke Chu, Harold Cuffe, Thomas Dee, Rajeev Dehejia, David Dorn, Chris Doucouliagos, Arthur Grimes, Nick Huntington-Klein, Kirabo Jackson, Nathan Kettlewell, Christine Mulhern, Martin Neugebauer, Bob Reed, Julia Rohrer, Roberto Weber, and seminar participants at Bocconi, the CESifo education meeting, CREST, University of New South Wales, University of Technology Sydney, the LEER conference KU Leuven, Royal Holloway, the University of Canterbury, the University of Melbourne, the University of St Gallen, the University of Western Australia, the University of Zürich, and Victoria University of Wellington. Interactive, country-specific results of this study are available at www.role-model-effects.com.

1. Introduction

Do students benefit from having same-sex teachers? The answer to this question matters for educational equity, equality of opportunity, and affirmative action policies. For example, if female STEM teachers are particularly good at teaching girls, hiring more of them could reduce sex-gaps in STEM performance. If male teachers are particularly good at teaching boys, hiring more of them could stop boys' performance decline in primary school.[1] Knowing about the size and generalizability of such same-sex teacher effects is vital for anyone interested in introducing impactful polices. If effects are small, new policies might not be cost effective. If effects are very heterogeneous, studies in one context might not be informative about effects in another context (List, 2020). In this paper we investigate the size and generalizability of same-sex teacher effects in education.

In the first part of our paper, we summarize the existing literature with a meta-analysis of same-sex-teacher effects on student performance in primary and secondary education. We identify 538 estimates from 24 studies and find a small average same-sex teacher effect of 0.030 standard deviations (SD) on grades and test scores. Although our meta-analysis provides a useful summary of the literature, it has two important shortcomings. First, the sign of the estimated average same-sex teacher effect is sensitive to how we correct for publication bias: some correction methods show small positive effects and others show small negative effects. Second, we cannot convincingly investigate the generalizability of effects because of differences in methodologies across studies. For example, no two studies use the same empirical strategy, econometric specification, or sample selection criteria. Recent studies have shown that such decisions can have large effects on estimates (e.g., Huntington-Klein et al., 2021; Breznau et al., 2022). It is therefore difficult to judge to what extent differences in same-sex teacher effect estimates reflect differences in empirical approaches or true heterogeneity.

[1] The idea that students benefit from having same-sex teachers has inspired many calls for policy interventions and recruitment initiatives around the world. For example, the OECD and World Bank have called for policies to attract more female STEM teachers to increase female representation in STEM studies and jobs (OECD 2012; World Bank 2020). UNICEF has identified the lack of female role models as a key contributor to girls' underperformance in STEM subjects (UNICEF 2020). Switzerland has launched the campaign "Men for primary schools" that aims to increase number of male primary school teachers with newspaper advertisements and taster days (Meyer, 2017). Norway has experimented with gender quotas for admission to teacher training programs to ensure a minimum percentage of male candidates. (Schaede and Mankki, 2022). Australia's "Males in Primary" initiative promotes teaching as a career choice for men. This includes scholarships specifically for male primary education students, mentorship programs, and promotional campaigns showcasing male teachers in primary schools. The US "Troops to Teachers" program encourages military veterans to retrain as teachers, with a particular focus on addressing the gender imbalance in primary schools (Nunnery et al., 2009). In England, the Training and Development Agency for Schools (TDA) undertook marketing campaigns to attract more men into teaching, highlighting the positive impact male teachers can have on students and the rewarding nature of the profession (Szwed, 2010).

In the second part of our paper, we overcome both shortcomings with our own study using data from 90 countries. We do not have to worry about publication bias because none of our estimates have been filtered through the publication process. We can use a consistent methodology to estimate effects for many, diverse countries. This approach allows us to rule out that differences in estimates are due to different empirical approaches.

To estimate same-sex teacher effects, we build a large-scale multi-country dataset. We combine science and math test scores for 4th and 8th grade students from the Trends in International Mathematics and Science Study (TIMSS) with literacy test scores of 4th grade students from the Progress in International Reading Literacy Study (PIRLS). Test scores in both studies are designed to be comparable between countries. We also have access to measures of students' job preferences, subject enjoyment, and subject confidence, allowing us to study effects beyond test scores. Our resulting dataset contains 3,047,752 children taught by 231,942 teachers in 105,916 primary and secondary schools across six continents. Our sample size is over 90 times the sample size of the median study in our meta-analysis.

To identify the causal same-sex teacher effects, we estimate a complementary set of fixed effects models that differ in their source of identifying variation and their key identifying assumptions. We start with a country fixed effects model, which serves as our baseline estimate with minimal controls. Beyond this base specification, we estimate effects with four sets of fixed effects: (1) school fixed effects, (2) classroom fixed effects, (3) student fixed effects, and (4) student and teacher fixed effects. The gradual inclusion of more-restrictive fixed effects makes concerns about omitted variables increasingly implausible. In our most restrictive specification, we exploit that the same student has a female math teacher and a male science teacher (or vice versa) while additionally holding constant unobserved fixed teacher characteristics across students. All our fixed effects specifications deliver virtually identical results. From the least to the most restrictive specification, the point estimates hardly change, while the R^2 increases from 0.38 to 0.96. The consistency of our estimates together with the stark increase in R^2 show that omitted variables bias is unlikely to drive our results.[2]

The results of our multi-country study show very small average same-sex teacher effects of 0.015 SD on test scores. Across all specifications, the 99% confidence intervals allow us to rule out effects smaller than 0.009 and larger than 0.022 SD. We see no sizable heterogeneity across subject matter, student characteristics, or teacher characteristics. Along all these dimensions, estimated effects are always positive but never exceed 0.019 SD. We see

[2] When restricting our sample to countries with institutional random assignment of students to classrooms, we find very similar results for all our outcomes of interest. These results are further evidence that omitted variables bias does not threaten the validity of our identification strategy.

larger average same-sex teacher effects on job preferences (0.064 SD), subject confidence (0.050 SD), and subject enjoyment (0.089 SD).

We probe the generalizability of same-sex teacher effects by investigating whether they are positive in most settings. If they are not, we explore in which kinds of settings effects may be generally positive. In the language of philosophy of science, we study whether positive same-sex teacher effects are a phenomenon (a stable and general feature of the world) and what the boundary conditions of this phenomenon are (in which kinds of contexts it holds) (Bogen and Woodward, 1998; Eronen and Bringmann, 2021; Busse et al., 2017).[3]

We address those questions in two steps. In our first step, we estimate same-sex teacher effects for many diverse contexts. We operationalize a "context" as a country-education level-outcome combination and estimate effects for all possible contexts. Differences in estimates between contexts also reflect sampling error. Even if the true same-sex teacher effects would be identical across all contexts, we would expect differences in estimates by chance alone. In our second step, we therefore use meta-analysis methods to account for sampling error and estimate country-level distributions of the true same-sex teacher effects for each combination of outcome and education level (e.g., the distribution of effects on test scores in primary schools across all countries).

Our results reveal an interesting difference between primary and secondary education. In primary education, same-sex teacher effects are *not* generally positive. For example, our results suggest that effects on test scores are positive in only half of the countries in our analysis (and negative in the other half). This result shows that hiring more primary school teachers is not a reliable policy for stopping boys' performance decline. For subject enjoyment, same-sex teacher effects are positive in 96% of countries, whereas for subject confidence, effects are positive in only 69% of countries.

In secondary education, same-sex teacher effects are near universally positive for all outcomes and countries. For test scores, the true same-sex teacher effects are tiny and effectively of the same magnitude for all countries. For non-test score outcomes, same-sex teacher effects are positive in more than 95% of the countries. For these outcomes we see more variation between countries. For example, same-sex teacher effects on job preferences are larger in rich and gender-equal countries. These larger effects are consistent with previous work

[3] An example of a phenomenon is "human males are taller than human females." This phenomenon refers to *average* sex differences in height (the distributions of male and female height overlap) (Roser et al., 2021). It is also a statement that holds for most groups but not for all groups. Finding individual exceptions does not "disprove" the phenomenon. One boundary condition of this phenomenon specifies the age range in which this phenomenon holds. For example, it holds for adults but *not* for children of all ages because girls aged 10–13 are generally taller than boys of the same age (girls' growth spurts start earlier) (Britannica, 2024).

highlighting that female role models can inspire students to follow in their footsteps in France and the United States (Breda et al., 2020; Mansour et al., 2022; Carrell et al., 2010; Kofoed and McGovney, 2019; Bettinger and Long, 2005). However, our results also suggest that such effects would be weaker in poorer and less gender-equal countries.

We contribute to the literature in three ways. First, our meta-analysis provides a comprehensive summary of the literature on same-sex teacher effects on performance. This is particularly important for a literature that has inspired many gender-targeted recruitment campaigns and calls for gender quotas (see footnote 1). Without this summary, researchers and policy makers risk being swayed by individual studies that are widely publicized because they happen to find large effects. Our meta-analysis provides convincing evidence that effects on performance are, on average, very small. Our meta-analysis also reveals that—based on the existing literature—the average same-sex teacher effect is so small that most existing studies would not have been able to reliably detect it. Second, our multi-country study vastly expands the scope of the literature on same-sex teacher effects. We produce well-identified estimates for 90 countries including 55 countries in which these effects have not yet been studied. We go beyond student performance and study same-sex teacher effects on students' job preferences, subject enjoyment, and subject confidence—all of which are outcomes that policy makers may find important on their own. Third, we explore the generalizability of same-sex teacher effects. To do this, we use meta-analytical methods to explicitly account for sampling error and model the distribution of true same-sex teacher effects across settings. Overall, we provide the most extensive evidence on same-sex teacher effects to date.

Our paper relates to several studies that have investigated generalizability with either multi-context studies or meta-analyses (e.g., Aaronson et al., 2021; Altmejd et al., 2021; Dudek et al., 2020; Falk et al., 2018; Meager, 2019; Vivalt, 2020; Wößmann and West, 2006). Both approaches have shortcomings. The former suffers from sampling error while the latter suffers from publication bias and differences in methods between studies. We overcome these shortcomings by analyzing estimates from a multi-context study with meta-analysis methods, combining the advantages of both approaches. Our approach is a particularly useful tool for mature literatures that have not managed to converge, of which the same-sex teacher effects literature is a prime example. Even after 24 studies and our meta-analysis, we still do not know the sign of the average effect or how much effects vary by context. Such literatures often remain in purgatory, where results are simply described as mixed and inconsistencies are either ignored or handwaivingly attributed to differences in settings. Large-scale, multi-setting studies along with meta-analysis methods provide a way out of this dilemma.

Our paper also relates to an emerging literature on external validity. This literature grapples with the fact that all estimates come from a specific setting and investigate when and how such "local" estimates allow us to learn something about an effect in a different context. For example, Angrist and Fernandez-Val (2013) and Bisbee et al. (2017) show how to reweight instrumental variable estimates from one context to predict a treatment effect in another context (see also Dehejia et al., (2021) for a similar approach for natural experiments). List (2020) examines why interventions that appear to be effective in one context fail to be effective when implemented at scale and identifies five reasons: non-representative samples, false positives, unscalable ingredients, cost traps, and negative spillovers. Andrews and Oster (2019) propose a way to bound external validity bias driven by non-representative experimental samples. All these studies are interested in how evidence from one "reference setting" translates into effects in one "target setting." In this paper, by contrast, we are interested in the distribution of the true effects and whether effects are generally positive.

2. A Meta-Analysis on Same-Sex Teacher Effects

2.1 What Are Same-Sex Teacher Effects?

We follow the existing literature and define the same-sex teacher effect as the premium of having a same-sex teacher—on top of the general effect of having a female or male teacher (e.g., Hoffmann and Oreopoulos, 2009; Lim and Meer, 2017). Such same-sex teacher effects are typically estimated with variations of the following regression model:

$$Outcome = \beta_0 + \beta_1 Female\ Student + \beta_2 Female\ Teacher + \beta_3 Female\ Student \times Female\ Teacher + u. \quad (1)$$

In this model, β_3 captures the same-sex teacher effect. A positive same-sex teacher effect could be driven by female students benefitting more from female teachers than male students as well as male students benefitting more from male teachers than female students. This effect is distinct from sex differentials in teacher effectiveness. For example, there would be no same-sex teacher effect if girls and boys benefit equally from having a female teacher. However, there would be a positive same-sex teacher effect if girls benefit more than boys from having a female teacher.

Although we follow the literature by calling β_3 a same-sex *teacher* effect, we note that this effect could also be driven by the behavior of students. For example, we could also observe same-sex teacher effects because students behave differently with teachers of their own sex.

Several studies have estimated same-sex teacher effects on career choices and performance in tertiary education. These studies usually find positive effects. For example,

Carrell et al. (2010) show positive same-sex teacher effects on the probability of students' taking math and science classes and the probability of graduating with a STEM degree. Mansour et al. (2022) follow up on these students and show positive same-sex teacher effects on the probability of obtaining a STEM master's degree and working in a STEM occupation. Porter and Serra (2020) show that exposure to female economists increases female students' probability of majoring in economics. Neumark and Gardecki (1998) find that female doctoral students with female mentors graduate faster without having worse placements. Hoffmann and Oreopoulos (2009) exploit within-student and within-instructor variation and find only small same-sex teacher effects of, at most, 0.05 SD on grades and 1.2 percentage points lower probability of dropping a class. These effects are not present for math and science instructors and disappear when the authors include student fixed effects.

In this paper, our focus is on same-sex teacher effects on student performance in primary and secondary education. We summarize the same-sex teacher effect estimates from previous studies with a meta-analysis.

2.2 A Meta-Analysis on Same-Sex Teacher Effects in Primary and Secondary Education

For our meta-analysis, we identified 24 studies on same-sex teacher effects on grades and test scores in primary and secondary education.[4] The median study investigates same-sex teacher effects with 10,196 observations from one country. From these studies we extract all 538 same-sex teacher effect estimates from the main text of the papers and their appendices. These estimates either stem from estimations of variations of Equation (1) or were obtained by combining coefficients from split sample regressions estimating the effect of having a female teacher (compared to a male teacher) separately for girls and boys (see Appendix A for more details on how we construct those estimates and their standard errors). To make estimates comparable, we ensure all estimates and standard errors are measured in standard deviations of the outcome of each study. We do this by dividing estimates and standard errors by the standard deviation of the outcome in all studies that did not report their estimates in standardized units. We describe our preregistration and data collection in greater detail in Appendix A. In this section, we focus on describing the results.

[4] These are Ammermüller and Dolton (2006), Antecol et al., (2015), Bhattacharya et al. (2022), Buddin and Zamarro (2008), Carrington et al. (2008), Coenen and Klaveren (2016), Dee (2007), Eble and Hu (2020), Escardíbul and Mora (2013), Evans (1992), Gong et al. (2018), Hermann and Diallo (2017), Holmlund and Sund (2008), Hwang and Fitzpatrick (2021), Lee et al. (2019), Lim and Meer (2017), Lim and Meer (2020), Lindahl (2007), Mulji (2016), Muralidharan and Sheth (2016), Neugebauer et al. (2011), Rakshit and Sahoo (2020), Xu and Li (2018), Xu (2020). See Table A1 for some summary statistics about these studies.

Our included estimates cover many different settings: 238 use data from Europe, 187 come from Asia, 94 come from North America, and 19 come from Africa; 153 are based on data from primary education, 375 are based on secondary education, and 10 use both; 57 estimates come from settings that use experimental methods with an explicit random manipulation of the student–teacher assignment, and the remaining 481 estimates come from settings with naturally occurring variation in classroom assignment; 37 estimates of same-sex teacher effects are on grades, and 501 are on test scores. Many of these estimates are not precise enough to reliably detect small effects. The median ex-post minimum detectable effect size (MDE) is 0.129 SD (calculated for 95% confidence and 80% power by multiplying the standard error by 2.8 (see e.g., Chabé-Ferret 2022; Ch. 7).

We summarize all 538 estimates using a three-level random effects model (O'Connell et al., 2022).[5] This model allows true same-sex teacher effects to differ by study and accounts for the dependence of estimates within each study. By fitting the distribution of the same-sex teacher effect point estimates and accounting for their uncertainty (as measured by their standard errors), this approach produces estimates of the distribution of underlying true same-sex teacher effects. We estimate the three-level random effects model via restricted maximum likelihood and apply the Hartung–Knapp adjustment. This adjustment incorporates estimate uncertainty in the estimation of the standard deviation in the distribution of same-sex teacher effects (Harrer et al., 2021, Ch. 4). Applying this procedure, we estimate the average same-sex teacher effect to be 0.030 SD with a standard error of 0.013 (p-value = 0.0194).[6]

Note the vast increase in power to detect same-sex teacher effects once we combine studies. Our combined estimates imply a minimum detectable average same-sex teacher effect of 0.036 SD, which is 3.6 times smaller than the median MDE among the estimates in our meta-analysis (0.036 SD versus 0.129 SD). Only 79 of the 538 point estimates would have had enough statistical power to detect the average same-sex teacher effect of 0.030 SD.

[5] Following the recommendation in Irsova et al. (2023), we summarize all estimates (discussed in this section) as well as a subset of potentially more reliable estimates (discussed in Section 2.4). Appendix Figure A2 shows funnel plots for all same-sex teacher effects estimates and their standard errors.

[6] One might be concerned that the estimated average same-sex teacher effect of 0.030 SD is mainly driven by the point estimates of a few studies that happen to contribute many precise estimates. To check whether this is the case, we record the weight of each point estimate (i.e., how much it contributes to the calculation of the overall average effect) and calculate the sum of the weights of the point estimates for each study. The sum of the weights at the study level never exceeds 4.77%, which shows that no individual study has an outsized effect on the estimated average same-sex teacher effect. We also explore alternative models to summarize all estimates. A random effect model that does not account for the dependence of estimates within-study yields an average same-sex teacher effect of 0.034 SD (std. err. = 0.003) and a standard deviation of 0.050. Using the fixed effects model that assumes the true same-sex teacher effect is the same for all studies, our estimate of the same-sex teacher effect is 0.010 SD (std. err. = 0.0004).

The estimate of the standard deviation of the distribution of the true same-sex teacher effect is 0.058 SD. Leveraging the assumption that the true same-sex teacher effects are normally distributed, we take the estimates of the mean and standard deviation to infer that 70% of true same-sex teacher effects are positive and 30% of true same-sex teacher effects are negative ($1 - \Phi\left(-\frac{0.030}{0.058}\right) = 0.7$). This distribution also reveals that 36.5% of same-sex teacher effects are larger than 0.05 SD and 8.4% are smaller than –0.05 SD. This estimated heterogeneity is substantial and suggests it is important to find out in which settings same-sex teachers help or hurt student performance.

We explore what drives the heterogeneity in same-sex teacher effects using four separate meta-regressions that includes as moderators: (1) whether studies use experimental or quasi-experimental variation, (2) the continent where the studies were conducted, (3) whether they analyze data from elementary, secondary school students, or a mix of both, and (4) whether they use test scores or grades as outcomes (see Table A2 in Appendix). Our results show no meaningful difference between estimates of same-sex teacher effects using experimental or quasi-experimental methods nor between estimates based on test scores or grades. However, we see some evidence of geographic heterogeneity. Compared to same-sex teacher estimates from Africa, same-sex teacher effects estimates are 0.051 SD smaller in Asia, 0.053 SD smaller in Europe, and 0.128 SD smaller in North America, with the difference between Africa and North America being statistically significant at the 5% level. We also find evidence that same-sex teacher effects are significantly smaller in primary education than they are in secondary education (–0.007 SD versus 0.051 SD).

It is unclear to what extent this heterogeneity reflects differences in true same-sex teacher effects across continents and levels of education rather than differences in study methods. No two studies in our meta-analysis use the same methodology. Two recent studies have shown that even seemingly innocent differences in methodology can have large effects on estimates. Huntington-Klein et al. (2021) and Breznau et al. (2022) apply the "many analysts" approach in which many researchers are given the same dataset and asked to answer the same research question. Both studies report many differences in methodological decisions between researchers and substantial variation in point estimates. Those findings suggest that our estimated standard deviation of the true same-sex teacher effect of 0.058 SD reflects differences in methods.

2.3 Do Same-Sex Teacher Effects Studies Suffer From Publication Bias?

Publication bias could affect our estimated average same-sex teacher effect of 0.030 SD. For example, researchers could be more likely to report specifications that show positive same-sex teacher effects, studies that show positive and significant same-sex teacher effects (either by chance or *p*-hacking) may be more likely to be written up, or reviewers and editors could behave more favorably toward studies that show positive effects. We will use all 538 main estimates to probe the existence of publication bias with two approaches.

In our first approach, we focus on discontinuities around *z*-scores of 1.64, 1.96, and 2.58—the critical values for statistical significance at the 10%, 5% , and 1% levels (see Brodeur et al. 2016). Appendix Figure A3 shows no evidence of heaping at the right side of these critical values. In our second approach, we estimate the relationship between estimated effect sizes and the precision of the estimate. If there is publication bias favoring positive same-sex teacher effect estimates, we would expect more-imprecise estimates to be larger.

We apply there are three popular ways to estimate the relationship between effect sizes and statistical precision. First, we regress the effect size on the ex-post MDE using a standard least squares estimator and test whether the slope coefficient of the MDE is positive. Second, we perform the precision effect test (PET) (Stanley and Doucouliagos 2014). Similar to the MDE regressions, this test consists of regressing the effect size on the standard error, and it tests for significance of the slope. The key difference from the MDE regressions is that observations in the precision effect regressions are weighted by the inverse of the estimated variance of the estimates. This test therefore gives more weight to more-precise estimates. Third, we perform Egger's test (Egger et al., 1997). This test consists of regressing *z*-scores on the inverse of the standard error. In contrast to the other two tests, the Egger's test shows evidence of publication bias if the *constant* is statistically significant (see Harrer et al., 2021, Ch. 9). In all three regressions, we account for the dependence of estimates within the same study by clustering at the study level.[7]

All three tests show evidence of publication bias. The estimated effect size significantly increases with the size of the MDEs (*p*-value < 0.001, see Appendix Figure A4). When we remove three outlier estimates from Ammermüller and Dolton (2006), the relationship between effect sizes and their respective MDEs remains similar but is no longer statistically significant

[7] We cannot correct for the mechanical dependence between effect size and standard error (Pustejovsky and Rodgers 2019) because the inputs required for this correction are generally not reported in the included studies. However, this correction is likely to be small because it shrinks with the model's degrees of freedom, and most estimates in our meta-analysis have samples many orders of magnitude larger than the typical study in fields where this correction is used (see e.g., Bierwiaczonek and Kunst, 2021; Kalén et al., 2021).

(*p*-value = 0.927).[8] The PET and Egger's test results also indicate the presence of publication bias regardless of whether the outlier estimates are included (all *p*-values for these tests are smaller than 0.001).

2.4 How Do Publication Bias Corrections Affect Our Estimate?

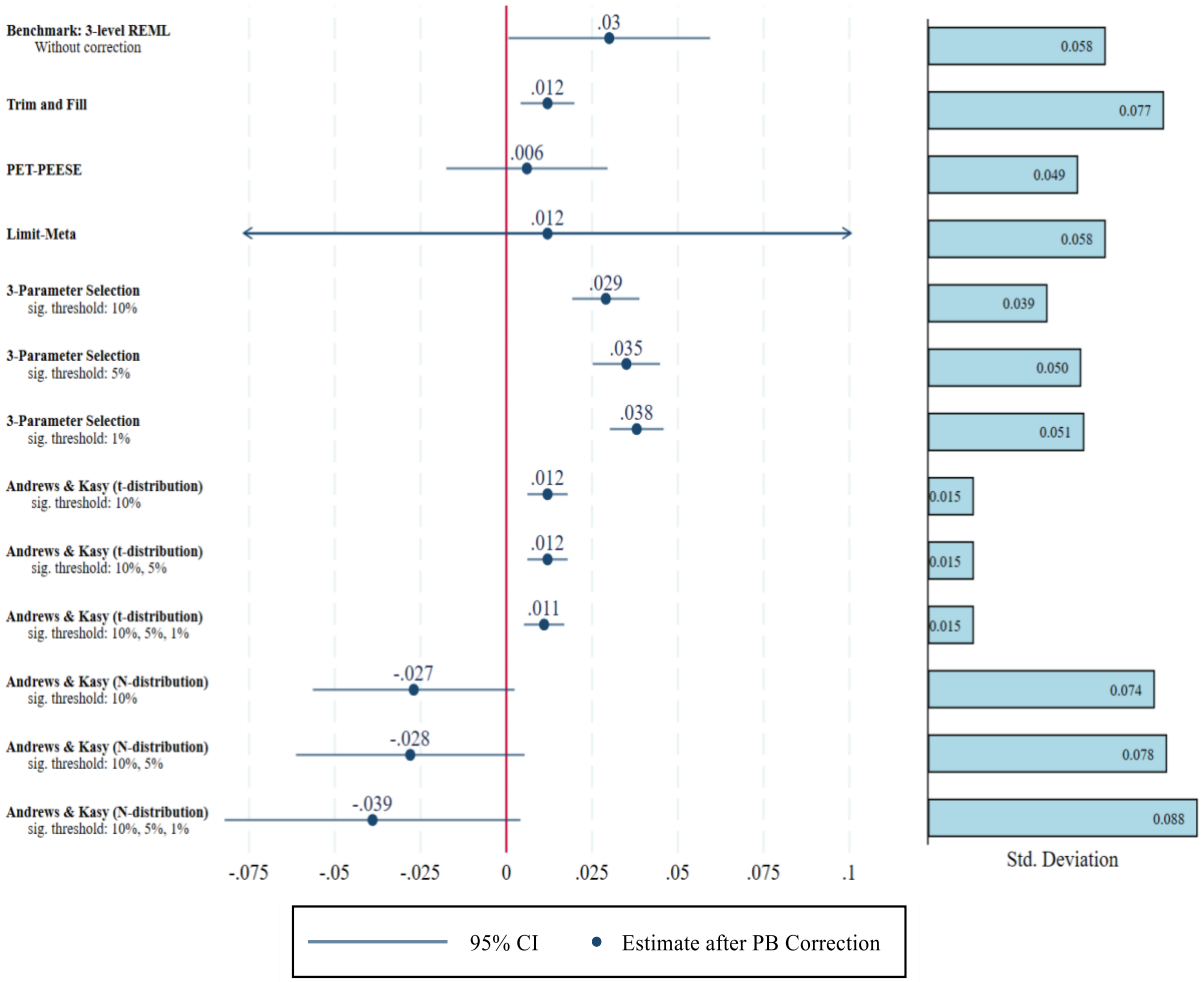
Figure 1 shows the estimated average same-sex teacher effect and estimated standard deviation of the true same-sex teacher effect after applying 12 of the most popular publication bias correction procedures. Trim and fill, PET-PEESE, and limit-meta focus on correcting for publication bias by using information from more-precisely estimated effects in the analysis to quantify and account for potential publication bias present in less precisely estimated effects. The methods of three-parameter selection and Andrews and Kasy (2019) focus on correcting for publication bias by modeling the probability that an estimate is published based on its sign and significance at conventional significance thresholds.

Figure 1 shows that the different procedures deliver broadly similar effect sizes. Corrected same-sex teacher estimates range between –0.039 and 0.038 SD. As expected, corrected estimates are generally of lower magnitude. Five out of the 12 corrected estimates are no longer statistically significantly different from zero at the 5% significance level. Trim and fill, PET-PEESE, and limit-meta reduce the same-sex teacher estimate to roughly a third to one half of the three-level random effect estimate. The three-parameter selection models do not change the same-sex teacher estimate much, varying between 0.029 and 0.038 SD depending on which significance threshold is assumed to drive publication bias. However, the Andrews and Kasy (2019) corrections show a curious pattern. When the underlying effects are assumed to follow a *t*-distribution, the effects shrink to around 0.012 SD, but assuming an underlying normal distribution of true effect yields negative corrected estimates, ranging between –0.027 and –0.039 SD.[9] The estimated standard deviations are also broadly similar between the different methods, ranging from 0.015 SD to 0.088 SD.

[8] These outliers are same-sex teacher effect estimates of 1.15, 2.07, and 0.92 SD with MDEs of 14.10, 15.19, and 19.13 SD, respectively. These estimates are very large and imprecise compared to the other estimates included in our meta-analysis.

[9] The Andrews and Kasy (2019) publication bias corrections are known to be quite noisy, so it is perhaps unsurprising that this method produces a relatively wide range of estimates and the only negative estimates of the average effect.

Figure 1: Same-Sex Teacher Effect Estimates After Correcting for Publication Bias



Notes: All estimated mean effects and estimated standard deviations are in the unit of standard deviation of the outcome variable. As a benchmark, the 3-level restricted maximum likelihood (REML) shows the estimated same-sex teacher effect without correcting for publication bias as shown and described in Section 2.2. All other estimates apply different publication bias corrections. Trim and Fill: We use the inverse variance method for pooling estimates. We use the REML method to estimate the variance and apply the Knapp–Hartung adjustment to account for the uncertainty in the estimation of the between-study heterogeneity. PET-PEESE: we use estimates from the Precision-Effect Test (PET) model rather than from the Precision-Effect Estimate with Standard Errors (PEESE) model because the intercept in the PET model is not statistically significantly different from zero at the 5% level (p-value = 0.3055) using one-sided *t*-test. We use the inverse variance method for pooling estimates. We use the REML method to estimate the variance and apply the Knapp–Hartung adjustment to account for the uncertainty in the estimation of the between-study heterogeneity. Limit-Meta: Uses 3-level REML as input. In the figure, the confidence intervals of this estimate were cut for readability reasons; the lower bound is –0.373 and the upper bound is 0.397. 3-Parameter Selection: We use 0.05, 0.025, and 0.01 as jumps in the publication probability function. REML estimator of the standard deviation of the effect size and the standard deviation of the effect size. Andrews and Kasy: We use the Andrews and Kasy (2019) correction method, assuming the effects are either *t*-distributed or normally distributed. We estimate separate corrections for cutoffs at the 0.05, 0.05, and 0.025, and 0.05, 0.025, and 0.01 significance levels for both positive and negative effects. We allow the probability of publication bias to be asymmetric. We produce an estimate using Kasy's App: https://maxkasy.github.io/home/metastudy. Other correction methods: Andrews and Kasy (2019)'s non-parametric GMM method did not produce a useful corrected estimate due to singularity issues. We also tried various continuous selection models assuming underlying beta, half-normal, and logistic publication probability distributions, which also did not yield useful estimates due to non-convergence issues. The bars on the right show the estimated standard deviation of the true same-sex teacher effects. Table A3 shows more details on the estimates shown in this figure.

In Appendix A we show alternative meta-analysis estimates using the set of "most controlled" estimates within each study, defined as those from model specifications using the largest number of control variables and narrowest within-group variation. From this alternative

meta-analysis, we also exclude "first difference" estimates, defined as effects of same-sex teachers on test score or grade *gains* (i.e., the difference between test scores or grades at two points in time for each student). This latter restriction only affects one estimate from Dee (2007). Our resulting subset of most-controlled estimates includes 297 estimates. This alternative meta-analysis produces very similar estimates, with an average same-sex teacher effect estimate of 0.032 SD (std. err. = 0.020) and a standard deviation of 0.060 SD. We also see: (1) similar effect heterogeneity, though with less statistical precision to detect differences; (2) little graphical evidence of publication bias in *z*-scores histograms and funnel plots; (3) more-conclusive evidence for publication bias on MDE plots and related tests; and (4) similar (though generally more muted) publication-bias corrected effects. See Tables A4 and A5 and Figures A5, A6, and A7 for these results.

We have shown that there is substantial heterogeneity in same-sex teacher effect estimates and evidence of publication bias. Depending on how we correct for publication bias, we find small positive effects or small negative average effects. Taken together, these estimates suggest same-sex teacher effects are small, but we cannot conclusively determine the sign of the average effect. Our meta-analysis is also not conclusive about the heterogeneity of same-sex teacher effects. The estimated standard deviations suggest substantial heterogeneity in effects between settings. However, meta-analysis methods struggle to distinguish between heterogeneity due to differences in true effects and due to differences in methodology.

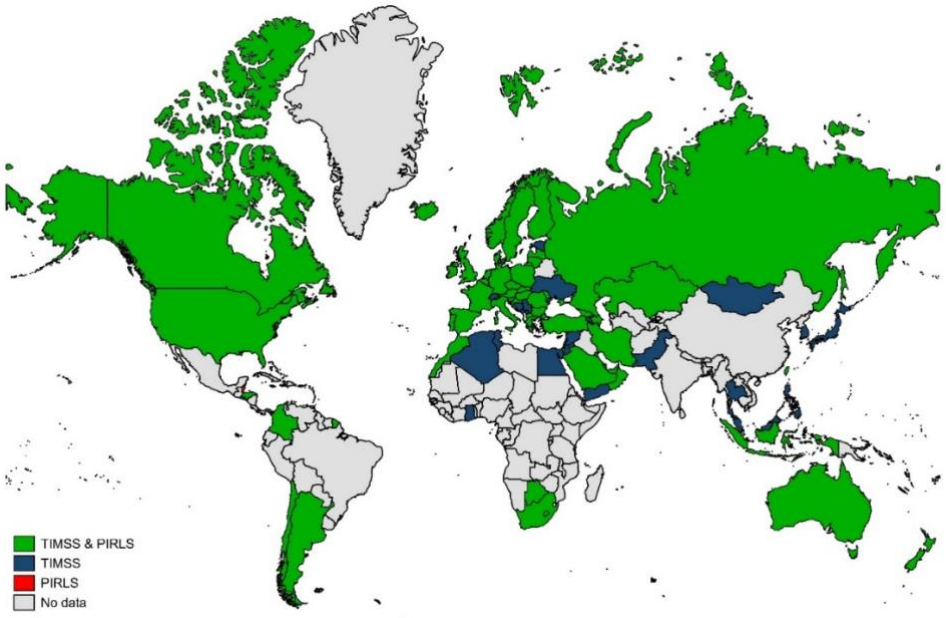
In theory, meta-regressions can tease out the effect of differences in methodology. In practice, this is challenging for three reasons. First, there are just too many methodological differences between studies. We have 24 different studies and researchers made more than 24 decisions in each study in terms of how to code their variables, how to restrict their sample, which outliers to remove, and which controls to include (Huntington-Klein et al., 2021; Breznau et al., 2022). Second, we cannot rule out that methodological differences are correlated with other factors (e.g., the context of the study) that affect the outcome. Third, while methodological differences would inflate our estimate of the standard deviation of the true same-sex teacher effects, the presence of publication bias would likely shrink it. In the presence of both these issues we cannot determine whether our estimates overstate or understate the variation in true same-sex teacher effects across studies. To better understand the heterogeneity of same-sex teacher effects, we therefore need comparable estimates from many settings that are free of publication bias. To be able to obtain those estimates, we build a large multi-country dataset.

3. Our Multi-Country Dataset

To estimate same-sex teacher effects we combine data from TIMSS and PIRLS. Both studies are administered by the International Association for the Evaluation of Educational Achievement (IEA), which specializes in administering education assessments that allow for international comparisons. TIMSS measures the skills and knowledge in mathematics and sciences of 4th graders (9- to 10-year-old children) and 8th graders (13- to 14-year-old children). PIRLS measures the reading skills of 4th graders.

For both studies, we use all waves up to December 2021, which is when we finished our data collection. These are seven waves of TIMSS (1995, 1999, 2003, 2007, 2011, 2015, and 2019) covering 86 different countries and four waves of PIRLS (2001, 2006, 2011, and 2016) covering 64 different countries. In Appendix B, we describe how we combine the data and the observations we had to exclude due to survey implementation issues. After these exclusions, we are left with 703 country-study-grade wave combinations from 90 countries covering 1995–2019. Figure 2 shows which countries were included in at least one wave for each study.

Figure 2: Countries for Which Data Are Available from TIMSS, PIRLS, or Both



Notes: The countries in red are those for which we have only PIRLS data. These are Trinidad and Tobago, Belize, Luxembourg, and Macao. They are hard to see on the map because all are small countries.

The data collection and study design are very similar for TIMSS and PIRLS. Both studies are centrally organized by the IEA and conducted by a national research coordinator in each country. The national research coordinators randomly select schools in their country and classes within these schools. We describe the details of this two-stage stratified random sampling design in Appendix B. Within the selected schools and classes, the national research

coordinator administers tests to students as well as surveys to students and teachers. We use these tests to measure students' ability in a subject and data from the surveys to identify the sex of the teacher and the student as well as several student and teacher characteristics that we use for our heterogeneity analyses. The complete surveys as well as much more background information on TIMSS and PIRLS are available at https://timssandpirls.bc.edu/.

The tests are designed by IEA experts with the goal of measuring reading skills, math skills, and science skills, as well as allowing for comparison of students' skills across countries. Each test is translated into the local language and these translations are checked to ensure that questions retain the original meaning and level of difficulty. All test booklets are marked by coders who are hired by the national research coordinator and trained by the IEA. During the marking, the coders do not see the names of the students. The quality of the marking is checked in two ways. First, a sample of tests within each country is marked by two coders independently. Second, a sample of tests from different countries are marked by coders who speak the pertinent languages. For example, coders who speak German and English are asked to mark tests of English and German students. The consistency of marking is very high. Within and across countries, coders agree whether a question is correct in more than 90% of cases. Appendix Table B3 shows sample questions from PIRLS and TIMSS test booklets.

Our main outcomes are math, science, and reading test scores, each measured as the average of five plausible test score values for each student and topic. In Appendix B we provide more details on the construction and use of these plausible values. In addition to test scores, we use three further outcomes: (1) students' job preferences, which captures their interest in choosing a job related to a subject; (2) students' enjoyment of a subject; and (3) students' confidence in a subject. We take these measures from the surveys in which students were shown several statements and asked how much they agree with them on a 4-point scale ranging from "Disagree a lot" to "Agree a lot." We measure job preferences by students' agreement with statements like, "*I would like a job that involved using mathematics."* We measure subject enjoyment and subject confidence by students' agreement to statements like, "*I enjoy reading"* and "*Reading is easy.*" Each of the statements references the specific course a student took. For example, students who took a general science class would be shown the statement, "*I enjoy learning science,"* whereas students who took a biology course would be shown, "*I enjoy learning biology."* The statements measuring subject enjoyment and subject confidence were included for all students in both studies. The statement measuring job preferences was only shown to 8th grade students in TIMSS. Table 1 shows the wording of the statements and in which studies they were included.

Table 1: Measurement of Job Preferences, Subject Enjoyment, and Subject Confidence

Subject	Study	Grade	Question item
Panel A: Job Preferences			
Math	TIMSS	8	I would like a job that involved using mathematics.
Science	TIMSS	8	I would like a job that involved using science.
Panel B: Subject Enjoyment			
Math	TIMSS	4 & 8	I enjoy learning mathematics.
Science	TIMSS	4 & 8	I enjoy learning science.
Reading	PIRLS	4	I enjoy reading.
Panel C: Subject Confidence			
Math	TIMSS	4 & 8	I usually do well in mathematics.
Science	TIMSS	4 & 8	I usually do well in science.
Reading	PIRLS	4	I usually do well in reading.

Notes: This table shows the item wording for the questions measuring job preferences, subject confidence, and subject enjoyment. The job preference and subject confidence questions are preceded by the text, "How much do you agree with these statements about [mathematics/science/biology]?" The subject enjoyment questions are preceded by the text, "How much do you agree with these statements about learning [mathematics/science/biology]?" Each statement is then followed by a block of questions that include our chosen question on job preferences, subject confidence, and subject enjoyment. Agreement is measured on a 4-point scale with labeled answers "Agree a lot," "Agree a little," Disagree a little," and "Disagree a lot."

In the raw data, PIRLS and TIMSS include observations at the student–teacher level. If students have multiple teachers for a given subject, the test scores are shown multiple times in the data. This happens in roughly 10% of the raw data, and particularly often for science. For example, in some schools, science is taught in two separate courses (e.g., biology and physics) by two distinct teachers, but students take only one science test in TIMSS, which captures material from both classes. Estimating same-sex teacher effects with this data structure would assign a higher weight to students who were taught by multiple teachers. To avoid this problem, we collapse our data at the student-assessment level, which leaves us with one observation per student in PIRLS and two observations for students in TIMSS—one for math and one for science. For students with multiple teachers in any one subject, teacher sex then becomes the share of female teachers in that subject. For example, for a student taught by one male and one female teacher in science, "female teacher" would take the value of 0.5.

4. Empirical Strategy

To measure same-sex teacher effects on test scores, we estimate the following regression model:

$$Score_{isj} = \beta_1 Female\ Student_i + \beta_2 Female\ Teacher_j + \beta_3 Female\ Student_i \times Female\ Teacher_j + \gamma' X_{isj} + u_{isj}, \quad (2)$$

where $Score_{isj}$ is the test score of student i in subject s that is taught by teacher j. $Female\ Student_i$ is a dummy variable indicating the sex of the student, $Female\ Teacher_j$ is the share of female teachers in subject s (which is equivalent to a dummy variable when students have only one teacher in subject s), and $Female\ Student_i \times Female\ Teacher_j$ is an interaction term of these two variables. X_{isj} is a vector of control variables that differ by specification, and u_{isj} is the error term. The same-sex teacher effect is captured by β_3, which shows the additional premium or penalty from having a same-sex teacher, on top of the general effect of having a female teacher. We estimate Eq. (2) via ordinary least squares regressions (OLS) and cluster our standard errors at the classroom level following the criteria outlined in Abadie et al. (2023).[10]

For the standardization of our dependent variables, we take advantage of the fact that the TIMSS and PIRLS tests scores are designed to be comparable across countries and over time and are standardized to have means of 500 and standard deviations of 100 in their first waves (see Appendix B). To interpret our results in terms of "global" standard deviations, we therefore standardize the test scores by subtracting 500 and dividing by 100. Although we describe our empirical strategy in terms of test scores, we also estimate same-sex teacher effects on job preferences, subject enjoyment, and subject confidence. We standardize each of these variables to have means of zero and standard deviations of one in our base dataset (see Appendix B). This approach allows us to interpret our results in terms of "global" standard deviations in these outcomes too.

In cases in which students have one teacher per subject, OLS estimates of β_3 are analogous to a "difference-in-difference" estimator (see Muralidharan and Sheth, 2016). Without any additional control variables, $\hat{\beta}_3$ is equal to the girl–boy difference in test scores of students taught by a female teacher minus the girl–boy difference of students taught by a male teacher. In the absence of omitted variable bias, the first difference would capture a same-sex teacher effect (e.g., female teachers being better at teaching girls than boys) *and* sex differences in student ability (e.g., girls being more able than boys) for students taught by female teachers. The second difference would capture a same-sex teacher effect (e.g., male teachers being better at teaching boys) *and* sex differences in student ability (e.g., girls being more able than boys) for students taught by male teachers. If sex differences in student ability are the same for female and male teachers, $\hat{\beta}_3$ isolates the same-sex teacher effect.

[10]Abadie et al. (2023) distinguish between clustered sampling and clustered treatments. In our case, the treatment $Female\ Student_i \times Female\ Teacher_j$ has no clear clustered structure, but our data can be described as a small sample of the population of classrooms in grades 4 and 8 in participating countries. For these kinds of settings, Abadie et al. (2023) recommend clustering at the sampling level, which is in our case is the classroom.

For students who are taught by multiple teachers in the same subject (e.g., they have two science teachers), the same-sex teacher coefficient captures the additional premium or penalty from having same-sex teachers *in all courses related to a subject* (e.g., all science courses), on top of the general effect of having female teachers *in all courses related to that subject.*

Besides the same-sex teacher effect, $\hat{\beta}_3$ could also capture biases from omitted variables. One instance of how this would happen is if sex differences in subject-specific ability are correlated with the number of female teachers. For example, the girl–boy difference in science ability might be larger than the girl–boy difference in math ability, and there might be more female science teachers than female math teachers. In this scenario, the fact that we observe female teachers more often in subjects in which girls are particularly able would lead to a positive bias of our same-sex teacher estimate. We address this concern by holding average sex differences in subject-specific ability constant: in all specifications, X_{isj} includes dummy variables for the test subject (e.g., science) and female student by test subject interaction terms (e.g., *Female Student* × *science*).

A related concern is that sex differences in teaching ability are correlated with the number of girls in a classroom. For example, female science teachers might be more effective than male science teachers and there might be more girls in science courses. This type of sorting would also lead to an upward bias in our same-sex teacher estimate. We address this concern by holding average sex differences in subject-specific teaching ability constant. In all specifications, X_{isj} includes female teacher times test subject interaction terms (e.g., *Female Teacher* × *science*).

Other threats to identification stem from systematic differences in student ability and teaching effectiveness due to non-random assignment of students to teachers. We therefore exclude observations from single-sex schools and single-sex classrooms within schools. We address remaining concerns about non-random sorting of students and teachers by estimating specifications with the following five sets of fixed effects: (1) country fixed effects, (2) school fixed effects, (3) classroom fixed effects, (4) student fixed effects, and (5) student and teacher fixed effects. We further estimate results for a subsample of countries that have an institutional mandate of random assignment and show that average effects in these countries are very similar to our overall results (see Appendix Table B5).

Preferred specification—student fixed effects and teacher fixed effects: In our preferred specification, we include student fixed effects and teacher fixed effects. In this specification,

we use *within-student, across-subject variation* to hold constant all student characteristics that are the same across subjects. For example, we exploit that the same student may have a female science teacher and a male math teacher (or vice versa). By also including teacher fixed effects we address one main concern: that more-effective teachers could be assigned to a higher share of students of their own sex.

This specification imposes several additional restrictions on our estimation sample. Most importantly, it requires us to drop data from PIRLS because this study only has data from one subject per student. The specification also requires us to exclude students who were taught only by teachers of one sex and students who had the same share of female teachers in both math and science (e.g., 50% of female teachers in all math courses and 50% of female teachers in all science classes). Finally, we are also forced to exclude rare instances in which teachers taught students who were either all girls or all boys. Note that in this specification the coefficients on the female student dummy and female teacher dummy are not identified because these variables are perfectly colinear with student and teacher fixed effects.

Our identifying assumption for this specification is that *within students* and *within teachers*, $FemaleStudent_i \times FemaleTeacher_j$ is unrelated to unobserved variables affecting students' test scores.

Credibility of causal effects: Our preferred specification addresses many intuitive concerns about sources of bias. Any omitted factors that systematically affect students or teachers of one sex are addressed by the inclusion of student and teacher fixed effects. For example, our estimates would not be biased by test designs that favor girls or school principals who are more supportive of male teachers. Student fixed effects also eliminate any bias caused by students who are more able in general (in both math and science) from being more likely to be assigned to a same-sex teacher. We also do not have to be concerned about typical sex differences in subject-specific student and teacher ability because X_{isj} includes subject main effects and interactions with the sex of students and teachers. Thus, no bias would be introduced if students are more likely to be assigned to same-sex teachers in subjects in which they are generally more able.

The most likely source of bias that remains is if deviations from average sex differences in subject-specific ability are correlated with teacher sex.[11] For example, our estimates would

[11] One can always think of implausible sources of bias like external TIMSS coders favoring girls but only when they were taught by female teachers. This source of bias is highly unlikely because coders do not observe students' sex nor do they know the sex of the teacher.

be biased if girls who have a particularly high science ability—compared to the average sex difference in science ability—are more likely to be assigned to a female science teacher.

We are not concerned about this type of incidental sorting because any residual sorting of concern would also have to be related to the sex match of teachers and students. For example, one can imagine that girls in one classroom are particularly good in science because they live in a neighborhood with a charismatic veterinarian who passionately teaches girls about animal biology. However, such a neighborhood characteristic would only bias our estimates if these girls were also more likely to be assigned to a female teacher in their science class.

We are also not concerned about any reassignment in response to student and teacher characteristics for two reasons. First, we believe explicit changes to classrooms or teacher assignments are rare. Second, for these changes to bias our estimates, they would have to be related to both the sex difference of subject-specific ability and to the sex of the teacher. We find this implausible. For example, while it is possible that male science teachers are more likely to be assigned to classrooms with many male troublemakers, it is *not* plausible that these troublemakers are also particularly bad in science *compared* to math.

Summary statistics of estimation samples: Table 2 shows summary statistics of our least restrictive estimation sample (using country fixed effects) and the most restrictive estimation sample (using student and teacher fixed effects).

In our country fixed effects sample, we have data from up to 3,047,752 different students. Students are on average 11.4 years old, 10% of them are foreign born, 75% speak the test language at home, and 38% have at least one parent with a university degree. For these students, we observe 1,453,989 math scores, 1,421,602 science scores, and 759,789 reading scores. We also observe 202,406 teachers; 71% of them are female, they have on average 16.5 years of teaching experience, and 30% have a bachelor's degree or higher.

In our preferred specification sample, we observe 568,346 different students who are, on average, 13.4 years old. The increase in average age from our least restrictive sample is driven by the exclusion of PIRLS, which only contains data on 4th graders. In addition to the increase in age, the students have similar characteristics on average. For example, 9% are foreign born (compared to 10% in our country fixed effects sample), 73% speak the test language at home (compared to 75%), and 36% have at least one parent with a university degree (compared to 38%). For these students, we observe 565,196 math scores and 560,622 science scores. However, we do see some differences in our teacher characteristics. The 49,018 teachers in this sample are less likely to be female (54% versus 71%) and more likely to be

more than 40 years old (84% versus 69%), are less likely to have majored in education (60% versus 71%), and are more likely to teach in their area of expertise (89% versus 75%).

Table 2: Summary Statistics for Our Most and Least Restrictive Estimation Samples

	Country FE sample		Most restrictive (preferred) specification sample			
	N	Mean	N	Mean	Female	Male
Student characteristics:						
Female	3,047,752	0.49	568,346	0.49	1	0
Age (years)	3,037,107	11.4	566,236	13.4	13.4	13.4
Foreign-born	2,270,763	0.10	533,194	0.09	0.08	0.09
25+ books at home	2,942,553	0.58	555,067	0.54	0.56	0.53
Speaks test language at home	2,899,132	0.75	549,548	0.73	0.73	0.73
Parent(s) have university degree	923,878	0.38	389,209	0.36	0.35	0.37
Teacher characteristics:						
Female	202,406	0.71	52,574	0.54	1	0
Experience (years)	198,316	16.5	51,650	15.9	15.1	15.9
40+ years old	201,949	0.69	52,473	0.83	0.59	0.59
Bachelor's degree or higher	196,279	0.30	50,728	0.34	0.29	0.27
Majored in education	171,313	0.71	43,621	0.60	0.64	0.62
Teaches field of expertise	135,745	0.75	45,835	0.89	0.88	0.86
Outcomes in math:						
Math test scores	1,453,989	485	565,196	484	483	486
Confident in math	1,414,575	3.00	551,331	2.96	2.91	3.01
Enjoys math	1,405,166	2.98	547,694	2.93	2.90	2.96
Wants a job involving math	922,028	2.53	395,258	2.54	2.44	2.63
Outcomes in science:						
Science test scores	1,421,602	482	560,622	482	480	485
Confident in science	1,386,829	3.05	548,918	3.02	2.98	3.06
Enjoys science	1,383,653	3.09	547,801	3.05	3.01	3.08
Wants a job involving science	907,777	2.57	390,955	2.57	2.52	2.61
Outcomes in reading:						
Reading test scores	759,789	513				
Confident in science	737,130	3.47				
Enjoys science	736,038	3.36				

Notes: This table shows the number of observations and means for our country fixed effects sample and our preferred estimation sample. "N" refers to unique students when describing student characteristics, unique teachers when describing teacher characteristics, and unique student-by-subject-matter combinations when describing math, science, and reading outcomes. The country fixed effects sample consists of up to 3,047,752 unique students, 202,406 unique teachers, 105,916 unique schools, and 144,372 unique classrooms from 90 countries. The preferred estimation sample consists of 568,346 unique students, 52,573 unique teachers, 22,004 unique schools, and 26,137 unique classrooms from 82 countries.

Overall, these statistics show two things. First, we have many observations, even for our most restrictive, preferred estimation sample. Second, the characteristics of the students and especially the teachers included in our samples differ by specification. These differences can drive differences in point estimates if, for example, same-sex teacher effects vary by student and teacher age. In our main analysis, we therefore report two estimates for each set of fixed effects: one that retains the largest possible estimation sample and one that holds the same sample constant across all fixed effects specifications.

5. Average Same-Sex Teacher Effects on Test Scores and Non-Test Score Outcomes

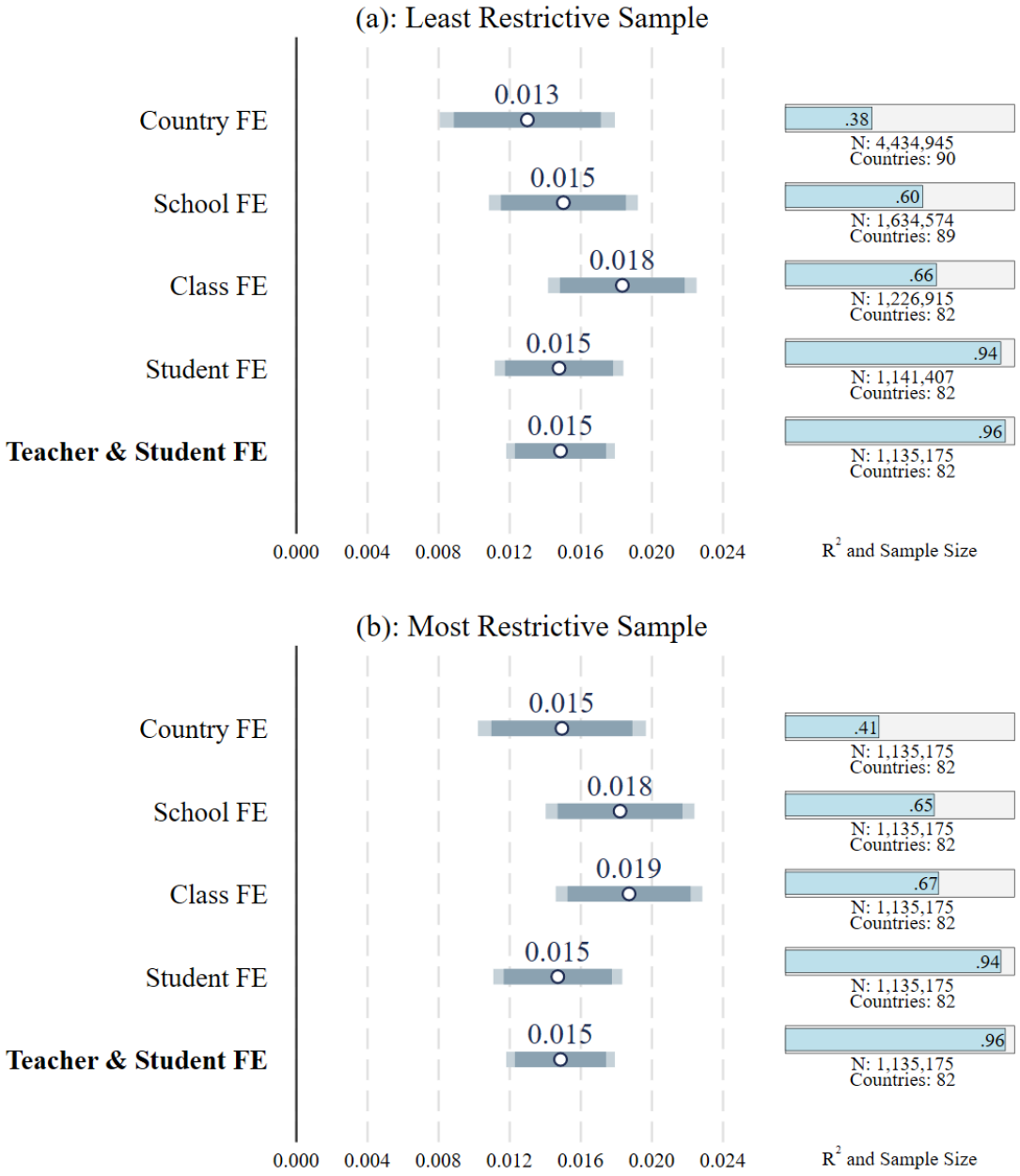
Same-sex teacher effects on test scores: Figure 3(a) shows same-sex teacher effect estimates with different sets of fixed effects where we keep the largest possible estimation sample for each specification. In our least restrictive specification with country fixed effects, our estimation sample consists of 4,434,945 observations from 3,047,752 students for whom we have math, science, or reading test scores. In this specification the R^2 is 0.38, and the estimated same-sex teacher effect is 0.013 SD. As we include more-restrictive fixed effects, the R^2 increases substantially but our point estimates barely change. In our preferred specification, we include student and teacher fixed effects. Including these fixed effects reduces our estimation sample to 1,135,175 observations from 568,346 students for whom we have math and science test scores and increases the R^2 to 0.96. This specification shows a precisely estimated same-sex teacher effect of 0.015 SD.

To check to what extent the small changes in point estimates are driven by differences in the estimation sample, Figure 3(b) shows estimates that keep the sample constant at the 1,135,175 observations we use in our preferred specification. With our smaller and more restrictive sample, we see somewhat larger point estimates in the country and school fixed effects specifications (0.015 SD and 0.018 SD). However, our conclusions remain the same. No matter the sample restrictions or the included fixed effects, we see a highly statistically significant same-sex teacher effect of around 0.015 SD. The 99% confidence intervals allow us to rule out effects smaller than 0.006 and larger than 0.024 SD for all same-sex teacher estimates shown in Figure 3.

The fact that effects are remarkably stable across specifications in Figure 3 is evidence that our causal identification strategy is strong. The first-order driver of endogeneity in our setting is sex-based student and teacher sorting. This sorting could happen at the school, classroom, or subject level. The degree of endogeneity created by this sorting depends on how much students, parents, and teachers can influence the student-teacher sex match. For example, if students and teachers can heavily sort into schools (because they have strong preferences and freedom to choose) but cannot sort into classes within schools (because classroom assignment is alphabetical for students and random for teachers), the school-level endogeneity would be much larger than the classroom-level endogeneity. Consequently, moving from school fixed effects to classroom fixed effects should bring large changes in our estimates. However, since our estimates barely move as we include increasingly narrow sets of fixed effects, this means

that either: (1) all the sorting happens at a narrower level than we observe and none of it happens at broader levels (which we find implausible), or that (2) sex-based sorting never introduced a first-order bias in the first place (which we find much more likely). [12]

Figure 3: Same-Sex Teacher Effects—Test Scores



Notes: This figure shows estimated same-sex teacher effects from regressions of standardized test scores on a $\text{FemaleStudent}_i \times \text{FemaleTeacher}_j$ interaction term, a set of other control variables (see Section 4), and different sets of fixed effects (as indicated to the left of the vertical line). The inclusion of different fixed effects imposes different sample restrictions. For example, estimating specifications with student fixed effects requires us to limit our sample to students for whom we observe two test scores. Panel (a) shows same-sex teacher effect estimates from specifications that use the largest possible estimation sample. Panel (b) shows estimates with one consistent estimation sample as imposed by our preferred teacher and student fixed effects specification (see Section 4). Appendix Table B4 shows the corresponding regression table. Horizontal bars show 90% and 95% confidence intervals that are based on standard errors clustered at the classroom level.

[12] Appendix Table B5 shows that if we restrict our sample to countries with institutional random assignment, we find very similar results for all outcomes of interest.

The estimated same-sex teacher effect in our analysis is half the size of the average same-sex teacher effect estimate from our meta-analysis (0.015 SD compared to 0.030 SD). It is hard to say what drives this difference. It could be differences in true effects, differences in methodologies, or publication bias. While meta-analysis estimates are hard to interpret, our analysis is more transparent. By holding the methodology constant and reducing concerns about publication bias, we get a better sense of what is and, more importantly, what is not driving our average same-sex teacher estimate.[13]

A same-sex teacher effect of around 0.015 SD is small. It represents a 1.5 point increase on the TIMSS or PIRLS tests. It is also small compared to the predicted effect of other demographic characteristics in our data. For example, the predicted effect of having at least one university-educated parent on test scores (0.605 SD) is 40 times larger than our estimated same-sex teacher effect, and the predicted effect of speaking the test language at home (0.636 SD) is 42 times larger than our same-sex teacher effect.[14]

Our same-sex teacher effect estimate is also small compared to estimates of teacher value-added and teacher experience. For example, the estimate of Chetty et al. (2014) of a one standard deviation increase in teacher value-added (VA) on students' math test scores (0.149 SD) is ten times as large as our same-sex teacher effect. Clotfelter et al. (2006)'s estimate of having a teacher with 12+ years of experience instead of a rookie teacher on math scores (0.113 SD) is eight times larger. Hanushek et al. (2005)'s estimate of having a teacher with six-plus years of experience instead of a rookie teacher (0.12 SD) is eight times larger.

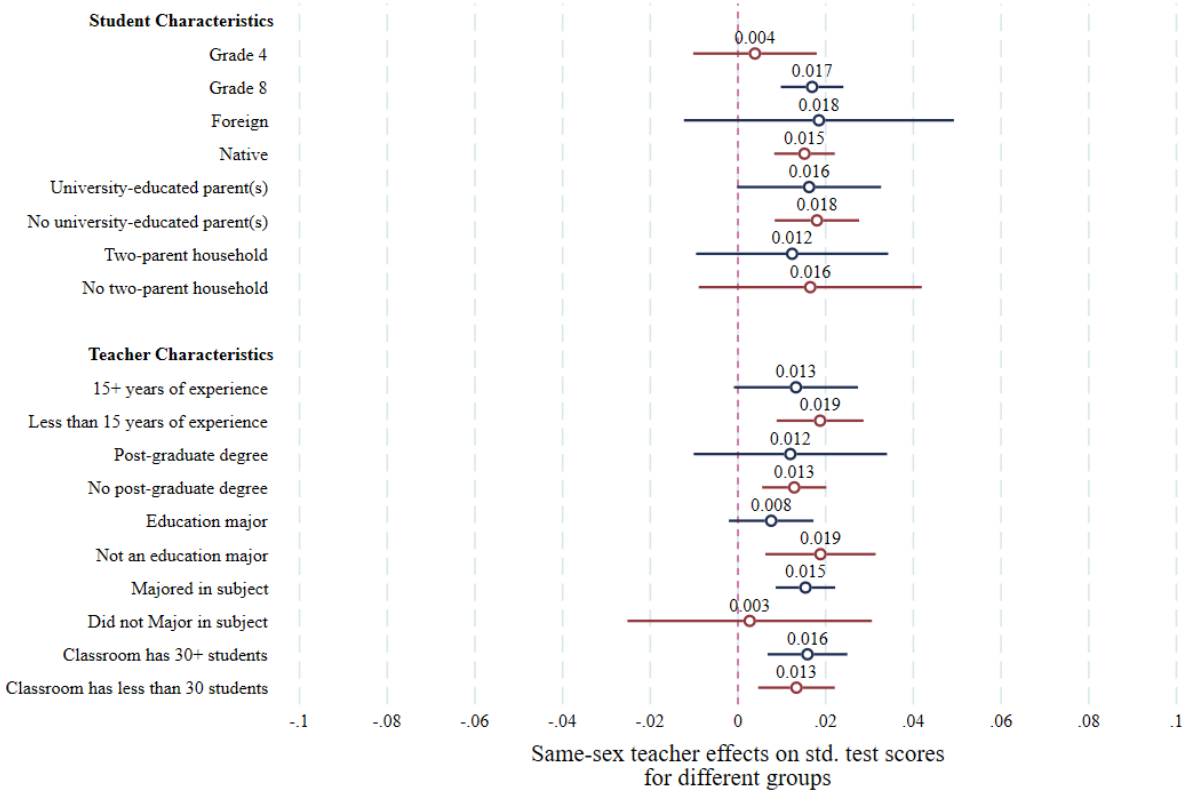
We explore the heterogeneity of same-sex teacher effects on test scores by subject, student characteristics, and teacher characteristics (see Appendix B for more details). We see somewhat larger same-sex teacher effects in math than in science (0.0188 SD versus 0.0117 SD) and statistically insignificant same-sex teacher effects in reading (0.0026 SD). The estimated same-sex teacher effects along the 18 student and teacher characteristics we consider are also all positive and range between 0.003 SD for teachers who did not major in the subject they are teaching to 0.019 SD for teachers with less than 15 years of teaching experience (see Figure 4). Overall, we observe minimal variation in the effects based on subjects, student

[13] We stress the importance of holding the methodology constant to ensure that the methodology does not vary at the same time as the setting—something that is unfortunately usually unavoidable in meta-analyses. However, holding the methodology constant does *not* mean researchers should avoid exploring how methodological choices affect their results. In our study, we intentionally show same-sex teacher effects estimates with very different samples (ranging from 1,135,175 to 4,434,945 observations) and very different empirical specifications (ranging from country fixed effects to student and teacher fixed effects). The stability of our results across this wide range of empirical approaches gives us confidence that our results are not an artifact of arbitrary methodological choices.

[14] These predicted effects are based on bivariate regression of test scores on: (1) a dummy indicating that at least one of the student's parents is university educated, or (2) a dummy variable indicating that the student speaks the test language at home.

characteristics, and teacher characteristics. These findings indicate that the small average effects do not conceal large same-sex teacher effects for any particular subgroup.

Figure 4: Student- and Teacher-Level Heterogeneity for Test Scores Estimates



Notes: This figure shows estimated same-sex teacher effects from regressions of standardized test scores on a $\text{FemaleStudent}_i \times \text{FemaleTeacher}_j$ interaction term, student fixed effects, teacher fixed effects, as well as other control variables from our preferred specification (see Section 4) for the different subsamples indicated on the left of the figure. Table B1 shows the corresponding regression table. Horizontal lines show 95% confidence intervals that are based on standard errors clustered at the classroom level.

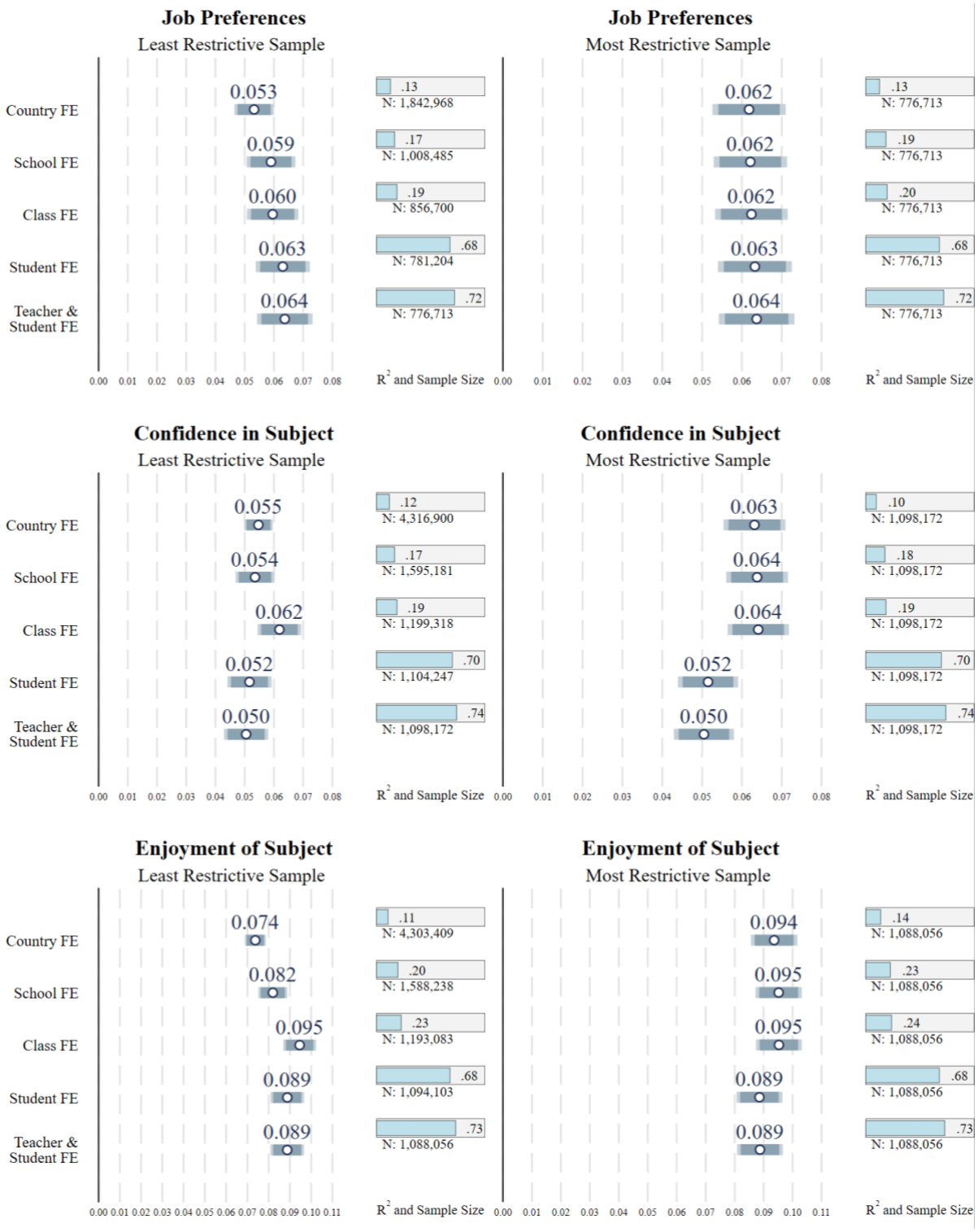
Same-sex teacher effects beyond test scores: Teachers' influence on their students may go beyond test scores. Same-sex teachers may also inspire students to follow in their footsteps and to make similar educational or occupational choices (Carrell et al., 2010; Card et al. 2022; Mansour et al., 2022). They may also affect students' confidence or how much they enjoy a subject. To test for such effects, we estimate same-sex teacher effects using the same set of fixed effects that we used for our test score analysis.

Figure 5 shows same-sex teacher effect estimates for non–test score outcomes. We keep the largest possible estimation sample for each specification in the left column and show estimates for the consistent sample of our most restrictive specification in the right column. Our results show that the estimated same-sex teacher effect on job preferences in our preferred specification (0.064 SD) is substantially larger than for test scores (0.015 SD). We further find same-sex teacher effects of similar magnitudes on subject confidence (0.050 SD) and on

subject enjoyment (0.089 SD). As for test scores, our results are very similar regardless of our sample restrictions or included fixed effects.

Although we do not have data on students' actual job choices, we find it plausible that these could also be affected. Moore and Burrus (2019) show there is a strong relationship between intention to choose a STEM major or a career as measured in secondary education and the subsequent choice of STEM majors and careers. Teachers who affect students' stated job preferences, their confidence, and their enjoyment of a subject may also affect their career trajectory by, for example, influencing which subjects the students choose in high school and university. Such effects on job choices would also be consistent with findings from previous studies. For example, Mansour et al. (2022) study the impact of professors at the United States Air Force Academy and find same-sex teacher effects on receiving a STEM master's degree and working in a STEM occupation. Similarly, Kofoed and McGovney (2019) study mentors at the U.S. Military Academy and find same-sex teacher effects on choosing their mentor's occupation.

Figure 5: Same-Sex Teacher Effects—Job Preferences, Subject Enjoyment, and Confidence



Notes: This figure shows estimated same-sex teacher effects from regressions of standardized job preferences on a $\text{FemaleStudent}_i \times \text{FemaleTeacher}_j$ interaction term, a set of other control variables (see Section 4), and different sets of fixed effects (as indicated on the left). We exclude eight countries because of missing data on job preferences from the first row (Algeria, Azerbaijan, Bosnia and Herzegovina, El Salvador, Honduras, Poland, Mongolia, and Yemen). The inclusion of different fixed effects imposes different sample restrictions. For example, estimating specifications with student fixed effects requires us to limit our sample to students for whom we observe two test scores. Figures in the left column show same-sex teacher effect estimates from specifications that use the largest possible estimation sample. Figures from the right column show estimates with one consistent estimation sample as imposed by our preferred teacher and student fixed effects specification (see Section 4). Appendix Table B4 shows the corresponding regression table. Horizontal bars show 95% and 90% confidence intervals that are based on standard errors clustered at the classroom level.

6. Going Beyond Global Average Effects

In the previous section we have shown *average* same-sex teacher effects across many countries. However, such "global" average effects are rarely of interest in practice. Policy makers and researchers wanting to build on our results typically want to know about effects in one specific context. For example, the organizers of Australia's "Males in Primary" initiative want to know whether they can expect positive same-sex teacher effects in Australian primary schools. Or, researchers trying to understand sex-differences in STEM college applications in Ireland might want to know about same-sex teacher effects on STEM performance in Irish secondary schools (Delaney and Devereux 2019).

We help those policy makers and researchers in two ways. First, we use country-level estimates and meta-analysis methods to explore the generalizability of same-sex teacher effects. This section shows in which kinds of contexts we should expect positive same-sex teacher effects. Second, we estimate the best linear unbiased predictions (BLUPs) of same-sex teacher effects for many specific contexts. Those BLUPs help people interested in one of the many contexts included in our analysis.

On the generalizability of same-sex teacher effects: We probe the generalizability of same-sex teacher effects with data from many diverse contexts. If we find positive effects for all of them, we can be confident that same-sex teacher effects are generally positive. If effects differ meaningfully between contexts, we may still uncover that effects are generally positive in one kind of context. For example, we may find that same-sex teacher effects are generally positive for one outcome or one level of education.

Empirically, we could address the question of generalizability by estimating same-sex teacher effects in many contexts and inspecting the estimates. However, each estimate also reflects sampling error. Even if same-sex teacher effects are universally positive, some estimates could be negative due to chance alone. We therefore take advantage of meta-analysis methods that allow us to account for sampling error and estimate the distribution of the true country-level same-sex teacher effects. In particular, we estimate the following random effects model (see Borenstein et al., 2010)

$$\hat{\beta}_{3c} = \theta + \zeta_c + \epsilon_c, \qquad (3)$$

where $\hat{\beta}_{3c}$ is the same-sex teacher effect estimate for country c, θ is the average of the true same-sex teacher effect of all countries included in our analysis (the "grand mean"), ζ_c is the country-specific deviation from the average same-sex teacher effect, and the term ϵ_c shows the difference between the true effect and estimate for country *c* due to sampling error. We

assume that the true country-level same-sex teacher effects, β_{3c}, are normally distributed with a variance of τ^2, and that ϵ_c is normally distributed with a mean of zero and a variance of $se(\hat{\beta}_{3c})^2$.

$$\beta_{3c} \sim N(\theta, \tau^2) \quad (4)$$

$$\epsilon_c \sim N\left(0,\ se(\hat{\beta}_{3c})^2\right) \quad (5)$$

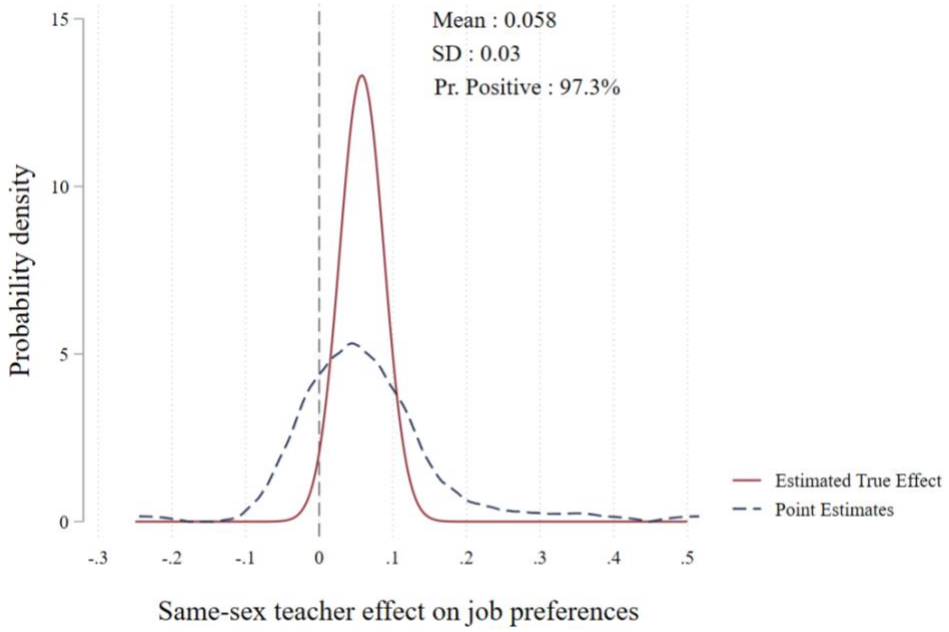
We estimate our key parameters of interest, θ and τ^2, via restricted maximum likelihood, using the coefficient estimates $\hat{\beta}_{3c}$ as best estimates for a country's true same-sex teacher effect and their standard errors to estimate the variance of the sampling error. In contrast to typical meta-analyses, we can do this without having to worry about differences in methodologies between estimates and publication bias.

Figure 6 illustrates our approach for same-sex teacher effects on job preferences in secondary education. In blue is the kernel density of the distribution of country-level same-sex teacher effects estimates on job preferences in secondary education. This distribution partly reflects sampling error. In red is the narrower estimated distribution of the same-sex teacher effects described in Equation (4). This distribution is normal by assumption, with a fitted mean of 0.058 SD and a fitted standard deviation of 0.030 SD. We can further leverage the normality assumption and infer that the same-sex teacher effects on job preferences are positive in 97% of the countries.

We estimate the distributions of the true same-sex teacher effects for all possible education level–outcome combinations (e.g., primary education test scores).[15] The credibility of these estimates depends on how reasonable the assumption is that the true effects are normally distributed. We test this assumption following Jackson and Mackevicius (2024) and show in Appendix B that the normality assumption is reasonable for all four outcomes and all grade-outcome combinations.

[15] We show the country-level estimates for all outcomes and their standard errors—the input for this analysis—in Table B6 in the Appendix.

Figure 6: Densities of the Same-Sex Teacher Effect Estimates and the Fitted Distribution of True Same-Sex Teacher Effects for Job Preferences

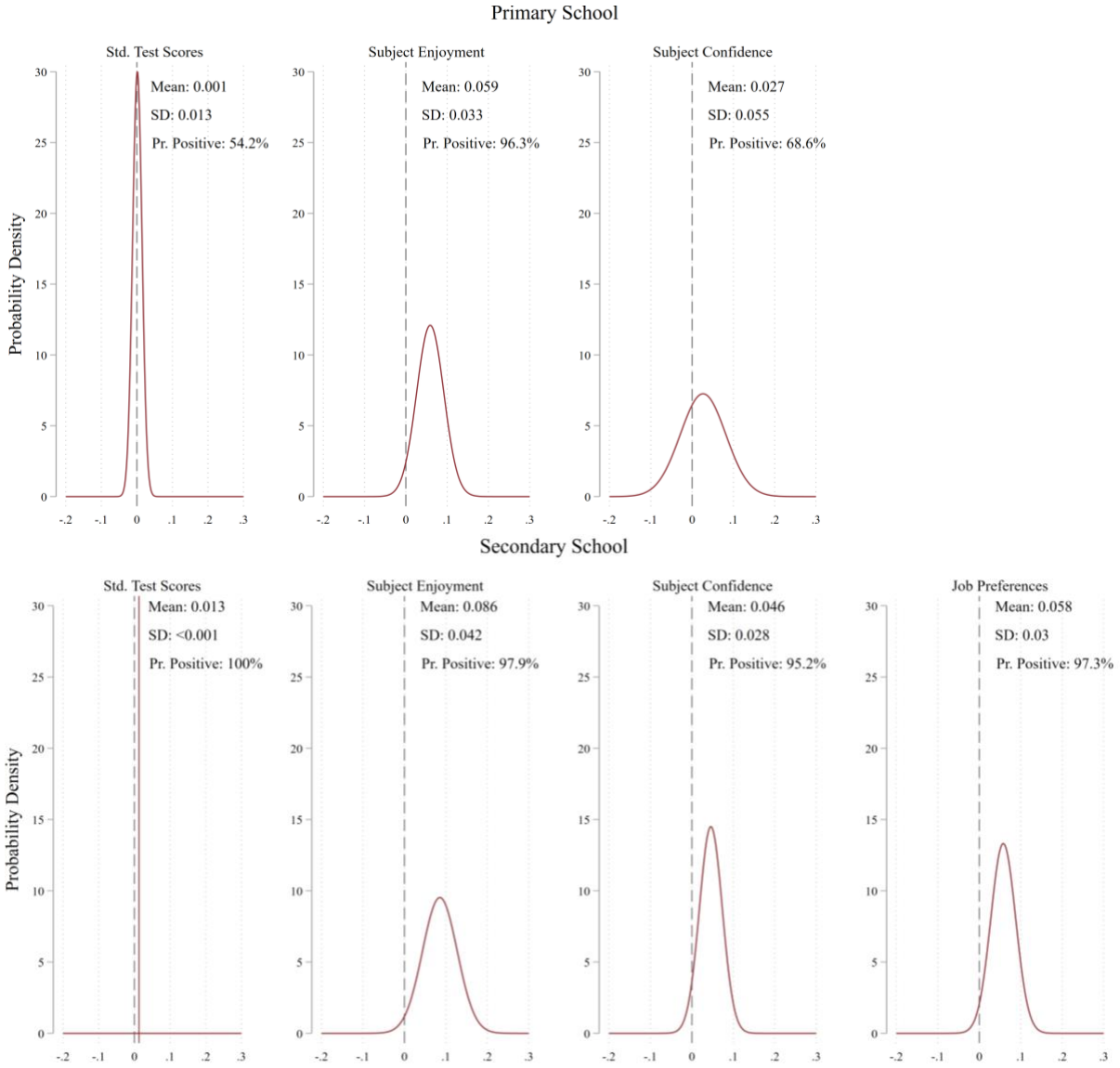


Notes: This figure shows a kernel density estimate of the distribution of the 71 country-level estimated same-sex teacher effects on job preferences (blue line), which uses a bandwidth of 0.03. The figure also shows the density of a normal distribution with mean 0.058 and standard deviation 0.030 (red line). These are the estimated parameters for the true same-sex teacher effects derived from the 71 country-level estimates using a random effects meta-analysis estimated with restricted maximum likelihood.

Figure 7 shows the estimated distributions of same-sex teacher effects for all combinations of education levels and outcomes. These distributions reveal interesting differences between primary and secondary education. In primary education, the distributions differ meaningfully by outcome. For test scores, we see a narrow distribution (standard deviation 0.013 SD) centered around 0.001 SD. This tiny estimated mean is consistent with our estimated global average same-sex teacher effects in primary education of 0.004 SD (see Figure 4) and our meta-analysis estimate for primary education of –0.007 SD (see Table A2). However, knowing the distribution allows us to go beyond these average effects. For example, we can infer that same-sex teacher effects are positive for only 46% of the countries in our sample (leaving 54% with negative effects). For subject enjoyment, the distribution is wider (standard deviation 0.033 SD) and centered around 0.059 SD, suggesting that same-sex teacher effects are positive in 96% of countries. For subject confidence, the distribution is widest (standard deviation 0.055 SD) and centered around 0.026 SD, suggesting effects are positive in 69% of countries.

Taken together, our analysis suggests that same-sex teacher effects in primary education are *not* generally positive. Policy makers should be aware that hiring more male primary school teachers to stop boys' performance decline (relative to girls) may backfire and produce small negative effects in some settings.

Figure 7: Distributions of Same-Sex Teacher Effects in Primary and Secondary Education



Notes: This figure shows the estimated country-level distribution of same-sex teacher effects for all available education level and outcome combinations. Those distributions are normal (by assumption) with means and standard deviations shown in the each subfigure. We estimate the means and standard deviations with restricted maximum likelihood. "Per. Positive" indicates the estimated percentage of countries for which the true same-sex teacher effect is positive. We display the estimated distribution of effects on test scores in secondary school with a vertical line at 0.013 because the estimated standard deviation is so small (<0.001 SD) that our estimates suggest that the true effects are effectively the same for all countries.

In contrast, same-sex teacher effects appear to be near universally positive in secondary education. For test scores, the estimated average same-sex teacher effect is 0.013 SD, which is similar to our estimated global average of 0.017 SD (see Fig, 4) but smaller than our meta-analysis estimate of 0.051 SD (see Table A2). The standard deviation of this distribution is tiny (<0.001 SD), which implies that effects are positive and effectively equal to the mean effect of 0.013 SD for all countries included in our analysis. We see more variation for non-test score

outcomes. The distributions are centered around 0.041 SD for subject enjoyment, around 0.027 SD for subject confidence, and 0.058 SD for job preferences. For all of these outcomes, we estimate that same-sex teacher effects are positive in more than 95% of the countries.

Taken together, these results for secondary education suggest that hiring more female teachers to increase girls' performance is unlikely to backfire but also will not have large effects. However, the effects for non-test score outcomes are larger and near universally positive, suggesting that hiring more female teachers might still be a worthwhile policy.

BLUPs of same-sex teacher effects: A BLUP is a weighted average of the overall mean estimate and a country-specific estimate of same-sex teacher effects. The weighting considers that country-level estimates also reflect sampling error. With a standard error of zero, the BLUP would be equal to the country-level estimate. The larger the standard error of a country-level estimate, the more weight is assigned to the overall mean in the BLUP estimation for that country.

We construct empirical Bayes estimates of these BLUPs by adapting the formula of Jackson and Mackevicius (2024, p. 422). Formally, we construct the BLUP for country c, as

$$\hat{\beta}_{3cBLUP} = w\hat{\theta} + (1 - w)\hat{\beta}_{3c} \,,$$

where $\hat{\theta}$ is the estimated average same-sex teacher effect (see Eq. 4), $\hat{\beta}_{3c}$ is the same-sex teacher effect estimate for country c, and the weight $w = \hat{\sigma}_c^2/(\hat{\sigma}_c^2 + \hat{\tau}^2)$ is a function of the squared standard error of the country-level estimate ($\hat{\sigma}_c^2$) and the estimated variance of same-sex teacher effects ($\hat{\tau}^2$), ensuring that more-precise country-level estimates receive more weight.

We report the BLUPs for all combinations of education level and outcome in an interactive map on our dedicated study website https://www.role-model-effects.com/. This website provides a useful tool for policy makers interested in one specific context. For example, the organizers of the "Males in Primary" initiative can see that our best estimates for same-sex teacher effects in Australian primary schools are 0.002 SD for test-scores, 0.071 SD for subject enjoyment, and 0.060 SD for subject confidence. These estimates suggest that hiring more male primary school teachers in Australia will not stop boys from falling behind academically but may improve how they feel about school.

7. What Explains Country-Level Heterogeneity in Same-Sex Teacher Effects?

We have shown that same-sex teacher effects on test scores do not vary much between countries. For non–test score outcomes, in contrast, we find meaningful heterogeneity. In this section we will focus on explaining country-level differences in same-sex teacher effects on one policy-relevant outcome that varies markedly between countries: students' job preferences. In Appendix B, we also show these results for same-sex teacher effects on subject confidence and enjoyment.

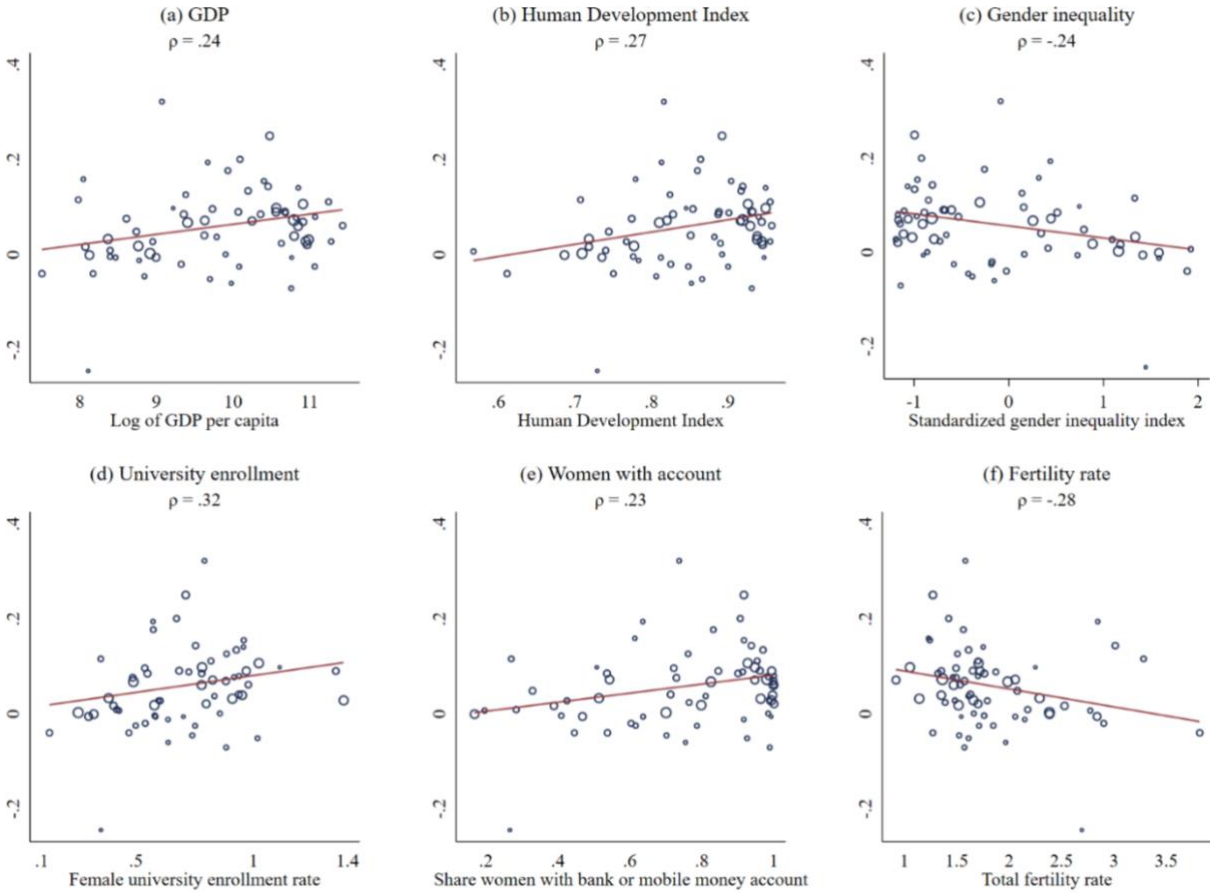
We explore country-level heterogeneity in same-sex teacher effects on job preferences in two ways. First, we show a series of scatterplots that relate the size of country-level estimates to country-level observable characteristics. These plots show the estimated same-sex teacher effects on job preferences on the y-axis and a given characteristic, for example, GDP per capita, on the x-axis. For brevity, we detail how we measure these characteristics in the figure notes. These scatterplots allow us to visually inspect the relationship between those two variables.

Second, we use meta-regressions to estimate separate same-sex teacher effects on job preferences for countries above and below the median for a given characteristic (e.g., above- and below-median GDP per capita). To do this, we use country-level estimates and their standard errors as inputs and estimate separate bivariate random effects meta-regressions. In each model the single regressor is a dummy that indicates whether a country is above the median for a given characteristic. In these specifications, the coefficient on the intercept identifies the estimated same-sex teacher effect for below-median countries, and we get the estimated same-sex teacher effect for above-median countries by adding this coefficient and the coefficient on the regressor. We discuss those estimates in the text and show the corresponding regressions in Table B7 in the appendix. Using both approaches, we explore whether same-sex teacher effects are related to a country's economic development, gender inequality, or sex differences in math and science performance.

Economic development. Same-sex teacher effects may be smaller in less developed countries where job choices are typically more constrained by necessity and tradition. For example, children expected to work on the family farm or in the family business might have fewer opportunities to enter STEM occupations. We use two measures for economic development: GDP per capita and the Human Development Index (HDI). Panels (a) and (b) of Figure 8 show that same-sex teacher effects on job preferences are positively related to the log of a country's GDP per capita and a country's HDI. Our meta-regressions confirm these results. Same-sex teacher effects estimates are significantly larger in countries with above-median GDP per

capita (0.800 SD compared to 0.0338 SD) and in countries that have an above-median HDI (0.796 SD compared to 0.0326 SD).

Figure 8: Same-Sex Teacher Effects in Job Preferences and Country-Level Correlates



Notes: These panels show the bivariate relationships between the estimated same-sex teacher effects on standardized job preferences shown in Figure 7 (on the y-axes) and different country-level characteristics (on the x-axes). ρ shows the Pearson's correlation coefficient between the two variables; the line shows a fitted least squares regression line. The size of each circle in the plot is dependent on the inverse of the standard error of the estimate, showing larger circles for more-precisely estimated effects. The characteristic shown in Panel (a) is log GDP per capita from 2019, which is taken from the World Bank World Development Indicators 2019. This characteristic is not available for Palestine, Scotland, Syria, and Taiwan. The characteristic shown in Panel (b) is the Human Development Index in 2017 computed by the United Nations (UN) as a composite measure of a country's average life expectancy at birth, years of schooling, and expected years of schooling, and the gross national income per capita in PPP terms. This characteristic is not available for Palestine, Scotland, and Taiwan. The characteristic shown in Panel (c) is the Gender Inequality Index (GII) from the Human Development Report 2020 published by the UN. The GII is calculated using this formula: GII = $\sqrt[3]{\text{Health} * \text{Empowerment} * \text{LFPR}}$ where Health is computed as Health = $(\sqrt{\frac{10}{\text{MMR}} * \frac{1}{\text{ABR}}}+1)/2$ where MMR is maternal mortality rate and ABR is the adolescent birth rate. Empowerment is computed as Empowerment =$(\sqrt{\text{PR}_\text{F} * \text{SE}_\text{F}} + \sqrt{\text{PR}_\text{M} * \text{SE}_\text{M}})/2$ where PR_F is the share of parliamentary seats held by women, and PR_M is the share of parliamentary seats held by men. SE_F is share of the female population with at least some secondary education, and SE_M is the share of the male population with at least some secondary education. The GII is standardized to have a mean of zero and a standard deviation of 1 for the included countries. LFPR is computed as the mean of male and female labor force participation rates: LFPR = $\frac{\text{LFPR}_\text{F}+\text{LFPR}_\text{M}}{2}$. The GII is missing for Hong Kong, Palestine, Scotland, and Taiwan. The characteristic shown in Panel (d) is the female university enrollment rate in 2016/17. The female university enrollment rate is computed as the ratio of total female enrollment in tertiary education, regardless of age, to the female population of the age group that officially corresponds to the tertiary level of education. The data are taken from the Gender Data Portal of the World Bank. This characteristic is available for all countries except for Japan, Lebanon, Palestine, Scotland, Taiwan, Turkey, Ukraine, and the United Arab Emirates. The characteristic in Panel (e) is the share of the female population aged 15+ who owned a bank account or mobile money account in 2017. The data are taken from the Gender Data Portal of the World Bank. This characteristic is not available for Iceland, Palestine, Scotland, and Taiwan. The characteristic shown in Panel (f) is the total fertility rate in 2019. The data are taken from the Gender Data Portal of the World Bank. This characteristic is not available for Palestine, Scotland, and Taiwan.

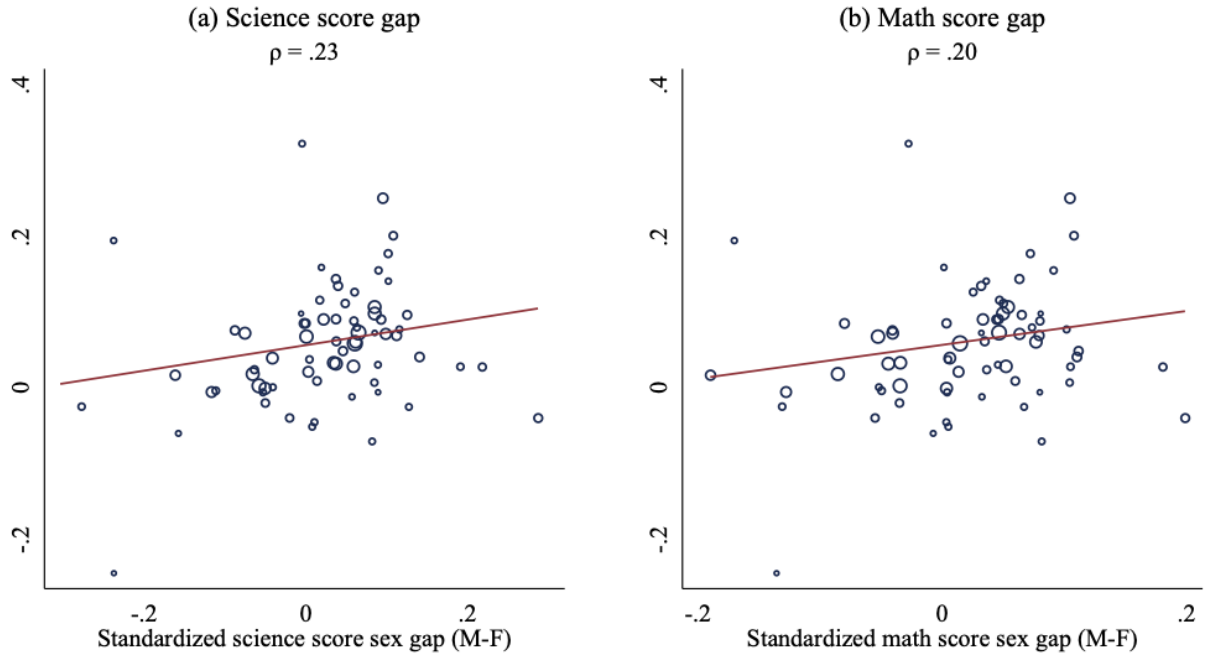
Gender inequality. Same-sex teacher effects might be stronger in gender-unequal countries where women face systemic barriers to education and the workplace. Or same-sex teacher effects might be stronger in gender-equal countries in which people are more aware of the remaining gender gaps. We measure gender inequality using the Gender Inequality Index from the United Nations Human Development Report (2020). This index is based on five measures: female secondary education completion, female labor force participation, share of parliamentary seats held by women, maternal mortality, and teenage birth rates.

Figure 8 (c) shows that same-sex teacher effects are smaller in more gender-unequal countries. Our regressions confirm these results: the estimated same-sex teacher effects are significantly smaller for above-median gender-inequality countries (0.0342 SD versus 0.0786 SD). Figure B2 in the Appendix shows that this relationship is driven by same-sex teacher effects being larger in countries where more women complete secondary education, in countries with lower maternal mortality, and in countries with lower teenage birth rates.

University enrollment, access to bank account, fertility rate. We also consider three additional measures of women's circumstances in a country: women's university enrollment, the share of women who have access to a bank account, and the fertility rate. Panels (d), (e), and (f) of Figure 8 show that same-sex teacher effects are larger in countries in which women have higher university enrollment, more access to bank accounts, and fewer children. Regressions confirm these results. We see significantly higher same-sex teacher effects in countries with above-median female university enrollment (0.0789 SD versus 0.0417 SD), above-median share of women who have access to a bank account (0.0769 SD versus 0.0314 SD), and significantly *lower* same-sex teacher effects in countries with above-median fertility rates (0.0436 SD versus 0.0753 SD).

Sex gaps in math and science test scores. Same-sex teacher effects on job preferences might depend on the differences in boys' and girls' ability in math and science. For example, in countries where boys outperform girls in math, girls might see having a female math teacher as evidence that girls can do well in math and might therefore be more open to choosing a career that requires this subject. The same logic would predict that in countries where girls outperform boys in math, boys' job preferences would be more influenced by having a male teacher.

Figure 9: Same-Sex Teacher Effects on Job Preferences and Test Score Gaps between Boys and Girls



Notes: This figure shows the relationship between the estimated same-sex teacher effects on standardized job preferences shown in Figure 7 and the standardized sex gap (M–F) in science (Panel a) or math (Panel b). The size of each circle in the plot is dependent on the inverse of the standard error of the estimate, showing larger circles for more-precisely estimated effects. These gaps are computed as the country mean of the standardized science/math score of boys minus the country mean of the standardized science/math score of girls. ρ shows the Pearson's correlation coefficient between the two variables; the line shows a fitted least squares regression line. Both panels contain data for all 71 countries for which we have same-sex teacher effects on job preferences.

Figure 9 shows that same-sex teacher effects are larger in countries with larger performance gaps in favor boys for science and math. We also estimate separate same-sex teacher effects for countries with above and below median boy–girl performance gaps. These regressions confirm our previous results. The estimated same-sex teacher effects are significantly larger in countries with above-median science and math test-score gaps (science: 0.0776 SD versus 0.0339 SD, math: 0.0818 SD versus 0.0407 SD).

Heterogeneity of same-sex teacher effects on subject enjoyment and confidence. The heterogenous same-sex teacher effects on subject enjoyment and subject confidence broadly mirror the pattern for same-sex teacher effects on job preferences. We show in Appendix B that same-sex teacher effects on subject enjoyment and subject confidence are larger in developed countries and smaller in countries with high gender inequality (see Tables B8 and B9, and Figures B3 and B4 in the appendix). More generally, we see same-sex teacher effects on these two outcomes are correlated with same-sex teacher effects on job preferences. The correlation between same-sex teacher effects on job preference and same-sex teacher effects on enjoyment is 0.50. The correlation between same-sex teacher effects on job preferences and

same-sex teacher effects on confidence is 0.31. In countries where same-sex teachers have a stronger effect on students' job preferences, we also see stronger same-sex teacher effects on how much students enjoy a subject and how confident they feel about it.

Putting everything together. We have shown that same-sex teacher effects on job preferences are larger in countries that are more developed, are more gender equal, in which women are more likely to go to university and to have a bank account, have fewer children, and in which girls perform worse than boys on science and math tests. These results paint a clear picture of the type of countries in which we should expect to find larger same-sex teacher effects on job preferences. For example, even though we do not have data on job preferences from India, we would expect only small same-sex teacher effects for this outcome as India is a poor and relatively gender-unequal country.

Understanding which environmental factors cause differences in same-sex teacher effects is difficult because we lack exogenous variation for these factors. However, the patterns we find are consistent with some explanations that can be tested using additional studies. One of these explanations is that larger same-sex teacher effects on job preferences are caused by girls being outperformed by boys in technical subjects and women having the opportunity to choose the job they want (e.g., because they live in a richer country, expect to go to university, or have fewer children). In these circumstances, having a female science teacher may be powerful in showing that girls can do jobs that involve science.[16]

8. Conclusion

There is a widespread belief that teachers are particularly good at teaching students of their own sex. Educators, politicians, and NGOs have therefore called for hiring more female teachers to boost girls' performance in math and science and to motivate girls to enter STEM professions. Similarly, there have been calls for hiring more male teachers in primary school to stop boys from falling behind. Our paper provides evidence about when such policies are likely to be effective.

We have shown that same-sex teacher effects on performance are, on average, small, whereas average effects on job preferences, subject confidence, and subject enjoyment are larger. We have also shown that effects differ meaningfully between primary and secondary

[16] Note that this pattern suggests that same-sex teacher effects are driven by girls' interaction with female teachers. In principle, we could also see stronger same-sex teacher effects in countries in which boys lag behind girls and can choose the job they want. However, it might be that same-sex teachers matter less for boys as there is no lack of examples of successful men in technical fields.

education. In primary education, same-sex teacher effects vary widely between countries and outcomes. For example, our results suggest that effects on test scores are negative in half of the countries. These results show that hiring more male primary school teachers to stop boys from falling behind may not be effective or may even backfire. In secondary education, same-sex teacher effects appear to be near universally positive. Effects on performance are positive and tiny for all countries. Effects on non–test score outcomes are generally positive, on average larger, and show more variation. For example, we see that same-sex teacher effects on job preferences are particularly large in rich and gender-equal countries. These results suggest that hiring more female STEM teachers may help to reduce occupational segregation—especially in rich and gender-equal countries.

Besides establishing these policy-relevant results, our paper demonstrates how to probe the generalizability of an effect. By showing that same-sex teacher effects are negative for some outcomes, and in some countries, we have established that such effects are *not* universal. Our rich dataset has also allowed us to show in which contexts same-sex teacher effects are generally positive (secondary education) and in which contexts effects are more mixed (primary education).

At the core of our approach is producing estimates from many diverse contexts and evaluating those estimates. This approach has important similarities to the traditional practice of assessing generalizability by waiting for research to accumulate and then reviewing the evidence. For example, research on the returns to education has consistently shown positive effects across many diverse contexts (Gunderson and Oreopolous, 2020; Patrinos and Psacharopoulos, 2020). These findings have led labor economists to agree that education generally yields positive returns.

However, this traditional practice of establishing generalizability has important shortcomings. Publication bias may lead to a distorted picture. Sampling error may lead to opposite-signed results, especially if studies are underpowered. Differences in methods between studies make it difficult to disentangle true and "artificial" heterogeneity.

How much we should worry about those shortcomings depends on the strength of the signal of the true effect relative to the noise from the publication process. For example, if the return to education is large and can be found in most contexts, the traditional approach is probably fine. However, for many other literatures, effects are smaller, less consistent, and often estimated with relatively small samples. In such situations, waiting for studies to accumulate and summarizing the results is often not enough. A better approach would be to

conduct a high-powered, multi-context study and take advantage of meta-analysis methods. We have shown how this can be done.

References

Aaronson, D., Dehejia, R., Jordan, A., Pop-Eleches, C., Samii, C., & Schulze, K. (2021). The effect of fertility on mothers' labor supply over the last two centuries. *The Economic Journal*, 131(633), 1–32.

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2023). When should you adjust standard errors for clustering? *The Quarterly Journal of Economics*, 138(1), 1–35.

Altmejd, A., Barrios-Fernández, A., Drlje, M., Goodman, J., Hurwitz, M., Kovac, D., Mulhern, C., Neilson C., & Smith, J. (2021). O brother, where start thou? Sibling spillovers on college and major choice in four countries. *The Quarterly Journal of Economics*, 136(3), 1831–1886. DOI: https://doi.org/10.1093/qje/qjab006

Ammermüller, A., & Dolton, P. (2006). Pupil-teacher gender interaction effects on scholastic outcomes in England and the USA. *ZEW – Centre for European Economic Research Discussion Paper No. 06–060*.

Andrews, I., & Kasy, M. (2019). Identification of and correction for publication bias. *American Economic Review*, 109(8), 2766–2794. DOI: https://doi.org/10.1257/aer.20180310

Angrist, I. D., & Fernandez—Val, I. (2013, May). ExtrapoLATE-ing: External Validity and. In Advances in Economics and Econometrics: Volume 3, Econometrics: Tenth World Congress (Vol. 51, p. 401). Cambridge University Press.

Antecol, H., Eren, O., & Ozbeklik, S. (2015). The effect of teacher gender on student achievement in primary school. *Journal of Labor Economics*, 33(1), 63–89. DOI: https://doi.org/10.1086/677391

Barro, R., & Lee, J. (2018). Barro-Lee educational attainment data. DOI: http://www.barrolee.com/

Bettinger, E. P., & Long, B. T. (2005). Do faculty serve as role models? The impact of instructor gender on female students. *American Economic Review*, 95(2), 152–157. DOI: https://doi.org/10.1257/000282805774670149

Bhattacharya, S., Dasgupta, A., Mandal, K., & Mukherjee, A. (2022). Identity and learning: A study on the effect of student-teacher gender matching on learning outcomes. *Research in Economics*, 76(1), 30–57. DOI: https://doi.org/10.1016/j.rie.2021.12.001

Bierwiaczonek, K., & Kunst, J. R. (2021). Revisiting the integration hypothesis: Correlational and longitudinal meta-analyses demonstrate the limited role of acculturation for cross-cultural adaptation. *Psychological Science*, 32(9), 1476-1493.

Bietenbeck, J., & Collins, M. (2023). New evidence on the importance of instruction time for student achievement on international assessments. *Journal of Applied Econometrics*.

Bogen, J., & Woodward, J. (1988). Saving the phenomena. *The Philosophical Review, 97(*3), 303–352.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*(2), 97–111.

Breda, T., Jouini, E., Napp, C., & Thebault, G. (2020). Gender stereotypes can explain the gender-equality paradox. *Proceedings of the National Academy of Sciences*, 117(49), 31063-31069.

Breznau, N., Rinke, E. M., Wuttke, A., Nguyen, H. H., Adem, M., Adriaans, J., & Van Assche, J. (2022). Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proceedings of the National Academy of Sciences*, *119*(44), e2203150119.

Britannica, The Editors of Encyclopaedia. "Boys' and girls' height curves." *Encyclopedia Britannica*, 2024, https://www.britannica.com/science/human-development/Boys-and-girls-height-curves. (retrieved on: April 22, 2024).

Brodeur, A., Lé, M., Sangnier, M., & Zylberberg, Y. (2016). Star wars: The empirics strike back. *American Economic Journal: Applied Economics*, *8*(1), 1–32.

Buddin, R., & Zamarro, G. (2008). Teacher quality, teacher licensure tests, and student achievement. RAND Education Working Paper WR-555-IES.

Busse, C., Kach, A. P., & Wagner, S. M. (2017). Boundary conditions: What they are, how to explore them, why we need them, and when to consider them. *Organizational Research Methods, 20*(4), 574-609

Card, D., Domnisoru, C., Sanders, S. G., Taylor, L., & Udalova, V. (2022). The impact of female teachers on female students' lifetime well-being. *NBER Working Paper Series*, 30430, https://www.nber.org/papers/w30430

Carrell, S. E., Page, M. E., & West, J. E. (2010). Sex and science: How professor gender perpetuates the gender gap. *Quarterly Journal of Economics*, 125(3), 1101–1144. DOI: https://doi.org/10.1162/qjec.2010.125.3.1101

Carrington, B., Tymms, P., & Merrell, C. (2008). Role models, school improvement and the "gender gap"—do men bring out the best in boys and women the best in girls? *British Educational Research Journal*, 34(3), 315–327. DOI: https://doi.org/10.1080/01411920701532202

Chabé-Ferret, S. (2023). *Statistical Tools for Causal Inference* (Ver. 2023-01-17). The Social Science Knowledge Accumulation Initiative (SKI). https://chabefer.github.io/STCI/

Chang, S., Cobb-Clark, D. A., & Salamanca, N. (2022). Parents' responses to teacher qualifications. *Journal of Economic Behavior & Organization*, 197, 419–446.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, 104(9), 2593–2632.

Clotfelter, C. T., H. F. Ladd, J. L. Vigdor. (2006) Teacher–student matching and the assessment of teacher effectiveness. *Journal of Human Resources*, 41(4), 778–820. DOI : https://doi.org/10.3386/w11936

Coenen, J., & van Klaveren, C. (2016). Better test scores with a same-gender teacher? *European Sociological Review*, 32(3), 452–464. DOI: https://doi.org/10.1093/esr/jcwOl2

de Gendre, A., Feld, J. and Salamanca, N. (2024). Re-examining the relationship between patience, risk-taking, and human capital investment across countries. *Journal of Applied Econometrics*.

Dee, T. S. (2007). Teachers and the gender gaps in student achievement. *The Journal of Human Resources*, 42(3), 528–554. DOI: https://doi.org/10.3368/jhr.XLII.3.528

Delaney, J. M., & Devereux, P. J. (2019). Understanding gender differences in STEM: Evidence from college applications. *Economics of Education Review*, *72*, 219–238.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3), 177–188. DOI: https://doi.org/10.1016/0197-2456(86)90046-2

Dudek, T., Brenøe, A. A., Feld, J., & Rohrer, J. M. (2022). No evidence that siblings' gender affects personality across nine countries. *Psychological Science*, 33(9), 1574–1587.

Eble, A., & Hu, F. (2020). Child beliefs, societal beliefs, and teacher-student identity match. *Economics of Education Review*, 77. DOI: https://doi.org/10.1016/j.econedurev.2020.101994

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, 315(7109), 629–634.

Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science*, *16*(4), 779–788.

Escardíbul, J.-O., & Mora, T. (2013). Teacher gender and student performance in mathematics. Evidence from Catalonia (Spain). *Journal of Education and Training Studies*, 1(1), 39–46. DOI: https://doi.org/10.11114/jets.v1i1.22

Evans, M. O. (1992). An estimate of race and gender role-model effects in teaching high school. *The Journal of Economic Education*, 23(3), 209–217. DOI: https://doi.org/10.1080/00220485.1992.10844754

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., & Sunde, U. (2018). Global evidence on economic preferences. *The Quarterly Journal of Economics*, 133(4), 1645-1692.

Gong, J., Lu, Y., & Song, H. (2018). The effect of teacher gender on students' academic and noncognitive outcomes. *Journal of Labor Economics*, 36(3), 743–778. DOI: https://doi.org/10.1086/696203

Goulas, S., Griselda, S., & Megalokonomou, R. (2022). Comparative advantage and gender gap in STEM. *Journal of Human Resources*, 0320-10781R2.

Gunderson, M., & Oreopolous, P. (2020). Returns to education in developed countries. In *The Economics of Education* (pp. 39–51). Academic Press.

Hanushek, E. A., Kain, J. F., O'Brien, D. M., & Rivkin, S. G. (2005) The market for teacher quality. *NBER Working Paper*, 11154. DOI: https://doi.org/10.3386/w11154

Harrer, M., Cuijpers, P., Furukawa, T. A., & Ebert, D. D. (2021). Doing meta-analysis with R: A hands-on guide. *Boca Raton, FL and London: Chapman & Hall/CRC Press.* ISBN 978-0-367-61007-4.

Hermann, Z., Diallo, A. (2017): Does teacher gender matter in Europe? Evidence from TIMSS data. *Budapest Working Papers on the Labour Market*, No. BWP – 2017/2. ISBN: 978–615–5594–86–1

Hoffmann, F., & Oreopoulos, P. (2009). A professor like me: The influence of instructor gender on college achievement. *Journal of Human Resources*, 44(2). DOI: https://doi.org/10.3368/jhr.44.2.479

Holmlund, H., & Sund, K. (2008). Is the gender gap in school performance affected by the sex of the teacher? *Labour Economics*, 15(1), 37–53. DOI: https://doi.org/10.1016/j.labeco.2006.12.002

Huntington-Klein, N., Arenas, A., Beam, A., Bertoni, M., Bloem, J. R., Burli, P., Chen, N.,

Grieco, P., Ekpe, G., Pugatch, T., Saavedra, M., & Stopnitzky, Y. (2021) The influence of hidden researcher decisions in applied microeconomics. *Economic Inquiry*. 59(3), 944-960.

Hwang, N., & Fitzpatrick, B. (2021). Student-teacher gender matching and academic achievement. *AERA Open*, 7. DOI: https://doi.org/10.1177/23328584211040058

Irsova, Z., Doucouliagos, H., Havranek, T., & Stanley, T. D. (2023). Meta-analysis of social science research: A practitioner's guide. *Journal of Economic Surveys*.

Jackson, K.C. and Mackevicius, C. (2024) What impacts can we expect from school spending policy? Evidence from evaluations in the U.S. *American Economic Journal: Applied Economics*. 16(1), 412-446.

Kalén, A., Bisagno, E., Musculus, L., Raab, M., Pérez-Ferreirós, A., Williams, A. M., & Ivarsson, A. (2021). The role of domain-specific and domain-general cognitive functions and skills in sports performance: A meta-analysis. *Psychological Bulletin*, 147(12), 1290.

Kofoed, M., & McGovney, E. (2019). The effect of same-gender or same-race role models on occupation choice: Evidence from randomly assigned mentors at West Point. *Journal of Human Resources*, 54(2), 430–467.

Lee, J., Rhee, D.-E., & Rudolf, R. (2019). Teacher gender, student gender, and primary school achievement: Evidence from ten francophone African countries. *The Journal of Development Studies*, 55(4), 661–679. DOI: https://doi.org/10.1080/00220388.2018.1453604

Lim, J., & Meer, J. (2017). The impact of teacher-student gender matches random assignment evidence from South Korea. *Journal of Human Resources*, 52(4), 979–997. DOI: https://doi.org/10.3368/jhr.52.4.1215-7585R1

Lim, J., & Meer, J. (2020). Persistent effects of teacher-student gender matches. Journal of Human Resources, 55(3), 809–835. DOI: https://doi.org/10.3368/jhr.55.3.0218-9314R4

Lindahl, E. (2007). Gender and ethnic interactions among teachers and students—evidence from Sweden. Institute for Labour Market Policy Evaluation Working Paper No. 2007:25.

List, J. (2020). Non est disputandum de generalizability? A glimpse into the external validity trial. No. w27535. *National Bureau of Economic Research*.

Mansour, H., Rees, D. I., Rintala, B. M., & Wozny, N. N. (2022). The effects of professor gender on the postgraduation outcomes of female students. *ILR Review*, 75(3), 693–715. DOI: https://doi.org/10.1177/0019793921994832

Meager, R. (2019). Understanding the average impact of microcredit expansions: A Bayesian hierarchical analysis of seven randomized experiments. *American Economic Journal: Applied Economics*, *11(*1), 57-91.

Meyer, A. (2017). Les hommes qui enseignent à l'école primaire. *Haute École Pédagogique - BEJUNE*

Moore, R., & Burrus, J. (2019). Predicting STEM major and career intentions with the theory of planned behavior. *The Career Development Quarterly*, 67(2), 139–155.

Mulji, N. (2016). The role of teacher gender on students' academic performance. *Department of Economics, Lund University Libraries.*

Muralidharan, K., & Sheth, K. (2016). Bridging education gender gaps in developing

countries: The role of female teachers. *Journal of Human Resources*, 51(2), 269–297. DOI: https://doi.org/10.3368/jhr.51.2.0813-5901R1

Neugebauer, M., Helbig, M., & Landmann, A. (2011). Unmasking the myth of the same-sex teacher advantage. *European Sociological Review*, 27(5), 669-689.

Neumark, D., & Gardecki, R. (1998). Women helping women? Role model and mentoring effects on female Ph.D. students in economics. *Journal of Human Resources*, 33(1), 220–246.

Nunnery, J., Kaplan, L., Owings, W. A., & Pribesh, S. (2009). The effects of Troops to Teachers on student achievement: One state's study. *NASSP Bulletin, 93*(4), 249-272.

O'Connell, A. A., McCoach, D. B., & Bell, B. A. (Eds.). (2022). Multilevel Modeling Methods with Introductory and Advanced Applications. IAP.

OECD (2012), Closing the Gender Gap: Act Now. *OECD Publishing*. DOI: http://dx.doi.org/10.1787/9789264179370-en

Patrinos, H. A., & Psacharopoulos, G. (2020). Returns to education in developing countries. In *The Economics of Education* (pp. 53–64). Academic Press.

Park, H., Behrman, J. R., & Choi, J. (2013). Causal effects of single-sex schools on college entrance exams and college attendance: Random assignment in Seoul high schools. *Demography*, 50(2), 447-469. DOI: https://doi.org/10.1007/s13524-012-0157-1

Porter, C., & Serra, D. (2020). Gender differences in the choice of major: The importance of female role models. *American Economic Journal: Applied Economics*, 12(3), 226–254.

Pustejovsky, James E., and Melissa A. Rodgers. (2019). Testing for funnel plot asymmetry of standardized mean differences. *Research Synthesis Methods,* 10 (1): 57–71.

Rakshit, S., & Sahoo, S. (2021). Biased teachers and gender gap in learning outcomes: Evidence from India. *IZA Discussion Paper No. 14305.*

Roser, M., Appel, C., & Ritchie, H. (2013). Human height. *Our world in data*. https://ourworldindata.org/human-height (retrieved on: 09.01.2024)

Schaede, U., & Mankki, V. (2022). Quota vs. quality? Long-term gains from an unusual gender quota CESifo Working Paper No. 9811. SSRN: https://ssrn.com/abstract=4150133

Stanley, T. D., Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, 5, 60–78. DOI: https://doi.org/10.1002/jrsm.1095

Szwed, C. (2010). Gender balance in primary initial teacher education: Some current perspectives. *Journal of Education for Teaching*, *36*(3), 303-317.

UNICEF (2020). Mapping gender equality in STEM from school to work. *UNICEF Office of Global Insight and Policy Report*. https://www.unicef.org/globalinsight/media/1361/file (retrieved on: 15.08.2022, 12:45)

UNICEF (2020). Towards an equal future: Reimagining girls' education through STEM. *UNICEF Education Section Programme Division*. https://www.unicef.org/media/84046/file/Reimagining-girls-education-through-stem-2020.pdf (retrieved on: 15.08.2022, 12:45)

Vivalt, E. (2020). How much can we generalize from impact evaluations?. *Journal of the European Economic Association*, *18*(6), 3045-3089.

Wang, C. C., & Lee, W. C. (2020). Evaluation of the normality assumption in meta-analyses.

American Journal of Epidemiology, 189(3), 235–242.
Wößmann, L., & West, M. (2006). Class-size effects in school systems around the world: Evidence from between-grade variation in TIMSS. *European Economic Review*, 50(3), 695–736.
World Bank (2020). The equality equation: Advancing the participation of women and girls in STEM. https://openknowledge.worldbank.org/bitstream/handle/10986/34317/Main-Report.pdf?sequence=1&isAllowed=y (retrieved on: 15.08.2022, 13:00)
Xu, D., & Li, Q. (2018). Gender achievement gaps among Chinese middle school students and the role of teachers' gender. *Economics of Education Review*, 67, 82–93. DOI: https://doi.org/10.1016/j.econedurev.2018.10.002
Xu, R. (2020). "When boys become the second sex": The new gender gap among Chinese middle school students. *The Yale Undergraduate Research Journal*, 1(1).

Appendix – Same-Sex Teacher Effects

Appendix A: Supplementary Information about the Meta-Analysis: Data Collection

Research team: The data collection was carried out by a team of four predoctoral researchers (Anna Valyogos, Matt Bonci, Timo Haller, and Ana Bras) under the supervision of Ulf Zölitz at the University of Zürich.

Databases and keywords: For our meta-analysis data collection, we searched Google Scholar, Web of Science (WoS), as well as preregistered trials at https://www.socialscienceregistry.org/ (AEA RCT registry), cos.io, and https://researchregistry.com/. We used the search term combinations "same-sex, role model, test," "same-sex, role model, grade," "gender, role models, test," "gender, role models, grade," "same gender, teacher, role model, test," "same gender, teacher, role model, grade," "same gender, instructor, role model, test," and "same gender, instructor, role model, grade."

Process: Using the above-mentioned keyword combinations, we searched the results from the first ten pages of Google Scholar, the first 100 results from WoS, the first 200 results of CoS, and all results from the other two preregistered webpages. We did not use any date restrictions and included both peer-reviewed and non-peer–reviewed studies. For Google Scholar, WoS, and CoS, we scrapped data using the corresponding APIs, while for the Social Science Registry and Research Registry, we performed manual downloads. Using these processes, we identified a total of 5,277 potential same-sex teacher studies.

Next, we removed duplicates (keeping the latest version) within and across the five data sources, thus narrowing our dataset to 4,150 studies. After that, we dropped all studies preregistered on the Social Science Registry that matched our keywords but failed to include test scores or grades among their primary outcomes. We further filtered these results from SSR on study status, keeping only projects classified as complete and offering available results. After these preprocessing steps, we manually screened the title, abstract, and where necessary, the introduction of the remaining 1,838 studies and excluded those that did not match all preregistered inclusion criteria. We then performed full-text assessments of 174 articles to identify point estimates. Next, we removed all studies that did not allow us to calculate *standardized* same-sex teacher effects and standard errors (e.g., if they did not report the

standard deviation of the outcome). This left us with 24 studies reporting at least one same-sex teacher effect.

To avoid overlooking studies that did not use our keyword combinations, we identified studies that had more than 50 citations. For these highly cited papers, we collected the top ten papers that cited these seminal studies using the "cited by" functionality on Google Scholar. Through this process, we identified 130 additional potential same-sex teacher studies. Of those 130 studies, none reported a same-sex teacher effect. That left our final sample of 24 studies shown in Table A1. Figure A1 summarizes the data collection using the PRIMSA flow chart.

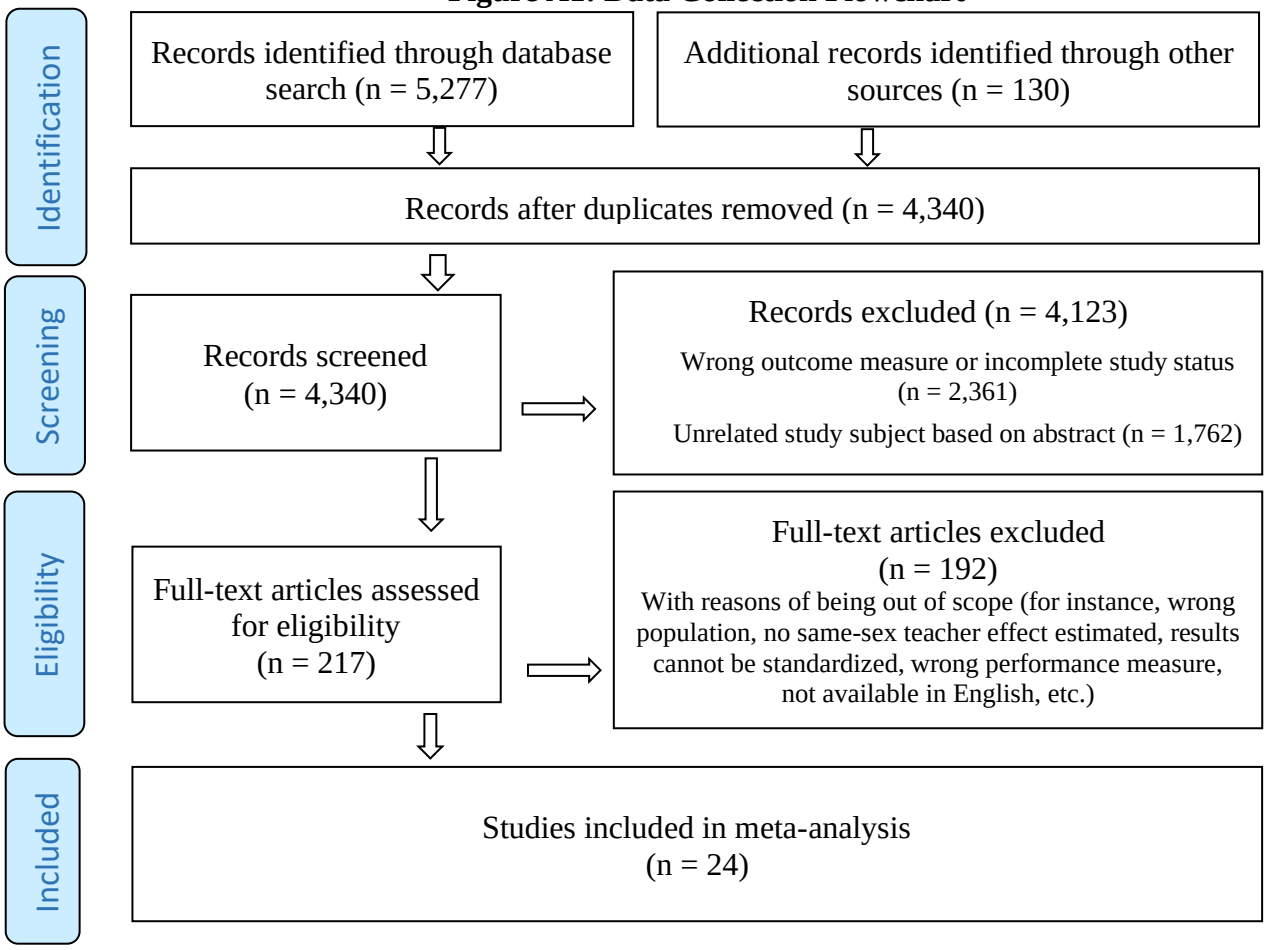
Table A1: List of Included Studies

Study	Journal	Country	Grades	N
Ammermüller & Dolton (2006)	ZEW Discussion Papers	England, USA	G4, G8	1,377
Antecol et al. (2015)	Journal of Labor Economics	USA	G1 to G5	1,624
Bhattacharya et al. (2020)	Research in Economics	India	G9	1,760
Buddin & Zamarro (2008)	RAND Education working papers	USA	G2 to G5	760,550
Carrington et al. (2008)	British Educational Research Journal	UK	G6	8,978
Coenen & Klaveren (2016)	European Sociological Review	Netherlands	G3 to G5	902
Dee (2007)	Journal of Human Resources	USA	G8	10,074
Eble & Hu (2020)	Economics of Education Review	China	G7, G9	7,977
Escardíbul & Mora (2013)	Journal of Education and Training	Spain	G9 to G12	2,073
Evans (1992)	Journal of Economic Education	USA	High school	1,251
Gong et al. (2018)	Journal of Labor Economics	China	G7, G9	18,202
Hermann & Diallo (2017)	Budapest Working Papers on the Labour Market	20 European countries	G8	3,244
Holmlund & Sund (2008)	Labour Economics	Sweden	High school	42,624
Hwang & Fitzpatrick (2021)	American Educational Research Association	USA	G3 to G8	650,036
Lee et al. (2019)	Journal of Development Studies	10 African countries	G6	17,801
Lim & Meer (2017)	Journal of Human Resources	South Korea	G9	24,231
Lim & Meer (2020)	Journal of Human Resources	South Korea	G7 to G12	9,026
Lindahl (2007)	IFAU – Working Papers	Sweden	G9	223,246
Mulji (2016)	Lund University Student Papers	Tunisia	G8	9,898
Muralidharan & Seth (2016)	Journal of Human Resources	India	G1 to G5	235,022

Neugebauer et al. (2011)	European Sociological Review	Germany	G4	2,269
Rakshit & Sahoo (2020)	Journal of Development Economics	India	G9	6,920
Xu & Li (2018)	Economics of Education Review	China	G7 to G9	7,472
Xu (2020)	The Yale Undergraduate Research Journal	China	G7, G9	18,996

Note: The sample size *N* refers to the median sample size for all coefficients of same-sex teacher effects on student performance in a given study.

Figure A1: Data Collection Flowchart



Coding: From each of the 24 studies, we recorded all same-sex teacher effects estimates on grades or test scores and their standard errors from the main paper and appendix. Besides recording these estimates and standard errors as they were reported in the paper, we standardized those estimates and standard errors that were not yet standardized by dividing them by the standard deviation of the outcome. In five out of 24 studies—Ammermüller and Dolton (2006), Dee (2007), Hermann and Diallo (2017), Hwang and Fitzpatrick (2021), and Neugebauer et al. (2011)—there were at least some same-sex teacher estimates that had to be reconstructed from separate regressions for girls and boys. Typically, these were separate

regressions of outcomes for boys and girls on a female teacher dummy. In these instances, we recovered the same-sex teacher effect as the difference between the female teacher effect for girls and the female teacher effect for boys. Recovering the standard error for this difference is impossible without making further assumptions. However, by assuming a zero covariance between both estimates, we recovered the standard error of the same-sex teacher effect as the square root of the sum of squared standard errors of the female teacher effect for girls and the female teacher effect for boys.

Furthermore, for each estimate we recorded the following information: study ID; citation (APA); abstract; link to publication (DOI or PDF); citation count as of November 25, 2022 (same as indicator for 100+ citations); main outcome (test score or grade); number of observations; effect size; standard error as reported; effect size in std. dev; standard error in std. dev; subject; estimation method; country; level of education; identification of main analysis; identifying variation in the main specification; first year of measurement; last year of measurement; coefficient specification type (the coefficient's type of interaction); subsample of students; single subject; most controlled estimate; heterogeneous effect (same as subsample of students); heterogeneity type (if the coefficient is from a subsample; for example, gender, single versus multiple teachers, native versus foreign students); included in appendix; model/table (the exact table/column location of the estimate); fixed effects; controls; comments.

For each paper, we classified one or multiple estimates as "most-controlled estimates." A study's most-controlled estimates are defined as those from the model specifications with the largest number of control covariates. For example, between an estimate that controls for student fixed effects and another that controls for student and teacher fixed effects, the latter is the most controlled. To define the most-controlled estimates, we also considered the level of within-group variation used, with smaller within-subgroup variation being more controlled. For example, between two estimates, one using school fixed effects and one using classroom fixed effects, the latter would be considered the most controlled. All our most-controlled estimates are still those targeting β_3 from Equation (1), either directly or from combining coefficients from split sample regressions on boy and girl outcomes separately. Finally, we added an updated citation count extracted from Google Scholar on November, 25, 2022, for all studies included in our final sample.

Consistency check: After the conclusion of the data collection, two predoctoral researchers not involved in the initial coding randomly selected five studies and replicated the data

collection. Any ambiguities identified through this process were resolved in discussions with a co-author on this project. We recorded whether a study was checked for consistency, whether inconsistencies were found, and how they were resolved. Out of 106 replicated estimates, we found two inconsistent estimates and three inconsistent standard errors, yielding an error rate of 4.72%. These false values were corrected in the base dataset. In addition to replicating the data collection for five studies, all estimates in the remaining 19 studies were cross-checked by a different research assistant. Any ambiguities identified through this process were resolved in discussions with a co-author on this project. This yielded an error rate of 7.65% and false values were corrected in the base dataset.

Preregistration and deviations

We preregistered our meta-analysis on osf.org. The complete preregistration is available at https://osf.io/rx2yv/. We adjusted the search process and the analysis as we learned more or encountered problems. In this section we record how we deviated from our preregistration and why.

Preregistered search terms: *"We will use the key words 'Same-sex role models,' 'same-sex teacher,' 'gender role model,' 'teacher gender,' 'instructor gender,' 'female instructor,' 'male instructor,' 'female teacher,' and 'male teacher' and require that the study must also mention either the word 'test-score' or 'grade.'"*

> **Things we did differently:** We used the following search terms: "same-sex role models," "same-sex teacher," "gender role model," "teacher gender," "instructor gender," "female instructor," "male instructor," "female teacher," and "male teacher" and a mention of either the word "test-score" or "grade" in each case and instead queried all sources using the eight keyword combinations outlined in section B1. We received many duplicate studies and therefore substituted "female" and "male" with "gender." We also received many irrelevant studies when not including "role model." We therefore restructured the search terms by linking preregistered key words. This led to an overall smaller, but more effective, set of search terms. Moreover, when querying the Research Registry, we also used "gender, role model" as an additional keyword combination, since our original search returned extremely few results for this source.

Preregistered description of initial search process: *"The RA will first identify studies by searching for the predetermined search terms in all the search platforms mentioned above. On Google Scholar, the RA will limit the search to the first 10 pages for each keyword."*

Things we did differently: We adapted our search process to the various functionalities offered by the data sources. For the AEA Registry, searching for our keywords proved unfeasible. We therefore downloaded all available data from the platform instead. Then, we used a Python script to filter for our keyword combinations in this downloaded metadata. Specifically, we required that at least one of our keywords appear within some subset of the "Title," "Abstract," "Intervention," and "Experimental design details" columns. This method resembles how the search would have presumably worked given a built-in search functionality, so we took these steps to imitate the preregistration as closely as possible.

In addition to limiting the search from Google Scholar to the first ten pages for each keyword group, we also limited the number of results looked at from WoS and CoS to the first 100 and first 200 results, respectively. Without such restrictions our data collection would have become intractable.

Preregistered removal of duplicates: *"Following this initial search, the RA will remove any duplicate studies and screen the titles and abstracts in accordance with the above criteria. At this stage, the RA will record studies that do not clearly fall outside the domain of our criteria in a spreadsheet. In cases of doubt, the RA will not exclude the study at this stage."*

Things we did differently: Duplicate removal across the five different sources was often challenging; therefore, some duplicates were only identified and dropped during the first screening stage. Our final sample is unaffected by this deviation; it only implies that the same article might have been screened multiple times.

In the initial screenings, if either the title or abstract of a study were unavailable or offered insufficient details, RAs extended their focus beyond the preregistration and read the introduction of the study as well. Again, this deviation had no impact on our final sample of relevant articles—it only influenced the stage at which an unrelated study was excluded.

Preregistered recording of information from initial screening: "*For each study that survives this initial screening, the RA will record the following information:*

1. *Date of search*
2. *Citation (APA)*
3. *Link to publication (DOI or pdf)"*

Things we did differently: We only collected the citation (in APA format) and publication link (DOI or pdf) for those studies that passed the full-text screening stage. We made this deviation for efficiency reasons, as significantly more studies than we had anticipated (1,838 studies altogether) passed the initial screening stage and required full-text assessment by the RAs.

Preregistered coding: *The RA will take a closer look at the studies recorded in the prescreened spreadsheet. If studies do not meet our three inclusion criteria, the RA will add why the studies should be excluded to the spreadsheet. To resolve ambiguities, the RA will consult with one of the co-authors on this project. For studies that do meet our criteria, the RA will add the following information to the spreadsheet:*

1. *Type of main outcome (Test score or grade)*
2. *Number of observations for main results*
3. *Record one main effect, as identified by authors. For this effect, record:*
    a. *Effect size as reported*
    b. *Standard error as reported*
    c. *Effect size in standard deviations of the outcome*
    d. *Standard error in standard deviations of the outcome*
    e. *Subject (e.g., math)*
    f. *Country where the study takes place*
    g. *Level of education (e.g., grade 8)*
4. *Identification of main analysis (e.g. experiment, natural experiment, observational)*
5. *Identifying variation in the main specification (e.g., between students, within schools, within classrooms)*
6. *Data first year of measurement*
7. *Data last year of measurement*
8. *Indicator for 100+ citations.*

If there is not one clear main effect, the RA will record multiple effect sizes from the main specification. For example, if a study shows separate same-sex teacher effects from three different countries but not one joint same-sex teacher effect from all countries, we will record all three country-level same-sex teacher effects."

Things we did differently: Instead of coding only a few main estimates, we coded *all* same-sex teacher estimates from each relevant study's main text and appendix. We decided to expand the data collection to all estimates to be more thorough. In addition to the information listed above, we also recorded the following information: citation

count as of November 25, 2022 (same as indicator for 100+ citations), main outcome (test score or grade), number of observations, effect size, standard error as reported, effect size in std. dev, subject, country, level of education, identification of main analysis, identifying variation in the main specification, first year of measurement, and last year of measurement.

Five out of 24 studies had at least some same-sex teacher estimates that had to be reconstructed from separate regressions for girls and boys. In these instances, we recovered the same-sex teacher effect as the difference between the female teacher effect for girls and the female teacher effect for boys (see above). We also recovered the standard error of the difference as the square root of the sum of squared standard errors of the boy and girl estimates.

Preregistered identification of overlooked studies: *"To avoid overlooking studies, the RA will go through all papers in the spreadsheet with more than 100 citations and use Google Scholar to (1) check for citing studies and (2) check for related articles using the Google Scholar embedded functionality. Any relevant study identified through this secondary search will be coded as described in step 2."*

Things we did differently: We extended our data collection by using 50 as our minimum citation cut-off instead of the preregistered threshold of 100. We decided to lower this requirement as we found fewer relevant studies than expected and noticed that numerous studies had a citation count above 50 but below 100. We leveraged Google Scholar's "citing studies" functionality but not its "related articles" option to check for potentially overlooked studies because the API returned these results more readily.

Preregistration of main effect recording:

"We will report:

- *Meta regression results using all studies in our spreadsheet. We will estimate this model using a random-effect (RE) meta-regression. We will use the DerSimonian and Laird (1986) method to estimate the weights unless this becomes analytically impracticable; else will use more standard restricted maximum likelihood methods. This estimate will be produced using the meta regress, random(dlaird) functionality in the Stata meta-analysis command suite."*

Things we did differently: We decided to estimate a three-level random effects model (Harrer et al., 2021, Ch. 10). This model allows for true same-sex teacher effects to differ by study and accounts for the dependence of same-sex teacher effect estimates within each study. We estimated it via the restricted maximum likelihood and applied the Hartung–Knapp adjustment.

Preregistration of publication bias correction:

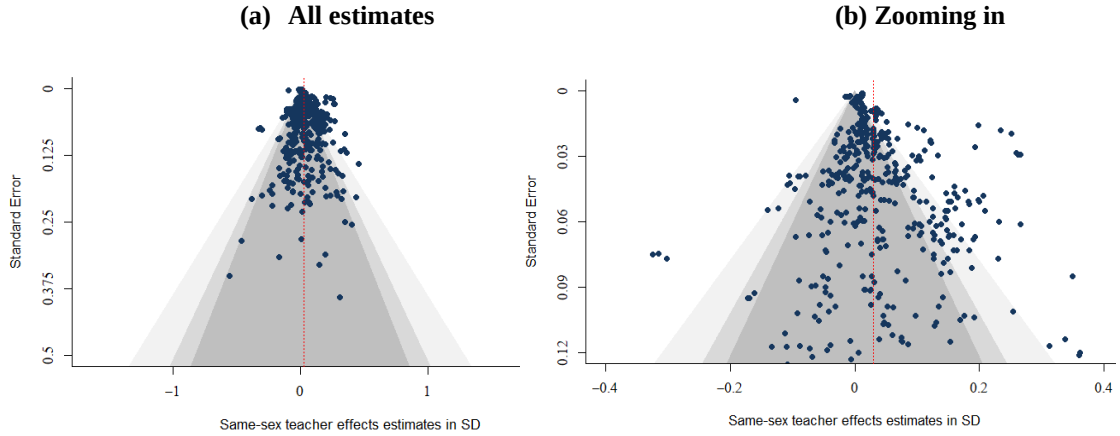
"We will report:

- *Estimates of the probability of publication for negative and significant results, negative and insignificant results, and positive and insignificant results (all relative to probability of publishing positive and significant results which is normalized to 1), as well as the estimate of the mean "latent study" same-sex teacher effect (μ) corrected for publication bias. These estimates will be produced using Andrews and Kasy (2019) method and estimated with a 1.96 cutoff for p(.) assuming that the latent effects are normally distributed."*

Things we did differently: In addition to the analysis in our pre-analysis plan we also implemented the 11 other publication bias correction methods shown in Table A3. We deviated from the pre-analysis plan because we found that results were quite sensitive to the exact correction method used.

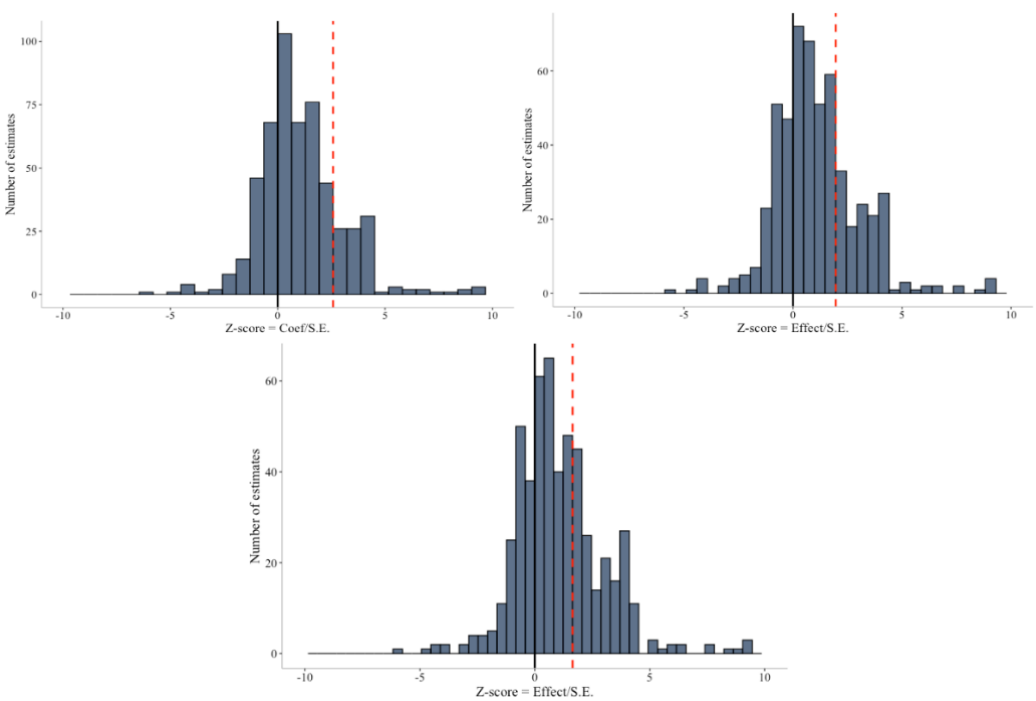
Meta-Analysis Supplementary Results

Figure A2: Funnel Plot of All Same-Sex Teacher Effect Estimates



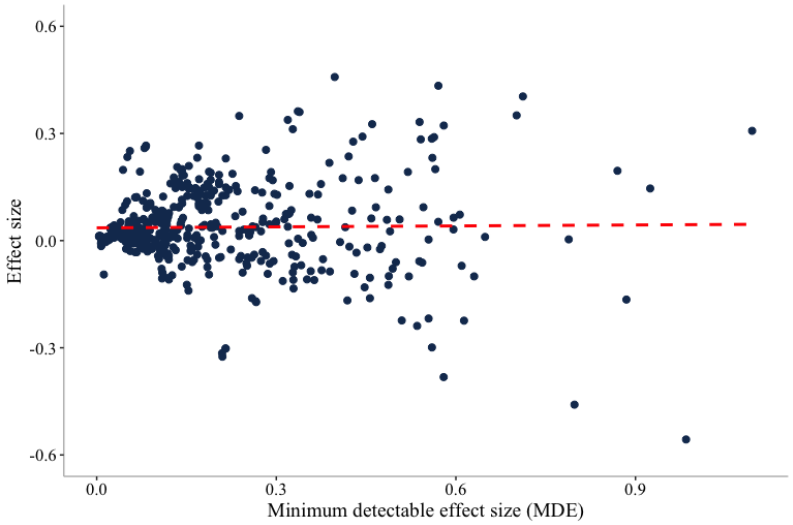
Notes: This figure shows a scatterplot of 535 same-sex teacher effects estimates from all 24 studies on the x-axis, with their standard error on the y-axis. To increase readability, this figure excludes three outlying same-sex teacher estimates of size 1.15, 2.07, and 0.92 SD with a standard error of 5.03, 5.42 and 6.83 SD, respectively. The gray shaded areas mark the traditional thresholds for statistical significance with 90%, 95%, and 99% confidence. The vertical dotted line marks our estimated average same-sex teacher effect of 0.030 SD.

Figure A3: *Z*-score Distribution with Critical Value with 90%, 95%, and 99% Two-Sided Critical Values Marked



Notes: The figures show z-scores of 534 same-sex teacher effects estimates from all 24 studies. These are all z-scores except for four outlier values (with z-scores of –22.39, 12.61, 12.68, and 12.79), which we excluded to make the figure more readable. The top, middle, and bottom figures include vertical dashed lines at 2.576, 1.960, and 1.645. These are the critical values for a two-sided test of statistical significance based on the Normal distribution with 90%, 95%, and 99% confidence. The top, middle, and bottom histograms use a bin width of 0.645, 0.490, and 0.410 to facilitate the detection of heaping at the relevant significance thresholds.

Figure A4: Minimum Detectable Effect Size (MDE) of Same-Sex Teacher Estimates



Notes: The red dashed line shows the linear regression fit between all 538 same-sex teacher effect estimates (y-axis) and their corresponding ex-post MDE size (x-axis). Each dot represents one same-sex teacher effect estimate. To increase readability, this figure excludes three outlying same-sex teacher estimates of size 1.15, 2.07, and 0.92 SD with MDEs of 14.10, 15.19, and 19.13 SD. The slope of the dashed line is 0.079, with a standard error of 0.003 clustered at the study level. Excluding the three outliers not shown in the figure yields a slope of 0.009 with a standard error of 0.097.

Table A2: Meta-Regression of Same-Sex Teacher Effect Estimates

Panel A: Identification (base = *Experimental*)				
	Coef.	Std. err.	95% CI	
Intercept	-0.009	(0.042)	-0.091	0.073
Observational/Natural experiment	0.043	(0.044)	-0.043	0.129
Panel B: Continent (base = *Africa*)				
	Coef.	Std. err.	95% CI	
Intercept	0.094	(0.043)	0.009	0.179
Asia	-0.051	(0.048)	-0.146	0.044
Europe	-0.053	(0.049)	-0.148	0.043
North America	-0.128	(0.050)	-0.226	-0.031
Panel C: School level (base = *Secondary*)				
	Coef.	Std. err.	95% CI	
Intercept	0.051	(0.016)	0.019	0.083
Primary	-0.058	(0.024)	-0.106	-0.011
Both	-0.047	(0.025)	-0.097	0.002
Panel D: Outcome (base = *Grades*)				
	Coef.	Std. err.	95% CI	
Intercept	-0.008	(0.037)	-0.081	0.065
Test scores	0.041	(0.037)	-0.032	0.113
Panel E: Single 3-LM Regression				
	Coef.	Std. err.	95% CI	
Intercept	0.137	(0.104)	-0.067	0.341
Identification (base = Experimental)				
Observational/Natural experiment	-0.063	(0.068)	-0.1967	0.070
Continent (base = Africa)				
Asia	-0.096	(0.067)	-0.229	0.036
Europe	-0.033	(0.066)	-0.162	0.096
North America	-0.144	(0.067)	-0.276	-0.013
School level (base = Secondary)				
Primary	-0.094	(0.033)	-0.159	-0.03
Both	-0.083	(0.034)	-0.149	-0.017
Outcome (base = Grades)				
Test scores	0.068	(0.041)	-0.013	0.149
Test for significance of all moderators (*p*-value)	0.001			
Test for residual heterogeneity (*p*-value):	<0.0001			
Variance components (τ)				
Between studies	0.0068			
Within studies	0.0003			

Notes: Coefficients from a series of three-level meta-regressions of same-sex teacher effects estimates on grades and test scores, estimated using the meta package in R. Our sample contains all 538 same-sex teacher estimates from 24 studies. The three levels account for nested interdependence while pooling information of individual participants into the various same-sex teacher effects in primary studies (level 1), pooling all same-sex teacher effects in each primary study (level 2), and pooling primary study same-sex teacher effects into an overall same-sex teacher effect (level 3). Panels A, B, C, and D produce bivariate regressions for each of the categories of interest, whereas Panel E shows coefficients for a single 3-LM Regression with all categories of interest as independent variables. All moderators are coded at the primary study level. Standard errors are in parentheses.

Table A3: Same-Sex Teacher Effect Meta-Analysis Estimates Corrected for Publication Bias

Estimation method	Significance threshold for selection	Average effect	Std. err.	95% Confidence Interval		Standard deviation of effect
3-level REML	-	0.030	(0.013)	0.005	0.055	0.058
Trim and Fill	-	0.012	(0.004)	0.004	0.020	0.077
PET-PEESE	-	0.006	(0.012)	-0.017	0.029	0.049
Limit-Meta	-	0.012	(0.197)	-0.373	0.397	0.058
3-Parameter Selection	10%	0.029	(0.004)	0.021	0.038	0.049
3-Parameter Selection	5%	0.035	(0.005)	0.026	0.044	0.050
3-Parameter Selection	1%	0.038	(0.004)	0.029	0.047	0.051
Andrews & Kasy (t)	10%	0.012	(0.003)	0.006	0.018	0.015
Andrews & Kasy (t)	10%, 5%	0.012	(0.003)	0.006	0.018	0.015
Andrews & Kasy (t)	10%, 5%, 1%	0.011	(0.003)	0.005	0.017	0.015
Andrews & Kasy (N)	10%	-0.027	(0.015)	-0.056	0.002	0.074
Andrews & Kasy (N)	10%, 5%	-0.028	(0.017)	-0.061	0.005	0.078
Andrews & Kasy (N)	10%, 5%, 1%	-0.039	(0.022)	-0.082	0.004	0.088

Notes: As a benchmark, the 3-level restricted maximum likelihood (REML) shows the estimated same-sex teacher effect without correcting for publication bias as shown and described in Section 2.2. All other estimates apply different publication bias corrections. Trim and Fill: We use the inverse variance method for pooling estimates. We use the REML method to estimate the variance and apply the Knapp–Hartung adjustment to account for the uncertainty in the estimation of the between-study heterogeneity. PET-PEESE: we use estimates from the Precision-Effect Test (PET) model rather than from the Precision-Effect Estimate with Standard Errors (PEESE) model because the intercept in the PET model is not statistically significantly different from zero at the 5% level (p-value = 0.3055) using one-sided *t*-test. We use the inverse variance method for pooling estimates. We use the REML method to estimate the variance and apply the Knapp–Hartung adjustment to account for the uncertainty in the estimation of the between-study heterogeneity. Limit-Meta: Uses 3-level REML as input. 3-Parameter Selection: We use 0.05, 0.025, and 0.01 as jumps in the publication probability function. REML estimator of the standard deviation of the effect size and the standard deviation of the effect size. Andrews and Kasy: We use the Andrews and Kasy (2019) correction method, assuming the effects are either t-distributed or normally distributed. We estimate separate corrections for cutoffs at the 0.05, 0.05, and 0.025, and 0.05, 0.025, and 0.01 significance levels for both positive and negative effects. We allow the probability of publication bias to be asymmetric. We produce estimates using Kasy's App: https://maxkasy.github.io/home/metastudy. Other correction methods: Andrews and Kasy (2019)'s non-parametric GMM method did not produce a useful corrected estimate due to singularity issues. We also tried various continuous selection models assuming underlying beta, half-normal, and logistic publication probability distributions, which also did not yield useful estimates due to non-convergence issues.

Table A4: Same-sex Teacher Effect Estimates Corrected for Publication Bias, Using the "Most Controlled" Set of Estimates

Estimation method	Significance threshold for selection	Average effect	Standard error	95% Confidence Interval		Standard deviation of effect
3-level REML	-	0.031	(0.014)	0.005	0.060	0.06
Trim and Fill	-	0.007	(0.004)	-0.002	0.015	0.057
PET-PEESE	-	0.004	(0.001)	-0.015	0.023	0.035
Limit-Meta	-	0.031	(0.161)	-0.285	0.347	0.060
3-Parameter Selection	10%	0.024	(0.005)	0.015	0.034	0.037
3-Parameter Selection	5%	0.033	(0.006)	0.022	0.044	0.041
3-Parameter Selection	1%	0.035	(0.006)	0.024	0.047	0.042
Andrews & Kasy (t)	10%	0.007	(0.002)	0.003	0.011	0.007
Andrews & Kasy (t)	10%, 5%	0.007	(0.001)	0.005	0.009	0.007
Andrews & Kasy (t)	10%, 5%, 1%	0.008	(0.001)	0.006	0.010	0.007
Andrews & Kasy (N)	10%	-0.017	(0.014)	-0.044	0.010	0.056
Andrews & Kasy (N)	10%, 5%	-0.011	(0.015)	-0.040	0.018	0.065
Andrews & Kasy (N)	10%, 5%, 1%	-0.012	(0.019)	-0.049	0.025	0.086

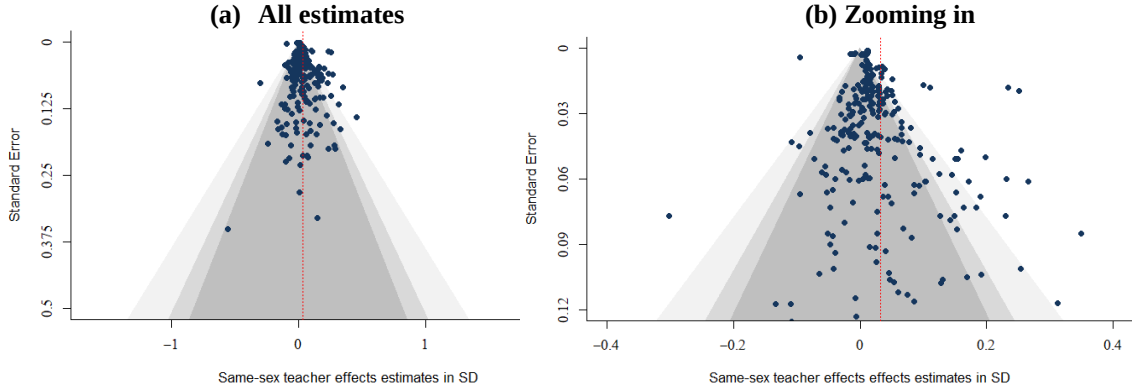
Notes: The "most controlled" estimates are defined as those from model specifications using the largest amount of control covariates and narrowest within-group variation. From these estimates we additionally exclude "first difference" estimates, defined as effects of same-sex teachers on test score or grade gains (i.e., the difference between test scores or grades at two points in time for each student). This latter restriction only affects one estimate from Dee (2007). Our resulting subset of most controlled estimates includes 297 estimates from our 24 selected studies. As benchmark, 3-level restricted maximum likelihood (REML) shows the estimated same-sex teacher effect without correcting for publication bias as shown and described in Section 2.2. All other estimates apply different publication bias corrections. Trim and Fill: We use the inverse variance method for pooling estimates. We use the REML method to estimate the variance and apply the Knapp–Hartung adjustment to account for the uncertainty in the estimation of the between-study heterogeneity. PET-PEESE: we use estimates from the Precision-Effect Test (PET) model rather than from the Precision-Effect Estimate with Standard Errors (PEESE) model because the intercept in the PET model is not statistically significantly different from zero at the 5% level (p-value = 0.3055) using one-sided *t*-test. We use the inverse variance method for pooling estimates. We use the REML method to estimate the variance and apply the Knapp–Hartung adjustment to account for the uncertainty in the estimation of the between-study heterogeneity. Limit-Meta: Uses 3-level REML as input. 3-Parameter Selection: We use 0.05, 0.025, and 0.01 as jumps in the publication probability function. REML estimator of the standard deviation of the effect size and the standard deviation of the effect size. Andrews and Kasy: We use the Andrews and Kasy (2019) correction method, assuming the effects are either t-distributed or normally distributed. We estimate separate corrections for cutoffs at the 0.05, 0.05, and 0.025, and 0.05, 0.025, and 0.01 significance levels for both positive and negative effects. We allow the probability of publication bias to be asymmetric. We produce estimate using Kasy's App: https://maxkasy.github.io/home/metastudy. Other correction methods: Andrews and Kasy (2019)'s non-parametric GMM method did not produce a useful corrected estimate due to singularity issues. We also tried various continuous selection models assuming underlying beta, half-normal, and logistic publication probability distributions, which also did not yield useful estimates due to non-convergence issues.

Table A5: Meta-Regression of Same-Sex Teacher "Most Controlled" Estimates

Panel A: Identification (base = *Experimental*)				
	Coef.	Std. err.	95% CI	
Intercept	-0.003	(0.047)	-0.095	0.088
Observational/Natural experiment	0.039	(0.049)	-0.057	0.136
Panel B: Continent (base = *Africa*)				
	Coef.	Std. err.	95% CI	
Intercept	0.097	(0.041)	0.017	0.177
Asia	-0.042	(0.046)	-0.132	0.048
Europe	-0.078	(0.047)	-0.170	0.014
North America	-0.112	(0.048)	-0.207	-0.018
Panel C: School level (base = *Secondary*)				
	Coef.	Std. err.	95% CI	
Intercept	0.047	(0.018)	0.012	0.082
Primary	-0.040	(0.028)	-0.095	-0.014
Both	-0.026	(0.029)	-0.083	0.003
Panel D: Outcome (base = *Grades*)				
	Coef.	Std. err.	95% CI	
Intercept	0.009	(0.042)	-0.074	0.093
Test scores	0.025	(0.043)	-0.060	0.109
Panel E: Single 3-LM Regression				
	Coef.	Std. err.	95% CI	
Intercept	0.077	(0.090)	-0.101	0.255
Identification (base = Experimental)				
Observational/Natural experiment	0.010	(0.057)	-0.103	0.122
Continent (base = Africa)				
Asia	-0.051	(0.054)	-0.157	0.055
Europe	-0.072	(0.053)	-0.175	0.032
North America	-0.112	(0.054)	-0.218	-0.006
School level (base = Secondary)				
Primary	-0.027	(0.034)	-0.093	0.040
Both	-0.012	(0.035)	-0.081	0.056
Outcome (base = Grades)				
Test scores	0.024	(0.046)	-0.066	0.113
Test for significance of all moderators (p-value)	0.001			
Test for residual heterogeneity (p-value):	<0.0001			
Variance components (τ)				
Between studies	0.0068			
Within studies	0.0003			

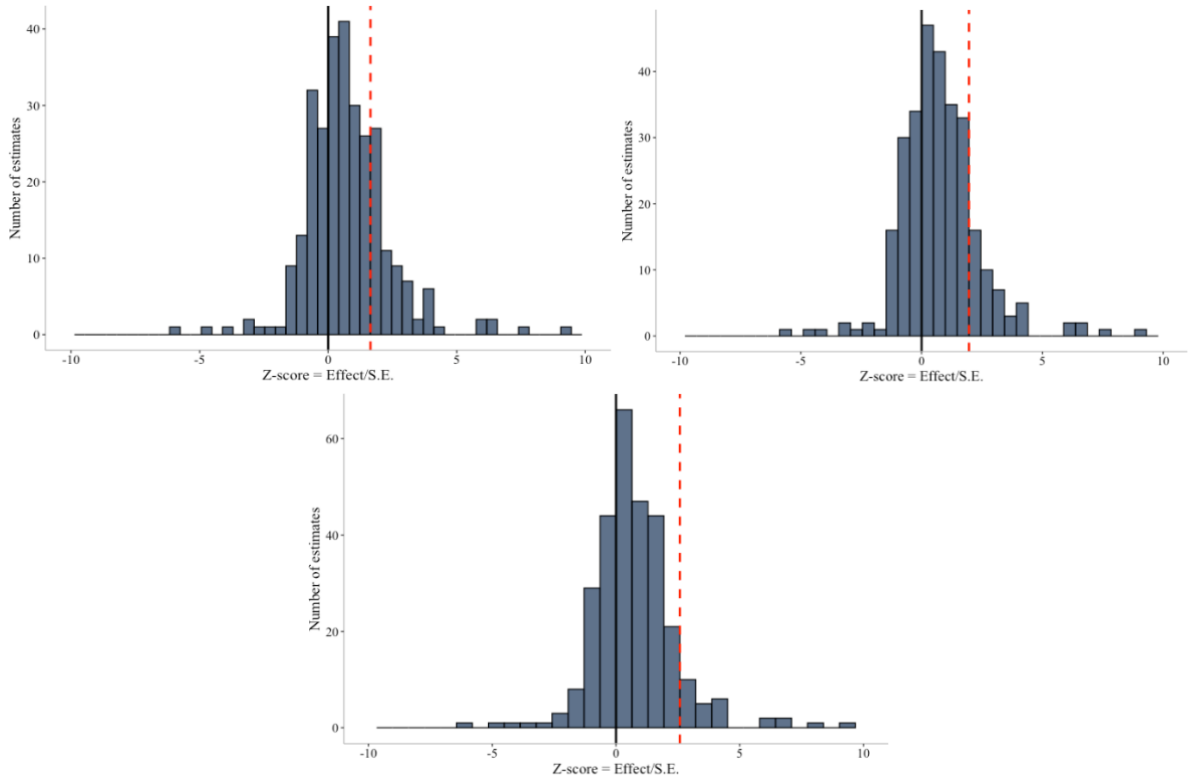
Notes: Coefficients from a series of three-level meta-regressions of same-sex teacher effects estimates on grades and test scores, estimated using the meta package in R. Our sample contains the 297 most-controlled same-sex teacher effects estimates from all 24 studies. The three levels account for nested interdependence while pooling information of individual participants into the various same-sex teacher effects in primary studies (level 1), pooling all same-sex teacher effects in each primary study (level 2), and pooling primary study same-sex teacher effects into an overall same-sex teacher effect (level 3). Panels A, B, C, and D produce bivariate regressions for each of the categories of interest, whereas Panel E shows coefficients for a single 3-LM Regression with all categories of interest as independent variables. All moderators are coded at the primary study level. Standard errors are in parentheses.

Figure A5: Funnel Plot of "Most Controlled" Same-Sex Teacher Effect Estimates



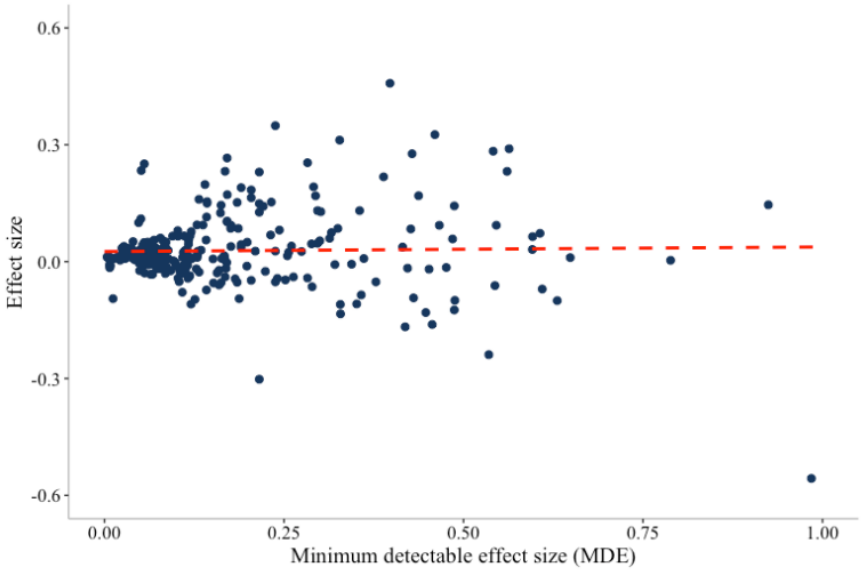
Notes: This figure shows a scatterplot of the 296 most-controlled same-sex teacher effects estimates from all 24 studies on the x-axis, with their standard error on the y-axis. To increase readability, this figure excludes one outlying same-sex teacher estimate of 2.07 SD with a standard error of 5.42 SD. The gray shaded areas mark the traditional thresholds for statistical significance at the 10%, 5%, and 1% level. The vertical dotted line marks our estimated average same-sex teacher effect of 0.032 SD in this sample.

Figure A6: *Z*-score Distribution of Most-Controlled Same-Sex Teacher Effect Estimates, with 90%, 95%, and 99% Two-Sided Critical Values Marked



Notes: This figure shows z-scores of the 297 most-controlled same-sex teacher effects estimates from all 24 studies. The top, middle, and bottom figures include vertical dashed lines at 2.576, 1.960, and 1.645. These are the critical values for a two-sided test of statistical significance based on the Normal distribution with 90%, 95%, and 99% confidence. The top, middle, and bottom histograms use a bin width of 0.645, 0.490, and 0.410 to facilitate the detection of heaping at the relevant significance thresholds.

Figure A7: Minimum Detectable Effect Size (MDE) Plot Most-Controlled Estimates



Notes: This red dashed line shows the linear regression fit between the 297 most-controlled same-sex teacher effects estimates from all 24 studies (y-axis) and their corresponding ex-post MDE size (x-axis). Each dot represents one same-sex teacher effect estimate. To increase readability, this figure excludes one outlying same-sex teacher estimate of size 2.07 SD with an MDE of 15.19 SD. The slope of the dashed line is 0.132, with a standard error of 0.005 clustered at the study level. Excluding the outlier not shown on the figure yields a slope of 0.012 with a standard error of 0.085.

Appendix B: Multi-Country Study Supplementary Material

PIRLS and TIMSS Sampling

TIMSS and PIRLS use the same two-stage stratified random sampling design and similar questionnaires of students, parents, teachers, and school principals. In each wave, each country's national research coordinator first samples roughly 150 to 200 schools and interviews school principals. In the second stage, they randomly sample one to three classrooms in the target grade (respectively 4th grade for PIRLS, and 4th or 8th grade for TIMSS) within each selected school, depending on school size.[17] Each cross-section and country-specific sample is representative of children in the survey target grade. The target sample size per country and wave is 5,000 children; however, countries often decide to use more children in their sample. The target response rate is 85% of schools, 95% of classrooms, and 85% of children in classrooms; country survey teams use an additional sample of replacement schools, classrooms, or students whenever those response rates are below target.

PIRLS and TIMSS Plausible Values

The test answers for each student are transformed into estimates of a student's subject-specific ability. For each student, the IEA calculates five *plausible values* per subject. These are different estimates of the student's latent subject-specific ability based on their answers. Each of the five sets of plausible values is standardized by setting the unweighted mean of all countries that participated in TIMSS 1995 to 500 points and setting their standard deviation to 100. To enable measurement of trends over time, achievement data from later TIMSS assessments (e.g., TIMSS 2011) were transformed to these same metrics. This was done by concurrently scaling the data from each successive assessment with the data from the previous assessment—a process known as concurrent calibration—and applying linear transformations to place the results from each successive assessment on the same scale as the results from the previous assessment (see TIMSS 2019 Technical Report, Chapter 11, p. 558). To simplify our analysis, we use the average of all five plausible values for each student as our main outcome variable. For simplicity, and following other studies that have worked with these data, we use

[17] Some countries did not identify 4th and 8th grades as adequate target grades. In England and New Zealand, children begin primary school at an early age. Therefore, these countries administered the TIMSS 4th grade assessment in the fifth year of schooling. The TIMSS 8th grade assessment for England and New Zealand was administered in the ninth year of schooling. Norway chose to assess its 5th and 9th grades to obtain better comparisons with Sweden and Finland. To provide a better match with the demands of the assessments, South Africa and Turkey administered the test to 5th and 9th graders (see TIMSS 2019 Technical Report, Chapter 9, p. 196).

the term "students' test score" to refer to the average of these five values. Previous work using TIMSS and PIRLS data shows that regression analysis results are generally robust to this simplification (e.g., de Gendre, et al., 2024; Bietenbeck and Collins, 2023).

Construction of Base Dataset Using PIRLS and TIMSS

The base dataset contains all available data at the student-assessment level after removing duplicate observations and removing observations from country-study-grade-wave combinations that suffered implementation issues. We construct this base dataset by first merging the student and teacher data for each study, wave, and country (e.g., TIMSS 1999 Armenia) and appending all country files per study wave (e.g., all TIMSS 1999). At this point, we systematically prepared and harmonized our variables of interest in each study-wave file to ensure that all variables in our estimation sample were comparable across waves and across TIMSS and PIRLS. We then appended all study-wave files into one large file per study (e.g., TIMSS), before appending the TIMSS and PIRLS files.

In total, we excluded 19 out of 731 country-study-grade-wave combinations because of survey implementation issues. We excluded two country-grade-wave cases with empty student or teacher background files, such that student or teacher sex cannot be recovered; this issue occurred in Bulgaria and in South Africa for 8th grade in wave 1995. We also excluded 17 country-grade-wave combinations in which students could not be linked to their teachers and classroom. Those issues took place in the first wave of TIMSS 8th grade in 1995 and were due to miscoding in some schools of the key variable linking students to teachers. Those survey implementation issues are documented in the 1995 user guide as "implementation issues," and led to duplicate student observations with multiple test scores because student identifiers and student-teacher linking codes are miscoded in the Student-Teacher Linkage files (AST* and BST* files). We analyzed those files for all countries, grades, and waves. We confirmed the implementation issues reported for 12 country-grade-wave cases in the 1995 documentation, and we excluded entire country-waves from our analyses when issues affected 97% to 98% of student observations in 8th in those countries (Belgium Flanders, Cyprus, Czech Republic, Hungary, Iran, Israel, Latvia, Lithuania, New Zealand, Romania, Slovak Republic, and Slovenia). In addition, we also excluded five country-grade-wave cases from our analyses where we found evidence of similar implementation issues affecting more than 10% of student observations (Canada 4th and 8th grades, affecting 59% of observations; Germany 8th grade,

affecting 66% of observations; England 8th grade, affecting 18% of observations; and Belgium Flanders 8th grade, affecting 16% of observations).

For rarer instances of duplicate student observations affecting 0.05% 6.5% of student observations in TIMSS, we simply dropped student observations with duplicates. This concerns 15 country-grade-wave cases in 1995 (Australia 8th grade, Austria 8th grade, Colombia 8th grade, Cyprus 4th grade, Denmark 8th grade, Greece 4th and 8th grades, Israel 4th grade, Kuwait 4th grade, Portugal 4th grade, Sweden 8th grade, Switzerland 8th grade , United States 4th and 8th grades, and Scotland 8th grade) and two cases in 1999 (England 8th grade. affecting 3.5% of observations, and Finland 8th grade, affecting 0.01% of observations). We document all these exclusions in our Stata do-file. We will make this do-file as well as all our estimation do-files available to the public upon acceptance of the paper.

In the base dataset, job preference has a mean of 2.55 and a standard deviation of 1.02, subject enjoyment has a mean of 3.11 and a standard deviation of 0.91, and subject confidence has a mean of 3.12 and a standard deviation of 0.82. We use those means and standard deviations to standardize these three variables for our analysis (see Section 4).

Subject-, Student-, and Teacher-Level Heterogeneity

Subject heterogeneity: We test whether our results differ by subject by estimating same-sex teacher effects in separate samples for students' math, science, and reading scores with our school fixed effects specification. This analysis is not possible with more-restrictive fixed effects as these require within-subject variation by classroom or student. Our results show some subject heterogeneity (see Appendix Table B2). Same-sex teacher effects are somewhat larger in math than in science (0.019 SD compared to 0.012 SD) and statistically indistinguishable from zero for reading (0.003 SD). These differences in effects also explain why restricting our estimation sample leads to slightly larger same-sex teacher estimates: because we cannot include reading scores in our preferred specification, our sample is limited to subjects (math and science) for which we see larger same-sex teacher effects.

Student- and teacher-level heterogeneity: We test whether our results differ by student and teacher characteristics by estimating same-sex teacher effects using our preferred specification separately for different subsamples of students and teachers. Table B1 shows little heterogeneity along any of the dimensions we consider. All point estimates are small and precisely estimated.

Table B1: Student- and Teacher-Level Heterogeneity for Test Scores Estimates

			95% Confidence Interval		
	Average Effect	SE	LB	UB	N
Dependent variable: **Std. Test Scores**					
Student characteristics					
4th grade	0.0039	(0.0072)	-0.0102	0.0180	160,480
8th grade	0.0169	(0.0036)	0.0098	0.0240	1,451,717
Foreign	0.0185	(0.0157)	-0.0123	0.0492	106,478
Native	0.0152	(0.0035)	0.0083	0.0220	1,405,458
University-educated parent(s)	0.0162	(0.0084)	0.0003	0.0327	428,801
No university-educated parent(s)	0.0180	(0.0049)	0.0084	0.0276	732,604
Two-parent household	0.0124	(0.0112)	-0.0095	0.0340	213,632
No two-parent household	0.0165	(0.0130)	-0.0090	0.0420	113,013
Teacher characteristics					
15+ years of experience	0.0132	(0.0072)	-0.0009	0.0273	602,999
Less than 15 years of experience	0.0187	(0.0051)	0.0087	0.0287	599,835
Post-graduate degree	0.0119	(0.0112)	-0.0100	0.0339	303,810
No post-graduate degree	0.0128	(0.0037)	0.0055	0.0200	1,025,066
Education major	0.0076	(0.0049)	-0.0020	0.0172	555,223
Not an education major	0.0188	(0.0064)	0.0063	0.0313	452,290
Majored in subject	0.0154	(0.0035)	0.0085	0.0222	1,218,558
Did not major in subject	0.0027	(0.0142)	-0.0251	0.0305	44,515
Classroom has 30+ students	0.0158	(0.0046)	0.0068	0.0248	433,798
Classroom has less than 30 students	0.0133	(0.0045)	0.0045	0.0221	1,178,387

Notes: This table shows estimated same-sex teacher effects from regressions of standardized test scores and job preferences on a $\text{FemaleStudent}_i \times \text{FemaleTeacher}_j$ interaction term, student fixed effects, teacher fixed effects, as well as other control variables from our preferred specification (see Section 4) for the different subsamples indicated on the left of the table.

Table B2: Subject Heterogeneity

	Math	*Science*	*Reading*
Panel A	Std. Dep. Var.: **Test scores**		
Same-sex teacher effect	0.0188	0.0117	0.0026
	(0.0032)	(0.0037)	(0.0097)
Male–female score gap	0.033	0.053	-0.153
R^2	0.62	0.60	0.39
Countries	85	85	56
Observations	845,647	834,934	79,541
Panel B	Std. Dep. Var.: **Job preferences**		
Same-sex teacher effect	0.0465	0.0769	
	(0.0060)	(0.0069)	
Male–female score gap	0.198	0.096	
R^2	0.18	0.23	
Countries	72	71	
Observations	511,263	505,472	
Panel C	Std. Dep. Var**.: Subject enjoyment**		
Same-sex teacher effect	0.0687	0.0947	0.0238
	(0.0048)	(0.0050)	(0.0162)
Male–female score gap	0.078	0.087	-0.359
R^2	0.22	0.24	0.12
Countries	85	85	56
Observations	818,346	814,662	77,443
Panel D	Std. Dep. Var.: **Subject confidence**		
Same-sex teacher effect	0.0432	0.0687	-0.0024
	(0.0048)	(0.0050)	(0.0136)
Male–female score gap	0.133	0.102	-0.079
R^2	0.18	0.21	0.08
Countries	85	85	56
Observations	823,421	814,854	77,551

Notes: This table shows estimated same-sex teacher effects from regressions of the outcome variable shown in the first row of each panel on a $\text{FemaleStudent}_i \times \text{FemaleTeacher}_j$ interaction term and other control variables from our school fixed effects specification (see Section 4). In this specification we can identify same-sex teacher effects on test scores, subject enjoyment, and subject confidence in 89 countries and on job preferences in 72 countries. However, math and science test scores and data on enjoyment and confidence are not available in four countries (Belize, Luxembourg, Macao, and Trinidad and Tobago). There is also no identifying within-school variation in same-sex science teachers in Honduras in our estimation sample. Reading test scores are also not available in 33 of these 89 countries. Standard errors clustered at the classroom level are in parentheses.

Plausibility of the Normality Assumption

One important assumption behind our results in Section 6.3 is that the true same-sex teacher effects underlying our estimates are normally distributed. We test how plausible this assumption is following Jackson and Mackevicius (2024) by implementing tests of normality.[18] These tests take as input standardized country-level same-sex teacher effects, $\widehat{\theta}_c^S$, standardized as:

$$\widehat{\theta}_c^S = \frac{\widehat{\theta}_c - \widehat{\Theta}_{-c}}{\sqrt{\hat{\tau}^2 + se_c^2 + se_{\widehat{\Theta}_{-c}}^2}},$$

where $\widehat{\theta}_c$ and se_c^2 are the same-sex teacher effect and standard error estimates for country c, $\widehat{\Theta}_{-c}$ and $se_{\widehat{\Theta}_{-c}}^2$ are meta-estimates of the mean of all same-sex teacher effects excluding country c (obtained with the random effect model) and its standard error, and $\hat{\tau}^2$ is the random effects variance estimate of the true same-sex teacher effects, estimated using estimates from all countries. Under the null hypothesis that same-sex teacher effects are normally distributed, $\widehat{\theta}_c^S$ should be distributed standard normal. Building on this insight, our tests for normality take the form of: (1) a graphical Quantile-Quantile (Q-Q) plot of the quantiles of $\widehat{\theta}_c^S$ contrasted with the standard normal quantiles, and (2) a Shapiro–Wilk test where the null hypothesis is that the estimates $\widehat{\theta}_c^S$ are normally distributed.[19]

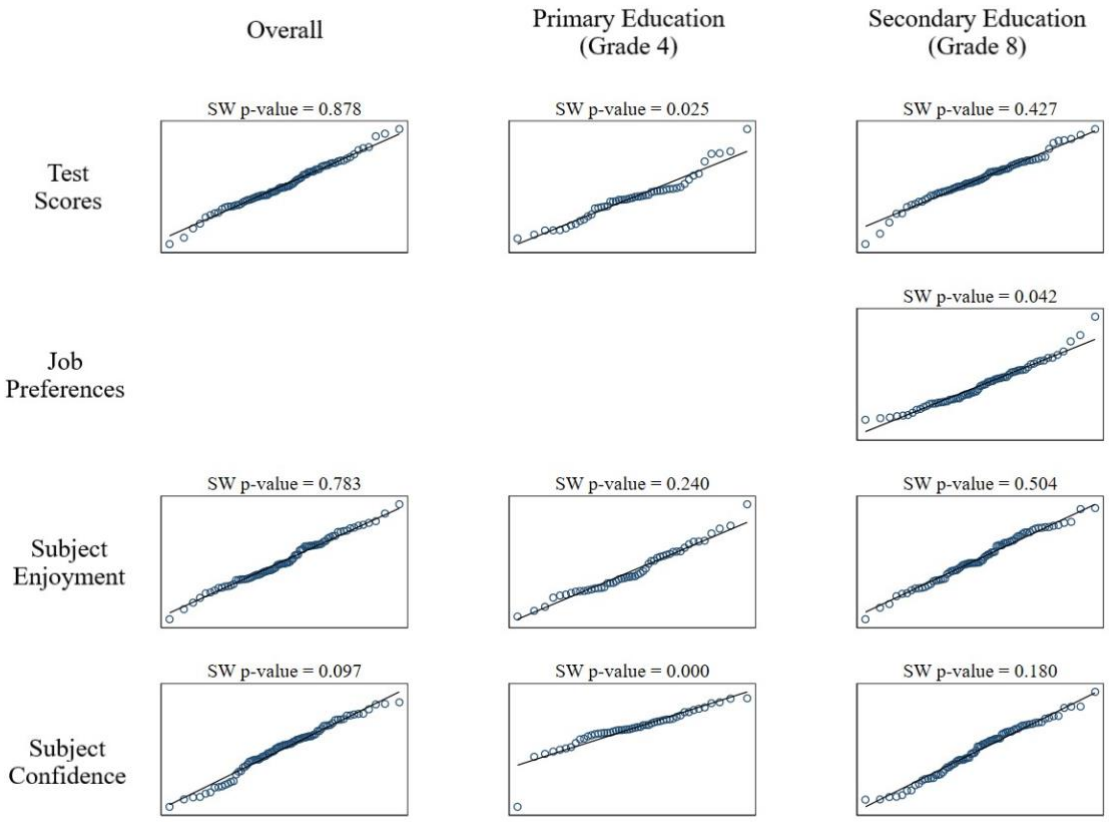
Figure B1 shows ten different Q-Q plots with standardized country-level same-sex teacher effects marked as circles and quantiles of the standard normal distribution as lines. The three columns in the figure correspond to data across grade levels and are, from left to right: "Grades 4 and 8 combined," "Grade 4 only," and "Grade 8 only." The four rows correspond to different student outcomes and are, from top to bottom: "Test scores," "Job preferences," "Subject Enjoyment," and "Subject Confidence." Above each figure we show the *p*-value of the Shapiro–Wilk test. When combining data from 4th and 8th grades, our country-level same-sex teacher effects estimates on test scores, subject enjoyment, and subject confidence are very

[18] This test is proposed in Wang and Lee (2020). Evaluation of the normality assumption in meta-analyses. *American Journal of Epidemiology*, 189(3), 235–242.

[19] Note, however, the statistical and conceptual limitations of the Shapiro–Wilk test for these exercises. Statistically, this and other tests for normality become more likely to reject their hull hypothesis as the sample size increases. This means that, with enough data and even tiny deviations from normality, the null of normality will always be rejected, which limits the value of the tests. Conceptually, the assumption of normality is a modeling choice which—like all models—is a simplification of reality. We *know* that the data is not normally distributed; the real question is whether assuming normality is a good enough approximation of reality for the purposes of our exercise. We therefore believe that the Q-Q plots are better suited to answer this question. For a more comprehensive discussion on these points, see Allen Downey's blog posts:
https://allendowney.blogspot.com/2013/08/are-my-data-normal.html
https://www.allendowney.com/blog/2023/01/28/never-test-for-normality/.

close to normally distributed. When using 4th grade only data, there is some evidence that large positive same-sex teacher effects on test scores and subject enjoyment are slightly more likely than what a normal distribution would predict. For subject confidence on 4th grade data, same-sex teacher effects look very close to normally distributed except for very extreme negative effects, which are far less likely than a normal distribution would predict. For 8th grade only data, the assumption of normality seems to hold closely for same-sex teacher effects on test scores, subject enjoyment, and subject confidence. Overall, the plots show that the assumption of normality is a reasonable modeling choice for all same-sex teacher effects distributions. All deviations from normality are small.

Figure B1: Q-Q Plots with Standardized Country-Level Same-sex Teacher Effects



Notes: This figure shows ten different Quantile-Quantile (Q-Q) plots for all country-level point estimates summarized in Table 4, where each circle plots quantiles of standardized country-level same-sex teacher effect point estimates (y-axis) and their corresponding quantile in the standard normal distribution (x-axis). The black 45-degree line is plotted as reference. SW *p*-value shows the *p*-value of the Shapiro–Wilk test that tests the null hypothesis of a normal distribution. The three columns correspond to data across grade levels and are, from left to right: "Grades 4 and 8 combined," "Grade 4," and "Grade 8." The four rows correspond to different student outcomes and are, from top to bottom: "Test scores," "Job preferences," "Subject enjoyment," and "Subject confidence." The same-sex teacher effect estimates are standardized as in Jackson and Mackevicius (2024).

Supplementary Tables and Figures on the Multi-Country Study

Table B3: Examples of Questions Used in PIRLS and TIMSS

Question	Answer	Percent Correct
According to the article, why did some people long ago believe in giants?	A correct response demonstrates understanding that people long ago believed in giants because they found huge bones/ skeletons/ fossils.	53%
Georgia wants to send letters to 12 of her friends. Half of the letters will need 1 page each and the other half will need two pages each. How many pages will be needed altogether?	Correct response: 18	34%
Bacteria that enter the body are destroyed by which type of cells? A. White blood cells B. Red blood cells C. Kidney cells D. Lung cells	Correct response: A	61%

Notes: This table shows three examples of test questions. The question in the first row was taken from PIRLS 2011 (https://nces.ed.gov/surveys/pirls/released.asp), the question in the second row was taken from the math for 4th graders test of TIMSS 2011 (https://nces.ed.gov/timss/released-questions.asp), and the question in the third row was taken from science for 8th graders of TIMSS 2011 (https://nces.ed.gov/timss/released-questions.asp). The third column shows the international average of the percentage of students who answered these questions correctly. The first question refers to a text entitled, "The Giant Tooth Mystery," which students had to read. After reading the text, students were asked why some people long ago believe in giants. Answers were coded as correct if they demonstrated "understanding that people long ago believed in giants because they found huge bones/skeletons/fossils." Fifty-three percent of students answered this question correctly. The second question asked students how many pages would be needed to write letters to 12 people if half of the letters will need one page each and the other half will need two pages each. Thirty-four percent of students answered this question correctly. The third question is a multiple-choice question asking about the type of cells that destroy bacteria that enter the body. Sixty-one percent of students answered this question correctly.

Table B4: Same-Sex Teacher Effects on All Outcomes

	Std. Test scores				
Least restrictive sample					
Same-sex teacher effect	0.0130	0.0150	0.0183	0.0148	0.0149
	(0.0025)	(0.0021)	(0.0021)	(0.0018)	(0.0016)
Fixed effects	Country	School	Classroom	Student	Student & Teacher
R-squared	0.38	0.60	0.66	0.94	0.96
Countries	90	89	82	82	82
Observations	4,434,945	1,634,574	1,226,915	1,141,407	1,135,175
Most restrictive sample					
Same-sex teacher effect	0.0149	0.0182	0.0187	0.0147	0.0149
	(0.0024)	(0.0021)	(0.0021)	(0.0019)	(0.0016)
Fixed effects	Country	School	Classroom	Student	Student & Teacher
R-squared	0.41	0.65	0.67	0.94	0.96
Countries	82	82	82	82	82
Observations	1,135,175	1,135,175	1,135,175	1,135,175	1,135,175
	Std. Job Preferences				
Least restrictive sample					
Same-sex teacher effect	0.0532	0.0590	0.0596	0.0630	0.0637
	(0.0034)	(0.0043)	(0.0045)	(0.0047)	(0.0048)
Fixed effects	Country	School	Classroom	Student	Student & Teacher
R-squared	0.13	0.17	0.19	0.68	0.72
Countries	72	72	72	71	71
Observations	1,842,968	1,008,485	856,700	781,204	776,713
Most restrictive sample					
Same-sex teacher effect	0.0618	0.0621	0.0624	0.0633	0.0637
	(0.0047)	(0.0047)	(0.0047)	(0.0047)	(0.0048)
Fixed effects	Country	School	Classroom	Student	Student & Teacher
R-squared	0.13	0.19	0.20	0.68	0.72
Countries	71	71	71	71	71
Observations	776,713	776,713	776,713	776,713	776,713
	Std. Confidence in Subject				
Least restrictive sample					
Same-sex teacher effect	0.0547	0.0535	0.0619	0.0516	0.0505
	(0.0025)	(0.0034)	(0.0038)	(0.0038)	(0.0039)
Fixed effects	Country	School	Classroom	Student	Student & Teacher
R-squared	0.12	0.17	0.19	0.70	0.74
Countries	90	89	82	82	82
Observations	4,361,900	1,595,181	1,199,318	1,104,247	1,098,172
Most restrictive sample					
Same-sex teacher effect	0.0632	0.0638	0.0641	0.0515	0.0505
	(0.0040)	(0.0040)	(0.0039)	(0.0039)	(0.0039)
Fixed effects	Country	School	Classroom	Student	Student & Teacher
R-squared	0.10	0.18	0.19	0.70	0.74
Countries	82	82	82	82	82
Observations	1,098,172	1,098,172	1,098,172	1,098,172	1,098,172
	Std. Enjoyment of Subject				
Least restrictive sample					
Same-sex teacher effect	0.0737	0.0820	0.0946	0.0888	0.0887
	(0.0025)	(0.0035)	(0.0039)	(0.0040)	(0.0040)
Fixed effects	Country	School	Classroom	Student	Student & Teacher
R-squared	0.11	0.20	0.23	0.68	0.73
Countries	72	72	72	71	71
Observations	4,303,409	1,588,238	1,193,083	1,094,103	1,088,056
Most restrictive sample					
Same-sex teacher effect	0.0936	0.0952	0.0953	0.0886	0.0887
	(0.0041)	(0.0041)	(0.0040)	(0.0040)	(0.0040)
Fixed effects	Country	School	Classroom	Student	Student & Teacher
R-squared	0.14	0.23	0.24	0.68	0.73
Countries	71	71	71	71	71
Observations	1,088,056	1,088,056	1,088,056	1,088,056	1,088,056

Notes: This table shows more details on the same-sex teacher effects estimates shown in Figures 3 and 4. The "same-sex teacher effect" in the table stems from a regressions of standardized test scores, job preferences, subject confidence, and subject enjoyment on a $\text{FemaleStudent}_i \times \text{FemaleTeacher}_j$ interaction term, a set of other control variables, and different sets of fixed effects. The inclusion of

different fixed effects imposes different sample restrictions (see Section 4). For example, estimating specifications with student fixed effects requires us to limit our sample to students for whom we observe two test scores. Thus, the table shows same-sex teacher effect estimates from specifications that use the least and most restrictive estimation sample. Standard errors clustered at the classroom level are in parentheses.

Table B5: Same-Sex Teacher Effects for Countries with Random Institutional Assignment

	(1)	(2)	(3)	(4)
	Std. Test scores	Std. Job preferences	Std. Subject confidence	Std. Subject enjoyment
Same-sex teacher effect	0.0211	0.0694	0.0858	0.105
	(0.0052)	(0.0135)	(0.0137)	(0.0147)
N	67,252	43,762	66,386	66,162
Adj. R2	0.881	0.459	0.472	0.387

Notes: This table shows same-sex teacher effects estimates for a subsample of countries with random institutional assignment. These countries are Greece (Goulas et al., 2022), South Korea (Park et al., 2013) and Taiwan (Chang et al., 2022). The "same-sex teacher effect" in the table stems from regressions of standardized test scores, job preferences, subject confidence, and subject enjoyment on a $\text{FemaleStudent}_i \times \text{FemaleTeacher}_j$ interaction term, and a set of other control variables, with student and teacher fixed effects (see Section 4). Standard errors clustered at the classroom level are in parentheses.

Table B6: Global Heterogeneity of Same-Sex Teacher Effects Estimates for All Outcomes

Outcome	Test Scores			Test Scores			Test Scores			Job Preferences		
	Overall			Primary Education (G4)			Secondary Education (G8)			Secondary Education (G8)		
Country	Average Effect	SE	N	Average Effect	SE	N	Average Effect	SE	N	Average Effect	SE	N
Algeria	0.009	0.017	6,664	N/A	N/A	N/A	0.009	0.017	6,664	N/A	N/A	N/A
Argentina	0.041	0.035	476	N/A	N/A	N/A	0.041	0.035	476	0.098	0.186	382
Armenia	0.003	0.028	13,820	0.019	0.070	1,040	-0.001	0.031	12,780	-0.004	0.069	8,346
Australia	0.010	0.010	28,061	0.002	0.032	1,731	0.012	0.010	26,330	0.029	0.023	22,970
Austria	0.017	0.042	6,054	0.134	0.042	380	-0.001	0.047	5,674	0.077	0.086	5,520
Azerbaijan	0.043	0.084	348	0.043	0.084	348	N/A	N/A	N/A	N/A	N/A	N/A
Bahrain	0.020	0.034	3,194	0.143	0.067	858	-0.037	0.037	2,336	-0.060	0.130	2,094
Belgium	-0.026	0.025	2,806	0.001	0.056	418	-0.035	0.028	2,388	-0.071	0.092	2,332
Bosnia and Herz.	0.021	0.025	6,752	N/A	N/A	N/A	0.021	0.025	6,752	N/A	N/A	N/A
Botswana	0.002	0.010	18,698	-0.002	0.055	1,352	0.003	0.010	17,346	-0.005	0.031	13,018
Bulgaria	-0.018	0.033	9,140	-0.101	0.066	174	-0.012	0.035	8,966	0.323	0.096	6,162
Canada	0.000	0.007	32,356	-0.024	0.023	2,570	0.004	0.007	29,786	0.061	0.024	23,802
Chile	0.004	0.012	20,734	-0.001	0.024	1,462	0.003	0.012	19,272	0.041	0.042	18,790
Colombia	0.000	0.016	9,476	0.016	0.047	1,310	-0.003	0.017	8,166	0.028	0.060	3,218
Cyprus	-0.011	0.012	27,888	-0.033	0.017	8,272	0.026	0.016	19,616	0.085	0.046	13,202
Czech Rep.	0.003	0.019	12,248	-0.049	0.026	1,288	0.025	0.023	10,960	-0.025	0.088	4,738
Denmark	0.007	0.018	7,148	0.022	0.015	5,608	-0.075	0.053	1,540	0.080	0.082	1,248
Egypt	0.022	0.021	8,538	N/A	N/A	N/A	0.022	0.021	8,538	0.116	0.064	6,718
El Salvador	0.012	0.013	4,534	0.000	0.000	180	0.019	0.013	4,354	N/A	N/A	N/A
England	0.004	0.010	15,483	-0.052	0.030	1,585	0.013	0.010	13,898	0.091	0.042	10,112
Estonia	0.056	0.051	4,378	N/A	N/A	N/A	0.056	0.051	4,378	0.001	0.089	4,356
Finland	0.006	0.011	16,163	0.022	0.023	1,206	0.005	0.013	14,957	0.039	0.026	14,591
France	-0.015	0.021	10,060	0.064	0.026	858	-0.023	0.023	9,202	0.088	0.058	8,756
Georgia	-0.009	0.025	8,188	-0.034	0.068	358	-0.008	0.027	7,830	-0.006	0.085	5,442
Germany	0.026	0.019	3,878	0.012	0.019	3,216	0.101	0.084	662	-0.006	0.163	654
Ghana	0.031	0.024	6,238	N/A	N/A	N/A	0.031	0.024	6,238	-0.040	0.049	4,164
Greece	0.032	0.024	8,630	N/A	N/A	N/A	0.032	0.024	8,630	0.090	0.044	8,208
Honduras	0.014	0.013	3,974	N/A	N/A	N/A	0.014	0.013	3,974	N/A	N/A	N/A
Hong Kong	0.019	0.008	34,144	-0.006	0.010	13,100	0.031	0.010	21,044	0.098	0.022	18,518
Hungary	0.025	0.011	34,136	-0.043	0.030	1,630	0.040	0.012	32,506	0.096	0.046	27,232
Iceland	0.004	0.044	1,806	-0.052	0.106	100	0.005	0.046	1,706	0.141	0.110	1,586
Indonesia	0.007	0.011	27,214	-0.010	0.031	1,110	0.007	0.011	26,104	0.033	0.023	20,262
Iran	0.063	0.072	324	N/A	N/A	N/A	0.063	0.072	324	-0.012	0.110	320
Ireland	0.052	0.022	6,210	N/A	N/A	N/A	0.052	0.022	6,210	0.112	0.057	5,872
Israel	0.026	0.017	11,596	0.000	0.000	48	0.025	0.017	11,548	0.144	0.047	9,672
Italy	-0.020	0.041	554	0.004	0.049	410	-0.110	0.039	144	0.251	0.034	140
Japan	0.019	0.008	40,674	0.001	0.008	10,842	0.028	0.010	29,832	0.073	0.016	25,184
Jordan	-0.023	0.058	668	N/A	N/A	N/A	-0.023	0.058	668	-0.245	0.183	586
Kazakhstan	0.026	0.027	11,090	0.000	0.000	112	0.020	0.028	10,978	-0.020	0.056	10,554
Kuwait	-0.031	0.085	1,024	-0.013	0.091	414	-0.019	0.190	610	0.516	0.288	530
Latvia	-0.013	0.037	5,608	0.000	0.000	108	-0.022	0.036	5,500	-0.051	0.087	5,400
Lebanon	0.019	0.014	19,824	N/A	N/A	N/A	0.019	0.014	19,824	0.049	0.045	14,560
Lithuania	0.001	0.021	20,982	0.000	0.000	58	0.002	0.021	20,924	0.038	0.074	16,282
Macedonia	0.009	0.019	16,012	N/A	N/A	N/A	0.009	0.019	16,012	0.076	0.045	14,796
Malaysia	0.022	0.009	22,700	N/A	N/A	N/A	0.022	0.009	22,700	0.068	0.020	19,436
Malta	0.009	0.017	3,324	0.003	0.017	1,740	0.034	0.050	1,584	0.469	0.234	376
Moldova	0.025	0.030	9,832	N/A	N/A	N/A	0.025	0.030	9,832	-0.040	0.058	9,294
Mongolia	-0.040	0.024	2,204	N/A	N/A	N/A	-0.040	0.024	2,204	N/A	N/A	N/A
Morocco	0.010	0.007	60,700	-0.002	0.013	15,614	0.020	0.009	45,086	-0.001	0.025	37,376
Netherlands	0.004	0.017	11,303	N/A	N/A	N/A	0.004	0.017	11,303	0.069	0.037	10,780
New Zealand	0.016	0.009	17,102	0.021	0.043	1,244	0.016	0.010	15,858	0.090	0.028	15,086
Norway	-0.001	0.013	11,430	0.003	0.026	2,548	-0.003	0.015	8,882	0.062	0.048	6,764
Oman	0.003	0.020	3,842	-0.012	0.036	1,366	0.010	0.024	2,476	0.195	0.112	2,040
Palestine	0.006	0.046	710	N/A	N/A	N/A	0.006	0.046	710	0.073	0.154	582
Philippines	-0.021	0.014	11,094	0.014	0.027	2,176	-0.029	0.015	8,918	0.017	0.036	8,460
Poland	0.008	0.023	2,656	0.008	0.023	2,656	N/A	N/A	N/A	N/A	N/A	N/A
Portugal	0.003	0.019	8,760	N/A	N/A	N/A	0.003	0.019	8,760	0.201	0.055	8,436
Qatar	0.019	0.016	5,006	0.000	0.031	1,522	0.028	0.018	3,484	-0.025	0.074	3,370
Romania	0.009	0.014	30,112	N/A	N/A	N/A	0.009	0.014	30,112	0.085	0.039	22,380
Russian Fed.	0.020	0.029	19,202	0.000	0.000	38	0.020	0.029	19,164	0.127	0.067	16,634
Scotland	0.035	0.021	8,078	-0.012	0.066	68	0.036	0.021	8,010	0.031	0.093	1,872
Serbia	0.064	0.024	12,402	N/A	N/A	N/A	0.064	0.024	12,402	-0.045	0.086	6,070
Singapore	0.008	0.007	42,232	0.008	0.013	12,064	0.009	0.008	30,168	0.032	0.022	26,598
Slovak Rep.	0.007	0.019	8,296	-0.034	0.019	1,770	0.014	0.025	6,526	0.177	0.061	6,398
Slovenia	0.023	0.017	17,630	N/A	N/A	N/A	0.023	0.017	17,630	0.135	0.047	13,510
South Africa	0.008	0.005	55,534	0.003	0.011	10,710	0.009	0.006	44,824	0.003	0.018	41,928
South Korea	0.028	0.012	14,036	-0.043	0.026	1,294	0.036	0.013	12,742	0.071	0.030	10,404
Spain	-0.004	0.016	9,298	-0.014	0.014	4,338	0.018	0.023	4,960	0.155	0.077	3,250
Sweden	0.003	0.011	24,166	-0.019	0.023	2,974	0.015	0.013	21,192	0.021	0.031	17,036
Switzerland	-0.005	0.036	3,250	N/A	N/A	N/A	-0.005	0.036	3,250	0.028	0.073	3,138
Syria	0.056	0.024	5,612	N/A	N/A	N/A	0.056	0.024	5,612	0.007	0.076	3,428
Taiwan	0.017	0.005	44,586	0.031	0.010	15,628	0.012	0.006	28,958	0.059	0.015	25,150
Thailand	0.003	0.009	23,156	-0.032	0.061	448	0.003	0.009	22,708	0.018	0.021	18,064
Tunisia	-0.002	0.011	20,010	0.003	0.029	1,986	0.022	0.010	18,024	0.009	0.053	13,562
Turkey	0.011	0.007	30,304	0.023	0.018	3,546	0.010	0.008	26,758	0.072	0.026	21,776
Ukraine	-0.017	0.035	7,354	N/A	N/A	N/A	-0.017	0.035	7,354	0.159	0.112	3,364
UAE	0.022	0.014	10,019	0.040	0.021	4,281	0.002	0.018	5,738	0.024	0.061	5,584
United States	0.013	0.006	49,652	0.001	0.014	4,578	0.014	0.007	45,074	0.107	0.022	34,250
Yemen	0.025	0.065	1,500	0.025	0.065	1,500	N/A	N/A	N/A	N/A	N/A	N/A

Table B6: (continued) Global Heterogeneity of Same-Sex Teacher Effects Estimates for All Outcomes

Outcome	Enjoyment			Enjoyment			Enjoyment		
Country	Overall Average Effect	SE	N	Primary Education (G4) Average Effect	SE	N	Secondary Education (G8) Average Effect	SE	N
Algeria	0.076	0.046	6,178	N/A	N/A	N/A	0.076	0.046	6,178
Argentina	0.117	0.122	438	N/A	N/A	N/A	0.117	0.122	438
Armenia	0.055	0.056	12,180	0.088	0.139	780	0.051	0.061	11,400
Australia	0.101	0.023	26,954	0.185	0.091	1,680	0.098	0.024	25,274
Austria	0.184	0.074	5,964	0.139	0.162	362	0.178	0.081	5,602
Azerbaijan	0.074	0.072	308	0.074	0.072	308	N/A	N/A	N/A
Bahrain	0.066	0.083	3,112	0.035	0.150	834	0.082	0.099	2,278
Belgium	-0.057	0.098	2,756	-0.151	0.257	406	-0.056	0.099	2,350
Bosnia and Herz.	0.091	0.096	6,314	N/A	N/A	N/A	0.091	0.096	6,314
Botswana	0.046	0.025	17,648	0.163	0.108	1,334	0.035	0.026	16,314
Bulgaria	0.270	0.084	8,424	0.111	0.098	174	0.282	0.091	8,250
Canada	0.137	0.022	31,294	0.036	0.070	2,494	0.146	0.024	28,800
Chile	0.053	0.035	20,150	-0.016	0.095	1,392	0.059	0.037	18,758
Colombia	0.023	0.031	8,516	0.038	0.095	1,170	0.022	0.033	7,346
Cyprus	0.072	0.032	26,914	0.085	0.053	7,940	0.122	0.037	18,974
Czech Rep.	0.050	0.061	12,114	0.065	0.150	1,264	0.044	0.058	10,850
Denmark	0.137	0.048	6,720	0.142	0.055	5,416	0.101	0.072	1,304
Egypt	0.029	0.057	7,622	N/A	N/A	N/A	0.029	0.057	7,622
El Salvador	0.040	0.058	4,192	0.000	0.000	170	0.073	0.058	4,022
England	0.049	0.034	15,051	-0.118	0.088	1,519	0.072	0.037	13,532
Estonia	0.087	0.093	4,306	N/A	N/A	N/A	0.087	0.093	4,306
Finland	0.025	0.029	15,955	-0.007	0.076	1,195	0.037	0.031	14,760
France	0.123	0.050	9,820	0.136	0.166	844	0.131	0.048	8,976
Georgia	0.102	0.059	7,724	0.054	0.115	306	0.106	0.064	7,418
Germany	0.032	0.070	3,452	0.002	0.075	2,804	0.085	0.202	648
Ghana	0.037	0.047	5,752	N/A	N/A	N/A	0.037	0.047	5,752
Greece	0.062	0.040	8,244	N/A	N/A	N/A	0.062	0.040	8,244
Honduras	0.133	0.065	3,818	N/A	N/A	N/A	0.133	0.065	3,818
Hong Kong	0.103	0.020	33,490	0.045	0.037	12,862	0.135	0.024	20,628
Hungary	0.211	0.036	33,690	0.277	0.104	1,580	0.198	0.039	32,110
Iceland	0.084	0.076	1,728	0.425	0.338	98	0.059	0.078	1,630
Indonesia	0.014	0.021	26,246	-0.034	0.078	1,048	0.019	0.021	25,198
Iran	0.241	0.119	318	N/A	N/A	N/A	0.241	0.119	318
Ireland	0.182	0.051	5,952	N/A	N/A	N/A	0.182	0.051	5,952
Israel	0.127	0.051	10,934	0.000	0.000	44	0.126	0.052	10,890
Italy	0.221	0.104	546	0.250	0.125	404	0.077	0.145	142
Japan	0.114	0.018	40,276	0.051	0.035	10,756	0.142	0.021	29,520
Jordan	-0.190	0.152	640	N/A	N/A	N/A	-0.190	0.152	640
Kazakhstan	0.035	0.045	10,848	0.000	0.000	112	0.034	0.047	10,736
Kuwait	0.305	0.126	930	0.244	0.089	382	0.420	0.297	548
Latvia	0.126	0.099	5,490	0.000	0.000	106	0.138	0.088	5,384
Lebanon	0.074	0.038	18,892	N/A	N/A	N/A	0.074	0.038	18,892
Lithuania	-0.003	0.056	20,500	0.000	0.000	58	0.000	0.057	20,442
Macedonia	0.176	0.047	14,848	N/A	N/A	N/A	0.176	0.047	14,848
Malaysia	0.079	0.019	22,372	N/A	N/A	N/A	0.079	0.019	22,372
Malta	0.024	0.068	3,224	0.006	0.065	1,684	0.075	0.188	1,540
Moldova	0.045	0.058	9,184	N/A	N/A	N/A	0.045	0.058	9,184
Mongolia	0.142	0.078	2,024	N/A	N/A	N/A	0.142	0.078	2,024
Morocco	-0.004	0.017	56,201	-0.003	0.023	13,732	-0.004	0.024	42,469
Netherlands	0.075	0.040	10,882	N/A	N/A	N/A	0.075	0.040	10,882
New Zealand	0.148	0.030	16,346	0.151	0.112	1,174	0.147	0.031	15,172
Norway	0.117	0.037	10,856	0.040	0.075	2,422	0.139	0.042	8,434
Oman	0.040	0.064	3,716	0.029	0.074	1,312	0.023	0.096	2,404
Palestine	0.264	0.252	674	N/A	N/A	N/A	0.264	0.252	674
Philippines	0.027	0.033	10,392	0.107	0.051	2,026	0.009	0.039	8,366
Poland	0.063	0.104	2,598	0.063	0.104	2,598	N/A	N/A	N/A
Portugal	0.145	0.037	8,504	N/A	N/A	N/A	0.145	0.037	8,504
Qatar	0.020	0.050	4,878	-0.047	0.086	1,426	0.052	0.061	3,452
Romania	0.114	0.033	28,996	N/A	N/A	N/A	0.114	0.033	28,996
Russian Fed.	0.119	0.048	18,694	0.000	0.000	38	0.121	0.048	18,656
Scotland	0.171	0.066	4,686	0.393	0.106	68	0.168	0.066	4,618
Serbia	0.062	0.055	11,786	N/A	N/A	N/A	0.062	0.055	11,786
Singapore	0.053	0.018	41,628	0.091	0.031	11,750	0.041	0.021	29,878
Slovak Rep.	0.139	0.055	8,034	0.122	0.089	1,700	0.125	0.067	6,334
Slovenia	0.164	0.041	17,416	N/A	N/A	N/A	0.164	0.041	17,416
South Africa	0.039	0.015	52,870	0.037	0.031	9,832	0.036	0.017	43,038
South Korea	0.121	0.031	13,940	0.002	0.087	1,286	0.138	0.033	12,654
Spain	0.087	0.046	8,948	-0.007	0.063	4,212	0.179	0.065	4,736
Sweden	0.079	0.027	23,396	0.100	0.067	2,880	0.074	0.029	20,516
Switzerland	0.076	0.069	3,188	N/A	N/A	N/A	0.076	0.069	3,188
Syria	-0.001	0.063	4,926	N/A	N/A	N/A	-0.001	0.063	4,926
Taiwan	0.108	0.018	43,978	0.121	0.038	15,332	0.121	0.020	28,646
Thailand	-0.026	0.019	22,782	-0.211	0.164	446	-0.022	0.019	22,336
Tunisia	0.056	0.035	18,676	0.118	0.057	1,784	0.035	0.043	16,892
Turkey	0.066	0.022	29,128	0.002	0.058	3,504	0.075	0.023	25,624
Ukraine	0.021	0.077	7,080	N/A	N/A	N/A	0.021	0.077	7,080
UAE	0.006	0.040	9,783	0.003	0.054	4,103	0.004	0.057	5,680
United States	0.128	0.018	47,690	0.006	0.048	4,356	0.141	0.020	43,334
Yemen	-0.080	0.165	1,048	-0.080	0.165	1,048	N/A	N/A	N/A

Table B6: (continued II) Global Heterogeneity of Same-Sex Teacher Effects Estimates for All Outcomes

Outcome	Confidence			Confidence			Confidence		
Country	Overall Average Effect	SE	N	Primary Education (G4) Average Effect	SE	N	Secondary Education (G8) Average Effect	SE	N
Algeria	-0.054	0.045	6,192	N/A	N/A	N/A	-0.054	0.045	6,192
Argentina	-0.027	0.153	410	N/A	N/A	N/A	-0.027	0.153	410
Armenia	-0.043	0.055	12,258	-0.102	0.129	798	-0.032	0.061	11,460
Australia	0.038	0.021	27,100	0.157	0.079	1,692	0.034	0.021	25,408
Austria	0.143	0.062	5,998	0.211	0.079	372	0.118	0.069	5,626
Azerbaijan	-0.125	0.174	306	-0.125	0.174	306	N/A	N/A	N/A
Bahrain	0.140	0.073	3,136	0.173	0.127	842	0.126	0.089	2,294
Belgium	0.049	0.100	2,750	-0.135	0.111	404	0.093	0.119	2,346
Bosnia and Herz.	0.231	0.110	6,450	N/A	N/A	N/A	0.231	0.110	6,450
Botswana	0.036	0.029	17,868	0.157	0.116	1,330	0.028	0.030	16,538
Bulgaria	0.102	0.071	8,584	-0.107	0.219	172	0.124	0.075	8,412
Canada	0.079	0.020	31,510	0.064	0.062	2,504	0.080	0.021	29,006
Chile	0.024	0.035	20,346	0.199	0.090	1,374	0.013	0.037	18,972
Colombia	-0.035	0.037	9,054	-0.029	0.160	1,212	-0.034	0.039	7,842
Cyprus	0.023	0.029	27,280	0.076	0.044	8,026	0.022	0.038	19,254
Czech Rep.	-0.031	0.048	12,134	-0.059	0.116	1,260	-0.019	0.049	10,874
Denmark	0.091	0.046	6,843	0.126	0.054	5,419	-0.070	0.065	1,424
Egypt	0.053	0.058	7,970	N/A	N/A	N/A	0.053	0.058	7,970
El Salvador	-0.075	0.046	4,340	0.000	0.000	170	-0.060	0.046	4,170
England	0.039	0.030	15,121	0.030	0.083	1,549	0.040	0.032	13,572
Estonia	0.071	0.089	4,344	N/A	N/A	N/A	0.071	0.089	4,344
Finland	-0.032	0.028	15,961	0.021	0.093	1,193	-0.037	0.030	14,768
France	0.065	0.046	9,866	0.052	0.158	842	0.080	0.045	9,024
Georgia	0.015	0.061	7,744	-0.039	0.210	308	0.019	0.065	7,436
Germany	0.054	0.064	3,462	0.008	0.064	2,814	0.267	0.215	648
Ghana	0.022	0.048	5,804	N/A	N/A	N/A	0.022	0.048	5,804
Greece	0.049	0.042	8,194	N/A	N/A	N/A	0.049	0.042	8,194
Honduras	0.097	0.056	3,820	N/A	N/A	N/A	0.097	0.056	3,820
Hong Kong	0.072	0.020	33,618	0.025	0.032	12,902	0.094	0.024	20,716
Hungary	0.089	0.033	33,860	0.063	0.081	1,604	0.095	0.036	32,256
Iceland	-0.071	0.067	1,748	-0.369	0.181	94	-0.057	0.070	1,654
Indonesia	0.019	0.020	26,622	0.035	0.069	1,048	0.018	0.021	25,574
Iran	-0.085	0.181	320	N/A	N/A	N/A	-0.085	0.181	320
Ireland	0.121	0.047	5,988	N/A	N/A	N/A	0.121	0.047	5,988
Israel	0.093	0.046	11,122	0.000	0.000	44	0.091	0.046	11,078
Italy	0.060	0.134	546	0.020	0.103	404	0.126	0.429	142
Japan	0.045	0.014	40,392	-0.002	0.027	10,772	0.067	0.017	29,620
Jordan	-0.142	0.095	640	N/A	N/A	N/A	-0.142	0.095	640
Kazakhstan	0.020	0.047	10,806	0.000	0.000	110	0.027	0.049	10,696
Kuwait	-0.054	0.109	930	0.011	0.129	396	-0.165	0.205	534
Latvia	-0.032	0.099	5,514	0.000	0.000	108	-0.003	0.087	5,406
Lebanon	0.008	0.034	18,948	N/A	N/A	N/A	0.008	0.034	18,948
Lithuania	-0.045	0.056	20,604	0.000	0.000	58	-0.042	0.056	20,546
Macedonia	0.037	0.050	14,978	N/A	N/A	N/A	0.037	0.050	14,978
Malaysia	0.062	0.024	22,478	N/A	N/A	N/A	0.062	0.024	22,478
Malta	-0.048	0.071	3,232	0.001	0.081	1,678	-0.206	0.156	1,554
Moldova	0.024	0.057	9,284	N/A	N/A	N/A	0.024	0.057	9,284
Mongolia	0.059	0.057	2,018	N/A	N/A	N/A	0.059	0.057	2,018
Morocco	-0.011	0.018	56,219	0.000	0.030	13,726	-0.013	0.024	42,493
Netherlands	0.050	0.037	10,965	N/A	N/A	N/A	0.050	0.037	10,965
New Zealand	0.087	0.025	16,470	0.276	0.124	1,206	0.073	0.025	15,264
Norway	0.046	0.033	10,924	-0.020	0.061	2,448	0.065	0.039	8,476
Oman	0.038	0.052	3,674	-0.150	0.079	1,280	0.151	0.064	2,394
Palestine	0.125	0.121	688	N/A	N/A	N/A	0.125	0.121	688
Philippines	0.031	0.033	10,552	0.016	0.067	2,002	0.038	0.037	8,550
Poland	0.102	0.094	2,572	0.102	0.094	2,572	N/A	N/A	N/A
Portugal	0.048	0.037	8,626	N/A	N/A	N/A	0.048	0.037	8,626
Qatar	-0.011	0.043	4,870	0.020	0.070	1,420	-0.026	0.052	3,450
Romania	0.092	0.034	29,248	N/A	N/A	N/A	0.092	0.034	29,248
Russian Fed.	0.055	0.048	18,810	0.000	0.000	34	0.059	0.048	18,776
Scotland	0.093	0.036	7,856	-0.155	0.214	68	0.096	0.036	7,788
Serbia	0.074	0.056	11,970	N/A	N/A	N/A	0.074	0.056	11,970
Singapore	0.053	0.019	41,788	0.057	0.036	11,790	0.054	0.022	29,998
Slovak Rep.	0.062	0.047	8,138	0.141	0.077	1,698	0.006	0.057	6,440
Slovenia	0.104	0.041	17,464	N/A	N/A	N/A	0.104	0.041	17,464
South Africa	-0.002	0.016	52,224	-0.044	0.033	9,606	0.008	0.018	42,618
South Korea	0.060	0.026	13,976	-0.131	0.078	1,278	0.086	0.027	12,698
Spain	0.103	0.039	9,096	0.014	0.048	4,214	0.190	0.055	4,882
Sweden	0.007	0.024	23,466	0.015	0.052	2,882	0.000	0.028	20,584
Switzerland	0.022	0.067	3,172	N/A	N/A	N/A	0.022	0.067	3,172
Syria	0.154	0.058	5,164	N/A	N/A	N/A	0.154	0.058	5,164
Taiwan	0.095	0.017	44,216	0.138	0.037	15,412	0.090	0.018	28,804
Thailand	0.001	0.019	22,886	-0.420	0.110	446	0.007	0.019	22,440
Tunisia	0.062	0.039	19,092	-0.019	0.092	1,810	0.064	0.045	17,282
Turkey	0.062	0.023	29,002	-0.011	0.058	3,446	0.072	0.024	25,556
Ukraine	0.077	0.071	7,162	N/A	N/A	N/A	0.077	0.071	7,162
UAE	0.082	0.042	9,759	0.058	0.065	4,095	0.088	0.053	5,664
United States	0.068	0.017	48,210	0.082	0.046	4,428	0.067	0.018	43,782
Yemen	0.100	0.176	1,132	0.100	0.176	1,132	N/A	N/A	N/A

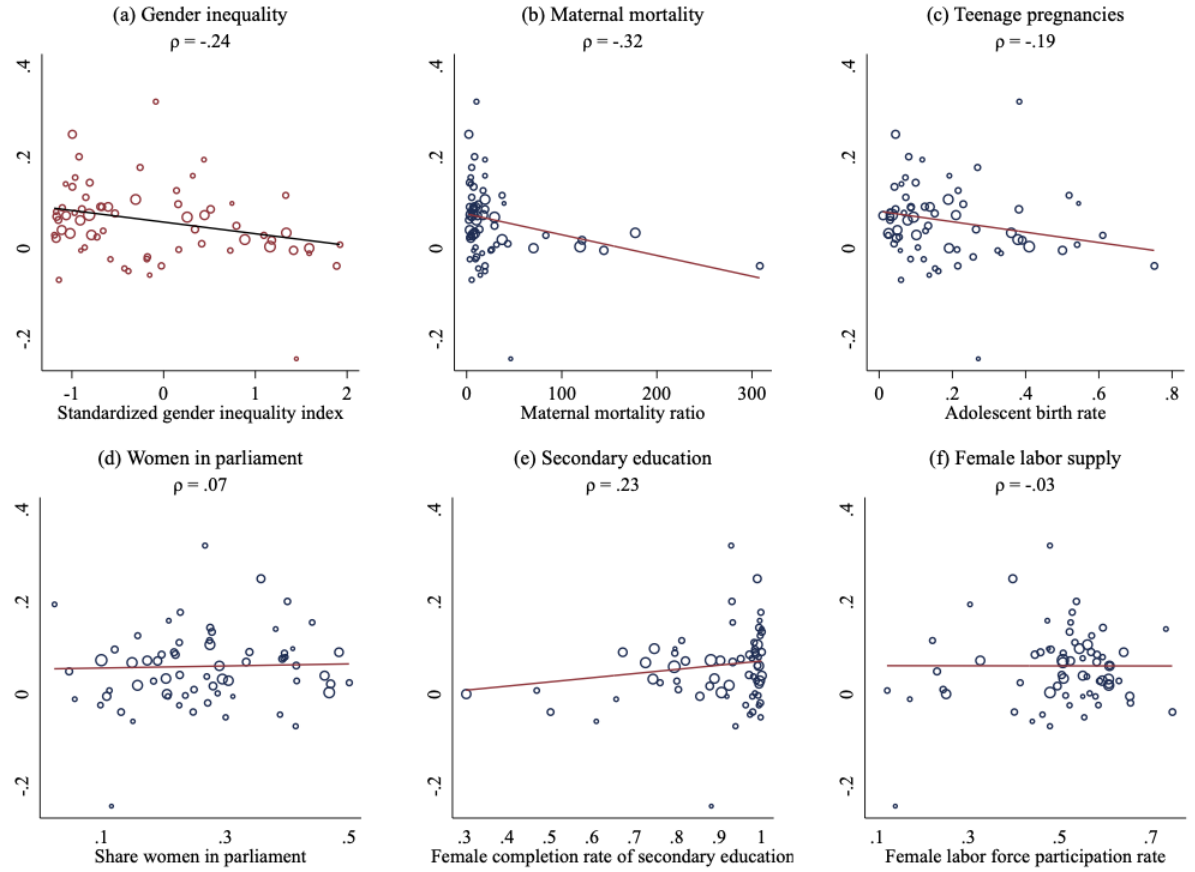
Notes: This table shows estimated same-sex teacher effects from regressions of standardized test scores, job preferences, subject enjoyment, and subject confidence on a $\text{FemaleStudent}_i \times \text{FemaleTeacher}_j$ interaction term, student fixed effects, teacher fixed effects, as well as other control variables from our preferred specification (see Section 4) for the different country subsamples indicated on the left of each panel. The smaller number of estimated same-sex teacher effects is due to missing data for Algeria, Azerbaijan, Bosnia and Herzegovina, El Salvador, Honduras, Mongolia, Poland, and Yemen.

Table B7: Global Heterogeneity for Same-Sex Teacher Effects in Job Preferences

Panel A Input: Same-sex teacher effects on job preferences	GDP per capita	Human Development Index	Gender Inequality Index	University enrollment
Intercept	0.0338	0.0326	0.0786	0.0372
	(0.0114)	(0.0113)	(0.0101)	(0.0123)
Above median	0.0462	0.0470	-0.0444	0.0417
	(0.0150)	(0.0148)	(0.0152)	(0.0167)
Countries	67	68	67	63
Panel B Input: Same-sex teacher effects on job preferences	Bank account ownership	Fertility	Science score M–F gap	Math score M–F gap
Intercept	0.0314	0.0753	0.0339	0.0407
	(0.0122)	(0.0108)	(0.0111)	(0.00969)
Above median	0.0455	-0.0317	0.0437	0.0411
	(0.0154)	(0.0157)	(0.0144)	(0.0142)
Countries	67	68	71	71

Notes: This table shows estimated same-sex teacher effects from separate meta-regressions to estimate separate same-sex teacher effects on job preferences for countries above and below the median for a given characteristic (e.g., above- and below-median GDP per capita). We use country-level estimates and their standard errors as inputs and estimate separate bivariate random-effect meta-regressions, where the single regressor is a dummy that indicates whether a country is above the median for a given characteristic. All regressions use the Hartung–Knapp adjustment. Standard errors are in parentheses.

Figure B2: Same-Sex Teacher Effects on Job Preferences and Gender Inequality



Notes: This figure shows bivariate relationships between the same-sex teacher effects estimates on standardized job preferences shown in Figure 7 (on the y-axes) and the Gender Inequality Index (GII) or the different measures contributing to the GII (on the x-axes). ρ shows the Pearson's correlation coefficient between the two variables; the line shows a fitted least squares regression line. The GII is calculated using this formula: GII = $\sqrt[3]{\text{Health} * \text{Empowerment} * \text{LFPR}}$ where Health = $(\sqrt{\frac{10}{\text{MMR}} * \frac{1}{\text{ABR}}} + 1)/2$, MMR is the maternal mortality ratio, and ABR is the adolescent birth rate. The MMR is defined by WHO as the number of maternal deaths over a certain period per 100,000 live births during the same period, and the ABR is defined as births per 10,000 female adolescents. Empowerment $= (\sqrt{\text{PR}_\text{F} * \text{SE}_\text{F}} + \sqrt{\text{PR}_\text{M} * \text{SE}_\text{M}})/2$ where PR_F and PR_M are the shares of parliamentary seats held by women and men, and SE_F and SE_M are the shares of the female/male population with at least some secondary education. LFPR is the mean of male and female labor force participation rates: LFPR = $\frac{\text{LFPR}_\text{F} + \text{LFPR}_\text{M}}{2}$. Data on the GII (Panel (a)) are taken from the Human Development Report 2020 published by the UN. The GII is not available for Palestine, Scotland, Syria, and Taiwan. The figure shows the standardized GII, which has a mean of zero and a standard deviation of 1 for the included countries. The measure shown in Panel (b)) is maternal mortality in 2015, which is taken from UN data. Data on maternal mortality are not available for Hong Kong, Palestine, Scotland, and Taiwan. The measure shown in Panel (c) is the ABR in 2017, which is taken from UN data. These data are not available for Hong Kong, Palestine, Scotland, and Taiwan. The measure shown on Panel (d) is the share of parliamentary seats held by women in 2020, which is taken from the Gender Data Portal of the World Bank. These data are not available for Hong Kong, Palestine, Scotland, and Taiwan. The measure shown in Panel (e) is the share of women with a secondary education in 2017, which is taken from UN data and Barro and Lee (2018). This measure is not available for Lebanon, Oman, Palestine, and Scotland. The measure shown in Panel (f) is the female labor force participation in 2020, which is taken from World Bank data. These data are not available for Macedonia, Palestine, Scotland, and Taiwan.

Table B8: Global Heterogeneity for Same-Sex Teacher Effects on Enjoyment

Panel A Input: Same-sex teacher effects on subject enjoyment	GDP per capita	Human Development Index	Gender Inequality Index	University enrollment
Intercept	0.0538	0.0533	0.0970	0.0573
	(0.0091)	(0.0090)	(0.0089)	(0.0095)
Above median	0.0451	0.0449	-0.0409	0.0435
	(0.0124)	(0.0123)	(0.0128)	(0.0135)
Countries	75	76	75	69
Panel B Input: Same-sex teacher effects on subject enjoyment	Bank account ownership	Fertility	Science score M–F gap	Math score M–F gap
Intercept	0.0603	0.0878	0.0557	0.0568
	(0.0100)	(0.0096)	(0.0095)	(0.0082)
Above median	0.0314	-0.0184	0.0409	0.0475
	(0.0133)	(0.0134)	(0.0124)	(0.0119)
Countries	75	76	79	79

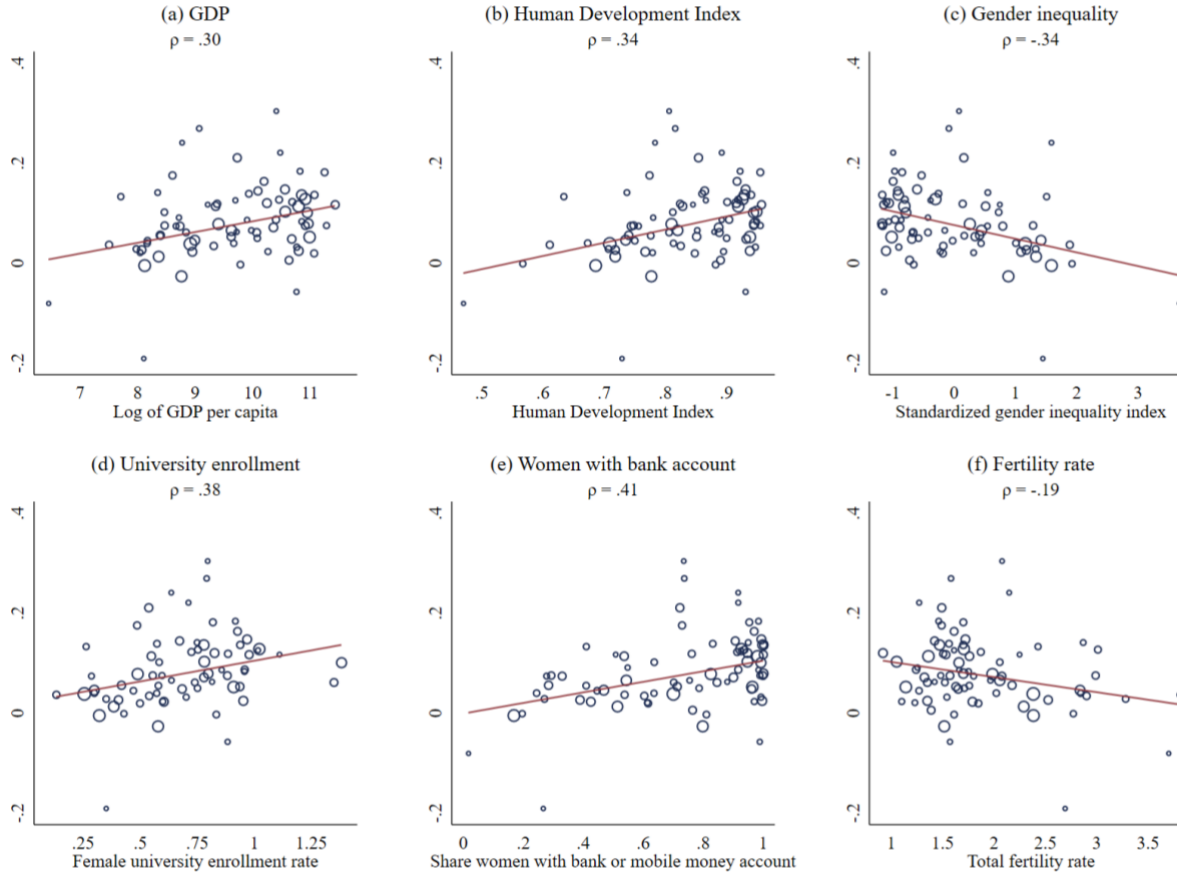
Notes: This table shows estimated same-sex teacher effects from separate meta-regressions to estimate separate same-sex teacher effects on subject enjoyment for countries above and below the median for a given characteristic (e.g., above- and below-median GDP per capita). We use country-level estimates and their standard errors as inputs and estimate separate bivariate random-effect meta-regressions, where the single regressor is a dummy that indicates whether a country is above the median for a given characteristic. All regressions use the Hartung–Knapp adjustment. Standard errors are in parentheses.

Table B9: Global Heterogeneity for Same-Sex Teacher Effects on Subject Confidence

Panel A Input: Same-sex teacher effects on subject confidence	GDP per capita	Human Development Index	Gender Inequality Index	University enrollment
Intercept	0.0229	0.0237	0.0509	0.0235
	(0.0073)	(0.0074)	(0.0069)	(0.0077)
Above Median	0.0288	0.0288	-0.0253	0.0290
	(0.0097)	(0.0099)	(0.0102)	(0.0107)
Countries	75	76	75	69
Panel B Input: Same-sex teacher effects on subject confidence	Bank account ownership	Fertility	Science score M–F gap	Math score M–F gap
Intercept	0.0264	0.0489	0.0204	0.0302
	(0.0080)	(0.0074)	(0.0074)	(0.0070)
Above Median	0.0238	-0.0171	0.0374	0.0263
	(0.0102)	(0.0104)	(0.0095)	(0.0101)
Countries	75	76	79	79

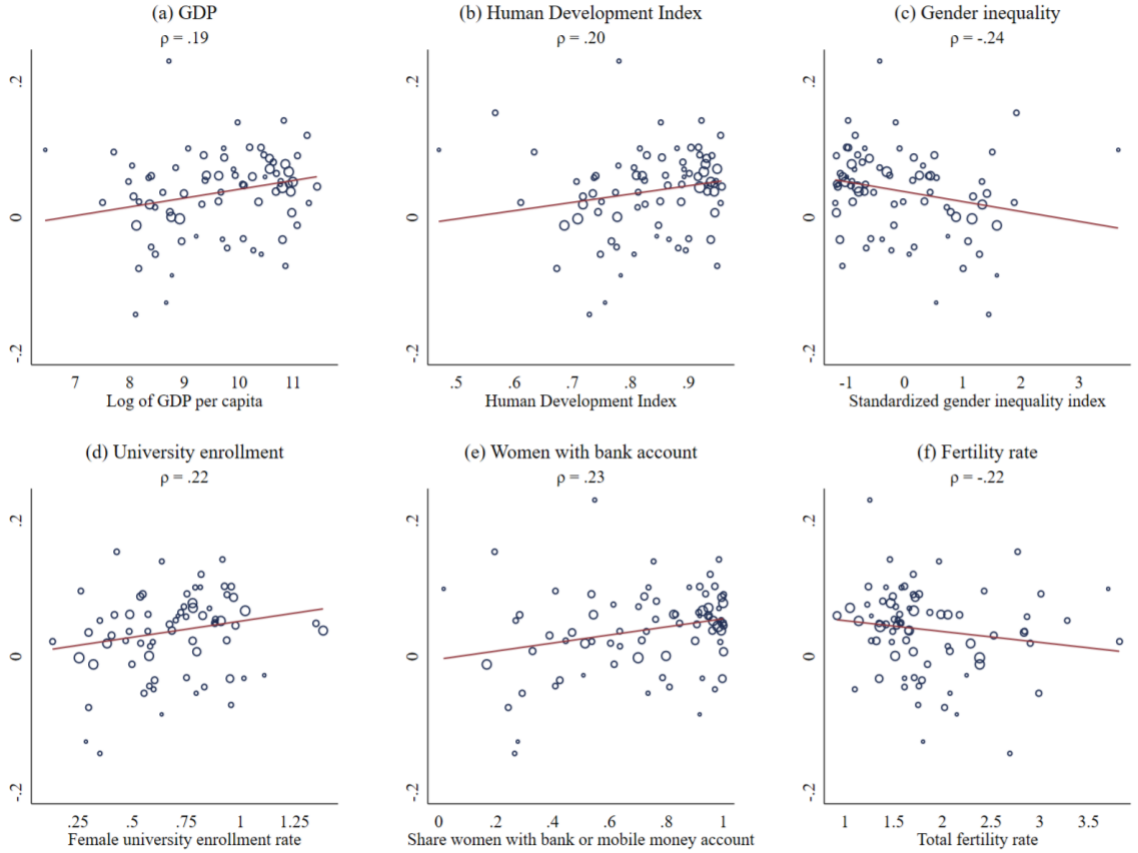
Notes: This table shows estimated same-sex teacher effects from separate meta-regressions to estimate separate same-sex teacher effects on subject confidence for countries above and below the median for a given characteristic (e.g., above- and below-median GDP per capita). We use country-level estimates and their standard errors as inputs and estimate separate bivariate random-effect meta-regressions, where the single regressor is a dummy that indicates whether a country is above the median for a given characteristic. All regressions use the Hartung–Knapp adjustment. Standard errors are in parentheses.

Figure B3: Same-Sex Teacher Effects on Subject Enjoyment and Country-Level Correlates



Notes: These panels show the relationship between the estimated same-sex teacher effects on standardized subject enjoyment shown in Table B6 and different country-level characteristics. ρ shows the Pearson's correlation coefficient between the two variables; the line shows a fitted least squares regression line. The size of each circle in the plot is dependent on the inverse of the standard error of the estimate, showing larger circles for more-precisely estimated effects. The characteristic shown in Panel (a) is log GDP per capita from 2019, which is taken from the World Bank World Development Indicators 2019. This characteristic is not available for Palestine, Scotland, Syria, and Taiwan. The characteristic shown in Panel (b) is the Human Development Index computed by the UN as a composite measure of a country's average life expectancy at birth, years of schooling and expected years of schooling, and the gross national income per capita in purchasing power parity (PPP) terms. This characteristic is not available for Palestine, Scotland, and Taiwan. The characteristic shown in Panel (c) is the standardized Gender Inequality Index (GII) from the Human Development Report 2020 published by the UN. The GII is calculated using this formula: GII = $\sqrt[3]{\text{Health} * \text{Empowerment} * \text{LFPR}}$ where Health is computed as Health = $(\sqrt{\frac{10}{\text{MMR}} * \frac{1}{\text{ABR}}}+1)/2$ where MMR is maternal mortality rate, and ABR is the adolescent birth rate. Empowerment is computed as Empowerment = $(\sqrt{\text{PR}_\text{F} * \text{SE}_\text{F}} + \sqrt{\text{PR}_\text{M} * \text{SE}_\text{M}})/2$, where PR_F is the share of parliamentary seats held by women, and PR_M is the share of parliamentary seats held by men. SE_F is the female population with at least some secondary education, and SE_M is the male population with at least some secondary education. LFPR is computed as the mean of male and female labor force participation rates: LFPR = $\frac{\text{LFPR}_\text{F}+\text{LFPR}_\text{M}}{2}$. The GII is missing for Hong Kong, Palestine, Scotland, and Taiwan. The characteristic shown in Panel (d) is the female university enrollment rate for 2016/17. The female university enrollment rate is computed as the ratio of total female enrollment in tertiary education, regardless of age, to the female population of the age group that officially corresponds to tertiary education. This rate can hence be larger than 1, for example, if the number of over-age women in tertiary education is large. The data are taken from the Gender Data Portal of the World Bank. This characteristic is available for all countries except Japan, Lebanon, Palestine, Scotland, Taiwan, Turkey, Ukraine, and the United Arab Emirates. The characteristic in Panel (e) is the share of women of the female population aged 15+ who owned a bank account or mobile money account in 2017. Data are taken from the Gender Data Portal of the World Bank. This characteristic is not available for Iceland, Palestine, Scotland, and Taiwan. The characteristic shown in Panel (f) is the total fertility rate in 2019. The data are taken from the Gender Data Portal of the World Bank. This characteristic is not available for Palestine, Scotland, and Taiwan.

Figure B4: Same-Sex Teacher Effects on Subject Confidence and Country-Level Correlates



Notes: These panels show the relationship between the estimated same-sex teacher effects on standardized subject confidence shown in Table B6 and different country-level characteristics. ρ shows the Pearson's correlation coefficient between the two variables; the line shows a fitted least squares regression line. The size of each circle in the plot is dependent on the inverse of the standard error of the estimate; larger circles show more-precisely estimated effects. The characteristic shown in Panel (a) is log GDP per capita from 2019, which is taken from the World Bank World Development Indicators 2019. This characteristic is not available for Palestine, Scotland, Syria, and Taiwan. The characteristic shown in Panel (b) is the Human Development Index computed by the UN as a composite measure of a country's average life expectancy at birth, years of schooling and expected years of schooling, and the gross national income per capita in PPP terms. This characteristic is not available for Palestine, Scotland, and Taiwan. The characteristic shown in Panel (c) is the standardized Gender Inequality Index (GII) from the Human Development Report 2020 published by the UN. The GII is calculated using this formula: GII = $\sqrt[3]{\text{Health} * \text{Empowerment} * \text{LFPR}}$, where Health is computed as Health = $(\sqrt{\frac{10}{\text{MMR}} * \frac{1}{\text{ABR}}} + 1)/2$, where MMR is maternal mortality rate and ABR is the adolescent birth rate. Empowerment is computed as Empowerment =$(\sqrt{\text{PR}_\text{F} * \text{SE}_\text{F}} + \sqrt{\text{PR}_\text{M} * \text{SE}_\text{M}})/2$, where PR_F is the share of parliamentary seats held by women, and PR_M is the share of parliamentary seats held by men. SE_F is the female population with at least some secondary education, and SE_M is the male population with at least some secondary education. LFPR is computed as the mean of male and female labor force participation rates: LFPR = $\frac{\text{LFPR}_\text{F} + \text{LFPR}_\text{M}}{2}$. The GII is missing for Hong Kong, Palestine, Scotland, and Taiwan. The characteristic shown in Panel (d) is the female university enrollment rate for 2016/17. The female university enrollment rate is computed as the ratio of total female enrollment in tertiary education, regardless of age, to the female population of the age group that officially corresponds to tertiary education. This rate can hence be larger than 1, for example, if the number of over-age women in tertiary education is large. The data are taken from the Gender Data Portal of the World Bank. This characteristic is available for all countries except for Japan, Lebanon, Palestine, Scotland, Taiwan, Turkey, Ukraine, and the United Arab Emirates. The characteristic in Panel (e) is the share of women of the female population aged 15+ who owned a bank or mobile money account in 2017. Data are taken from the Gender Data Portal of the World Bank. This characteristic is not available for Iceland, Palestine, Scotland, and Taiwan. The characteristic shown in Panel (f) is the total fertility rate in 2019. The data are taken from the Gender Data Portal of the World Bank. This characteristic is not available for Palestine, Scotland, and Taiwan.