University of Zurich[UZH]

URPP Equality of Opportunity

Same-Sex Role Model Effects in Education

Alexandra de Gendre
Jan Feld
Nicolás Salamanca
Ulf Zölitz

Equality of Opportunity Research Series #30
September 2023

University of Zurich[UZH]

URPP Equality of Opportunity

URPP Equality of Opportunity Discussion Paper Series No.30, September 2023

Same-Sex Role Model Effects in Education

Alexandre de Gendre
University of Melbourne
a.degendre@unimelb.edu.au

Jan Feld
Victoria University of Wellington
jan.feld@vuw.ac.nz

Nicolás Salamanca
University of Melbourne
n.salamanca@unimelb.edu.au

Ulf Zölitz
University of Zurich
ulf.zölitz@econ.uzh.ch

URPP Equality of Opportunity, University of Zurich, Schoenberggasse 1, 8001 Zurich, Switzerland
info@equality.uzh.ch, www.urpp-equality.uzh.ch

Same-Sex Role Model Effects in Education*

Alexandra de Gendre
(University of Melbourne)

Jan Feld
(Victoria University of Wellington)

Nicolás Salamanca
(University of Melbourne)

Ulf Zölitz
(University of Zürich)

Septemeber 2023

Abstract

We study same-sex role model effects of teachers with a meta-analysis and our own study of three million students in 90 countries. Both approaches show that role model effects on performance are, on average, small: 0.030 SD in the meta-analysis and 0.015 SD in our multi-country study. Going beyond test scores, our multi-country study documents larger average role model effects on job preferences (0.063 SD). To understand the universality of these effects, we estimate the distributions of country-level same-sex role model effects. Although role model effects on test scores appear universally small, we find substantial cross-country variation for job preferences, with larger effects in countries with larger gender gaps. These results are consistent with role models inspiring students to overcome gender stereotypes and pursue a STEM career. However, in countries with negligible gender gaps, role models do not seem to have this equalizing function.

Keywords: Same-sex role models, STEM, teachers, external validity, multi-country study, gender role models, standardized test scores, grades, job preferences, science, math, reading, 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, David Dorn, Chris Doucouliagos, Arthur Grimes, Nick Huntington-Klein, Nathan Kettlewell, Christine Mulhern, Martin Neugebauer, Bob Reed, Julia Rohrer, Roberto Weber, and seminar participants at Bocconi, the CESifo education meeting, CREST, 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

Role models could serve as a powerful policy tool to reduce inequality in education. For example, exposure to more female STEM teachers is commonly thought to increase women's STEM performance. Similarly, exposure to more male primary school teachers promises to reduce boys' underperformance. Although the idea of same-sex teachers boosting performance has inspired calls for policy interventions,[1] it is not clear whether role models deliver what they promise. Recently published studies have shown same-sex role model effects on student performance that are positive (Gong, Lu and Song 2018), insignificant (Andersen and Reimer 2019), and even negative (Antecol, Eren and Ozbekik 2015). Yet, no systematic evidence exists on the direction or magnitude of same-sex role model effects on student performance.

In the first part of our paper, we fill this gap with a meta-analysis. We identify 538 estimates from 24 studies and find an average same-sex role model effect of 0.030 standard deviations (SD) for grades and test scores in primary and secondary education. Although our meta-analysis provides a useful summary of the literature, it has two important shortcomings. First, the sign of the estimated average role model effect is sensitive to how we correct for publication bias, with some correction methods showing small positive effects and others showing small negative effects. Second, we cannot convincingly investigate heterogeneity in role model effects because of differences in methodology across studies. No two studies use the same empirical strategy, econometric specification, or sample selection criteria. Recent studies have shown that such seemingly innocuous 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 role model effect estimates reflect differences in empirical approaches or true heterogeneity.

Knowing the degree of heterogeneity of the true role model effects is important. With a small average effect, a large standard deviation of the true effect implies that role model effects are large and positive in some settings as well as large and negative in other settings. To go beyond the mean and better understand the heterogeneity of role model effects, we need an approach that allows us to explicitly hold the methodology constant.

In the second part of our paper, we estimate role model effects with a multi-country study that applies a consistent methodology to data from 90 countries. Our multi-country

[1] For example, UNICEF identified the lack of female role models as a key contributor to girls' underperformance in STEM subjects (UNICEF 2020). The OECD and World Bank have both called for policies to attract more female STEM teachers to increase female representation in STEM studies and jobs (OECD 2012; World Bank 2020).

approach has two key advantages. First, it uses a much larger sample size—over 90 times the sample size of the median study in our meta-analysis—which makes it possible to detect smaller average effects. This feature is particularly important when plausible effect sizes are small. Second, it allows us to investigate how stable results are across countries and what drives differences between countries.

To estimate role model 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). Our resulting dataset contains 3,047,752 children taught by 231,942 teachers in 105,916 primary and secondary schools across six continents.

Two key features make this combined dataset particularly useful to conduct a multi-country study on role model effects. First, test scores in this data are designed to be comparable between countries. This feature allows us to make fair cross-country comparisons of role model effects. Second, both datasets contain measures of students' subject enjoyment and subject confidence, and TIMSS also has data on job preferences. These outcome variables allow us to obtain evidence on role model effects that go beyond students' test scores.

To identify the causal effects of same-sex role models, 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 role model effects with four additional 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 but a male science teacher (or vice versa) while additionally holding constant unobserved teacher characteristics. All our fixed effects specifications deliver virtually identical results. From the least to the most conservative 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]

[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.

The results of our multi-country study show very small average same-sex role model effects on test scores of 0.015 SD. Across all specifications, the 99 percent confidence intervals allow us to rule out effects smaller than 0.009 and larger than 0.022 SD. However, teachers' influence on students might go beyond test scores (Jackson 2018). Teachers may inspire students to follow in their footsteps and to make similar educational or occupational choices (Carrell, Page and West 2010; Card et al. 2022). We therefore estimate role model effects for three non-test score outcomes. Here we observe larger effects. We see role model effects of, on average, 0.064 SD on students' preferences for working in a job that involves math or science. We also find role model effects of similar magnitude on subject enjoyment (0.089 SD) and subject confidence (0.050 SD).

We then go beyond these mean effects to answer a key question that has not yet been studied: How universal are same-sex role model effects? The universality of an effect is impossible to prove and easy to disprove. A universal effect is one that holds in all contexts, yet it is impossible to study all contexts. However, our rich dataset allows us to estimate role model effects in many contexts, making meaningful progress in assessing how universal these effects are.

When assessing the universality of role model effects in many contexts, it is particularly important to account for sampling error. Without doing that, we could confuse estimates that are negative by chance as evidence against a universally positive role model effect. We therefore use meta-analysis methods to estimate the distribution of the true role model effect, that is, the distribution of the estimates that is not due to sampling error. A key difference to traditional meta-analyses is that we apply meta-methods to our own estimates, which allows us to hold the methodology constant and means we do not have to worry about publication bias. In other words, we take advantage of meta-analysis methods while being able to "quality-control" the inputs. We find no meaningful cross-country variation in role model effects for test scores. In primary education, role model effects exceed 0.05 SD in only 1 percent of countries, while in secondary education, they never exceed this threshold. Put differently, same-sex role model effects on test scores appear universally small. In contrast, we do find meaningful cross-country variation for non-test score outcomes in both primary and secondary education. Role model effects for non-test score outcomes are near universally positive, with the magnitude of this positive effect differing depending on the setting and outcome. For instance, role model effects on job preferences exceed 0.05 SD in 60 percent of countries.

In the final part of the paper, we explore what drives the cross-country variation in role model effects on job preferences—a policy-relevant outcome that varies markedly between countries. We show that role model effects on job preferences are more pronounced in rich and gender-equal countries, where women are particularly underrepresented in STEM.

We connect our findings to the gender-equality paradox—the observation that countries that are more gender-equal have larger gender gaps in STEM performance and occupational representation. Our results are consistent with the idea that same-sex role models matter more in settings where sex gaps are larger: In countries with meaningful STEM gender gaps, same-sex role models appear to inspire students to overcome stereotypes and pursue careers in their teachers' subject matter. In countries with negligible gender gaps, however, same-sex role models do not have this equalizing function. Taken together, our results suggest that hiring more female teachers or increasing the exposure to female role models in other ways may be an especially useful policy tool to increase women's STEM representation in the United States, Canada, and many European countries.

We contribute to the literature in three important ways. First, our meta-analysis provides a comprehensive summary of the current state of the literature on same-sex role-model effects on performance. This is particularly important for a literature about an effect that has inspired many calls for policy changes. Without this summary, researchers and policy makers risk being swayed by individual studies that happen to find a large effect. Our meta-analysis provides convincing evidence that role model effects on performance are, on average, very small. Our meta-analysis reveals that—based on the existing literature—the average same-sex role model effect is *so* small that most existing studies would not have been able to accurately detect it. Second, our multi-country analysis vastly expands the scope of the same-sex role model literature. We produce well-identified role model effects estimates for 90 countries including 55 countries in which these effects have not yet been studied. We go beyond student performance and study role model effects on students' job preferences, subject enjoyment, and self-confidence, all of which are outcomes that policy makers may find important on their own. Third, we explore how universal role model effects are. To do this, we borrow from meta-analytical methods to explicitly model the distribution of same-sex role model effects across settings. Overall, our results provide the most extensive exploration of same-sex role model effects to date.

More broadly, our paper showcases the scientific benefit of multi-country studies that provide insights on how universal effects are. This approach is a particularly useful tool for mature literatures that have not managed to converge, of which the same-sex role model effect literature is a prime example. Even after 24 studies and our meta-analysis, we still do not know the sign of the true average role model effect or how much effects vary by context. Such literature often remains in purgatory, where the results are simply described as mixed and inconsistencies are either ignored or handwaivingly attributed to differences in settings. A multi-setting approach provides a way out of this dilemma.

Several recent papers also combine data from multiple countries and estimate credible causal effects. For example, Wößmann and West (2005) study the impact of class size on test scores in 11 countries using data similar to ours. They rule out meaningful class-size effects in nine out of 11 countries and provide suggestive evidence that benefits of class-size reductions are negatively correlated with countries' teacher salaries. Altmejd et al. (2021) study sibling spillovers on field-of-study choices in four countries and show that results are remarkably consistent across very different settings. Kleven et al. (2019) study how the arrival of a child affects earnings in six countries and provide evidence on the country-level determinants of the child penalty. We see this multi-setting approach as the natural progression of the credibility revolution in economics. Estimating causal effects in multiple settings helps us learn whether and why effects differ by context. By estimating the distribution of effects and discussing their universality we go beyond the state of the art and provide a new perspective on the generalizability of results.

In the remainder of this paper, we investigate the importance of same-sex role models in education. In the next section, we define same-sex role model effects and summarize the literature on these effects using a meta-analysis. In Section 3, we briefly discuss the benefits of analyzing one research question with data from multiple settings. We describe the data for our own analysis in Section 4 and our empirical strategy in Section 5. Sections 6 through 8 present our results. We show estimates of average role model effects in Section 6, estimate the distribution of role model effects in Section 7, and explore the factors that predict effect differences between countries in Section 8. Finally, we conclude in Section 9.

2. A Meta-Analysis on Role Model Effects

2.1 What are role model effects?

We follow the existing literature and define the same-sex role model effect as the premium of having a same-sex teacher—on top of the general effect of having a female or male teacher (Hoffmann and Oreopoulos 2009; Muralidharan and Sheth 2016; Lim and Meer 2017; Eble and Hu 2020). Such role model effects are typically estimated with variations of the following regression model:

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

In this model, β_3 captures the role model effect. A positive role model 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 role model effect if girls and boys benefit equally from having a female teacher. However, there would be positive role model effects if girls benefit more than boys from having a female teacher.

Although we follow the literature and call β_3 a role model effect, note that this effect could be driven by the behavior of teachers, students, or both. For example, we could observe role model effects because teachers use teaching styles that students of their own sex can more easily relate to. However, we could also observe role model effects because students behave differently with teachers of their own sex.

Several studies have estimated role model effects on career choices and performance in tertiary education. For example, Carrell, Page, and West (2010) show positive role model effects on the probability of 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 role model 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 by 90 percent. 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 role model 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 role model effects on student performance in primary and secondary education. We summarize the role model effects shown in previous studies with a meta-analysis.

2.2 A meta-analysis on role model effects in primary and secondary education

For our meta-analysis, we identified 24 studies on role model effects on grades and test scores in primary and secondary education.[3] The median study investigates role model effects with

[3] These studies are Ammermüller and Dolton (2006), Antecol, Eren and Ozbekik (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, Lu, and Song (2018), Hermann and Diallo (2017), Holmlund and Sund (2008), Hwang and Fitzpatrick (2021), Lee, Rhee, and Rudolf (2019), Lim and Meer (2017),

10,196 observations from one country. From these studies we extract all 538 role model 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 pre-registration and data collection in greater detail in Appendix A. In this section, we focus on describing the results.

Our included estimates cover many different settings: 238 use data from Europe, 187 from Asia, 94 from North America, and 19 from Africa; 153 are based on data from primary education, 375 from secondary education, and 10 from both; 57 estimates come from settings that use experimental methods with an explicit random manipulation of the student–teacher assignment, the remaining 481 estimates come from settings with naturally occurring variation in classroom assignment; 37 estimates of role model 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)—calculated for 95 percent confidence and 80 percent power by multiplying the standard error by 2.8 (see e.g., Chabé-Ferret, 2022; Ch. 7)—is 0.129 SD.

We summarize all 538 estimates using a three-level random effects model (Connell, McCoach, and Bell 2022).[4] This model allows true role model effects to differ by study and accounts for the dependence of estimates within each study. By fitting the distribution of the role model effect point estimates and accounting for their uncertainty (as measured by their standard errors), this approach also produces estimates of the distribution of underlying true role model 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 calculation of the standard deviation in the distribution of role model effects

Lim and Meer (2020), Lindahl (2007), Mulji (2016), Muralidharan and Sheth (2016), Neugebauer, Helbig and Landmann (2011), Rakshit and Sahoo (2020), Xu and Li (2018), Xu (2020).

[4] Appendix Figure A2 shows funnel plots for these role model effects and their standard errors.

(Harrer et al., 2021, Ch. 4). Applying this procedure, we estimate the average role model effect to be 0.030 SD with a standard error of 0.013 (p-value = 0.0194).[5]

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

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

We explore what drives the heterogeneity in role model effects using four separate meta-regressions that includes as moderators: (1) whether studies use experimental or quasi-experimental variation, (2) the continent where they were conducted, (3) whether they analyze data from elementary or secondary school students (or a mix of both), and (4) whether they use test scores or grades as outcomes (see Appendix Table A1). Our results show no meaningful difference between estimates of role model 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 role model estimates from Africa, role model 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 percent level. We also find evidence that role model effects are 0.058 SD

---

[5] One might be concerned that the estimated average role model 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 percent, which shows that no individual study has an outsized effect on the estimated average role model 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 role model effect of 0.034 SD (std. err. = 0.003) and a standard deviation of 0.050. Using the fixed effect model that assumes the true role model effect is the same for all studies, our estimate of the role model effect is 0.010 SD (std. err. = 0.0004).

smaller in primary education than they are in secondary education; this difference is also significant at the 5 percent level.

It is unclear to what extent this heterogeneity reflects differences in true role model 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. Their findings suggest that our estimated standard deviation of the true role model effect of 0.058 SD likely also reflects differences in methods.

2.3 Do role model effects studies show publication bias?

Publication bias could bias our estimated average role model effect of 0.030 SD. For example, researchers could be more likely to report specifications that show positive role model effects, studies that show positive and significant role model 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 percent, 5 percent, and 1 percent levels. 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 role model effect estimates, we would expect more-imprecise estimates to be larger.

There are three popular ways to estimate the relationship between effect sizes and statistical precision. We apply all three of them. First, we regress the effect size on the ex-post MDE using a standard least squares estimator. Second, we perform the precision effect test or 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 in 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.[6]

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 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 (*p*-value = 0.927).[7] 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). In the next section, we explore how our estimated average role model effect changes if we correct for publication bias.

2.4 How do publication bias corrections affect our estimate?

Figure 1 shows the estimated average role model effect and estimated standard deviation of the true role model 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.

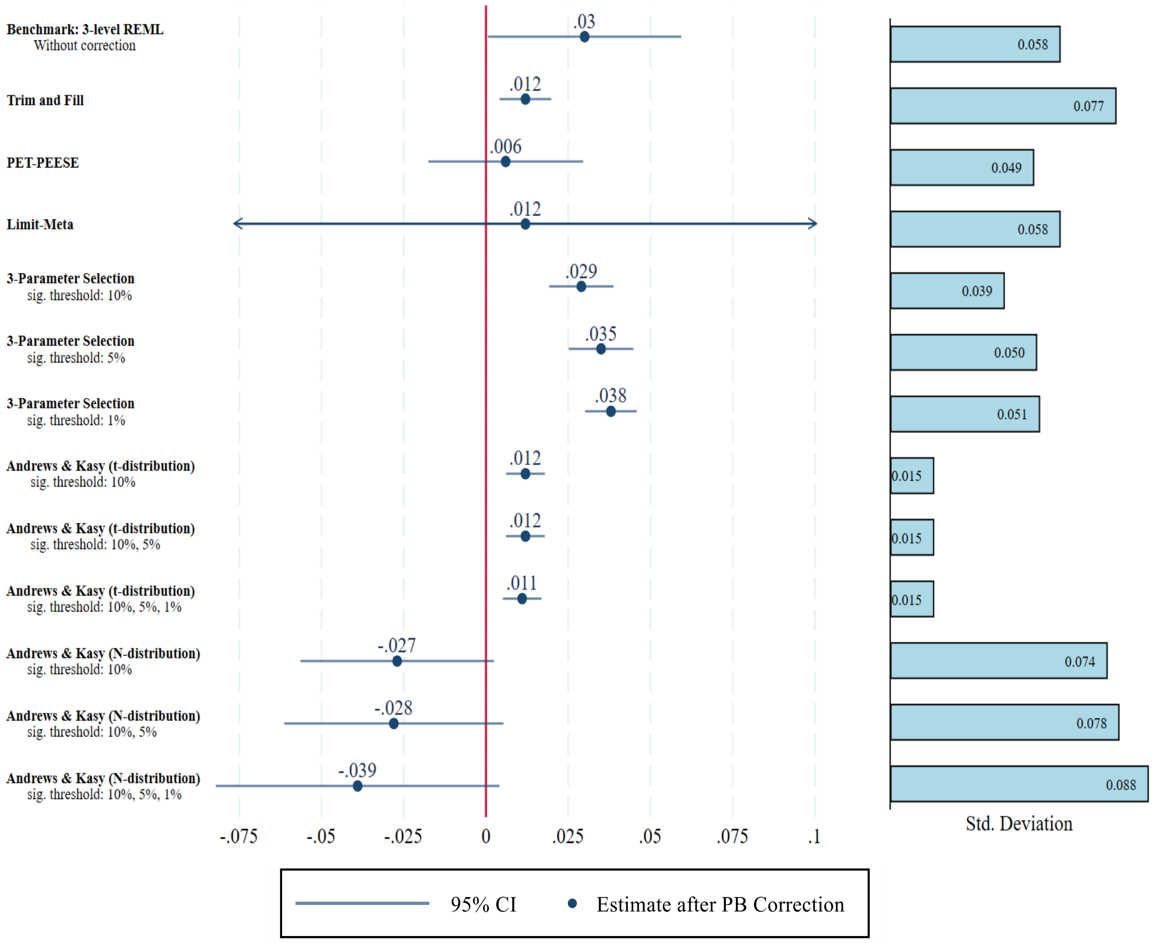
Figure 1 shows that the different procedures deliver broadly similar effect sizes. Corrected role model estimates range between –0.039 and 0.038 SD. Corrected estimates are generally of lower magnitudes, which is to be expected. Four out of the 12 corrected estimates are no longer statistically significantly different from zero at the 5 percent significance level. Trim and fill, PET-PEESE, and limit-meta reduce the role model estimate to roughly a third to half of the three-level random effect estimate. The three-parameter selection models do not

[6] We cannot correct for the mechanical dependence between effect size and standard error (Pustejovsky and Rodgers 2019) because the input 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).

[7] These outliers are role model 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.

change the role model estimate much, varying between 0.029 and 0.038 depending on which significance threshold is assumed to drive publication bias. The Andrews and Kasy (2019) corrections, however, 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. The estimated standard deviations are also broadly similar between the different methods ranging from 0.015 SD to 0.088 SD.[8]

Figure 1: Role Model Estimates After Correcting for Publication Bias



[8] 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 role models 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. The alternative meta-analysis produces very similar estimates, with an average role model 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 A3 and A4 and Figures A5, A6 and A7 for these results.

Notes: All estimated mean effects and estimated standard deviations are in the unit of standard deviation of the outcome variable. As benchmark, 3-level restricted maximum likelihood (REML) shows the estimated role model 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: Inverse variance method used for pooling estimates. REML estimator of the standard deviation of the effect size. Knapp–Hartung adjustment for the uncertainty in the between-study heterogeneity applied to the standard error of the effect size. PET-PEESE: Estimates from the PET model rather than from the precision-effect estimate with standard error (PEESE) model used because the one-sided *t*-test of intercept for the PET model does not reject the null hypothesis at the 5 percent level (*p*-value = 0.3055). Estimates weighted by their inverse variance. Assumption. Correction uses an REML estimator. 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 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 A2 shows more details on the estimates shown in this figure. The bars on the right show the estimated standard deviation of the true role model effects.

Taken together, we have shown that there is substantial heterogeneity in role model effect estimates and evidence of publication bias. Depending on how we correct for publication bias, we find small positive effects or small negative effects. Taken together, these estimates suggest role model 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 role model 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 too 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, for example, how to code their variables, how to restrict their sample, which outliers to delete, 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 role model 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 role model effects across studies. To better understand the heterogeneity of role model effects, we therefore need an approach that allows us to explicitly hold the methodology constant and that is free of publication bias.

3. The advantages of multi-setting analyses

We estimate role model effects by applying a consistent methodology to data from many countries.[9] This multi-country approach has the obvious advantage of increasing the sample size. The resulting increase in statistical power is particularly important if, as in our case, plausible effects are small. It also allows us to see how stable effects are across different settings and explore what drives differences in effect sizes.

Beyond these obvious advantages, our multi-country study has two more advantages over traditional meta-analyses. First, it allows us to apply the same methodology across data from different countries. For example, having access to individual-level data allows us to apply the same sample restrictions and include the same controls across different countries. Those seemingly innocuous methodological choices haven been shown to meaningfully affect estimates (Huntington-Klein et al. 2021; Breznau et al. 2022). Second, our approach alleviates concerns about publication bias. We do not have to worry about estimates disappearing in the publication process.

More generally, the approach of analyzing data from multiple settings with a consistent methodology is not new. There are several excellent studies that follow this approach. For example, Altmejd et al. (2021) study sibling spillovers on field-of-study choices using data from Chile, Croatia, Sweden, and the United States. The authors show siblings have a remarkably consistent impact on study choice across very different settings. Kleven et al. (2019) study how the arrival of a child affects women's and men's earnings in Austria, Denmark, Germany, Sweden, the United Kingdom, and the United States. The authors find that child penalties significantly differ by country and explore how countries' family policies and gender norms contribute to the size of child penalties in different settings. Dudek et al. (2022) combine 12 representative surveys from nine countries to estimate the effect of siblings' gender on personality. They find no meaningful effects on average and no meaningful heterogeneity across any of the 12 surveys.

Combining data from multiple settings does not have to come at the cost of estimating credible causal effects. In fact, many state-of-the-art methods can be applied to different contexts. For example, lab and field experiments can be conducted in different countries. Many

[9] Answering one research question with data from different settings is also done in mega-analyses. We have not found one agreed-upon definition of mega-analysis. Some researchers describe them as studies that re-analyze individual-level data from previous studies (e.g., Sung et al., 2014; Eisenhauer 2021). In contrast, the Global Trust Consortium (2017, p.2) has defined a mega-analysis as "the use of the largest possible number of observations of a phenomenon to quantify the strength of its correlates." While similar in spirit, our approach does not fit either of these descriptions.

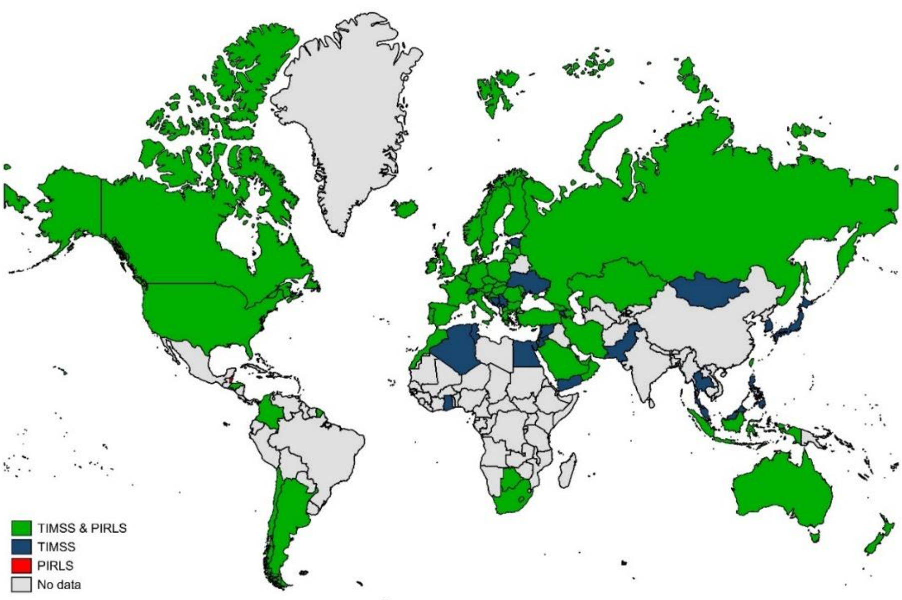
sources of exogenous variation apply in many contexts (e.g., siblings sex composition, increase in mandatory years of schooling). Today's availability of bigger and better data from a variety of settings makes it increasingly feasible to obtain many internally valid estimates.

4. Data

To estimate same-sex role model effects we build a large-scale multi-country dataset. We combine data from TIMSS and PIRLS for all available countries, waves, and education levels. TIMSS and PIRLS 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 as of 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 only have 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. Unless we specify otherwise, our description applies to both studies. 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 balancing tests and 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 they do not change the difficulty of the questions and retain the original meaning. 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 of 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 percent 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 specializing in 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 with students' agreement with statements like, "*I would like a job that involved using mathematics.*" We measure subject enjoyment and subject confidence with 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 therefore shown multiple times in the data. This happens in roughly 10 percent 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 only take one science test in TIMSS, which captures material from both classes. Estimating role model 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, teacher sex would take the value of 0.5.

5. Empirical Strategy

To measure the effect of same-sex role models on test scores, we estimate the following regression model:

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

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 only have 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 role model 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 Equation (2) via ordinary least squares regressions (OLS) and cluster our standard errors at the classroom level following the criteria outlined in Abadie et al. (2022).[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 role model 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

[10]Abadie et al. (2022) 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. (2017) recommend clustering at the sampling level, which is in our case is the classroom.

teacher. In the absence of omitted variable bias, the first difference would capture a role model 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 role model 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 role model effect.

For students who are taught by multiple teachers in the same subject (e.g., they have two science teachers), the role model 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 role model 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 role model 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 role model estimates. 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 B6).

Although we will present our main results with specifications showing all fixed effects, two specifications are particularly important for our analysis. First, the school fixed effects specification, commonly used in the existing literature (e.g., Ammermüller and Dolton 2006; Dee 2007; Lim and Meer 2017, 2020; Lee, Rhee, and Rudolph 2019), serves as a natural benchmark. The school fixed effects specification also represents the "most controlled" specification that still allows us to estimate separate effects for different subjects. Second, the student and teacher fixed effects specification allows us to rule out most concerns about omitted variable bias. We discuss both specifications in detail below.

School fixed effects specification: Parents choose their children's schools, either directly or indirectly, by choosing where to live. Similarly, teachers can influence in which schools they work. We address these concerns by including fixed effects for each school-by-grade-by-year combination (e.g., Marie Curie school, grade 4, 2012). For brevity, we refer to these fixed effects as *school fixed effects*.

For our school fixed effects specification, we exploit that *within* the same school, grade, and year, some students are assigned to female teachers and others to male teachers. For this reason, we additionally exclude schools that have no variation in teacher sex. These include schools with all female teachers or all male teachers and schools in which all classes have the same share of female teachers (e.g., all sampled classes have 50 percent female teachers). We also exclude observations for the rare remaining instances in which there is no variation in $Female\ Student_i \times Female\ Teacher_j$ at the school level. This can happen, for example, if all female teachers in a school only teach boys (e.g., the only female teacher in the school teaches chemistry to only the boys in a classroom). By actively excluding these instances, our regressions correctly report only observations that contribute to the identification of our effect of interest (see Miller, Shenhav, and Grosz 2021).

The identifying assumption for this specification is that *within a school*, our variable of interest—$FemaleStudent_i \times FemaleTeacher_j$—is unrelated to unobserved factors affecting students' test scores. This assumption would be violated if within schools, particularly high-ability girls were assigned to female teachers, particularly high-ability boys were assigned to male teachers, or both. We test the plausibility of our assumption by checking whether $Female\ Student_i \times Female\ Teacher_j$ is related to predetermined student characteristics that could be related to student ability. More specifically, we estimate versions of Equation (2) with school fixed effects where we replace the dependent variable with the following predetermined student characteristics: age in years, and three dummy variables indicating whether the student is foreign born, has at least one parent with a university degree, and lives in a two-parent household.

Our identifying assumption would also be violated if within schools, particularly effective teachers would be assigned to more students of their own sex. We test the validity of this assumption by checking whether $FemaleStudent_i \times FemaleTeacher_j$ predicts the following predetermined teacher characteristics that could be related to teaching effectiveness: years of teaching experience, and four dummy variables indicating whether the teacher is 40 years old or older, has a post-graduate degree, majored in education, or teaches in their field of expertise.

Table 2 shows the results of these balancing tests. Out of nine coefficients of interest, seven are tiny and statistically insignificant. We only see two statistically significant but tiny coefficients. First, the significant coefficient on student age shows that within schools, the girl–boy difference in age of students taught by a female teacher is 0.0094 years (three days) larger than the girl–boy age difference of students taught by a male teacher. In other words, students taught by a same-sex teacher are slightly older than students taught by an opposite-sex teacher. Second, the significant coefficient on teacher majored in education shows that within schools, the female teacher versus male teacher difference in the proportion of teachers who have majored in education is 0.35 percentage points larger for female students than for male students., which represents a miniscule 0.55 percent of the unconditional probability that a teacher majors in education in our sample. These are economically meaningless differences which, altogether, support the validity of our research design. In any case, these small imbalances do not affect our results in our preferred specification that includes student and teacher fixed effects.

Table 2: Balancing Tests

		Role model effect				
	Mean	Coef.	Std. err.	R-Squared	Countries	N
Student characteristics:						
Age (in years)	12.9	0.0094	(0.0019)	0.88	89	1,628,689
Foreign-born	0.09	-0.0008	(0.0008)	0.22	88	1,470,027
Parent(s) with university degree	0.38	-0.0009	(0.0015)	0.31	76	963,126
Two-parent household	0.66	-0.0018	(0.0023)	0.39	52	364,662

Teacher characteristics:						
40+ years old	0.76	-0.0017	(0.0021)	0.57	89	1,630,965
Experience (in years)	15.6	0.0297	(0.0283)	0.57	89	1,605,102
Has post-graduate degree	0.29	-0.0007	(0.0012)	0.64	89	1,571,995
Majored in education	0.63	0.0035	(0.0015)	0.60	86	1,301,828
Teaches field of expertise	0.86	0.0002	(0.0010)	0.67	82	1,273,757

Notes: This table shows results from regressions of predetermined student and teacher characteristics on a female student dummy, the share of female teachers, and the interaction of these two variables. The coefficient and standard error shown in the table are from this interaction term. The regressions additionally include the following controls: two subject matter dummies (science and math, base group: reading), interaction terms of all three subject dummies with the female student dummy (Female Student × science, Female Student × math, Female Student × reading), interaction terms of all three subject dummies with the female teacher dummy (Female Teacher × science, Female Teacher × math, Female Teacher × reading). The number of observations differs depending on the availability of data on predetermined characteristics. Appendix Table B4 replicates this balancing test for our preferred estimation sample. Standard errors clustered at the classroom level are in parentheses.

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 percent female teachers in all math courses and 50 percent 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 concerns people might intuitively have 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, test designs that favor girls and school principals who are more supportive of male teachers would not bias our estimates. 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, students being more likely to be assigned to same- sex teachers in subjects in which they are generally more able would not introduce any bias.

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 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 3 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, ten percent of them are foreign born, 75 percent speak the test language at home, and 38 percent 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 percent of them are

[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.

female, they have on average 16.5 years of teaching experience, and 30 percent 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 percent are foreign born (compared to 10 percent in our country fixed effects sample), 73 percent speak the test language at home (compared to 75 percent), and 36 percent have at least one parent with a university degree (compared to 38 percent). 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 percent versus 71 percent) and more likely to be more than 40 years old (84 percent versus 69 percent), are less likely to have majored in education (60 percent versus 71 percent), and are more likely to teach in their area of expertise (89 percent versus 75 percent).

Table 3: 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 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, role model 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.

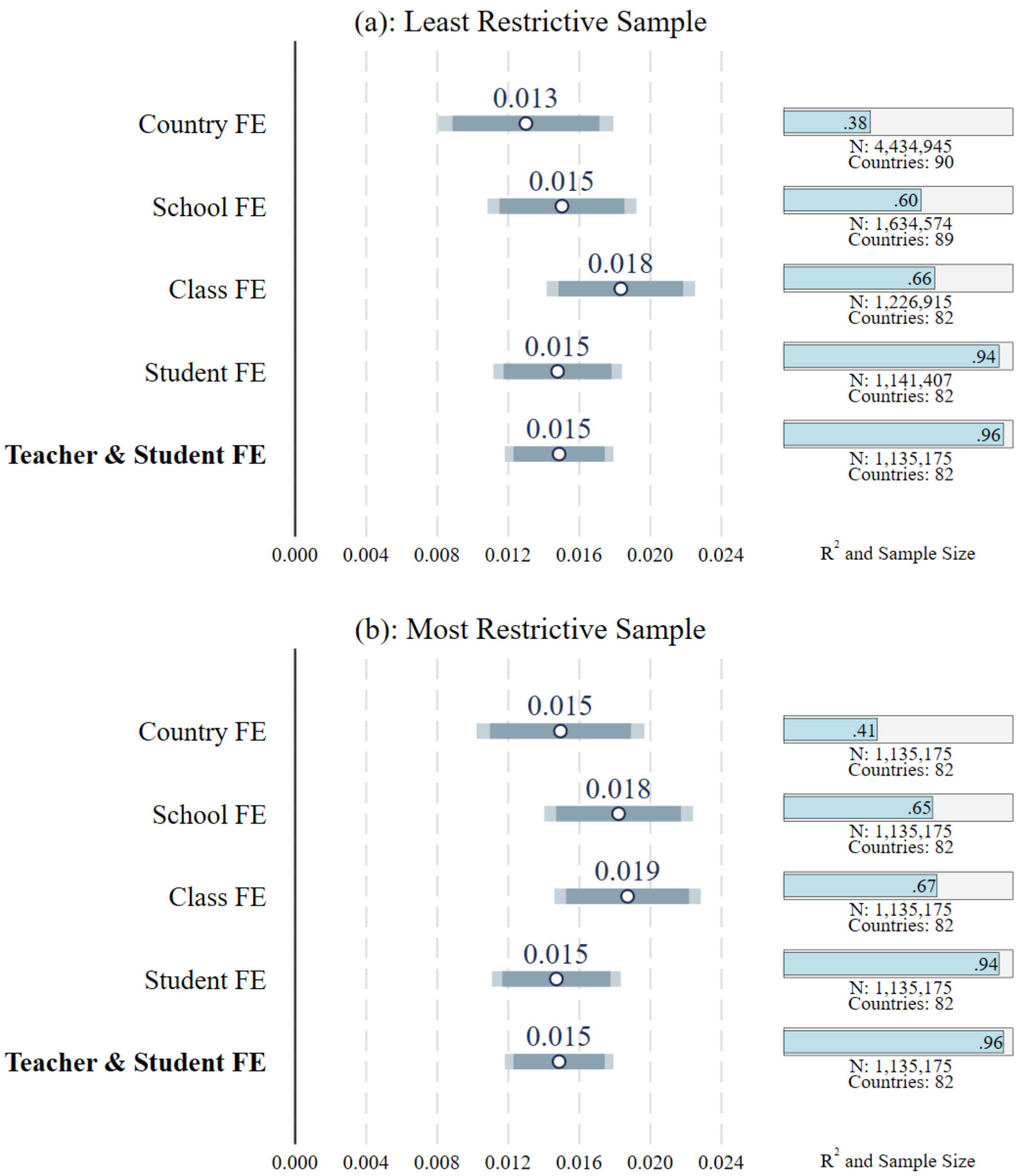
6. Average Role Model Effects on Test Scores and Non-test Score Outcomes

Role model effects on test scores: Figure 3(a) shows role model 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-squared is 0.38, and the estimated role model effect is 0.013 SD. As we include more-restrictive fixed effects, the R-squared increases substantially but our point estimates barely change. In our preferred specification, we include student and teacher fixed effects. The inclusion of 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-squared to 0.96. This specification shows a precisely estimated role model 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 role model effect of around 0.015 SD.[12]

Figure 3: Role Model Effects—Test Scores

[12] Appendix Table B6 shows when restricting our sample to countries with institutional random assignment, we find very similar results for all our outcomes of interest.



Notes: This figure shows estimated role model 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 5), 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 role model 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 5). Appendix Table B5 shows the corresponding regression table. Horizontal bars show 95 percent and 99 percent confidence intervals that are based on standard errors clustered at the classroom level.

The role model effect in our analysis is half of the size of the average role model 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. Although 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 role model estimate.[13]

A role model effect of around 0.015 SD is small. It represents a 1.5-point increase on the TIMSS or PIRLS tests. This effect is small compared to the predicted effect of other demographic characteristics in our data. For example, the predicted effect having at least one university-educated parent on test scores is 40 times as large as our estimated role model effect (0.605 SD) and the predicted effect of speaking the test language at home is 42 times larger than our role model effect (0.636 SD).[14] The 99 percent confidence intervals for these estimates allow us to rule out effects smaller than 0.008 and larger than 0.023 SD for all role model estimates shown in Figure 3.

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

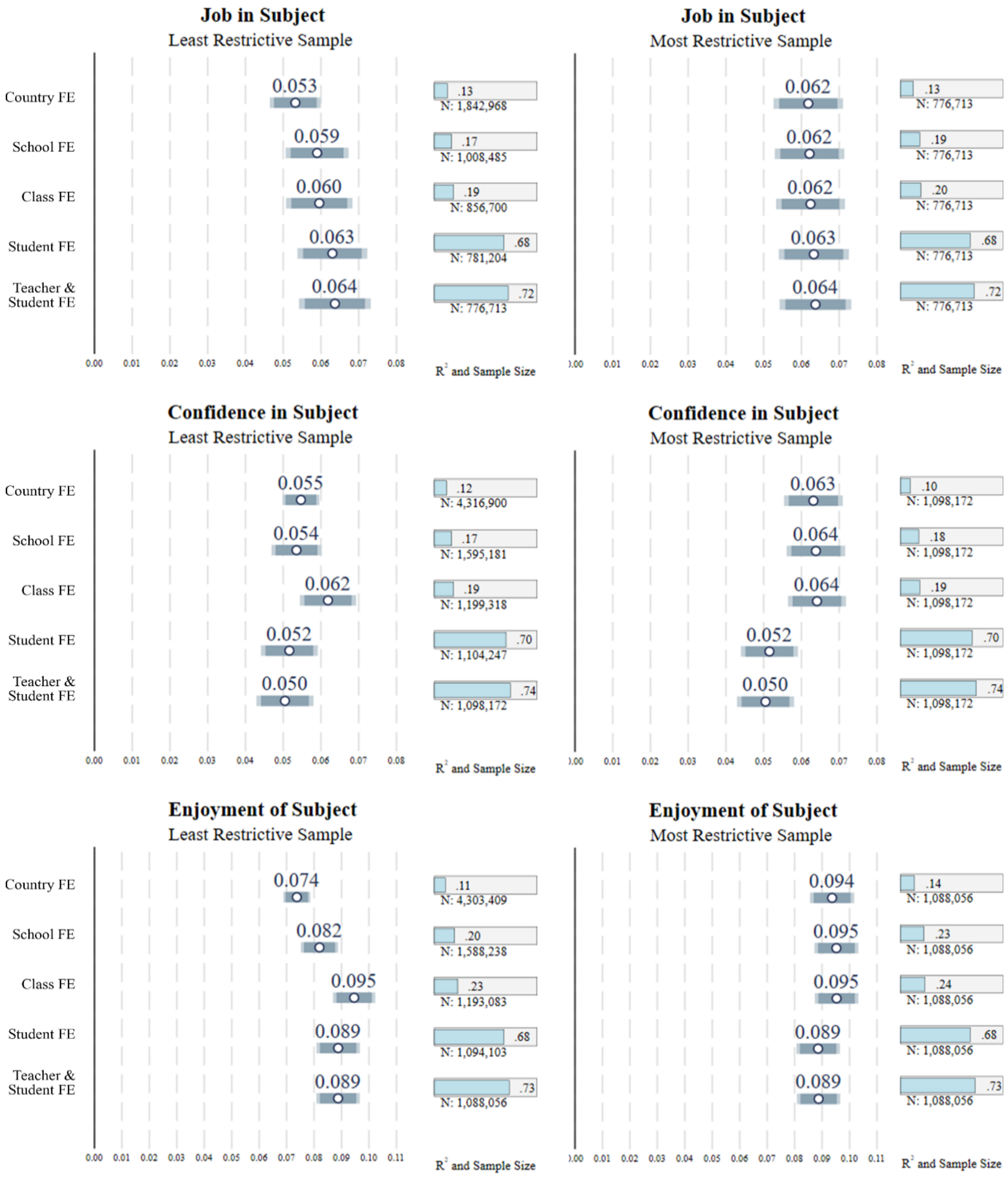
We explore heterogeneity of role model effects on test scores by subject, student characteristics, and teacher characteristics. Our results show larger role model effects in math than in science (0.0188 SD versus 0.0117 SD) and statistically insignificant role model effects in reading (0.0026 SD). Out of the 18 student and teacher characteristics we consider, only two show qualitatively different results to our estimated average effect. Role model effects are not statistically significant in grade 4 and for teachers who are not experts in their field (i.e., who did not major in the subject they are teaching). We show these heterogeneity results in Appendix B (Figure B1 and Tables B1 and B2).

---

[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 which is unfortunately often unavoidable in meta-analyses. However, holding the methodology constant does *not* mean researchers should avoid exploring how methodological choices affects their results. In our study, we intentionally show role model 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 artefact 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.

Figure 4: Role Model Effects—Job Preferences, Subject Enjoyment, and Confidence



Notes: This figure shows estimated role model 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 5), 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 role model 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 5). Appendix Table B5 shows the corresponding regression table. Horizontal bars show 95 percent and 99 percent confidence intervals that are based on standard errors clustered at the classroom level.

Role model effects beyond test scores: Teachers' influence on their students may go beyond test scores. Role models may also inspire students to follow in their footsteps and to make similar educational or occupational choices (Carrell, Page, and West 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 role model effects using the same set of fixed effects that we used for our test score analysis.

Figure 4 shows role model 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 same-sex role model effect for job preferences in our preferred specification (0.064 SD) is substantially larger than for test scores (0.015 SD). We further find role model 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. 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 (Moore and Burrus 2019). Teachers who affect students' stated job preferences, their confidence, as well as 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 role model 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 role model effects on choosing their mentor's occupation.

7. On the Universality of Same-Sex Role Model Effects

How universal are same sex-sex role model effects? The answer to this question is of central importance for policy and scientific progress. Policy makers want to know whether they can safely assume that role model effects are positive in their specific policy context. Similarly, as researchers, we want to know whether we should accept same-sex role model effects as a universal part of the education production function and build on this insight. Without exploring the universality of effects, we do not know to what extent our findings are applicable to other contexts.

We explore the universality of role model effects by testing whether and how they differ by context. Our rich dataset allows us to estimate role model effects in many different countries, in primary and secondary education, and for four different outcomes. Finding positive role model effects in all these contexts would give us confidence that such same-sex role model effects are nearly universal. In contrast, if we find that role model effects are positive in some settings and negative in others, we could rule out that role model effects are universally positive. Furthermore, finding that role model effects vary a lot by context would then open the possibility to exploring why that is the case.

Empirically, we could address the question of universality by estimating role model effects in many contexts and inspecting the estimates. However, each estimate also reflects sampling error. Even if role model 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 estimate the distribution of the true effect, that is, the distribution of the estimates that is not due to sampling error. More specifically, we estimate this distribution using country-level role model effects estimates from our most controlled specification and their standard errors as input in a random effects model.[15] In contrast to typical meta-analyses, we can do this without having to worry about differences in methodologies between estimates and publication bias.

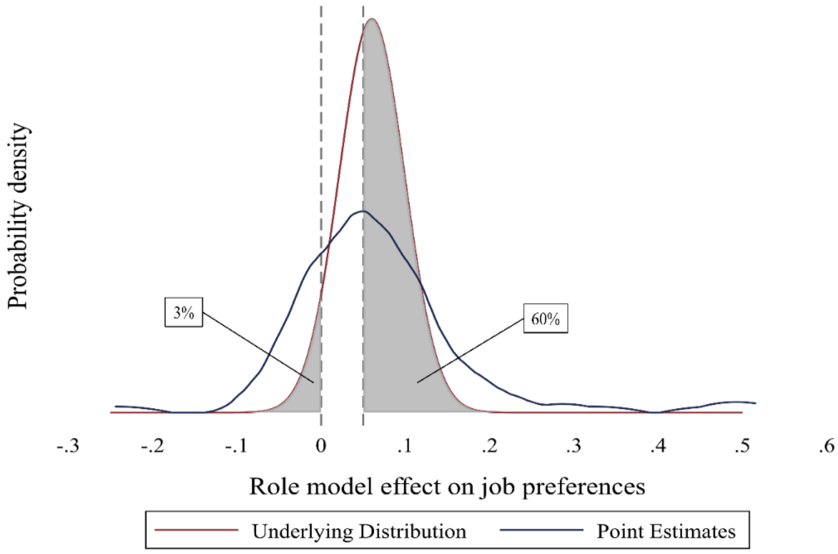
Figure 5 illustrates our approach for role model effects on job preferences. In blue is the kernel density of the distribution of country-level role model effects estimates on job preferences. This distribution partly reflects sampling error. In red is the narrower estimated distribution of the true role model effects. We obtained this distribution by using the country-level estimates and their standard errors as input in a random effects model, estimated using restricted maximum likelihood. The results of this estimation assume that the true role model effects are normally distributed, 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 true role model effects are negative in 3 percent of the countries and larger than 0.05 SD in 60 percent of the countries.

We estimate the distributions of the true role model effects for all four outcomes as well as all possible grade-outcome combinations (e.g., 4th grade test scores). The credibility of these

[15] All role model effect estimates and their standard errors used in this section are available on our dedicated study website https://www.role-model-effects.com/.

estimates depends on how reasonable the assumption is that the true effects are normally distributed. We test this assumption following Jackson and Mackevicius (2023) and show in Appendix B on pages 17–18 that the normality assumption is reasonable for all four outcomes and all grade-outcome combinations.

Figure 5: Densities of the Role Model Effect Estimates and the Fitted Distribution of True Role Model Effects for Job Preferences



Notes: This figure shows a kernel density estimate of the distribution of the 71 country-level estimated role model 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 role model effects derived from the 71 country-level estimates using a random effects meta-analysis estimated with restricted maximum likelihood. Vertical dashed lines at zero and 0.05 are also shown, for reference.

Table 4 summarizes key aspects of the estimated distributions of the true role model effects. Panel A shows estimates for the mean role model effects. In primary education (4^{th} grade), we see that the estimated mean of role model effects on test scores is almost exactly zero. The mean effect is similarly small for confidence (0.013 SD) but substantially larger for enjoyment (0.057 SD). In secondary education (8^{th} grade), the estimated mean effect is also tiny for test scores (0.013 SD) but larger for job preferences (0.058 SD), enjoyment (0.086 SD), and confidence (0.046 SD). Taken together, these results are consistent with our results from Section 6: on average, role model effects are small for test scores and larger for non-test score outcomes.

Panel B shows the estimated standard deviations of the true role model effects. These show very little variation of role model effects on test scores in our overall sample (0.0002 SD), in primary education (0.0226 SD), or in secondary education (0.0001 SD). These estimated standard deviations are much smaller than the estimated standard deviation of the true effect of 0.058 SD from our meta-analysis. This result is consistent with differences in methods between studies causing our meta-analysis to overestimate the standard deviation of the true effect. We generally see more variation in true effects for non-test score outcomes ranging from 0.026 SD (role model effect on confidence in overall sample) to 0.108 SD (effects on confidence in primary education).

Panel C shows the estimated share of countries in which role model effects are positive. These estimates reveal an interesting pattern. In primary education, the share of positive effects differs markedly by outcome. For test scores, we estimate that role model effects are positive for only 46 percent of the countries in our sample (leaving 54 percent with negative effects). Similarly, we estimate that role model effects on confidence are positive in merely 55 percent of countries. In contrast, role model effects on enjoyment are estimated to be positive in 95 percent of countries. In secondary education, results are more consistent. We estimate that virtually every country has positive role model effects on test scores, 97 percent of countries have positive role model effects on job preferences and enjoyment, and 95 percent of countries have positive role model effects on confidence. For all four outcomes, role model effects in secondary education appear to be near universally positive.

Panel D shows the share of countries in which we can expect a meaningfully positive effect, which we define as an effect larger than 0.05 SD. Overall, our estimates suggest that meaningfully positive effects on test scores are very rare (1 percent of countries in primary education and virtually no country in secondary education). In contrast, meaningfully positive effects are common for non-test score outcomes, ranging from 36 percent of countries (role model effects on confidence in primary education) to 80 percent (effects on enjoyment in secondary education).

Taking a careful look at the distribution of role model effects would be valuable for policy makers. For example, by studying the distribution we have learned that role model effects on test scores in primary education are negative in a substantial share of countries. This result suggests that policy makers should be aware that hiring more male primary school teachers to stop boys' performance decline can backfire and produce small negative effects in some settings. A closer look at the distribution of role model effects has also shown that effects on test scores in secondary education are universally positive, but also universally small. This result suggests that hiring more female teachers to increase girls' performance in secondary education is unlikely to backfire but also will not have large effects. However, the larger and nearly universal positive effects in secondary education for non-test score outcomes suggest that hiring more female teachers might still be a worthwhile policy. This policy promises effects on girls' job preferences, subject enjoyment, and confidence that are most likely positive and potentially meaningful in magnitude.

Table 4: The Distribution of Role Model Effects

Panel A: Average Mean Effect β	Overall	Primary Education (Grade 4)	Secondary Education (Grade 8)
Std. Test Scores	0.0106 [<0.0001]	-0.0021 [0.6717]	0.0130 [<0.0001]
Job Preferences	0.0578 [<0.0001]	N.A.	0.0578 [<0.0001]
Enjoyment	0.0810 [<0.0001]	0.0573 [<0.0001]	0.0855 [<0.0001]
Confidence	0.0435 [<0.0001]	0.0128 [0.5124]	0.0459 [<0.0001]
Panel B: Standard Deviation of Effect τ	Overall	Primary Education (Grade 4)	Secondary Education (Grade 8)
Std. Test Scores	0.0002	0.0226	0.0001
Job Preferences	0.0300	N.A.	0.0300
Enjoyment	0.0400	0.0354	0.0419
Confidence	0.0260	0.1082	0.0275
Panel C: Probability Effect Positive	Overall	Primary Education (Grade 4)	Secondary Education (Grade 8)
Std. Test Scores	1.0000	0.4627	1.0000
Job Preferences	0.9732	N.A.	0.9732
Enjoyment	0.9785	0.9472	0.9794
Confidence	0.9525	0.5471	0.9522
Panel D: Probability of Meaningful Effects (β > 0.05)	Overall	Primary Education (Grade 4)	Secondary Education (Grade 8)
Std. Test Scores	0.0000	0.0105	0.0000
Job Preferences	0.6033	N.A.	0.6033
Enjoyment	0.7809	0.5816	0.8019
Confidence	0.4012	0.3656	0.4401

Legend		
Panel A		β > 0.025
Panel B		τ > 0.025
Panel C		Prob > 95%
Panel D		Prob > 50%
Panel A-D		No data

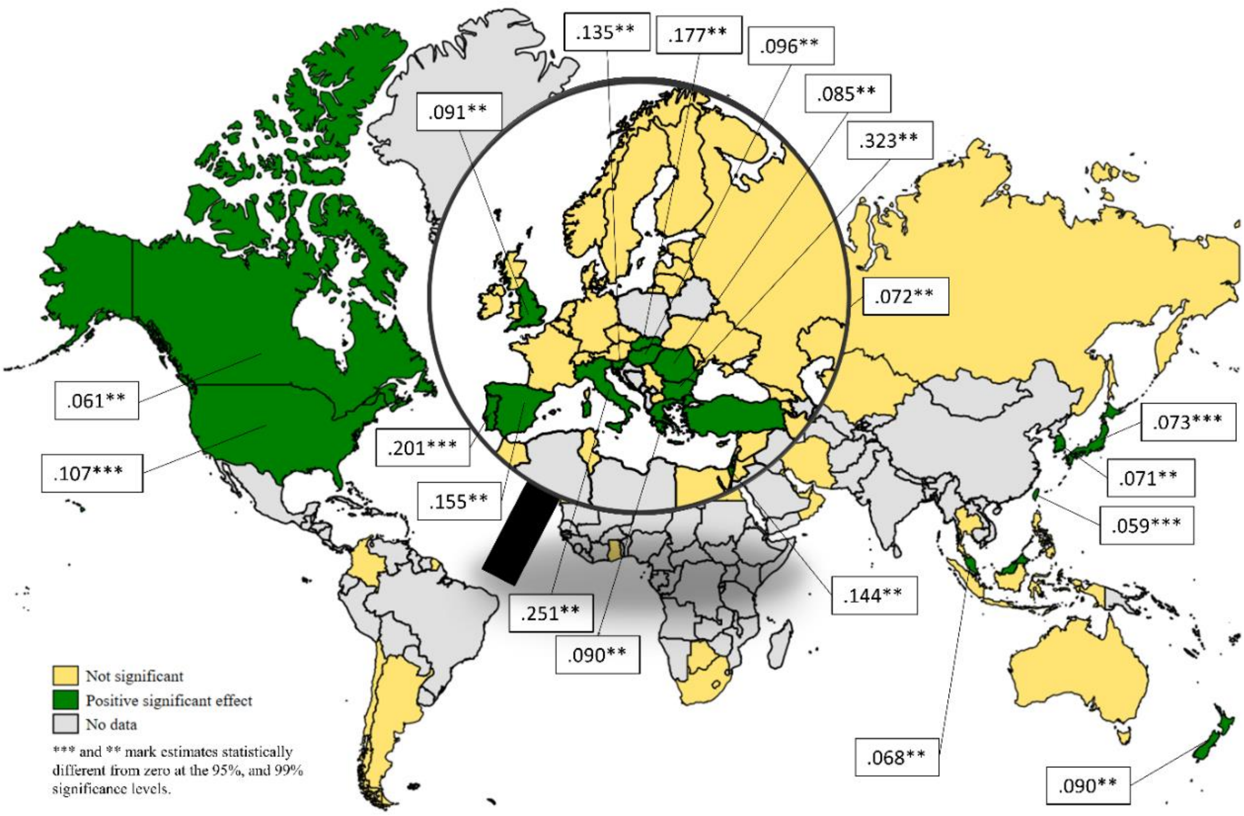
Notes: p-values are in square brackets. Job preferences are only measured in secondary education (grade 8). This is why, for job preferences, the "overall" results and "secondary education" results are identical.

8. What Explains Country-Level Heterogeneity in Role Model Effects?

We have shown that role model 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 heterogeneity in role model effects on one policy-relevant outcome that varies markedly between countries: students' job preferences. In Appendix B, we show these results for role model effects on subject confidence and enjoyment.

Figure 6 shows our country-level role model effects on job preferences. Our estimates range from –0.245 SD for Jordan to +0.516 SD for Kuwait; 21 estimates are positive and significant at the 5 percent level; 38 estimates are positive and insignificant; 16 estimates are negative and insignificant; and no estimates is negative and significant. The world map in Figure 6 shows positive and significant role model effects on job preferences in the United States (0.107 SD), Canada (0.061 SD), England (0.091 SD), Italy (0.251 SD), Spain (0.155 SD), Portugal (0.201 SD), Greece (0.090 SD), Malta (0.469 SD), Hungary (0.096 SD), Romania (0.085 SD), Bulgaria (0.323 SD), Slovak Republic (0.177 SD), Slovenia (0.135 SD), Israel (0.144 SD), Turkey (0.072 SD), Japan (0.073 SD), Malaysia (0.068 SD), South Korea (0.071 SD), Hong Kong (0.098 SD), Taiwan (0.059 SD), and New Zealand (0.090 SD).

Figure 6: Global Heterogeneity in Role Model Estimates —Job Preferences



The estimates in Figure 6, together with the overall average role model effect, are the essential components for producing the best guess of what the role model effect is in any given country. These best guesses are described by the Best Linear Unbiased Predictions (BLUPs) of the country-level role model effects on job preferences. BLUPs are our best guess of the true role model effect in any one country taking into account the country-level estimates and standard error, and our estimate of the between-country variation in role model effects. We construct Empirical Bayes estimates of these BLUPs adapting the formula of Jackson and Mackevicius (2023., p.8). Formally, we construct the BLUP for country c, $\hat{\theta}_c$, as

$$\hat{\theta}_c = w\hat{\bar{\beta}} + (1 - w)\hat{\beta}_c \,,$$

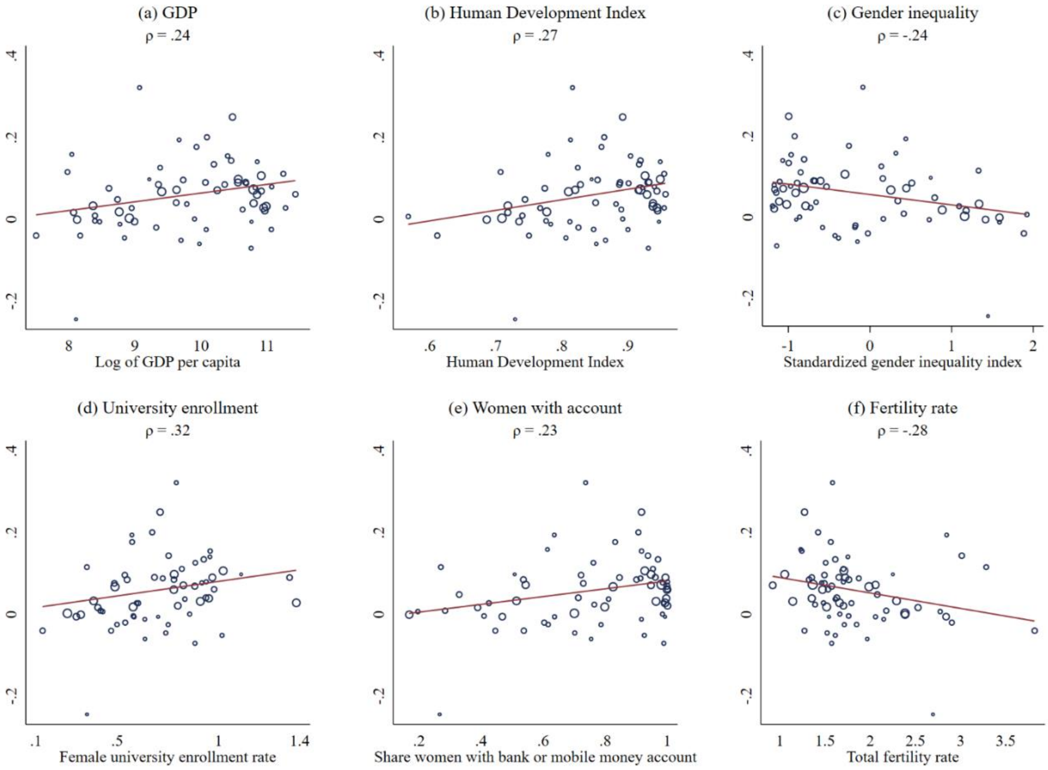
where $\hat{\hat{\beta}}$ is the average role model effect recovered from the meta-analysis on all country-level estimates, $\hat{\beta}_c$ is the role model effect estimate for country c and the weight $w = \hat{\sigma}_c^2/(\hat{\sigma}_c^2 + \hat{\tau}^2)$ is a function of the squared standard error of the country-level estimate ($\hat{\sigma}_c^2$) and the estimated variance of role model effects ($\hat{\tau}^2$), ensuring that more precise country-level estimates get more weight. We report the BLUPs for country-level role model effects on job preferences and all other outcomes in an interactive map on our dedicated study website https://www.role-model-effects.com/.

Finally, we explore country-level heterogeneity in role model 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 role model effect on job preferences on the y-axis and a given characteristic, for example, GDP per capita, on the x-axis. For brevity, we describe details on how we measured these characteristics in the respective figure notes. These scatterplots allow the reader to visually inspect the relationship between those two variables.

Second, we use meta-regressions to estimate separate role model 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-effect 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 role model effect for below-median countries, and we get the estimated role model 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 role model effects are related to a country's economic development, gender inequality, or sex differences in math and science performance.

Economic development. Role model 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). Figures 7 (a) and (b) show that role model effects on job preferences are positively related to the log of a country's GDP per capita and a country's HDI. Our regressions confirm these results. Role model effects are significantly larger in countries with above-median GDP per capita (0.0739 SD compared to 0.0502 SD) and in countries that have an above-median HDI (0.0746 SD compared to 0.0494 SD).

Figure 7: Role Model Effects in Job Preferences and Country-Level Correlates



Notes: These panels show the bivariate relationships between the estimated role model effects on standardized job preferences shown in Figure 6 (on y-axes) and different country-level characteristics (on 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: $GII = \sqrt[3]{Health * Empowerment * LFPR}$ where Health is computed as $Health = (\sqrt{\frac{10}{MMR} * \frac{1}{ABR}}+1)/2$ where MMR is maternal mortality rate and ABR is the adolescent birth rate. Empowerment is computed as $Empowerment = (\sqrt{PR_F * SE_F} + \sqrt{PR_M * SE_M})/2$ 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: $LFPR = \frac{LFPR_F+LFPR_M}{2}$. 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.

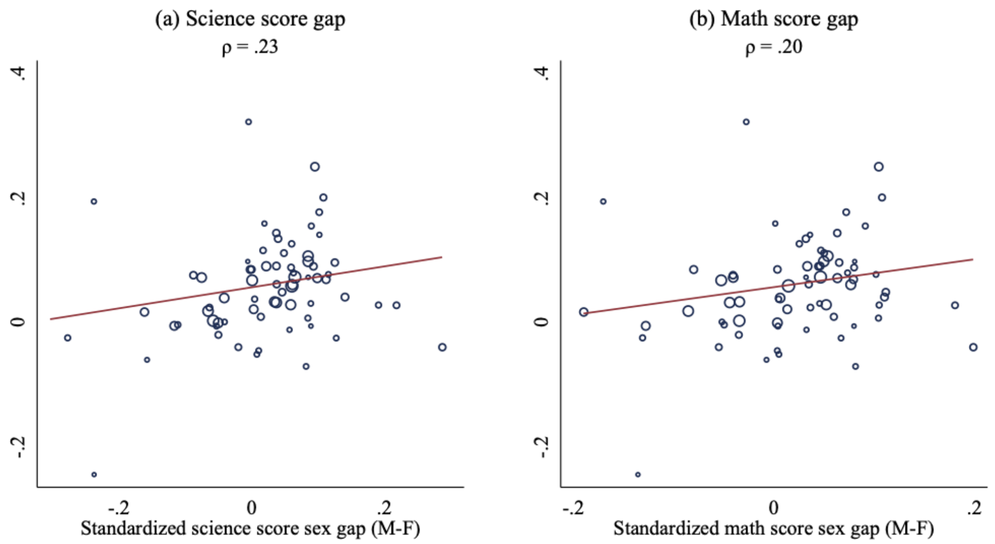
Gender inequality. Role model effects might be stronger in gender-unequal countries where women face systemic barriers to education and the workplace. Or role model effects might be stronger in gender-equal countries in which people are more aware of the remaining gender gaps. We measure gender inequality 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 7 (c) shows that role model effects are smaller in more gender-unequal countries. Our regressions confirm these results: the estimated role model effects are significantly smaller for above-median gender-inequality countries (0.0498 SD versus 0.0724 SD). Figure B3 in the appendix shows that this relationship is driven by role model 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. Figures 7(d) and 7(f) shows that role model effects are larger in countries in which women have higher university enrollment and fewer children. Regressions confirm these results. We see significantly higher role model effects in countries with above-median female university enrollment (0.0725 SD versus 0.0418 SD) and significantly *lower* role model effects in countries with above-median fertility rates (0.0540 SD versus 0.0739 SD). Figure 7 (f) suggests larger role model effects in countries where a higher proportion of women have access to a bank account. However, our regressions show the above-median compared to below-median difference is only significant at the 10 percent level (0.0704 SD versus 0.0514 SD).

Sex gaps in math and science test scores. Role model 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 8: Role Model Effects on Job Preferences and Test Score Gaps between Boys and Girls



Notes: This figure shows the relationship between the estimated role model effects on standardized job preferences shown in Figure 6 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 role model effects on job preferences.

Figure 8 shows that role model effects are larger in countries with larger performance gaps in favor boys for science and math. We also estimate separate role model effects for countries with above and below median boy–girl performance gaps. These regressions confirm our previous results. The estimated role model effect for above-median countries, where boys tend to outperform girls in science is 0.0939 SD and for below-median countries is 0.0371 SD.

Heterogeneity of role model effects on subject enjoyment and confidence. The heterogenous role model effects on subject enjoyment and subject confidence broadly mirror the pattern for role model effects on job preferences. We show in Appendix B that role model 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 B4, B5, B6 and B7 in the appendix). More generally, we see role model effects on these two outcomes are correlated with role model effects on job preferences. The correlation between role model effects on job preference and role model effects on enjoyment is 0.50. The correlation between role model effects on job preferences and role model effects on confidence is 0.31. In countries where role models have a stronger effect on students' job preferences, we also see stronger role model effects on how much students enjoy a subject and how confident they feel about it.

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

Understanding which environmental factors cause differences in role model 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 role model 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]

9. Conclusion

There is a widespread belief that the lack of same-sex role models exacerbates gender inequalities in education. Educators, politicians, and NGOs and have called for hiring more female teachers to boost girls' performance in math and science and to motivate girls to enter STEM jobs. Similarly, hiring more male teachers in primary school has become a policy target to stop boys from falling behind at that stage of education. Therefore, determining whether role model effects exist and how strong they is central for the design of policies that aim to increase representation through diversifying the teaching profession.

Our study provides comprehensive evidence on role model effects from a meta-analysis and our own multi-country analysis. We establish that role models have a negligible effect on performance. Our meta-analysis shows an average role model effect on students' performance in primary and secondary education of 0.030 SD. Our multi-country analysis finds an even smaller average role model effect on test scores of 0.015 SD. Furthermore, we show role model effects on test scores are small in most countries. We see larger average effects and more

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

variation for non-test score outcomes. For example, we find role model effects on job preferences of on average 0.064 SD, which are more pronounced in rich and gender-equal countries. Taken together, our results suggest that hiring more male teachers in primary school or more female teachers in STEM subjects will not close sex gaps in student performance. However, hiring more female STEM high school teachers promises to be an effective tool for reducing sex segregation in the labor market in rich and gender-equal countries.

In addition to establishing these policy-relevant results, our paper showcases the scientific benefits of answering one research question by combining data from multiple settings. This approach gave us enough statistical power to detect a statistically significant but tiny effect on test scores. Having data from many countries also allowed us to estimate the distribution of role model effects and thoroughly explore which settings show substantial role model effects for students' job preferences. In contrast to our meta-analysis, we could conduct this multi-country analysis without worrying about differences in methodology and publication bias. Moreover, combining our multi-country analysis using meta-analysis methods goes a long way toward answering the question of how universal same-sex role model effects are. With this approach we combine the best of both worlds—exploiting meta-analysis methods to distill the distribution of true effects from noisy estimates while using credible and comparable quasi-experimental estimates across many contexts as inputs.

We see studies that combine causal estimates from many settings as the next step in the credibility revolution in economics. As a discipline, we have become much better at producing credible causal estimates for one specific setting. We now see an increase in studies that apply the same methodological rigor to data from multiple settings and carefully explore whether and why effects differ by context. We hope this trend continues.

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. (2022). When should you adjust standard errors for clustering? *The Quarterly Journal of Economics*, 138(1), 1–35.

Abeler, J., Falk, A., & Kosse, F. (2021). Malleability of preferences for honesty. *IZA Discussion Paper No. 14304.*

Alan, S., Baysan, C., Gumren, M., & Kubilay, E. (2021). Building social cohesion in ethnically mixed schools: An intervention on perspective taking. *The Quarterly Journal of Economics*, 136(4), 2147–2194. DOI: https://doi.org/10.1093/qje/qjab009

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*.

Andersen, I. G., & Reimer, D. (2019). Same-gender teacher assignment, instructional strategies, and student achievement: New evidence on the mechanisms generating same-gender teacher effects. *Research in Social Stratification and Mobility*, 62, 100406. DOI: https://doi.org/10.1016/j.rssm.2019.05.001

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

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

Arel-Bundock, V., Briggs, R. C., Doucouliagos, H., Mendoza Aviña, M., & Stanley, T. D. (2022). Quantitative political science research is greatly underpowered, *I4R Discussion Paper Series No. 6*.

Asarta, C., Butters, R. B., & Thompson, E. (2013). The gender question in economic education: Is it the teacher or the test? *University of Delaware – Department of Economics*, Working Papers No. 13–12

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*.

Bisbee, J., Dehejia, R., Pop-Eleches, C., & Samii, C. (2017). Local instruments, global extrapolation: External validity of the labor supply–fertility local average treatment effect. *Journal of Labor Economics*, 35(S1), S99-S147.

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.

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

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.

Cho, I. (2012). The effect of teacher-student gender matching: Evidence from OECD countries. *Economics of Education Review*, 31(3), 54–67. DOI: https://doi.org/10.1016/j.econedurev.2012.02.002

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

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

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. DOI: https://doi.org/10.1177/09567976221094630

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.

Eisenhauer, J. G. (2021). Meta-analysis and mega-analysis: A simple introduction. *Teaching Statistics*, 43(1), 21-27.

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

Fairlie, R. W., Hoffmann, F., & Oreopoulos, P. (2014). A community college instructor like me: Race and ethnicity interactions in the classroom. *American Economic Review*, 104(8),

2567–2591. DOI: https://doi.org/10.1257/aer.104.8.2567
Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, 345(6203), 1502–1505. https://doi.org/10.1126/science.1255484
Gechter, M., Samii, C., Dehejia, R., & Pop-Eleches, C. (2018). Evaluating ex ante counterfactual predictions using ex post causal inference. *arXiv preprint* arXiv:1806.07016.
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.
Gust, S., Hanushek, E. A., & Wößmann, L. (2022). Global universal basic skills: Current deficits and implications for world development. *NBER Working Paper Series*, 30566. http://www.nber.org/papers/w30566.
Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., Brand, R., Brandt, M. J., Brewer, G., Bruyneel, S., Calvillo, D. P., Campbell, W. K., Cannon, P. R., Carlucci, M., Carruth, N. P., Cheung, T., Crowell, A., De Ridder, D. T. D., Dewitte, S., Zwienenberg, M. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, 11(4), 546–573. https://doi.org/10.1177/1745691616652873
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
IntHout, J., Ioannidis, J., Rovers M., & Goeman J. (2016). Plea for routinely presenting prediction intervals in meta-analysis. *BMJ Open,* 6(7). DOI: http://dx.doi.org/10.1136/bmjopen-2015-010247
Ioannidis, J. P., Stanley, T. D., & Doucouliagos, H. (2017). The power of bias in economics research. *Economic Journal,* 127(605). https://doi.org/10.1111/ecoj.12461

Jackson, C. K. (2018). What do test scores miss? The importance of teacher effects on non–test score outcomes. *Journal of Political Economy*, 126(5), 2072-2107.
Jackson, K.C. and Mackevicius, C. (2023) What impacts can we expect from school spending policy? Evidence from evaluations in the U.S. *American Economic Journal: Applied Economics*.
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.
Kleven, H., Landais, C., Posch, J., Steinhauer, A., & Zweimuller, J. (2019). Child penalties across countries: Evidence and explanations. *AEA Papers and Proceedings,* 109, 122–26. DOI: https://doi.org/10.1257/pandp.20191078
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.
Kvarven, A., Strømland, E., & Johannesson, M. (2020). Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*, 4(4), 423–434. DOI: https://doi.org/10.1038/s41562-019-0787-z
Lavy, V., & Sand, E. (2018). On the origins of gender gaps in human capital: Short- and long-term consequences of teachers' biases. *Journal of Public Economics*, 167, 263–279. DOI: https://doi.org/10.1016/j.jpubeco.2018.09.007
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
– (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.
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.
Miller, D. L., Shenhav, N. A., & Grosz, M. (2021). Selection into identification in fixed effects models, with application to Head Start. *Journal of Human Resources*, 0520-10930R1.
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.

Nixon, L., & Robinson, M. (1999). The educational attainment of young women: Role model effects of female high school faculty. *Demography*, 36(2), 185–194. DOI: https://doi.org/10.2307/2648107

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

Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2), 187–204. DOI: https://doi.org/10.1080/07350015.2016.1227711

Paredes, V. (2014). A teacher like me or a student like me? Role model versus teacher bias effect. *Economics of Education Review*, 39(C), 38–49. DOI: https://doi.org/10.1016/j.econedurev.2013

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

Paule, R. C., & Mandel, J. (1982). Consensus values and weighting factors. *Journal of Research of the National Bureau of Standards*, 87(5), 377–385.

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.

Pritchett, L., & Sandefur, J. (2014). Context matters for size: why external validity claims and development practice do not mix. *Journal of Globalization and Development*, 4(2), 161-197.

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.*

Raudenbush, S. W. (2009). Analyzing effect sizes: Random-effects models. In: Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). *The Handbook of Research Synthesis and Meta–Analysis*; 2nd edition, Russell Sage Foundation, 295–315.

Rohrer, J. M., Egloff, B., & Schmukle, S. C. (2015). Examining the effects of birth order on personality. *Proceedings of the National Academy of Sciences*, 112(46), 14224–14229. DOI: https://doi.org/10.1073/pnas.1506451112

Rothstein, D. S. (1995). Do female faculty influence female students' educational and labor market attainments? *ILR Review*, 48(3), 515–530. DOI: https://doi.org/10.1177/001979399504800310

Sidik, K., & Jonkman, J. N. (2019). A note on the empirical Bayes heterogeneity variance estimator in meta-analysis. *Statistics in Medicine,* 38(20), 3804–16. DOI: https://doi.org/10.1002/sim.8197

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, 35(1), 4–28. DOI: https://doi.org/10.1006/jesp.1998.1373

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

Stoet, G., & Geary, D. C. (2018). The gender-equality paradox in science, technology, engineering, and mathematics education. *Psychological Science*, 29(4), 581-593.

Sung, Y. J., Schwander, K., Arnett, D. K., Kardia, S. L., Rankinen, T., Bouchard, C., ... & Rao, D. C. (2014). An empirical comparison of meta-analysis and mega-analysis of individual participant data for identifying gene-environment interactions. *Genetic Epidemiology*, 38(4), 369-378.

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)

Veroniki, A. A., Jackson, D., Viechtbauer W., Bender R., Bowden, J., Knapp, G., Kuss, O., Higgins, J. P. T., Langan, D., & Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods*, 7, 55–79.

Viechtbauer, W. (2005). Bias and efficiency of meta–analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics,* 30(3), 261–93.

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.

Winters, M. A., Haight, R. C., Swaim, T. T., & Pickering, K. A. (2013). The effect of same-gender teacher assignment on student achievement in the elementary and secondary grades: Evidence from panel data. *Economics of Education Review*, 34(C), 69–75. DOI: https://doi.org/10.1016/j.econedurev.2013

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).