Do Same-Sex Teachers Affect Test Scores and Job Preferences?

A Super-Study and a Meta-Analysis on 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)

December 2022

Abstract

Previous studies provide contradicting evidence on same-sex role model effects in education. We resolve those contradictions with a meta-analysis and a super-study. Our meta-analysis summarizes 538 estimates, and our super-study provides new evidence from 90 countries and 3 million students. Both approaches show that role model effects on students' performance are small: 0.030 SD in the meta-analysis and 0.015 SD in the super-study. Moving beyond test scores, our super-study documents larger role model effects of 0.063 SD on job preferences, which are concentrated in rich and gender-equal countries. Furthermore, our results suggest that same-sex role model effects can be positive or negative for 4th grade students, but are near universally positive for 8th grade students.

Keywords: Same-sex role models, same-sex teacher, gender role model, student-teacher gender match, standardized test scores, grade, STEM, Science, Math, Reading

JEL classification: I21, I24, J24

---

* de Gendre: Department of Economics, The University of Melbourne and IZA. Feld: School of Economics and Finance, Victoria University of Wellington and IZA. Salamanca: Melbourne Institute: Applied Economics & Social Research, The University of Melbourne and IZA. Ulf Zölitz: University of Zurich, Department of Economics and Jacobs Center for Productive Youth and Child Development, CESifo, CEPR, IZA, Schönberggasse 1, 8001 Zurich, Switzerland 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 CE140100027 and CE200100025). Anna Valyogos, Francesco Serra, Matthew Bonci, Andrea Hofer, Timo Haller, Ana Bras, Albert Thieme and Madeleine Smith and provided outstanding research assistance. We received valuable comments from Luke Chu, Harold Cuffe, Nathan Kettlewell, Julia Rohrer, and seminar participants at Bocconi, University of Canterbury, University of Western Australia, University of Zurich and Victoria University of Wellington.

1. Introduction

Role models could serve as a powerful policy tool to reduce inequality in education. Exposure to more female STEM teachers, for example, promises to increase women's STEM performance and representation. Similarly, exposure to more male primary school teachers promises to stop boy's underperformance. While the idea of same-sex teachers boosting performance has inspired calls for policy interventions[1], it is not clear role model effects deliver what they promise. Recently published studies have shown role model effects on student performance that are positive (Gong et al., 2018), insignificant (Andersen and Reimer, 2019) and even negative (Antecol et al., 2015). While understanding the influence of same-sex teachers is a key issue, there exists no systematic evidence on the magnitude of these effects. There also exists no evidence on how universal these effects are from a global perspective and whether we should expect to find positive role model effects in most settings.

In the first part of our paper, we fill this gap with a meta-analysis of the existing literature. Our meta-analysis identifies 538 estimates from 24 studies and finds an average role model effect of 0.030 standard deviations (SD) for grades and test scores in primary and secondary education. After correcting for publication bias, we find even smaller role model effects, sometimes not distinguishable from zero. While our meta-analysis provides a useful summary of the role model effects literature, it has two important shortcomings. First, our meta-analysis is sensitive to how we correct for publication bias – effects disappear or even change sign depending on the method used. 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. It is therefore impossible to judge to what extent differences in role model effect

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

estimates reflect differences in empirical approaches versus true heterogeneity. To conclusively determine if and in which contexts role models matter, we need a different kind of study.

In the second part of our paper, we estimate role model effects with a super-study. Super-studies answer one research question by applying the same methodology to data from multiple settings. This increasingly popular approach in the social sciences has three key advantages. First, by combining data from multiple settings, super-studies allow for larger sample sizes which makes it possible to detect smaller average effects. This feature is particularly important when plausible effect sizes are small. Second, super-studies allow for testing how universal effects are across different contexts. Third, by holding the methodology constant super-studies make it easier to explore what explains differences in effects across contexts. Taken together, we think that the advantages of this approach are greatly underappreciated, and believe it deserves a distinct name.

We conduct a super-study by estimating role model effects with a consistent methodology using data from 90 countries. 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 super-study 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 study role model effects. First, test scores in this data are designed to be comparable between countries. This allows us to make cross-country comparisons of same-sex 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 effect 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 unobserved student and teacher characteristics constant. 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.

The results of our super-study show that the impact of same-sex role models on test scores is on average very small. The average role model effect on students' test scores is 0.015 SD. Across all specifications, the 99% confidence intervals allow us to rule out effects smaller than 0.009 and larger than 0.022 SD.

Focusing on this small average role model effect and ignoring the underlying distribution of effects would be misleading if there are many settings with meaningful positive or negative role model effects. To study how universal role model effects are and whether there are many settings with meaningful role model effects we draw on meta-analysis methods: We estimate the distribution of latent role model effects at the country level and find that role model

effects on test scores are near universally positive. Our estimates suggest that only in 2 percent of the countries we should expect negative role model effects. However, our analysis also shows that role model effects are small in all settings. We can rule out that they exceed 0.05 SD in any of the 90 countries (p-value<0.001).These results suggest that same-sex role models will do little to close performance gaps.

While same-sex role model effects on performance are very small, we see meaningful effects beyond test-scores. We find same-sex role model effects on students' preferences for working in a job that involves math or science of on average 0.064 SD. The distribution of the country level latent effect highlights that role model effects on job preferences are universally positive, but also vary depending on the setting. In 40 percent of countries latent estimates exceed 0.05 SD.

What country factors explain this global heterogeneity in how much role models affect job preferences? We show that effects are particularly pronounced in rich and gender equal countries, which also happen to be the countries where women are particularly underrepresented in STEM fields (see Breda et al. 2000 on the "gender equality paradox"). Same-sex role models may therefore be a particularly useful policy tool to increase women's representation in STEM in countries like the US, Canada, and Europe.

This paper makes three key contributions. First, our meta-analysis provides the first quantitative summary of the literature on same-sex role model effects on student performance. Our results uncover challenges to (1) correct for publication bias and to (2) compare role models effects across settings. Second, our super-study overcomes these challenges and provides new rigorous evidence on role mode effects on performance and job preferences using data from more than three million students. By holding the methodology constant and estimating the global distribution of role model effects we shed light on how universal role model effects are. Third, we introduce the term super-study and highlight the benefits of this

approach. While the credibility revolution in Economics has shifted attention to causal identification, we have turned a blind eye to the generalizability of results. This has created a situation where many literatures do not agree about effects, describe existing evidence as "mixed" or attribute inconsistent results to differences in settings without empirically engaging with the question of universality. Today's availability of bigger and better data from a variety of settings makes super-studies a powerful tool that allows for an empirically founded discussion of how universal effects are.

While our paper introduces the term super-study, several recent papers also combine data from a variety of settings and to estimate effects with a consistent methodology. In economics, Altmejd et al. (2021) conduct a super-study on sibling spillovers in study choices in Chile, Croatia, Sweden, and the United States. They find consistent results across these substantially different settings suggesting that sibling spillovers are a universal phenomenon that is not driven by institutional details. Kleven et al. (2019) conduct a super-study on how the arrival of a child affects women's and men's earnings in Austria, Denmark, Germany, Sweden, the UK and the US. Using the same empirical approach, econometric specification and sample selection criteria allows the authors to obtain comparable results and to explore how countries family policies and gender norms contribute to the substantial heterogeneity in child penalties in different settings. DellaVigna and Linos (2022) investigate the effectiveness of nudge interventions. Like our paper, they also compare estimates from a super-study with estimates from a meta-analysis. Their study covers 126 nudge interventions run by two large nudge units in the United States covering 12 million people. Their meta-analysis covers 74 estimates of similar effects published in the academic literature. All included effect sizes come from randomized controlled trials, ruling out endogeneity concerns that often consume economists' attention. Yet, effect sizes published in the academic literature are four times as

large as those in the field (8.4 vs. 1.4 percentage points). The authors show that most of this gap can be explained by publication bias exacerbated by low statistical power.

In psychology, super-studies recently settled three controversial debates. Hagger et al. (2016) show in a collaboration of 23 laboratories that there are no meaningful effects of ego depletion on subsequent cognitive performance. Similarly, two super-studies have combined data from multiple large representative surveys to show that neither birth order nor siblings sex affect personality (Rohrer et al., 2015; Dudek et al., 2022). For all scientific disciplines, super-studies are particularly valuable for research questions with a mature literature that disagrees about the effect size or struggles with the generalizability or replicability of findings.

In the remainder of this paper, we investigate the importance of same-sex role models in education. In the next section, we carefully define same-sex role model effects and summarize the literature on these effects using a meta-analysis. In Section 3, we explain more thoroughly what super-studies are and discuss their advantages over typical studies and meta-analyses. We describe the data for our super-study in Section 4 and our empirical strategy in Section 5. Section 6 shows the results of our super-study. We show role model effects estimates for our overall sample and separate estimates for each country. We also investigate what country level factors predict the size of role model effects. Finally, we conclude in Section 7 with a discussion of our results and a reflection on importance of super-studies in the social sciences.

2. Role model effects: What are they and what does the literature say?

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; Lim and Meer, 2017; Muralidharan and Sheth, 2015; Eble and Wu, 2020). Such role model effects are typically estimated with variations of the following regression model:

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

In this model, β_3 captures the role model effect. A positive role model effect could be driven by female students (compared to male students) benefitting more from female teachers, male students (compared to female students) benefitting more from male teachers, or both. 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.

Many studies have estimated role model effects on education and career choices. For example, Carrell et al. (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. (2021) 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%. Neumark and Gardecki (1998) find that female doctoral students with female mentors graduated faster without having worse placements.

Other studies have estimated role model effects on performance in tertiary education. Hoffmann and Oreopoulos (2009) exploit within-student and within-instructor variation and find only small same-sex role model effect of at most 0.05 SD on grades and 1.2 percentage lower probability of dropping a class. These effects are not visible for math and science instructors and disappear when the authors include student fixed effects. Dee (2007) takes a similar two-way teacher and student fixed-effects approach and finds same-sex role model effects on test scores of 0.05 SD. In this paper, we focus on role model effects on test scores and grades 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

We identified 24 studies for our meta-analysis on role model effects on grades and test scores in primary and secondary education (see Table B1 in the appendix for list of included studies). Most of these studies show multiple role model effects estimates. 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 from 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 B1 for more details on how we constructed those estimates and their standard errors). We describe our pre-registration and data collection in greater detail in Appendix B. In this section, we focus on describing the results.

Our included estimates cover many different contexts: 238 use data from Europe, 187 from Asia, 94 from North America, and 19 from Africa; 153 are based on data from primary education, and 375 from secondary education, and 10 from both; 57 estimates come from settings that use experimental methods (i.e., an explicit random manipulation of teacher

assignment to students), and the remaining 481 from settings with naturally occurring classroom assignment; 37 estimates of role model effects are on grades and 501 are on test scores.

We summarize all 538 estimates using a three-level random effects model (Connell, McCoach, and Bell, 2022).[2] This model allows true role model effects to differ by study and accounts for the dependence of role model effect estimates within each study. 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 (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 0.0128 (*p*-value=0.0X).[3] The 95% prediction interval ranges from -0.087 SD and 0.146 SD (Higgins et al., 2009; IntHout et al., 2016). This means that we can expect the effect of a similar future role model study (as those included in our meta-analysis) to lay between these two values with 95% probability.

The estimate of the standard deviation of the distribution of the true role model effect is 0.058 SD. We explore what drives this heterogeneity using a single meta-regression that includes as moderators whether studies use experimental or quasi-experimental variation, the continent where they take place, whether they analyze data from elementary or secondary school students (or a mix of both), and whether they take test scores or grades as outcomes (see Appendix Table A1). The differences described below are therefore accounting for correlated

[2] Appendix Figures A1 and A2 show funnel plots for these role model effects and their standard errors.

[3] 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 of studies which happen to contribute many precise estimates. To check if this is the case, we record the weight of each point estimate (i.e., how much it contributes to the calculation of the overall average effect) and calculate the sum of the weights of the point estimates for each study. The sum of the weights at the study level never exceeds 4.77%, which shows that no individual study has an outsized effect on the estimated average 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 which assumes that 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).

differences in all other moderators included. 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. We see, however, some evidence of geographic heterogeneity. Compared to role model estimates from Africa, role-model effects estimates are 0.096 SD smaller in Asia, 0.033 SD smaller Europe, and 0.144 SD smaller in North America, with the difference between Africa and North America being significant at the 5% level. We also find evidence that role models are 0.094 SD smaller in primary education than they are in secondary education; this difference is significant at the 1% level.

It is unclear to what extent these differences reflect differences in true role model effects across continents or level of education, or whether they are merely driven by differences in study methods. Furthermore, even random differences in methodology would lead to an overestimation of the standard deviation of the true role model effect. This means that our estimated standard deviation of the true role model effect of 0.058 SD likely also reflects differences in methods. The actual standard deviation of the true underlying role model effect is likely smaller.

2.3 Do role model effects studies show publication bias?

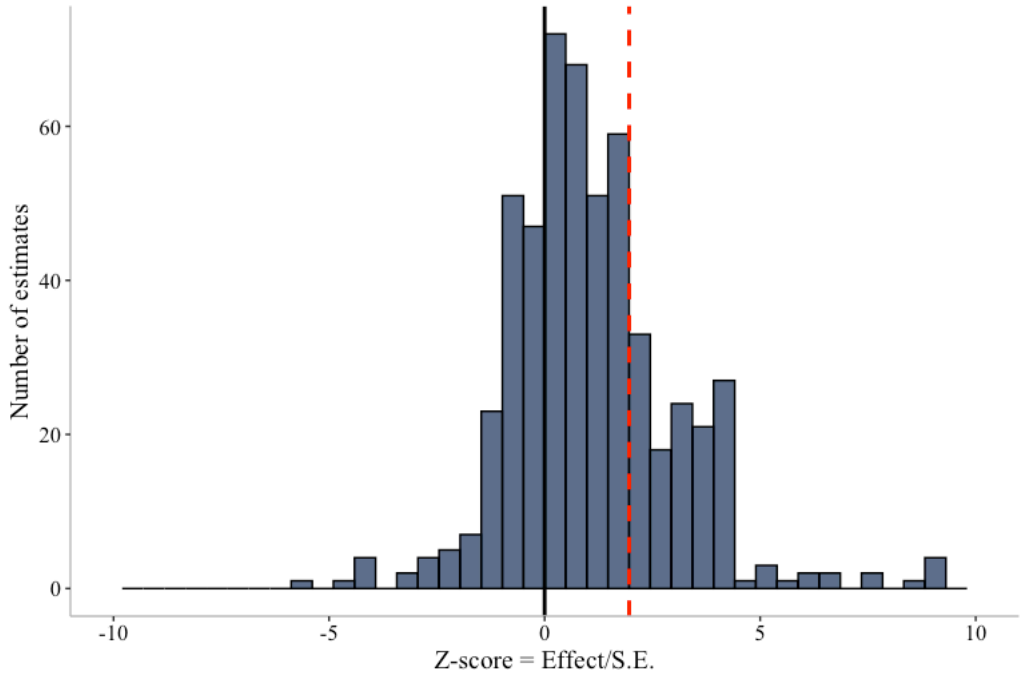
Publication bias could drive our average effect of 0.030 SD. It may be, for example, that researchers are more likely to report specifications that show positive role model effects, and studies that show positive and significant role model effects, either by chance or *p*-hacking, may be more likely to be written up. 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 1.96 – the critical value for statistical significance at the 5% level using two-sided tests (see Brodeur et al., 2016, Andrews and Kasy, 2019; DellaVigna and Linos, 2020). The intuition behind this test is that

in the absence of researchers actively trying to get statistically significant results (or editors and referees preferring significant results), we would expect *z*-scores just above 1.96 to be as likely as *z*-scores just below 1.96. In contrast, *p*-hacking and selective publication based on statistically significance would lead to substantially more *z*-scores just above 1.96.

Figure 1 plots the distribution of all *z*-scores with the vertical dashed line marking the 1.96 critical value. We see no evidence of heaping at the right of this critical value. If anything, there are more *z*-scores below the critical value. Appendix Figures A3 And A4 show similar plots highlighting the critical values for statistical significance at the 90% and 99% levels. These figures also show no evidence of publication bias.

Figure 1: *Z*-score Distribution with Critical Value at 95% Marked



Note: This figure shows *z*-scores of 534 role model effects estimates from all 24 studies. These are all z-scores except for 4 outlier values (with *z*-scores of -22.39, 12.61, 12.68. 12.79) which we excluded to make the figure more readable. The vertical dashed line marks 1.96, the two-sided test critical value of 95% for the normal distribution.

In our second approach, we estimate the relationship between estimated effect sizes and the precision of the estimate. The intuition of this approach is as follows: statistical precision limits the possibility of influencing the size of an estimate. With more precision, any *p*-hacking and selective publication can only have a small effect on the role model estimates that we

observe in the academic literature. In contrast, those forces can have a stronger influence on the estimated effect sizes with less precise estimates. If there is publication bias favoring positive role model effect estimates, we would therefore expect a positive relationship between effects sizes and statistical precision.

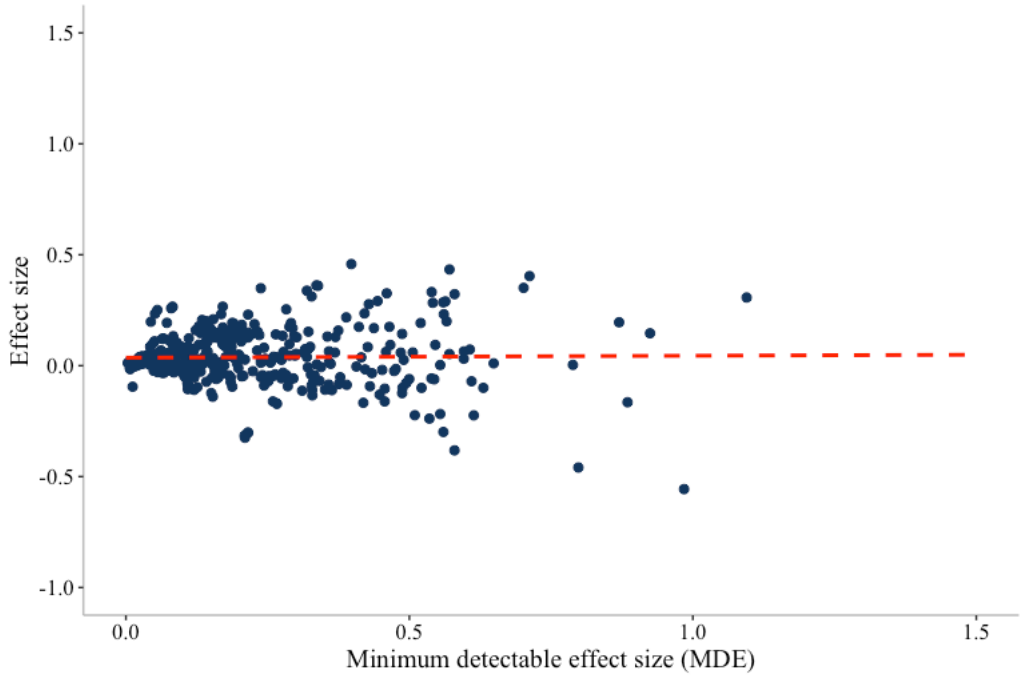
There are three popular ways to estimate the relationship between effect sizes and statistical precision. We apply all three of them. First, we regress the effect size on the ex-post minimum detectable effect size (MDE), which we calculate by multiplying the standard error by 2.8 (see Card and Krueger, 1995). Second, we perform the precision effect test (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 and therefore give more weight to more-precisely estimated effects. Third, we perform Egger's test (Egger et al., 1997). This test consists of regressing *z*-scores on the inverse of the standard error. In contrast to the other two tests, the Egger's test shows evidence of publication bias if the *constant* is statistically significant. In all three regressions, we account for the dependence of estimates within the same study by clustering at the study level.

Figure 2 shows that the estimated effect size significantly increases with the size of the MDEs (p-value < 0.001). This relationship remains positive but is no longer significant once we remove 3 outlier estimates[4] from Ammermüller and Dolton (2006) (p-value = 0.927). The precision effect test and Egger's test results 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).

[4] 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.

Taken together, these results suggest the presence of publication bias in the role model estimates included in our meta-analysis. In the next section, we explore how our estimated average role model effect changes if we correct for publication bias.

Figure 2: Minimum Detectable Effect Size (MDE) of All Role Model Estimates



Note: This figure shows the relationship between role model effect estimates (y-axis) and their corresponding ex-post MDE size (x-axis), calculated by multiplying the standard error by 2.8 (and thus assuming 80% power). Each dot represents a one role model effect estimate. To increase readability, this figure excludes 3 outlying role model estimates of size 1.15, 2.07 and 0.92 SD with MDEs of 14.10, 15.19 and 19.13 SD, respectively. The dashed line shows the linear fit from regressing point estimates on MDEs. The slope estimate from this regression is 0.079, with a standard error of 0.003 clustered at the study level. This regression was run with a sample of all 538 estimates including the outliers. Excluding those three outliers yields a slope of 0.009 and standard error of 0.097.

2.4 How do publication bias corrections affect our estimate?

Figure 3 shows the estimated average 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)

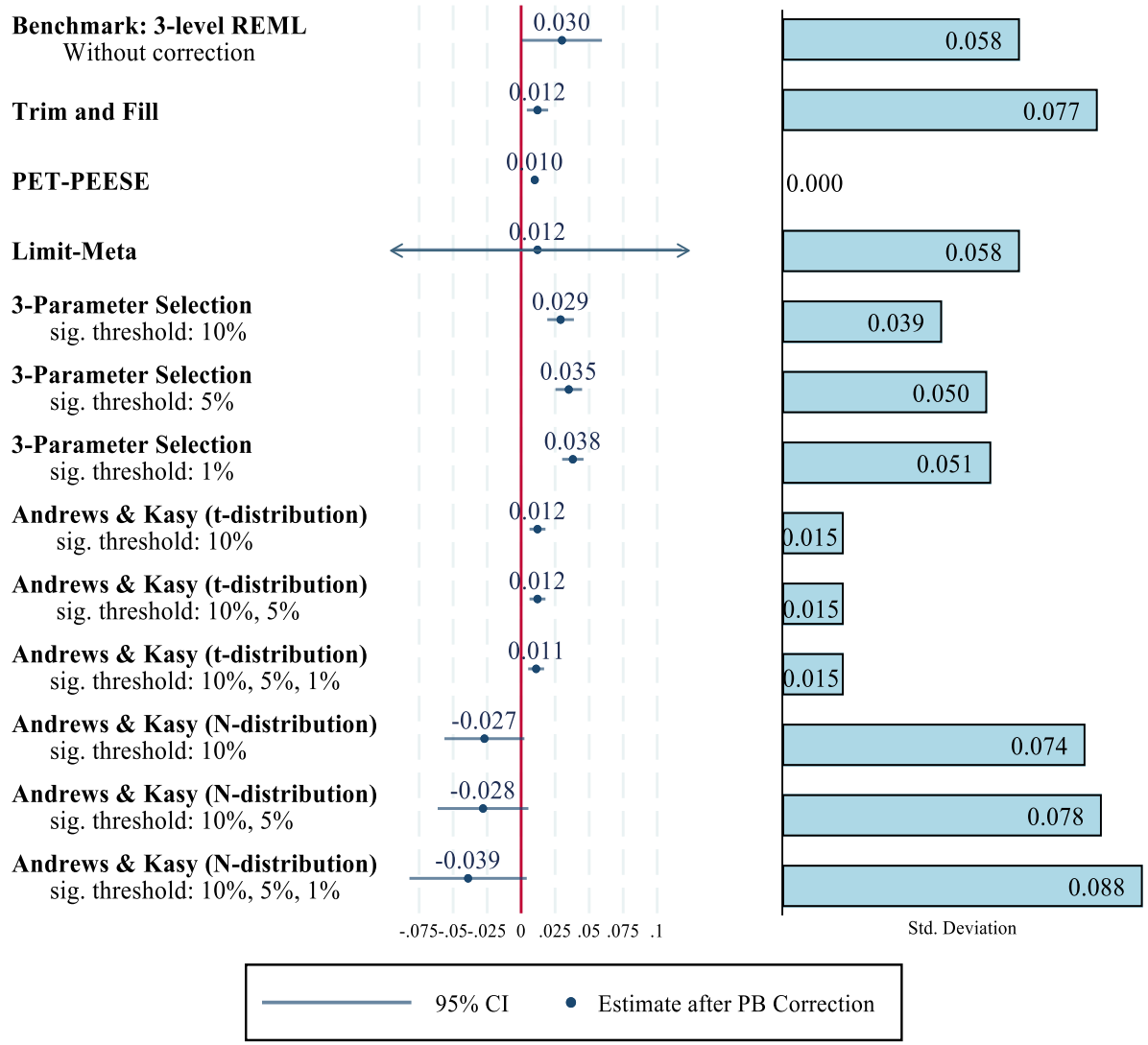
focus on correcting for publication bias by modeling the probability that an estimate is published based on its sign and significance at commonly used significance thresholds.

Figure 3 shows that the different procedures deliver substantially different 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% significance level. Trim and fill, PET-PEESE, and limit-meta reduce the role model estimate to roughly 1/3 to half of the three-level random effect estimate. The three-parameter selection models do not 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 -.027 and -.039 SD.

In Appendix B2 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 amount 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.[5]

[5] We also see similar effect heterogeneity, though with less statistical precision to detect differences; little graphical evidence of publication bias in *z*-scores histograms and funnel plots; more conclusive evidence on MDE plots and related tests; and similar (though generally more muted) publication-bias corrected effects. See Tables B2.1 and B2.2 and Figures B2.1 through B2.6 for these results.

Figure 3: Role Model Estimates After Correcting for Publication Bias



Note: As benchmark, 3-level REML shows the estimated role model effect without correcting for publications 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. Restricted maximum likelihood 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: We use estimates from the precision-effect test (PET) model rather than from the precision-effect estimate with standard error (PEESE) model because the one-sided *t*-test of intercept for the PET model does not reject the null hypothesis at the 5% level. Estimates weighted by their inverse variance. We use the *rma.uni()* function in *R* for implementing this method. Limit-Meta: Uses 3-level REML (see above) as input. The 95% Confidence intervals estimated by this method range from -0.373 SD to 0.397 SD. 3-Parameter Selection: We use 0.05, 0.025, and 0.01 as jumps in the publication probability function. Restricted maximum likelihood estimator of the standard deviation of the effect size and the standard deviation of the effect size. In the figure, the confidence intervals of this estimate were cut for readability reasons, the lower bound being -0.377 and the upper bound 0.396. Andrews and Kasy: We use 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, at the 0.05 and 0.025, and at the 0.05, 0.025 and 0.01 significance levels for both positive and negative effects. We allow the probability of publications bias to be asymmetric. We produce estimate using Kasy's App: https://maxkasy.github.io/home/metastudy. Other correction methods: Andrews and Kasy's (2019) 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, which is equal to zero for PET-PEESE by assumption.

Taken together, the results of our meta-analysis suggest that role model effects are small. We have also shown that there is substantial heterogeneity in role model effect estimates and suggestive evidence for publication bias. The different bias correction methods give a range of estimates. Some suggest small positive role model effects, others suggest effects very close to zero, and some even suggest negative effects. Based on these results, we find it unlikely that the true average role model effects are large and positive. However, we cannot rule out that role model effects are, on average, indistinguishable from zero or even negative. We also do know how role model effects differ by context. To find out if and in which contexts role model effects on students' performance exist, we need a different kind of study. We need a super-study.

3. What is a super-study?

Super-studies are a tool to help us understand how universal an effect is. Let us take same-sex role model effects as an example: Do role-models have a positive impact on students' performance in all or most settings? Knowing the answer to this question is of central importance for both science and policy.

One way to answer this question is to look at a typical study. We can see from our meta-analysis that the median study investigates role-model effects with 10,196 observations from one country. We think of this median study in terms of sample size and number of different countries included as typical for this literature.

A single typical study will give us limited evidence on the universality of role model effects for three reasons. First, the study might not be internally valid. Typical role-model effect studies vary widely in their research designs and the quality of their (quasi-)experiments. *P*-hacking and publication bias are also threats to internal validity. Second, the study might not be externally valid. For example, even if a typical study convincingly shows role-model effects

in the US, we do not know if these generalize to other countries. And third, typical studies often lack statistical power. Even if internally and externally valid, 10,196 observations might not be enough to uncover small role-model effects.

Our confidence in role-model effects being universal (or near universal) should increase with the number of typical studies which show role-model effects in different contexts. If role-model effects have been estimated in 25 studies and all of them show positive effects, we should have more confidence in role-model effects being universally positive. If these studies are conducted in different countries and institutional settings, our confidence in the universality of role-model effects should increase even more. In contrast, if all 25 studies show a mix of positive, negative, and null results, or if all studies were conducted in very similar settings, it could well be that role-model effects are very context dependent. This would also be important to know.

A meta-analysis is also an approach for assessing the universality of role model effects, yet meta-analyses face two large problems. First, they suffer from publication bias. Previous research has shown that null results are less likely to be written up (Franco, et al. 2014). This is a form of survivorship bias. If there are systematically missing estimates, meta-analyses can be misleading. Second, meta-analyses rely on studies conducted by many researchers who had to make many decisions in their analysis (e.g., which controls to include, which observations to drop) which can affect the estimates (Brenzau et al. 2022, Huntington-Klein et al., 2021). These methodological differences can be confused with differences in effects by context, making it difficult for meta-analyses to convincingly show heterogeneous effects.

Super-studies are a tool to find out if effects are (near) universal which addresses shortcomings from meta-analyses. Super-studies generate estimates from different contexts

while holding research methodology constant and avoiding the problem of systematically missing estimates.[6]

What makes a study "super" depends on the characteristics of a typical studies. To conduct a super-study, you need to undertake the following four steps.

1. **Plan a "typical" study**. Such a study is typical in terms of the sample size and context for a given literature.
2. **Extend this plan to multiple different settings**. In your extended plan consider if and how any data must be standardized to be more comparable across settings.
3. **Collect data from multiple settings**. This could mean combining multiple existing datasets or collecting new data in different settings. The more diverse and larger the number of settings the better. To deserve the title "super," a study should include data from at least three settings.
4. **Analyze the data jointly and separately for each setting**. Estimate one overall effect, one effect for each setting, and test whether the effect is stable across settings. If the effect is not stable and there is data from many settings, explore what explains this heterogeneity.

Let us apply these steps to our paper. In our study, we use data from two large-scale international comparison datasets, which gives us data from 90 countries and 4,434,945 observations. Our study is a super study because it contains 90 times the number of countries and more than 90 times the sample size of a typical study. We use this large and diverse sample

[6] In contrast to super-studies, megastudies answer different related research questions holding the methodology and context constant (Milkman et al., 2021). For example, the paper introducing megastudies tests 54 different interventions to increase gym visits in one setting. Besides changing the interventions, the researchers were using a consistent methodology (e.g., they measured the outcome in the same way). Megastudies are a great tool for finding new effective interventions. However, by only testing these interventions in one setting it is not clear to what extent these effects generalize. This is a question that can be answered with a super-study. We see megastudies and super-studies as complementary. Megastudies are particularly useful for finding promising interventions; super-studies are useful for testing to what extent the effects of these promising interventions are generalizable.

to estimate one overall role model effects, one role-model effect for each country, and explore whether and how these effects differ between countries.

The definition of super-study is explicitly benchmarked against what is typically done in a literature in terms of numbers of settings and sample sizes. It forces us to go beyond the norm to earn the adjective super. A study that uses a typical setting and sample size but investigates several heterogeneous effects across subgroups is not a super-study. Testing whether an effect differs, for example, by race, state, or decade is not enough.

This difference between typical and super-studies can also be seen in other contexts. Typical studies may use data from one household survey, use administrative data from one country, or run experiments in one laboratory or one field setting. Super-studies, in contrast, use data from multiple household surveys (e.g., Dudek et al., 2022), use administrative data from multiple countries (e.g., Altmejd et al. 2021), or have conducted the same experiment in multiple laboratories (e.g., Hagger et al. 2016).

In all super-studies, researchers try to apply the same methodology across different settings. For example, they estimate the same econometric models, apply the same sample restrictions, and follow the same experimental instructions in different contexts. However, no super-study uses the exact same methodology. Some changes are necessary to make data more comparable. For example, if household surveys measure an outcome using different scales, then researchers should harmonize these scales. When running the same experiment in different countries, researchers need to translate the instructions and pay their subjects in the local currency. Super-studies make these harmonization choices transparently and always keeping in mind the goal of wanting to generate estimates in different settings that are comparable.

Super-studies are not worthwhile or feasible for answering all questions. For example, some research questions are only about only one specific context and concerns about the generalizability of the findings are not part of the scientific discourse (e.g., Does Harvard

discriminate against Asians in their admissions?). Other studies exploit a unique natural experiment that might not be able to be replicated in other settings (e.g., the impact of the Mariel boatlift on local wages). Yet other literatures might have converged to a consensus without having to conduct super-studies. Moreover, super-studies are costly. It takes more time and money to combine data from many settings. Depending on the research question and state of the literature, the scientific return to these extra costs might be too low.

We therefore believe super-studies are particularly valuable for mature literatures that have failed to reach a consensus. To see why, think of two important reasons why consensus might have not been reached in a mature literature.

First, lack of consensus can occur in a literature if the effect in question is small or non-existent and where typical studies are underpowered. Using data from many settings, as a super-study would, leads to larger sample sizes and makes it easier to either detect small effects or rule out the existence of meaningful ones. Having larger samples also decreases concerns about *p*-hacking. With a sufficiently large sample, researchers will either find an effect if it exists or will find it easier to publish a convincing null result. While some typical studies have large enough samples (e.g., those using administrative data), two recent papers have shown that studies in economics and political science tend to have too small sample sizes (Ioannidis et al., 2017; Arel-Bundock et al., 2022). Especially if any plausible effect is small, there is an obvious benefit to increasing the sample size by using data from many settings.

Second, lack of consensus could happen in literatures where the effect in question is highly context specific yet the nature of this heterogeneity difficult to grasp because researchers use different methodologies. By using the same methodology for combining data from different

settings, super-studies allow researchers to better investigate how results differ across different contexts.[7]

While we believe that super-studies should play a more prominent role in research, we believe that typical studies will continue to make important contributions. For example, our super-study builds on previous studies that introduced the research question and postulated that there might be role model effects (e.g., Dee, 2007). We also build on discussions in previous studies on how to estimate and think about role model effects (e.g., Muralidharan and Seth, 2016). Our super-study would be much less interesting if there had not been so many previous studies. We stand on the shoulder of giants.

4. Super-study data

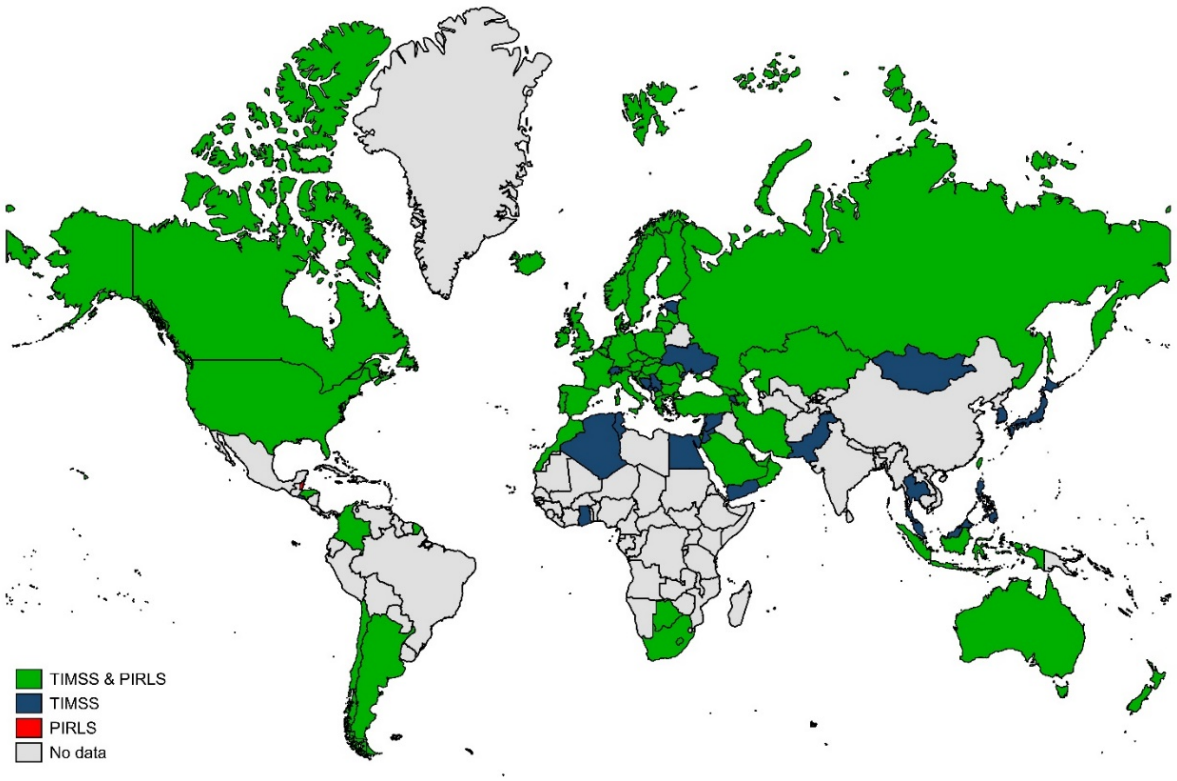
For our super-study on same-sex role model effects, we use data from TIMSS and PIRLS. Both studies were administered by the International Association for the Evaluation of Educational Achievement (IEA), which specializes in administering education assessments that allow for international comparisons. TIMSS measures the skills and knowledge in mathematics and sciences of 4th graders (nine- to ten-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 available 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. We describe how we combine the data and the observations we had to exclude due to data implementation issues in Appendix B. After these

[7] Super-studies are not limited to one methodology. They can estimate causal effect in different settings or, for example, simply document gaps (e.g., Gust et al. 2022). Not all super-studies are good. For example, one can apply the wrong methodology to multiple settings, producing universally biased estimates.

exclusions, we are left with 703 country-study-grade-wave combinations from 90 countries covering 1995–2019. Figure 4 shows which countries were included in at least one wave for each study.

Figure 4: Countries for which we have data from TIMSS, PIRLS or both



Note: The countries in red are those for which we only have data on PIRLS. These are Trinidad and Tobago, Belize, Luxembourg, and Macao. They are hard to see on the map since 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 sample 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% of cases. Appendix Table A2 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 C2 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 the 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 "Agree a lot" to "Disagree 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 the 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 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
Panel C: Subject Enjoyment			
Math	TIMSS	4 & 8	I enjoy learning mathematics
Science	TIMSS	4 & 8	I enjoy learning science
Reading	PIRLS	4	I enjoy reading

Note: This table shows the item wording for the questions measuring job preferences, subject enjoyment, and subject confidence. The job preference questions are preceded by the text "How much do you agree with these statements about [mathematics/science/biology]?". The 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 which 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 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 science teachers, teacher sex then becomes the share of female science teachers.[8]

5. Empirical Strategy

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

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

where $Score_{isj}$ is the test score of student i in subject s that is taught by teacher j. $Female\ Student_i$ is a dummy variable indicating the sex of the student, $Female\ Teacher_j$ is the share of female teachers in subject s (which is equivalent to a dummy variable when students 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 benefit from having a same-sex teacher, on top of the general effect of having a female teacher. We estimate Equation (1) via ordinary least squares regressions (OLS) and

[8] Appendix Table A9 shows that our main results are robust to excluding multiple-teacher classrooms.

cluster our standard errors at the classroom level following the criteria outlined in Abadie et al. (2017).[9]

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. To interpret our results in terms of "global" standard deviations, we therefore standardize the test scores by subtracting 500 and dividing by 100.

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 Seth, 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 equivalent test score difference of students taught by a male 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 benefit 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.*

[9]Abadie et al. (2017) 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.

Besides the role model effect, the role model estimate 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., FemaleStudent × 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., FemaleTeacher × 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 different sets of fixed effects.

Country fixed-effects specification: In our least restrictive specification, we estimate role model effects with country fixed effects. The identifying assumption for this specification is

that *within a country*, our variable of interest—$FemaleStudent_i \times FemaleTeacher_j$—is unrelated to unobserved factors affecting students' test scores. These estimates allow us to retain a large estimation sample and serve as a baseline for how results change with more-restrictive fixed effects.

School fixed-effects specification: Parents choose their children's schools, either directly or indirectly by choosing where to live. Similarly, teachers can influence which schools they work for. 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 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 courses have the same share of female teachers (e.g., all sampled courses have 50% 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).

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: Balancing Tests

		Role model effect				
	Mean	Coef.	Std.err.	*R*-Squared	Countries	Obs.
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) have 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

Note: 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 included 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 x science, female student x math, female student x reading), interaction terms of all three subject dummies with the female teacher dummy (female-teacher x science, female-teacher x math, female-teacher x reading). The number of observations differs depending on the availability of data on predetermined characteristics. Appendix Table A3 replicates this balancing test for our preferred estimation sample. Standard errors clustered at the classroom level are in parentheses. ***, **, and * mark estimates statistically different from zero at the 90%, 95%, and 99% significance levels.

Table 2 shows the results of these balancing tests. Out of ten coefficients of interest, eight 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% larger for female students than for male students. In our preferred specification, these small imbalances do not affect our results because we include student and teacher fixed effects.

Classroom fixed-effects specification: Another concern is that within schools there is systematic assignment of students and teachers that is not captured by our observed student and teacher characteristics. For example, it might be that within schools, female teachers are assigned to classrooms with particularly able girls and male teachers are assigned to particularly able boys. We address this concern by including classroom fixed effects.

Our classroom specification uses *within-classroom across-subject variation* in teacher sex to hold constant all average characteristics of students in a classroom, such as average student ability. For example, we exploit that the same classroom may have a female science teacher and a male math teacher. This set of fixed effects requires that we exclude observations from PIRLS, which only have one test score, and classrooms that have no variation in teacher sex or in $FemaleStudent_i \times FemaleTeacher_j$.

The identifying assumption for this specification is that *within classes*, $Female\ Student_i \times Female\ Teacher_j$ is unrelated to unobserved variables affecting students' test scores.

Student fixed-effects specification: Including student fixed effects has two advantages over including classroom fixed effects. First, it guards against biases caused by sorting of students within a classroom to different teachers. For example, high-ability girls in one classroom might be more likely to take biology, which is taught by a female teacher, whereas high-ability boys might be more likely to take chemistry, which is taught by a male teacher. Second, student fixed effects increase the precision of our estimates by greatly reducing the unexplained variation in test scores.

This specification requires excluding observations from students who have only one test score, students who were taught only by teachers of one sex, and students who were taught by two teaching teams with the same share of female teachers. We also note that the coefficient on the female student dummy is not estimated since this variable is perfectly colinear with the student fixed effects.

The identifying assumption for our student fixed effects specification is that *within students*, $Female\ Student_i \times Female\ Teacher_j$ is unrelated to unobserved factors affecting students' test scores.

Preferred specification—student fixed effects and teacher fixed effects: In our preferred specification, we include student fixed effects and teacher fixed effects. Adding teacher fixed effects addresses one main concern: that more-effective teachers could be assigned to a higher share of students of their own sex.

For this specification, the sample restrictions are almost identical to those of our student fixed-effects specification, with the additional exclusion of 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 since 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.

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 since 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.[10] 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.

[10] 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.

We are not concerned with 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 within their school.

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.

Finally, we also use our preferred specification—with the same independent variables—to estimate role model effects on subject enjoyment, subject confidence, and the preference for a job in given subject. We standardize each of these variables to have means of zero and standard deviations of one in our base dataset (see Appendix C.3). This approach allows us to interpret our results in terms of "global" standard deviations in these outcomes, too.

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 least restrictive sample, we have data from 3,002,411 students who are on average 11 years old. Ten percent of them are foreign born, 75% speak the test language at

home, and 38% have at least one parent with a university degree. For these students, we observe 3,675,156 math scores, 3,674,236 science scores, and 759,789 reading scores. We also observe 201,865 teachers of whom 71% are female, who have on average 16 years of teaching experience; 29% have a bachelor's degree or higher.

Table 3: Summary Statistics for Our Most and Least Restrictive Estimation Samples

	Country FE sample		Preferred specification sample			
	N	Mean	N	Mean	Female	Male
Student characteristics:						
Female	3,002,411	0.49	567,162	0.49	1	0
Age (years)	2,992,158	11.4	565,055	13.4	13.4	13.4
Foreign-born	2,229,873	0.10	532,059	0.09	0.08	0.09
25+ books at home	2,898,794	0.58	553,911	0.54	0.56	0.53
Speaks test language at home	2,856,006	0.75	548,401	0.73	0.73	0.72
Parent(s) have university degree	897,833	0.38	388,492	0.36	0.35	0.37
Teacher characteristics:						
Female	201,865	0.71	49,018	0.54	1	0
Experience (years)	197,769	16.5	48,154	16.0	15.3	15.9
40+ years old	201,408	0.69	48,925	0.84	0.59	0.59
Bachelor degree or higher	195,748	0.29	47,245	0.33	0.28	0.26
Majored in education	170,855	0.71	40,657	0.60	0.64	0.62
Teaches field of expertise	135,333	0.75	42,788	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				

Note: 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 in the first panel, unique teachers in the second panel, and unique student-by-subject-matter combinations in the third panel. The country fixed effects sample consists of 3,047,752 unique students, 231,942 unique teachers, 105,916 unique schools, 144,372 unique classrooms from 90 countries. The preferred estimation sample consists of 567,162 unique students, 49,018 unique teachers, 22,004 unique schools, 26,137 unique classrooms from 82 countries.

In our preferred specification sample, we observe 567,162 different students who are on average 13.6 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. Besides the increase in age, the students have similar characteristics on average. For example, 9% are foreign born (compared to 10% in our country fixed-effects sample), 73% speak the test language at home (compared to 75%), and 36% have at least one parent with a university degree (compared to 38%). For these students, we observe 565,196 math scores and 560,622 science scores. However, we do see some differences in our teacher characteristics. The 49,018 teachers in this sample are less likely to be female (54% vs. 73%) and more likely to be more than 40 years old (84% vs. 69%), are less likely to have majored in education (60% vs. 71%), and are more likely to teach in their area of expertise (89% vs. 79%).

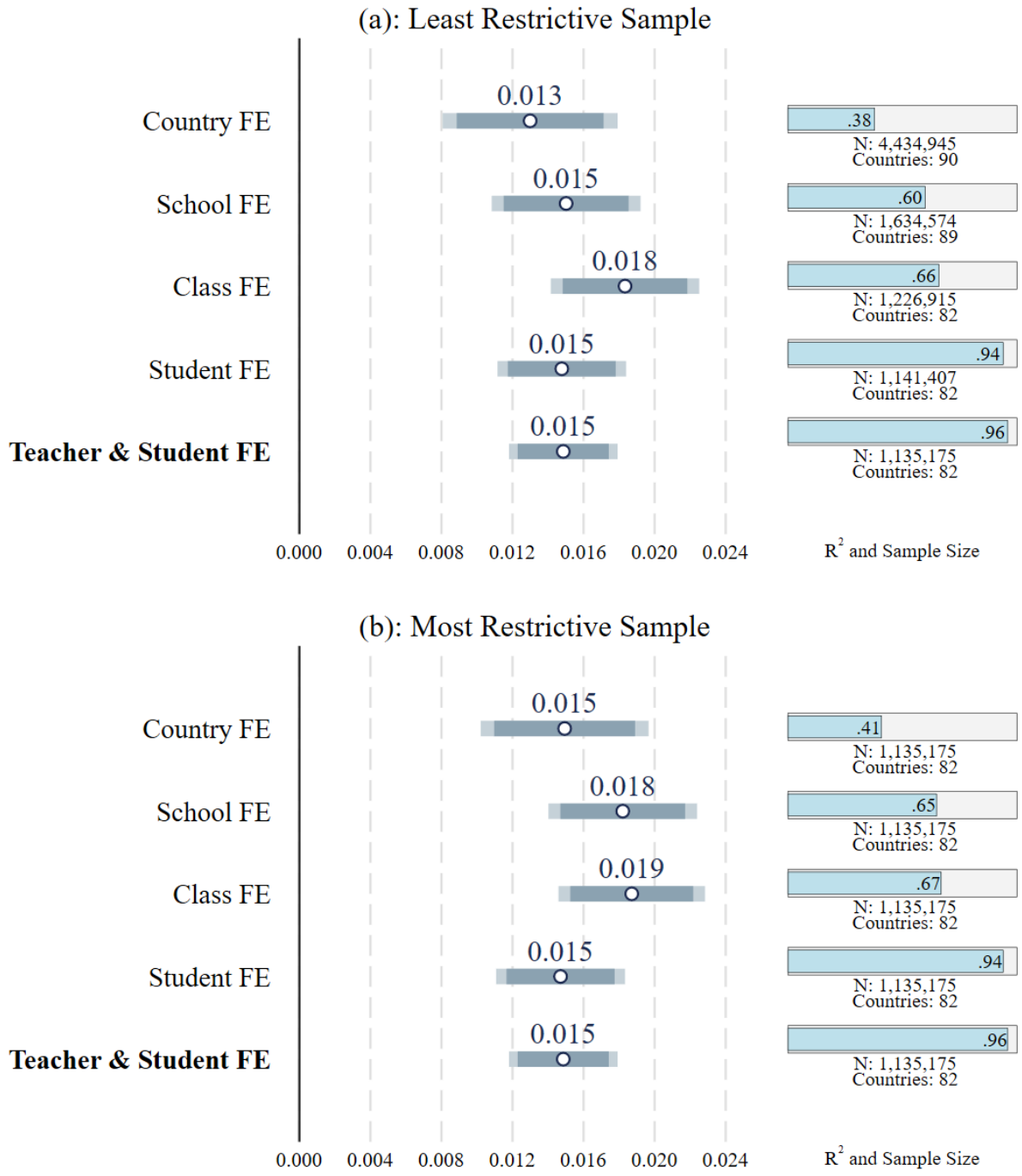
Overall, these statistics show two things. First, we have many observations, even for our most restrictive, preferred estimation sample. Second, the characteristics of the students and teachers included in our samples differ by specification. These differences can drive differences in point estimates 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 same sample constant across all fixed effects specifications.

6. Results

6.1 Role model effects on test scores

Figure 5(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

Figure 5: Role Model Effects—Test Scores



Note: This figure shows estimated role model effects from regressions of standardized test scores on a $FemaleStudent_i \times FemaleTeacher_j$ 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 A4 shows the corresponding regression table. Horizontal bars show 95% and 99% confidence intervals that are based on standard errors clustered at the classroom level.

the *R*-squared is 0.30 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 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, we provide estimates in which we keep the estimation sample constant across all specifications. Figure 5(b) shows estimates keeping 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. The 99% confidence intervals for these estimates allow us to rule out effects smaller than 0.009 and larger than 0.022 SD for all role model estimates shown in Figure 5. No matter the sample restrictions or the included fixed effects, we see a highly statistically significant role model effect of around 0.015 SD.

The role model effect in this super-study is smaller than 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. While meta-analysis estimates are hard to interpret, our super-study 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 role model estimate.

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).[11] Our role model effect estimate is also small compared to estimates of teacher value-added and teacher experience. For example, Chetty et al. (2014)'s estimate 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). Clotfelter et al. (2006) 's estimate of having a teacher with 12+ years of experience instead of a rookie teacher on math scores is eight times larger (0.113 SD). Hanushek et al. (2005) estimate of having a teacher with six-plus years of experience instead of a rookie teacher is eight times larger (0.12 SD).

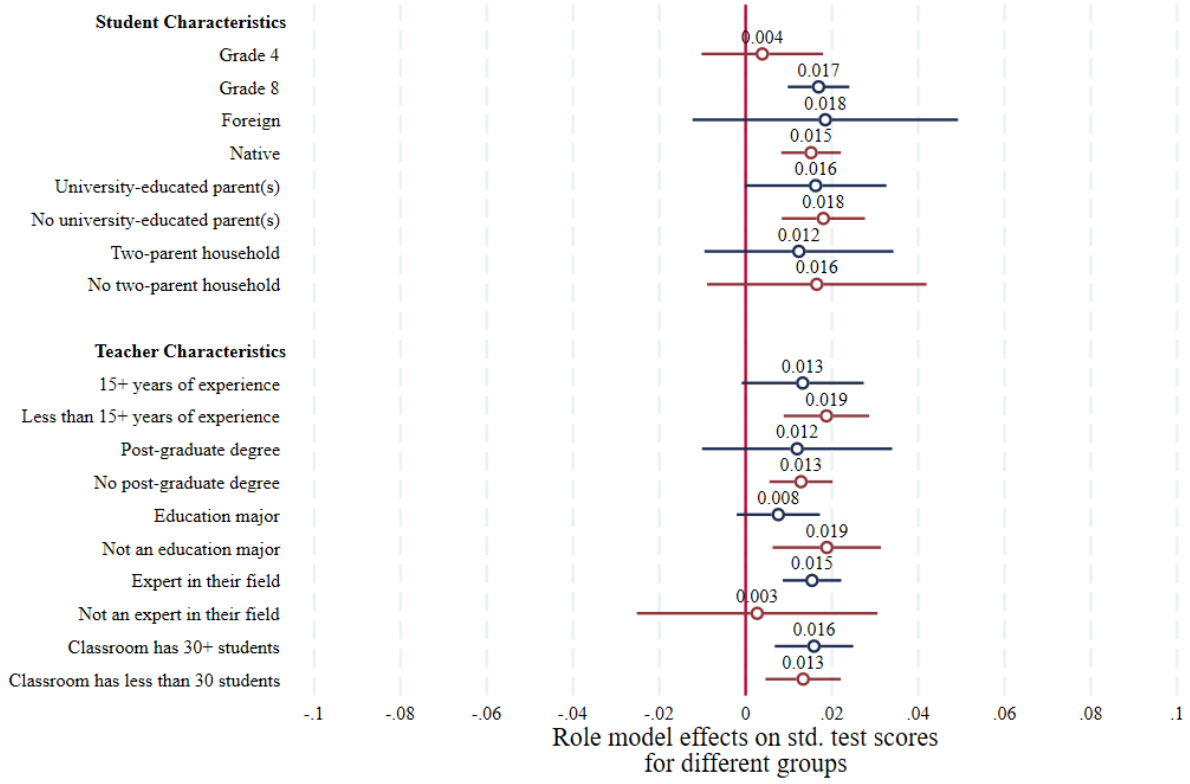
Subject heterogeneity: We test whether our results differ by subject by estimating role model effects in separate samples for students' math, science, and reading scores with our school fixed-effects specification. This analysis is not possible with more-restrictive fixed effects as these require within-subject variation by classroom or student. Our results show some subject heterogeneity (see Appendix Table A3). Role model effects are somewhat larger in math than in science (0.019 SD compared to 0.012 SD) and statistically indistinguishable from zero for reading (0.003 SD). These differences in effects also explain why restricting our estimation sample leads to slightly larger role model estimates: because we cannot include reading scores in our preferred specification, our sample is limited to subjects (math and science) for which we see larger role model effects.

[11] 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.

Student- and teacher-level heterogeneity: We test whether our results differ by student and teacher characteristics by estimating role model effects using our preferred specification separately for different subsamples of students and teachers. Figure 6 shows little heterogeneity along any of the dimensions we consider. All point estimates are small and precisely estimated.

We only see meaningful heterogeneity along two dimensions. Role model effects are larger in 8th grade compared to 4th grade (0.017 SD compared to 0.004 SD) and role model effects are larger for teachers who are experts in their field compared to those who are not (0.015 SD compared to 0.003 SD).

Figure 6: Student- and Teacher-Level Heterogeneity



Note: This figure shows estimated role model effects from regressions of standardized test scores on a $FemaleStudent_i \times FemaleTeacher_j$ interaction term, student fixed effects, teacher fixed effects, as well as other control variables from our preferred specification (see Section 5) for the different subsamples indicated on the left of the figure. Appendix Table A6 shows the corresponding regression table. Horizontal lines show 95% confidence intervals that are based on standard errors clustered at the classroom level.

6.2 Role model effects beyond test scores

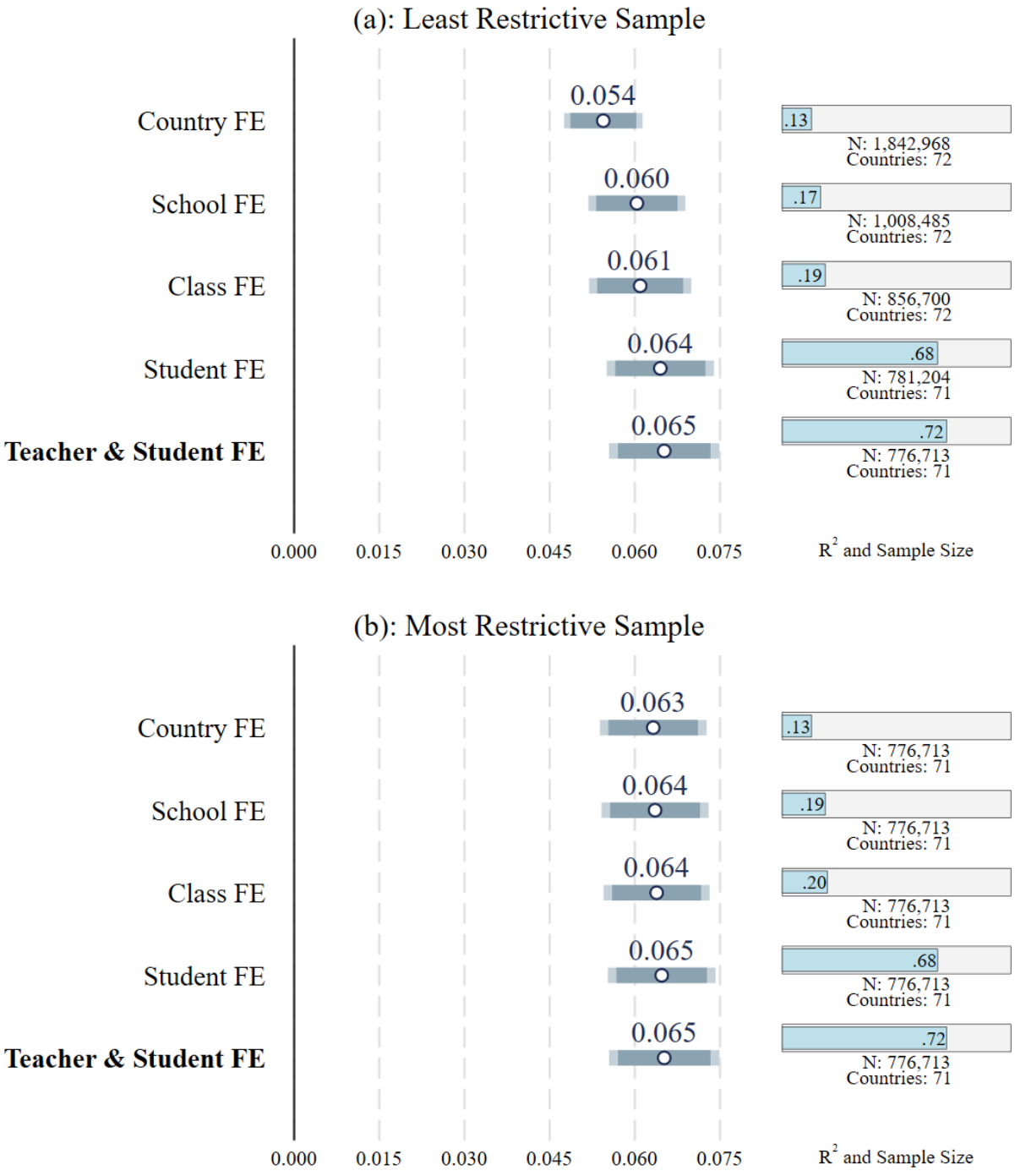
Teachers' influence on their students may go beyond test scores. Role models may also inspire students to follow in their footsteps and to make similar educational or occupational choices. They may also affect students' confidence and may affect how much students enjoy a subject. To test for such effects, we use the same set of fixed effects that we used for our test score analysis to estimate role model effects on job preferences

Figure 7 shows role model estimates for students' job preferences. We again keep the largest possible estimation sample for each specification in Panel (a) and show estimates for the consistent sample of our most restrictive specification in Panel (b). We see that the same-sex role model effect for job preferences in our preferred specification (0.065 SD) is substantially larger than for text scores (0.015 SD). As for test scores, our different fixed effect specifications yield very similar results.

We also use the identical set of controls from our preferred specification to estimate role model effects on two additional outcomes: subject confidence and subject enjoyment. Table 4 shows substantial effect for both outcomes. We see role model effects on subject enjoyment of 0.089 SD and role model effects on subject confidence of 0.051 SD.

While we do not have data on students' actual job choices, we find it plausible that these could also be affected. 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 chose in high school and university. Such effects on job choices would 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 a 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 a same-sex role model effect for choosing their mentor's occupation.

Figure 7: Role Model Effects—Job Preferences



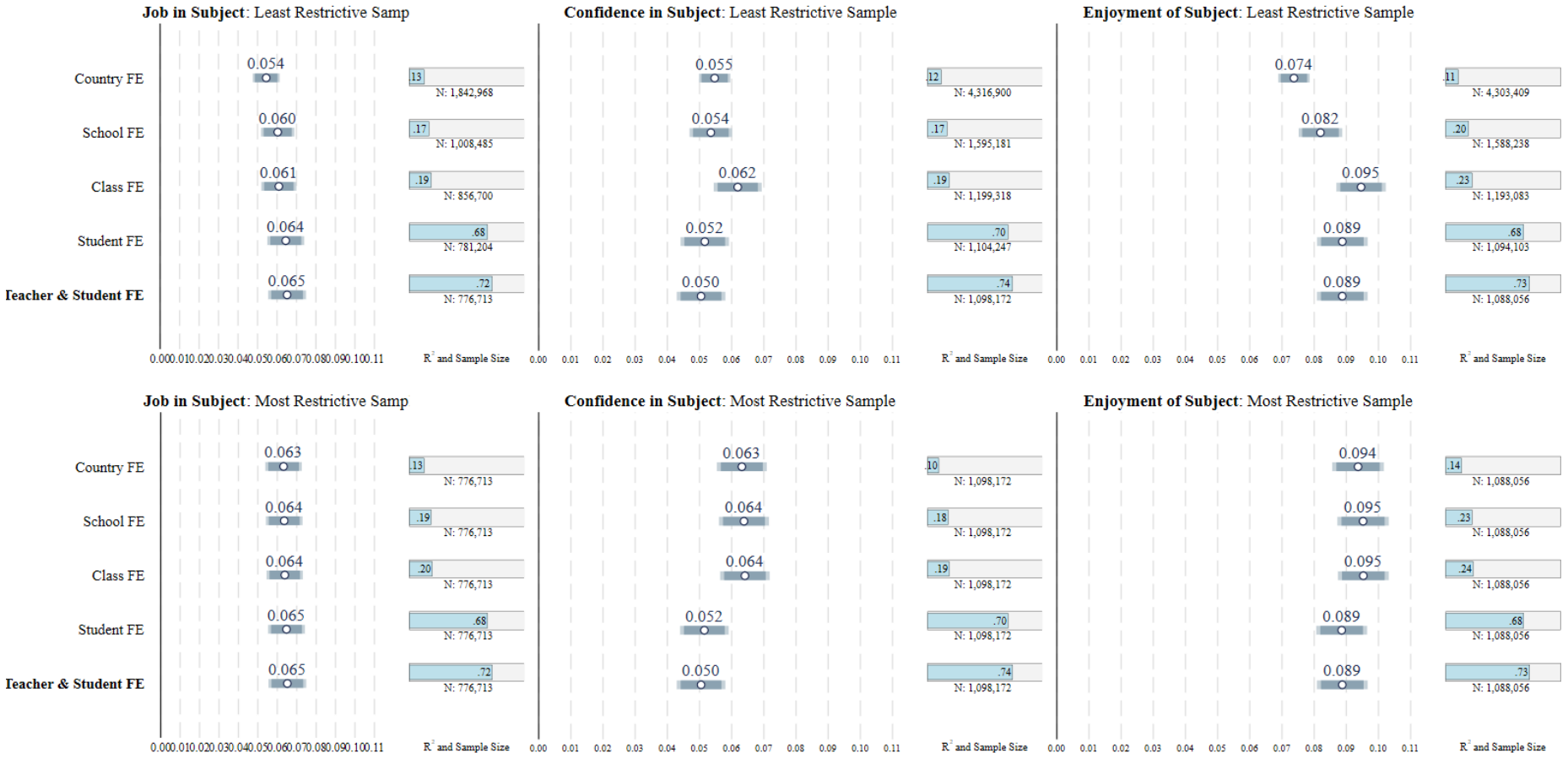
Note**:** This figure shows estimated role model effects from regressions of standardized job preferences on a $FemaleStudent_i \times FemaleTeacher_j$ interaction term, a set of other control variables (see Section 5), and different sets of fixed effects (as indicated to the left of the vertical line). We exclude eight countries with missing data on job preferences from panel (b) (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. 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 A4 shows the corresponding regression table. Horizontal bars show 95% and 99% confidence intervals that are based on standard errors clustered at the classroom level.

Table 4: Role Model Effects—Subject Enjoyment and Confidence

Std. Dependent Variable:	Enjoyment	Confidence
Role model effect	0.0887***	0.0505***
	(0.0040)	(0.0039)
R-Squared	0.73	0.74
Countries	82	82
Observations	1,088,056	1,098,172

Note: This figure shows estimated role model effects from regressions of subject enjoyment, and subject confidence on a $FemaleStudent_i \times FemaleTeacher_j$ interaction term, student fixed effects, teacher fixed effects, as well as other control variables from our preferred specification (see Section 5). See Table 3 for more details on the dependent variables. Standard errors clustered at the classroom level are in parentheses. ***, **, and * mark estimates statistically different from zero at the 1%, 5%, and 10% significance levels.

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



Note.

6.3 Global heterogeneity

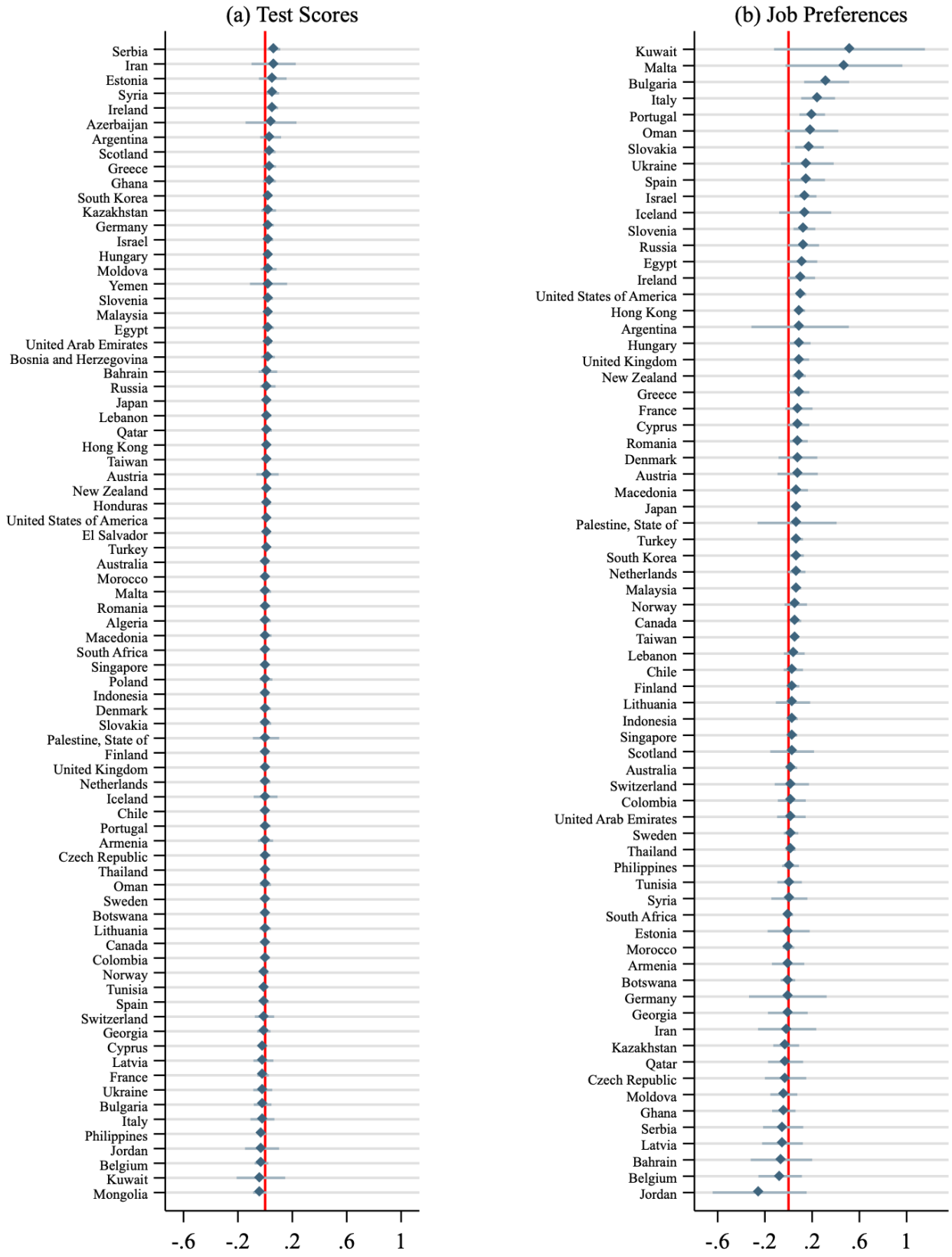
Our super-study approach allows us to use the same methodology to estimate separate role model effects in different countries. These country level estimates then allow us to test how universal role model effects are. In this section, we focus on the results on test scores and job preferences and, for brevity, show results for subject confidence and subject enjoyment in Appendix D.

6.3.1 Global heterogeneity in role model effects on test scores and job preferences

We start by estimating role model effects with our preferred set of controls separately for each country. Panel (a) of Figure 8 shows 79 country-specific role model estimates on test scores and their 95% confidence intervals.[12] These estimates range from –0.040 SD for Mongolia to +0.064 for Serbia; 10 estimates are positive and significant at the 5% level; 56 estimates are positive and insignificant; 17 estimates are negative and insignificant; no estimate is negative and significant. Figure 9 provides a world map showing that role model effects on test scores are positive and significant at the 5% level in the United States (0.013 SD), Ireland (0.052 SD),

[12] Because of multicollinearity, we lose three out of our 81 countries that only have one classroom per school left after applying our preferred specification restrictions (Albania, Pakistan, and Northern Ireland). Appendix Table A10 shows role model estimates for each of the 79 countries.

Figure 8: Role Model Effects by Country



Note: This figure shows estimated role model effects from regressions of standardized test scores (Panel a) or standardized job preferences (Panel b) on a $FemaleStudent_i \times FemaleTeacher_j$ interaction term, student fixed effects, teacher fixed effects, as well as other control variables from our preferred specification (see Section 5) for the different country-subsamples indicated on the left of each panel. Because of multicollinearity, we exclude three countries (Albania, Pakistan, and Northern Ireland) where there is only one classroom per school after applying our preferred specification restrictions. We also exclude eight countries with missing data on job preferences from panel (b) (Algeria, Azerbaijan, Bosnia and Herzegovina, El Salvador, Honduras, Poland, Mongolia, and Yemen). Panel (a) therefore shows 79 point estimates and Panel (b) shows 71 point estimates. Appendix Table A10 shows the corresponding regression table. Horizontal lines show 95% confidence intervals that are based on standard errors clustered at the classroom level.

Figure 9: Global Variation in Same-sex Role Model Effects—Test Scores

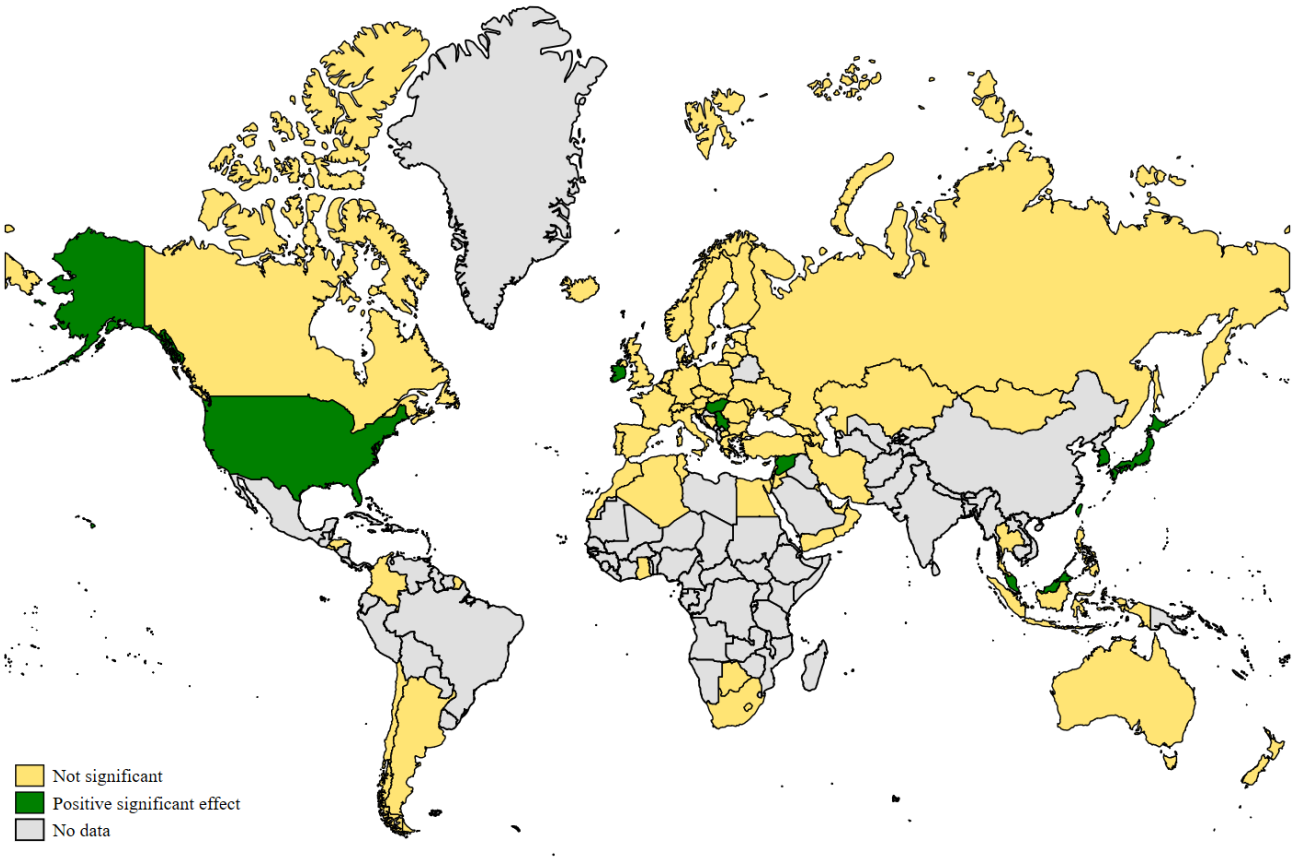
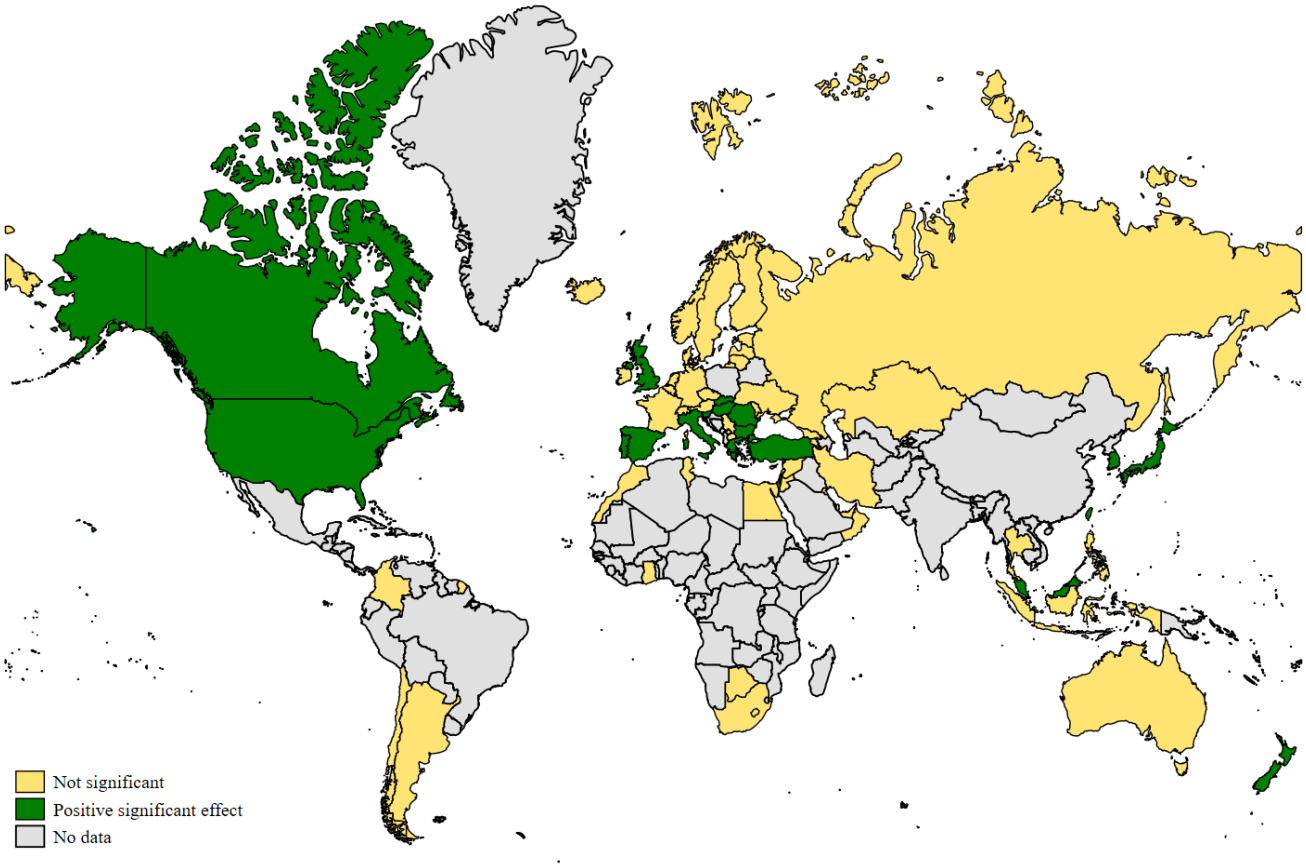


Figure 10: Global Variation in Same-sex Role Model Effects—Job Preferences



Hungary (0.025 SD), Serbia (0.064 SD), Japan (0.019 SD), Malaysia (0.022 SD), South Korea (0.028 SD), Syria (0.056 SD), Hong Kong (0.019 SD), and Taiwan (0.017 SD).

Panel (b) of Figure 8 shows 71 country-level estimates for role model effects for job preferences. These estimates range from –0.245 SD for Jordan to +0.516 SD for Kuwait; 21 estimates are positive and significant at the 5% level; 38 estimates are positive and insignificant; 16 estimates are negative and insignificant; and no estimates is negative and significant. Figure 10 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).

For both outcomes, the differences in point estimates can reflect true heterogeneity as well as sampling error. To estimate the degree of heterogeneity in role model effects, we estimate the standard deviation of the true effect sizes using random effects meta-regressions (see Section 2). For these regressions, we use the point estimates and standard errors shown in Figure 7 as input, treating all estimates as providing independent information since they are based on non-overlapping samples.

The results of these meta-regressions show that there is no discernible heterogeneity in role model effects on test scores. The estimated standard deviation of the true role model effects on test scores is 0.000 SD showing that essentially all differences in estimates shown in Figure 8 are driven by sampling error. In other words, for test scores, there is no meaningful heterogeneity of true effects among the included countries. Given the large number and diverse set of countries in our sample, we would be surprised if there are large role model effects on

math and science test scores in primary or secondary education in any country in the world. These role model effects appear to be universally small and positive.

We see larger heterogeneity among role model effects on job preferences. The estimated standard deviation of the underlying effect is 0.028 SD. The differences in point estimates shown in Figure 8 (b) do not only reflect sampling error, but also meaningful heterogeneity.

6.3.2 Global heterogeneity in role model effects on job preferences

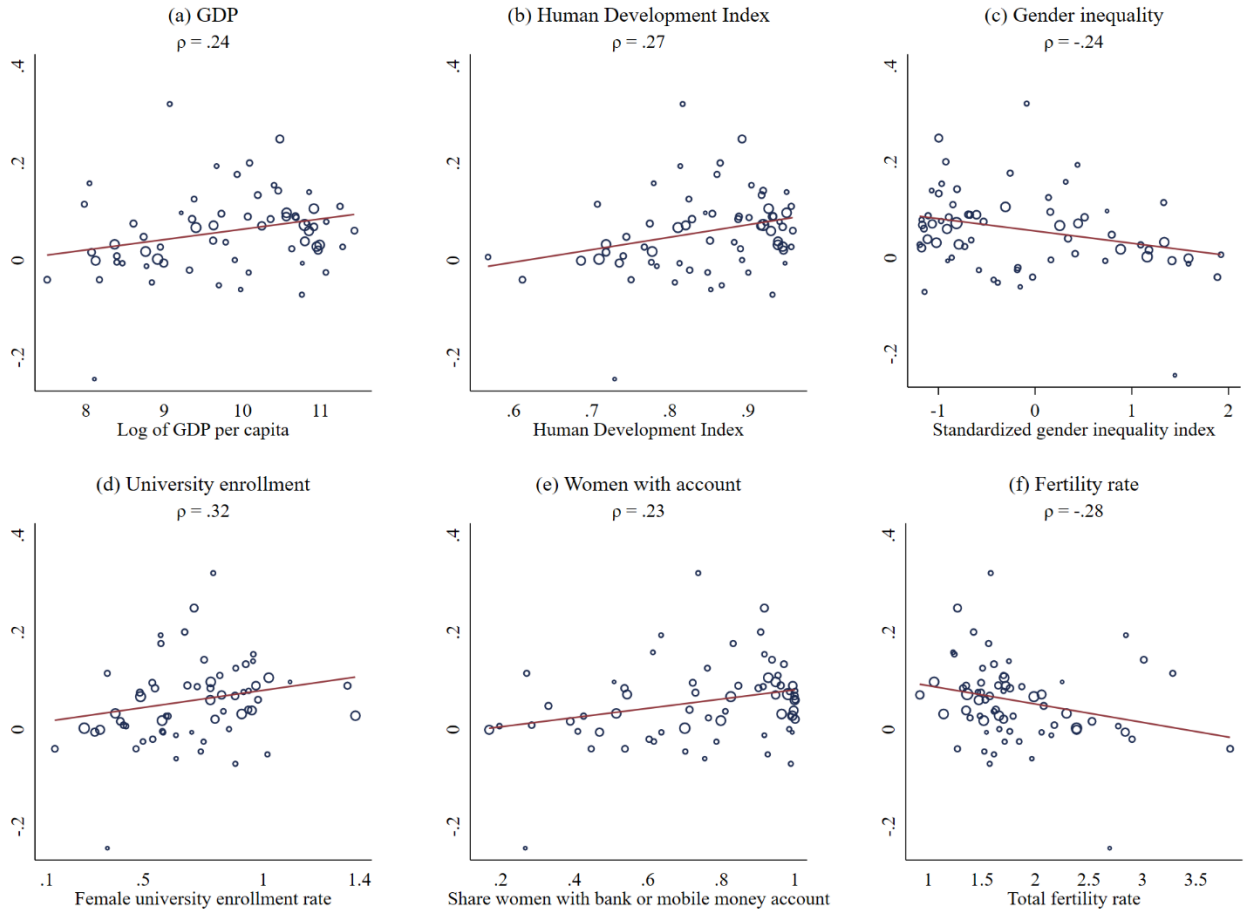
We explore country-level heterogeneity in role model effects on job preferences in two ways. First, we show a series of scatterplots that relate the size of role model effects to country-level observable characteristics. These plots show the estimated role model effect on job preferences as shown in Figure 8 (b) 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 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). We do this by regressing job preferences on all explanatory variables from our preferred specification as well as two additional variables: one dummy, which indicates a dummy variable for a country being above the median in a given characteristic, and one triple interaction term of this above-median dummy, a female student dummy, and the share of female teachers ($AboveMedian \times Female\ Student_i \times Female\ Teacher_j$. In these specifications, the coefficient on the $Female\ Student_i \times Female\ Teacher_j$ interaction term shows the estimated role model effect for below-median countries. By adding this coefficient and the coefficient on the new triple interaction term, we get the estimated role model effect

for above-median countries. We discuss those estimates in the text and show the corresponding regressions in Table A6 in the appendix.

Using both approaches, we explore whether role model effects depend on a country's economic development, gender inequality, or sex differences in math and science performance.

Figure 11: Role Model Effects in Job Preferences and Country-level Correlates



Note: These panels show the bivariate relationships between the estimated role model effects on standardized job preferences shown in Figure 8 (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 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 UN. 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 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 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.

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 10 (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 who have an above-median HDI (0.0746 SD compared to 0.0494 SD).

Gender inequality. Role model effects might be stronger in gender-unequal countries where women face systematic 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 (UN) 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 11 (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 vs. 0.0724 SD). Figure A7 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 enrolment, 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 11 (d) and (f) show 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 enrolment (0.0725 SD vs. 0.0418 SD) and significantly *lower* role model effects in countries with above-median fertility rates (0.0540 SD vs. 0.0739 SD). Figure 10 (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% level (0.0704 SD vs. 0.0514 SD).

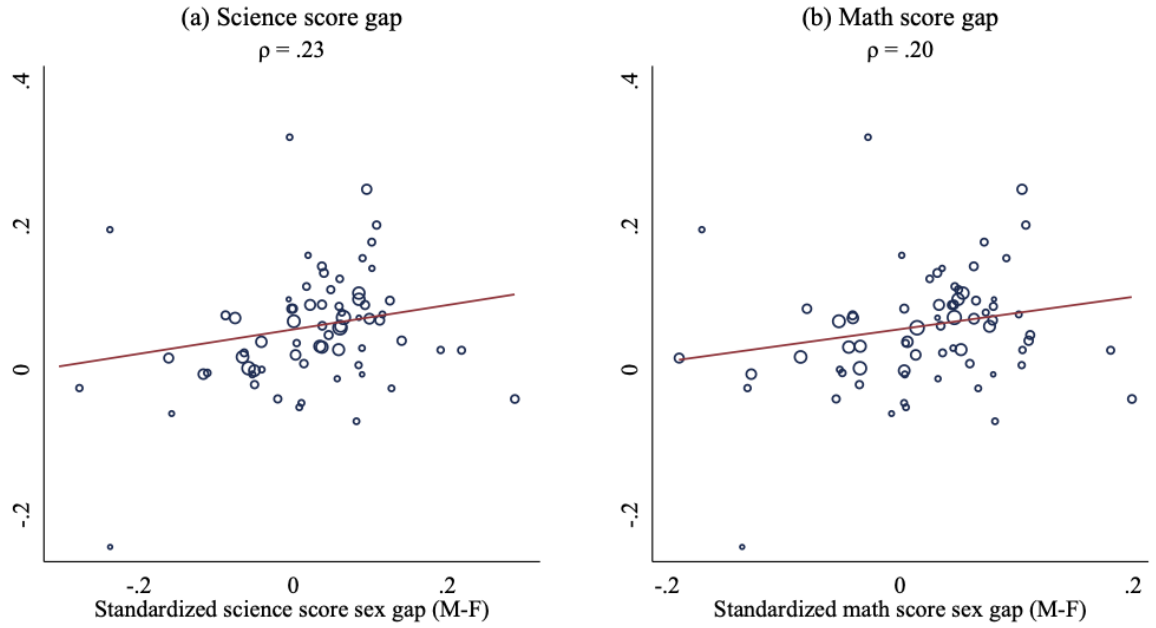
Sex gaps in math and science test scores. Role model effects on job preferences might depend on the differences in boys' and girls' ability in math and science. For example, in countries where boys outperform girls in math, girls might see having a female math teacher as evidence that girls can do well 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 11 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 mirrors the pattern for role model effects on job preferences. We show in Appendix D that role model effects on subject enjoyment and subject confidence are larger in developed counties and smaller in countries with high gender inequality (see Tables D1 and D2 in appendix).

Figure 12: Role Model Effects on Job Preferences and Test Score Gaps between Boys and Girls



Note: This figure shows the relationship between the estimated role model effects on standardized job preferences shown in Figure 7 and the standardized sex gap (M-F) in science (Panel a) or math (Panel b). 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.

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. Even though we do not have data on job preferences from India, we would expect no 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 show are consistent with some explanations that can further 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, and have fewer children). In these circumstances, having a female science teacher may be more powerful in showing that girls can do jobs that involves science.[13]

7. 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 to motivate girls to enter STEM

[13] 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 behind girls and can choose the job they want. However, it might be that role model matter less for boys as there is no lack of examples of successful men in technical fields.

jobs. Similarly, hiring more male teachers in elementary school has become a policy target to stop boys from falling behind at that stage of education. Whether role model effects exist and how strong they are, is therefore 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 a super-study. 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 super-study finds an even smaller role model effect on test scores of 0.015 SD. We see very little heterogeneity across student characteristics, across teacher characteristics, or between countries. Role model effects on test scores appear to be universally positive and small. Because teachers' influence might go beyond test scores, we test how role models affect students' job preferences. We find role model effects on job preferences (0.064 SD) are substantially larger than for test score (0.015 SD). These role model effects for job preferences are concentrated in developed and gender-equal countries. Taken together, our results show 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 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 super-studies. Combining data from multiple settings gave us enough statistical power to detect a statistically significant but tiny effect on test scores. Having data from 90 countries allowed us to thoroughly explore heterogeneity and to show in which settings substantial role model effects for students' job preferences exist. In contrast to meta-analysis, we could conduct this analysis without worrying about differences in methodology and systematically missing estimates.

A key benefit of super-studies is their ability to resolve open debates in the literature. Some literatures appear to produce a constant stream of conflicting results. These results could reflect that effects are highly context specific. There could also be no effect in any context and all results are driven by publication bias. Or, there could be an effect that is too small to detect with a typical study. In such a literature, adding one more typical study will not allow researchers to make meaningful progress. However, researchers can make a meaningful contribution by combining and analyzing data from multiple settings. Using this approach, recent super-studies have shown that younger siblings follow their older siblings' education choices (Altmejt et al., 2021), and neither birth order nor sibling sex has an effect on personality (Rohrer et al., 2015; Dudek et al., 2022).

While super-studies are not new, the name is. We hope it catches on. We believe this paper showcases the importance of super-studies for making scientific progress. We hope that this approach inspires researchers to conduct more super-studies that provide better answers to the fundamental research questions of our time. We also hope that when readers see "evidence from a super-study" they will know what to expect and find the approach valuable.

References

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., Doucouliagos, H., Aviña, M. M., & Stanley, T. D. (2022). Quantitative Political Science Research Is Greatly Underpowered. https://files.ca-

1.osf.io/v1/resources/7vy2f/providers/osfstorage/62c473cb7ddff522a39a6bf6?action=download&direct&version=1

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

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.

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

Carrell, S. E., Sacerdote, B. I., & West, J. E. (2013). From natural variation to optimal policy? The importance of endogenous peer group formation. *Econometrica,* 81(3), 855–882. DOI: https://doi.org/10.3982/ECTA10168

Carrington B., Tymms P., & Merrell C. (2008). Role models, school improvement and t‘e 'gender ’ap' – 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

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

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

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), pp. 778-820. OI :

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

DellaVigna, S., & Linos, E. (2022). RCTs to scale: Comprehensive evidence from two nudge units. *Econometrica*, 90(1), 81–116. DOI: https://doi.org/10.3982/ECTA18709

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*, 09567976221094630.

Eble, A., & Hu, F. (2020). Child beliefs, societal beliefs, and teacher-student identity match. *Economics of Education Review*, 77, 101994. 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.

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. OI : https://doi.org/10.1126/science.1255484

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

Gust, S., Hanushek, E. A., & Woessmann, L. (2022). Global universal basic skills: Current deficits and implications for world development (30566). *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., …, & Zwienenberg, M. (2016). A Multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 11(4), 546–573. DOI: https://doi.org/10.1177/1745691616652873

Hanushek, E. A. J. F. Kain, D. M' O'Brien, S. G. Rivkin. The market for teacher quality. *NBER Working Paper* (11154) (2005) OI :

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, Nick, Andreu Arenas, Emily Beam, Marco Bertoni, Jeffrey R. Bloem, Pralhad Burli, Naibin Chen, Paul Grieco, Godwin Ekpe, Todd Pugatch, Martin Saavedra, and Yaniv Stopnitzky. (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, 23328584211040056. 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

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

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

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

Milkman, K.L., D. Gromet, H. Ho, J. Kay, T. Lee, P. Pandiloski, Y. Park, A. Rai, M. Bazerman, J. Beshears, L. Bonacorsi, C. Camerer, E. Chang, G. Chapman, R. Cialdini, H. Dai, L. Eskreis-Winkler, A. Fishbach, J.J. Gross, A. Horn, A. Hubbard, S.J. Jones, D. Karlan, T. Kautz, E. Kirgios, J. Klusowski, A. Kristal, R. Ladhania, G.

Loewenstein, J. Ludwig, B. Mellers, S. Mullainathan, S. Saccardo, J. Spiess, G. Suri, J.H. Talloen, J. Taxer, Y. Trope, L. Ungar, K.G. Volpp, A. Whillans, J. Zinman, A.L. Duckworth (2021). Megastudies Improve the Impact of Applied Behavioural Science. *Nature*, (600), 478-483. DOI:

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

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

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

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

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.

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., …, 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.

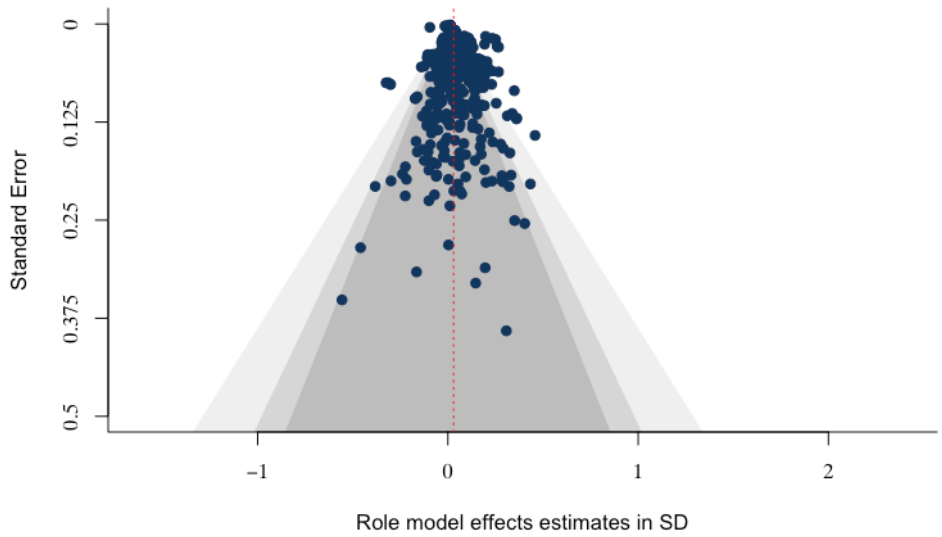
Winters, M. A., Haight, R. C., Swaim, T. T., & Pickering, K. A. (2013). The effect of same-gender teacher assignment on student achievement in the elementary and secondary grades: Evidence from panel data. *Economics of Education Review*, 34(C), 69–75. DOI: https://doi.org/10.1016/j.econedurev.2013

World Bank (2020). The Equality Equation: Advancing the Participation of Women and Girls in STEM. https://openknowledge.worldbank.org/bitstream/handle/10986/34317/Main-Report.pdf?sequence=1&isAllowed=y (retrieved on: 15.08.2022, 13:00)

Xu, D., & Li, Q. (2018). Gender achievement gaps among Chinese middle school students and the role of teachers' gender. *Economics of Education Review*, 67, 82–93. DOI: https://doi.org/10.1016/j.econedurev.2018.10.002

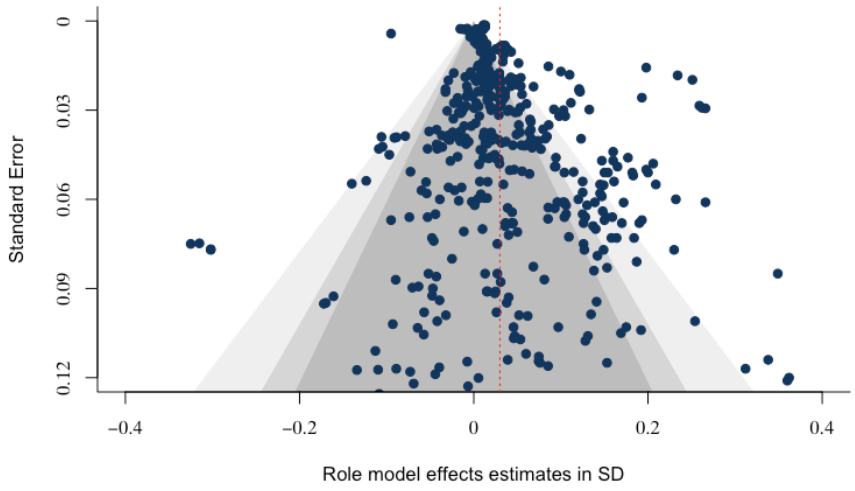
Xu, R. (2020). "When boys become the second sex": The new gender gap among Chinese middle school students. *The Yale Undergraduate Research Journal,* 1(1).

Appendix
Appendix A
Supplementary Tables and Figures
Figure A1: Funnel Plot of All Role Model Effect Estimates



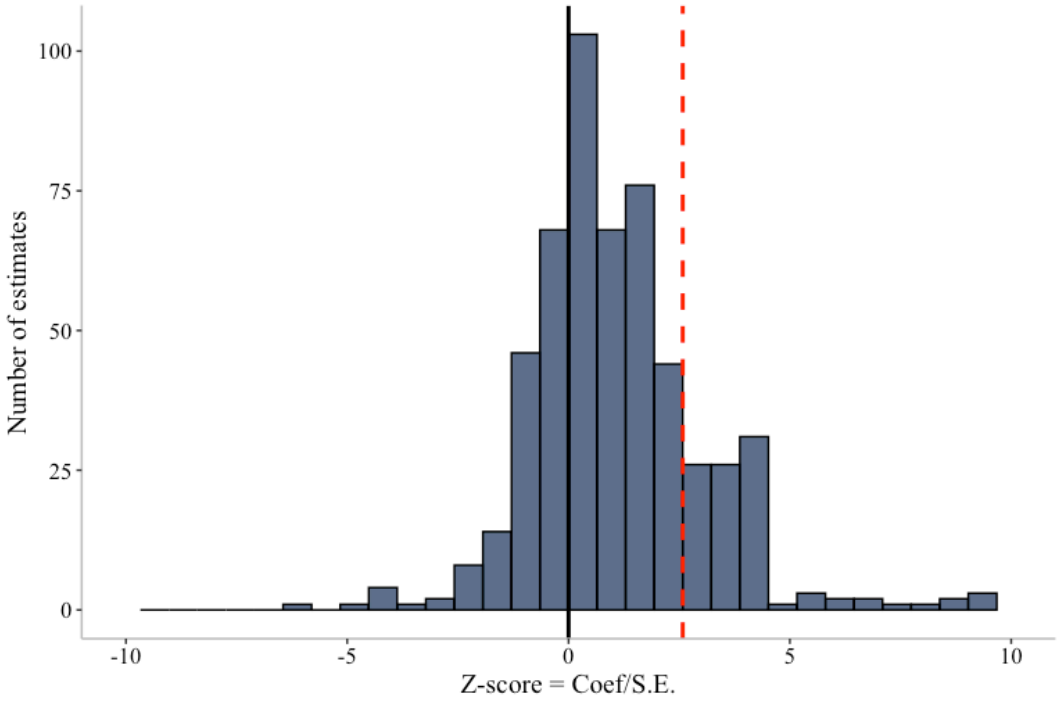
Note: This figure shows a scatterplot of all 535 role model effects estimates from all 24 studies on the x-axis, with their standard error on the y-axis. To increase readability, this figure excludes 3 outlying role model estimates of size 1.15, 2.07 and 0.92 SD with a standard error of 5.03, 5.42 and 6.83 SD respectively. The grey shaded areas mark the traditional thresholds for statistical significance with 90%, 95% and 99% confidence. The vertical dotted line marks our estimated average role model effect of 0.030 SD.

Figure A2: Zoom into Funnel Plot of All Role Model Effect Estimates



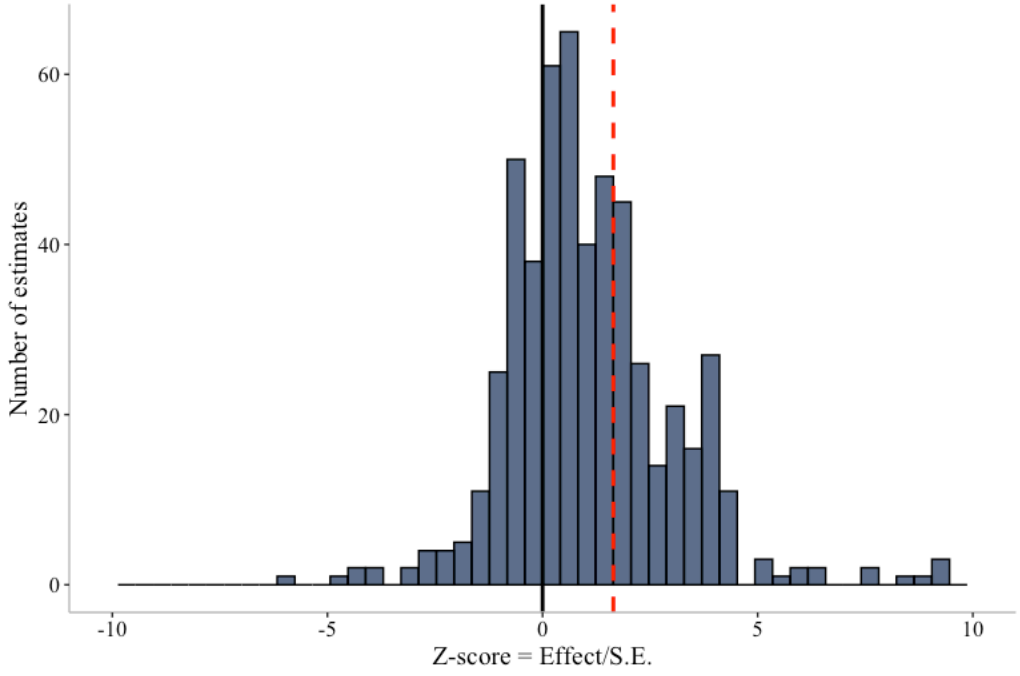
Note: This figure shows a zoomed-in subsection of the scatterplot in Figure A1 where most estimates are present. The grey shaded areas mark the traditional thresholds for statistical significance with 90%, 95% and 99% confidence. The vertical dotted line marks our estimated average role model effect of 0.030 SD.

Figure A3: *Z*-score Distribution of All Role Model Effect Estimates, with 99% Two-sided Critical Value Marked



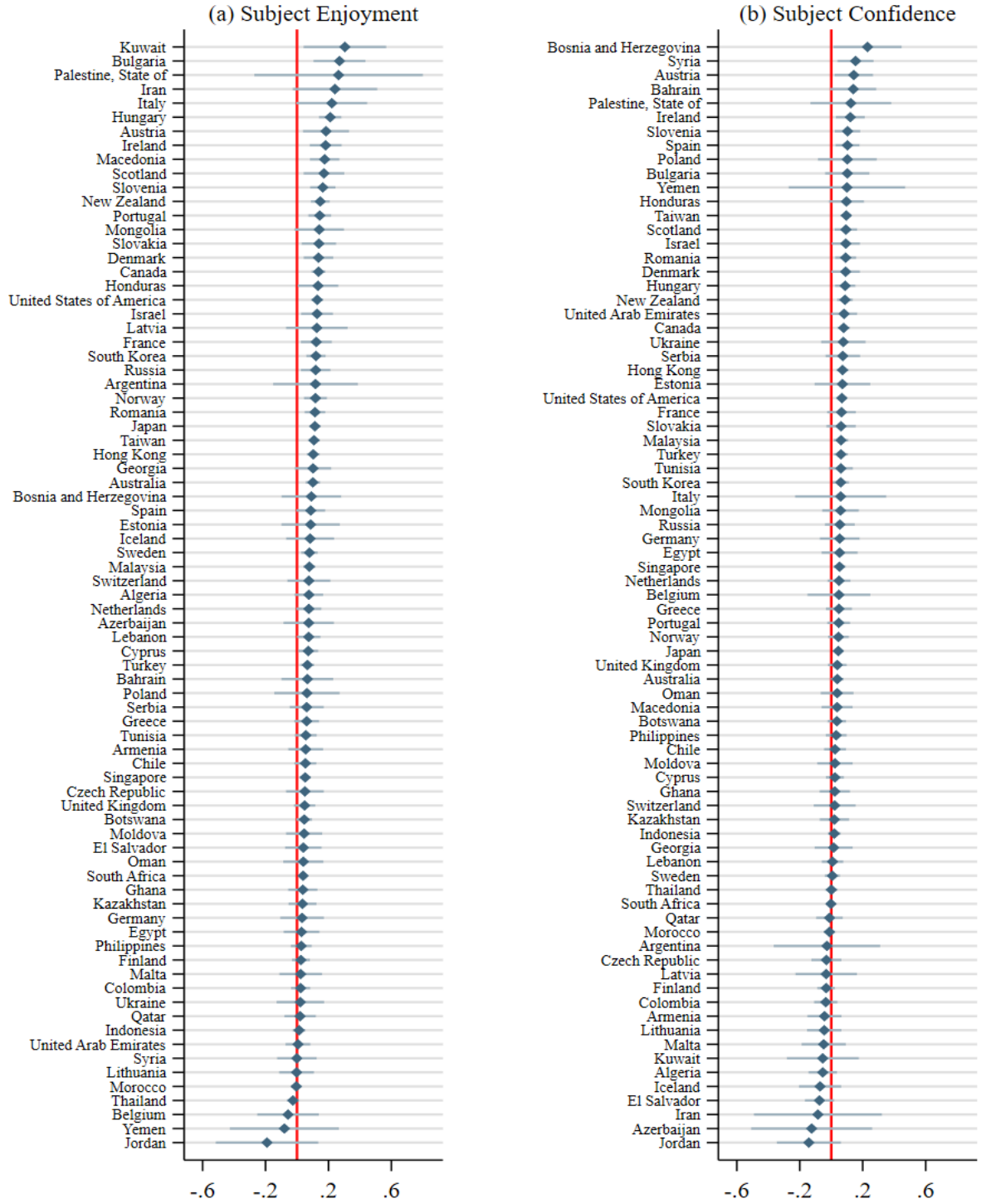
Note: This figure shows *z*-scores of 534 role model effects estimates from all 24 studies. These are all z-scores except for 4 outlier values (with *z*-scores of -22.39, 12.68, 12.79 and 12.61) which we excluded to make the figure more readable. The vertical dashed line marks 2.58, the two-sided test critical value of 99% for the normal distribution. The histogram uses a bin width of 0.645.

Figure A4: *Z*-score Distribution of All Role Model Effect Estimates, with 90% Two-sided Critical Value Marked



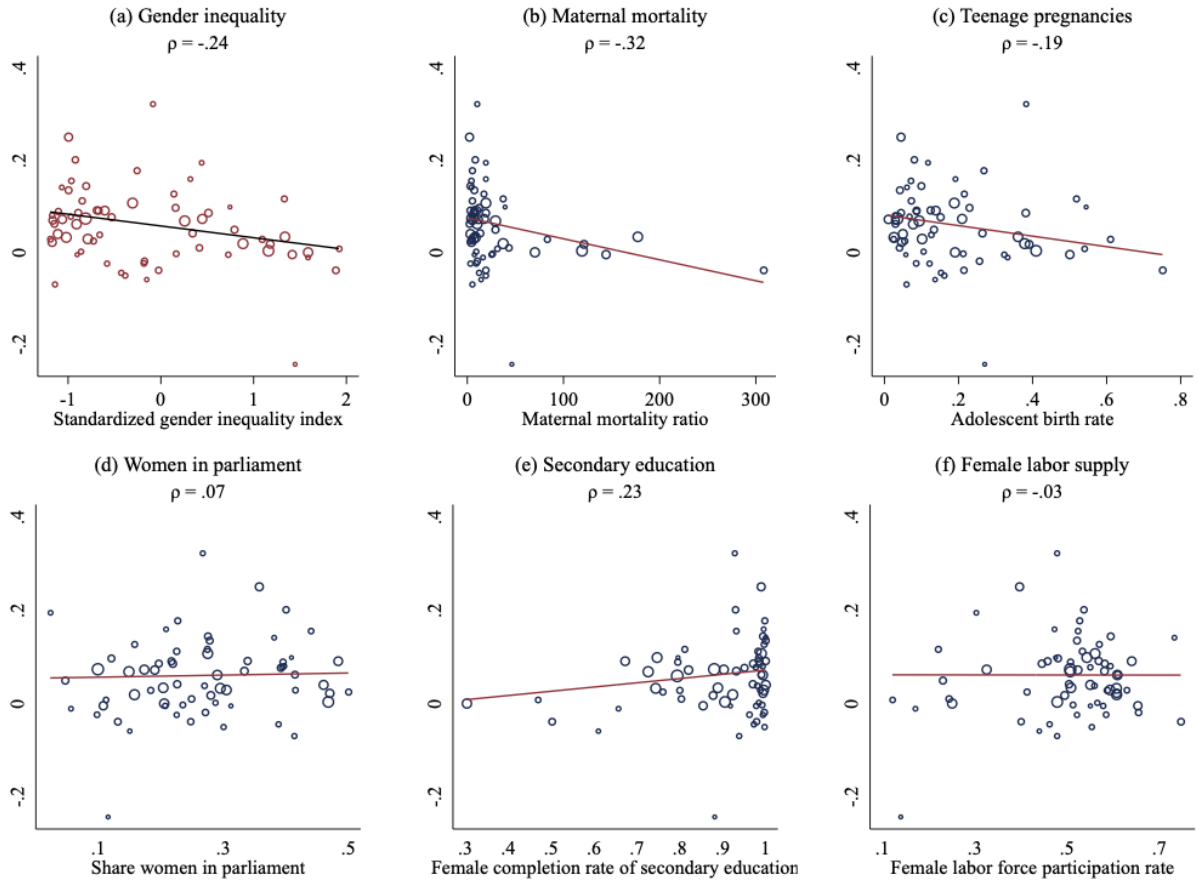
Note: This figure shows *z*-scores of 534 role model effects estimates from all 24 studies. These are all z-scores except for 4 outlier values (with *z*-scores of -22.39, 12.68, 12.79 and 12.61) which we excluded to make the figure more readable. The vertical dashed line marks 1.64, the two-sided test critical value of 90% for the normal distribution. The histogram uses a bin width of 0.41.

Figure A6: Role Model Effects by Country—Confidence and Enjoyment



Note: This figure shows estimated role model effects from regressions of standardized subject confidence (Panel a) or standardized subject enjoyment (Panel b) on a $FemaleStudent_i \times FemaleTeacher_j$ interaction term, student fixed effects, teacher fixed effects, as well as other control variables from our preferred specification (see Section 5) for the different country-subsamples indicated on the left of each panel. This figure shows 79 different role-model effects estimates on subject confidence and 79 different role-model estimates on subject enjoyment. Because of multicollinearity, we exclude three countries (Albania, Pakistan, and Northern Ireland) where there is only one classroom per school after applying our preferred specification restrictions. We also exclude eight countries for which no school meets our preferred specification sample criteria (Belize, Croatia, Kosovo, Luxembourg, Macao, Montenegro, Saudi Arabia, and Trinidad and Tobago). Horizontal lines show 95% confidence intervals that are based on standard errors clustered at the classroom level.

Figure A7: Role Model Effects on Job Preferences and Gender Inequality



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

Table A1: Meta-regression of Role Model Estimates

			95% CI	
	Coef.	Std.err.		
Intercept	0.137	(0.104)	-0.067	0.341
Variation (base = Experimental)				
Quasi-experimental	-0.063	(0.068)	-0.1967	0.070
Continent (base = Africa)				
Asia	-0.096	(0.067)	-0.229	0.036
Europe	-0.033	(0.066)	-0.162	0.096
North America	-0.144**	(0.067)	-0.276	-0.013
School level (base = Secondary)				
Elementary	-0.094***	(0.033)	-0.159	-0.03
Both	-0.083**	(0.034)	-0.149	-0.017
Outcome (base = Grades)				
Test scores	0.068	(0.041)	-0.013	0.149
Test for significance of all moderators (p-value)	0.001			
Test for residual heterogeneity (p-value):	<0.0001			
Variance components (τ)				
Between studies	0.0068			
Within studies	0.0003			

Note: Coefficients from a single three-level meta-regression of role model effects estimates on grades and test scores, estimated using the *meta* package in R. Our sample contains all 538 role model estimates from 24 studies. The three levels account for nested interdependence while pooling of information of individual participants into the various role model effects in primary studies (level 1), the pooling of all role model effects in each primary study (level 2), and the pooling of primary study role model effects into an overall role model effect (level 3). All moderators are coded at the primary study level. Standard errors are in parentheses. ***, **, and * mark estimates statistically different from zero at the 1%, 5%, and 10% significance levels.

Table A2: Role Model Effect Estimates Corrected for Publication Bias

Estimation method	Significance threshold for selection	Average effect	Standard error	95% Confidence Interval		Standard deviation of effect
3-level REML	-	0.030	(0.013)	0.005	0.055	0.058
Trim and Fill	-	0.012	(0.004)	0.004	0.020	0.077
PET-PEESE	-	0.010	(0.000)	0.009	0.011	0.000
Limit-Meta	-	0.012	(0.197)	-0.373	0.397	0.058
3-Parameter Selection	10%	0.029	(0.004)	0.021	0.038	0.049
3-Parameter Selection	5%	0.035	(0.005)	0.026	0.044	0.050
3-Parameter Selection	1%	0.038	(0.004)	0.029	0.047	0.051
Andrews & Kasy (t)	10%	0.012	(0.003)	0.006	0.018	0.015
Andrews & Kasy (t)	10%, 5%	0.012	(0.003)	0.006	0.018	0.015
Andrews & Kasy (t)	10%, 5%, 1%	0.011	(0.003)	0.005	0.017	0.015
Andrews & Kasy (N)	10%	-0.027	(0.015)	-0.056	0.002	0.074
Andrews & Kasy (N)	10%, 5%	-0.028	(0.017)	-0.061	0.005	0.078
Andrews & Kasy (N)	10%, 5%, 1%	-0.039	(0.022)	-0.082	0.004	0.088

Note: As benchmark, 3-level REML shows the estimated role model effect without correcting for publications 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. Restricted maximum likelihood 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: We use estimates from the precision-effect test (PET) model rather than from the precision-effect estimate with standard error (PEESE) model because the one-sided *t*-test of intercept for the PET model does not reject the null hypothesis at the 5% level. Estimates weighted by their inverse variance. We use the *rma.uni()* function in *R* for implementing this method. Limit-Meta: Uses 3-level REML (see above) as input. 3-Parameter Selection: We use 0.05, 0.025, and 0.01 as jumps in the publication probability function. Restricted maximum likelihood estimator of the standard deviation of the effect size and the standard deviation of the effect size. Andrews and Kasy: We use 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, at the 0.05 and 0.025, and at the 0.05, 0.025 and 0.01 significance levels for both positive and negative effects. We allow the probability of publications bias to be asymmetric. We produce estimate using Kasy's App: https://maxkasy.github.io/home/metastudy. Other correction methods: Andrews and Kasy's (2019) 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 A3: Examples of Questions Used in PIRLS and TIMSS

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

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

Table A4: Replication of Table 2 Balancing Tests for our Preferred Estimation Sample

		Role model effect				
	Mean	Coef.	Std.err.	R-Squared	Countries	Obs.
Student characteristics:						
Age (in years)	13.4	0.0117***	(0.0015)	0.83	82	1,134,443
Foreign-born	0.09	0.0008	(0.0005)	0.24	81	1,068,395
Parent(s) have university degree	0.36	-0.0011	(0.0012)	0.32	76	779,689
Two-parent household	0.65	0.0013	(0.0016)	0.43	51	243,296
Teacher characteristics:						
40+ years old	0.82	-0.0027	(0.0030)	0.60	82	1,136,294
Experience (in years)	15.8	0.0519	(0.0385)	0.62	82	1,117,368
Has post-graduate degree	0.30	-0.0004	(0.0016)	0.67	82	1,095,100
Majored in education	0.59	0.0037*	(0.0020)	0.64	78	934,042
Teaches field of expertise	0.89	-0.0003	(0.0012)	0.66	79	987,219

Note: This table shows results from replicating our balancing test shown in Table 2 for the estimation sample from our preferred student and teacher fixed effects specification (see Section 5). Standard errors clustered at the classroom level are in parentheses. ***, **, and * mark estimates statistically different from zero at the 1%, 5%, and 10% significance levels.

Table A5: Role Model Effects on Test Scores

	Std. Test scores				
Panel (a)					
Least restrictive sample					
Role model effect	0.0130***	0.0150***	0.0183***	0.0148***	0.0149***
	(0.0025)	(0.0021)	(0.0021)	(0.0018)	(0.0016)
Fixed effects	Country	School	Classroom	Student	Student & Teacher
R-squared	0.38	0.60	0.66	0.94	0.96
Countries	90	89	82	82	82
Observations	4,434,945	1,634,574	1,226,915	1,141,407	1,135,175
Panel (b)					
Most restrictive sample					
Role model effect	0.0149***	0.0182***	0.0187***	0.0147***	0.0149***
	(0.0024)	(0.0021)	(0.0021)	(0.0019)	(0.0016)
Fixed effects	Country	School	Classroom	Student	Student & Teacher
R-squared	0.41	0.65	0.67	0.94	0.96
Countries	82	82	82	82	82
Observations	1,135,175	1,135,175	1,135,175	1,135,175	1,135,175

	Std. Job Preferences				
Panel (c)					
Least restrictive sample					
Role model effect	0.0532***	0.0590***	0.0596***	0.0630***	0.0637***
	(0.0034)	(0.0043)	(0.0045)	(0.0047)	(0.0048)
Fixed effects	Country	School	Classroom	Student	Student & Teacher
R-squared	0.13	0.17	0.19	0.68	0.72
Countries	72	72	72	71	71
Observations	1,842,968	1,008,485	856,700	781,204	776,713
Panel (d)					
Most restrictive sample					
Role model effect	0.0633 ***	0.0636 ***	0.0639***	0.0647***	0.0652***
	(0.0048)	(0.0048)	(0.0047)	(0.0048)	(0.0050)
Fixed effects	Country	School	Classroom	Student	Student & Teacher
R-squared	0.13	0.19	0.20	0.68	0.72
Countries	71	71	71	71	71
Observations	776,713	776,713	776,713	776,713	776,713

Note: This table shows more details on the role-model effects estimates shown in Figures 5 and 7. The "role model effect" in the table stems from a regressions of standardized test scores (Panels a and b)or job preferences (Panels c and d) on a $FemaleStudent_i \times FemaleTeacher_j$ interaction term, a set of other control variables (see Section 5), and different sets of fixed effects. 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) and Panel (c) show role model effect estimates from specifications that use the largest possible estimation sample. Panel (b) and Panel (d) show estimates with one consistent estimation sample as imposed by our preferred teacher and student fixed effects specification (see Section 5). Standard errors clustered at the classroom level are in parentheses. ***, **, and * mark estimates statistically different from zero at the 1%, 5%, and 10% significance levels.

Table A6: Student- and Teacher-Level Heterogeneity for Test Scores Estimates

			95% Confidence Interval		
	Average Effect	SE	LB	UB	N
Dependent variable: **Std. Test Scores**					
Student Characteristics					
Grade 4	0.0039	(0.0072)	-0.0102	0.018	160,480
Grade 8	0.0169***	(0.0036)	0.0098	0.024	1,451,717
Foreign	0.0185	(0.0157)	-0.0123	0.0492	106,478
Native	0.0152***	(0.0035)	0.0083	0.022	1,405,458
University-educated parent(s)	0.0162*	(0.0084)	0.0003	0.0327	428,801
No university-educated parent(s)	0.0180***	(0.0049)	0.0084	0.0276	732,604
Two-parent household	0.0124	(0.0112)	-0.0095	0.034	213,632
No two-parent household	0.0165	(0.0130)	-0.009	0.042	113,013
Teacher Characteristics					
15+ years of experience	0.0132*	(0.0072)	-0.0009	0.0273	602,999
Less than 15+ years of experience	0.0187***	(0.0051)	0.0087	0.0287	599,835
Post-graduate degree	0.0119	(0.0112)	-0.01	0.0339	303,810
No post-graduate degree	0.0128***	(0.0037)	0.0055	0.02	1,025,066
Education major	0.0076	(0.0049)	-0.002	0.0172	555,223
Not an education major	0.0188***	(0.0064)	0.0063	0.0313	452,290
Expert in their field	0.0154***	(0.0035)	0.0085	0.0222	1,218,558
Not an expert in their field	0.0027	(0.0142)	-0.0251	0.0305	44,515
Classroom has 30+ students	0.0158***	(0.0046)	0.0068	0.0248	433,798
Classroom has less than 30 students	0.0133***	(0.0045)	0.0045	0.0221	1,178,387

Note: This table shows estimated role model effects from regressions of standardized test scores and job preferences on a $FemaleStudent_i \times FemaleTeacher_j$ interaction term, student fixed effects, teacher fixed effects, as well as other control variables from our preferred specification (see Section 5) for the different subsamples indicated on the left of the table.

Table A7: Subject Heterogeneity

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

Note: This figure shows estimated role model effects from regressions of the outcome variable shown in the first row of each panel on a $FemaleStudent_i \times FemaleTeacher_j$ interaction term and other control variables from our school fixed effects specification (see Section 5). In this specification we can identify role model effects on test scores, enjoyment and confidence in 89 countries, and on job preferences in 72 countries. However, math and science test scores and data on enjoyment and confidence are not available in four countries (Belize, Luxembourg, Macao, and Trinidad and Tobago). There is also no identifying within-school variation in same-sex science teachers in Honduras in our estimation sample. Reading test scores are also not available in 33 of these 89 countries. Standard errors clustered at the classroom level are in parentheses. ***, **, and * mark estimates statistically different from zero at the 1%, 5%, and 10% significance levels.

Table A9: Main Results Without Multiple-Teacher Classroom

Std. outcome:	**Test Scores**	**Job Preference**	**Subject Enjoyment**	**Subject Confidence**
Role model effect	0.0181***	0.0719***	0.0900***	0.0620***
	(0.0045)	(0.0077)	(0.0063)	(0.0069)

Note: This table shows estimated role model effects from regressions of the outcome shown on in the column headings on a $FemaleStudent_i \times FemaleTeacher_j$ interaction term, student fixed effects, teacher fixed effects, as well as other control variables from our preferred specification (see Section 5) *without classes that were taught by multiple teachers*. Standard errors clustered at the classroom level are in parentheses. ***, **, and * mark estimates statistically different from zero at the 1%, 5%, and 10% significance levels.

Table A10: Global Heterogeneity of Role Model Effects on Test Scores and Job Preferences

	Std. Test Scores				Std. Job Preferences		
Country	Average Effect	SE	N	Country	Average Effect	SE	N
Serbia	0.064	0.024	12,402	Kuwait	0.516	0.287	530
Iran	0.063	0.072	324	Malta	0.469	0.233	376
Estonia	0.056	0.051	4,378	Bulgaria	0.323	0.096	6,162
Syria	0.056	0.024	5,612	Italy	0.251	0.033	140
Ireland	0.052	0.022	6,210	Portugal	0.201	0.055	8,436
Azerbaijan	0.043	0.084	348	Oman	0.195	0.112	2,040
Argentina	0.041	0.034	476	Slovak Rep.	0.177	0.061	6,398
Scotland	0.035	0.021	8,078	Ukraine	0.159	0.112	3,364
Greece	0.032	0.024	8,630	Spain	0.155	0.077	3,250
Ghana	0.031	0.024	6,238	Israel	0.144	0.047	9,672
South Korea	0.028	0.012	14,036	Iceland	0.141	0.110	1,586
Kazakhstan	0.026	0.027	11,090	Slovenia	0.135	0.047	13,510
Germany	0.026	0.019	3,878	Russian Fed.	0.127	0.067	16,634
Israel	0.026	0.017	11,596	Egypt	0.116	0.064	6,718
Hungary	0.025	0.011	34,136	Ireland	0.112	0.057	5,872
Moldova	0.025	0.030	9,832	United States	0.107	0.022	34,250
Yemen	0.025	0.065	1,500	Hong Kong	0.098	0.022	18,518
Slovenia	0.023	0.017	17,630	Argentina	0.098	0.185	382
Malaysia	0.022	0.009	22,700	Hungary	0.096	0.046	27,232
Egypt	0.022	0.021	8,538	England	0.091	0.042	10,112
UAE	0.022	0.014	10,019	New Zealand	0.090	0.028	15,086
Bosnia and Herz.	0.021	0.025	6,752	Greece	0.090	0.044	8,208
Bahrain	0.020	0.034	3,194	France	0.088	0.058	8,756
Russian Fed.	0.020	0.029	19,202	Cyprus	0.085	0.046	13,202
Japan	0.019	0.008	40,674	Romania	0.085	0.039	22,380
Lebanon	0.019	0.014	19,824	Denmark	0.080	0.081	1,248
Qatar	0.019	0.016	5,006	Austria	0.077	0.086	5,520
Hong Kong	0.019	0.008	34,144	Macedonia	0.076	0.045	14,796
Taiwan	0.017	0.005	44,586	Japan	0.073	0.016	25,184
Austria	0.017	0.042	6,054	Palestine	0.073	0.154	582
New Zealand	0.016	0.009	17,102	Turkey	0.072	0.026	21,776
Honduras	0.014	0.013	3,974	South Korea	0.071	0.030	10,404
United States	0.013	0.006	49,652	Netherlands	0.069	0.037	10,780
El Salvador	0.012	0.013	4,534	Malaysia	0.068	0.020	19,436
Turkey	0.011	0.007	30,304	Norway	0.062	0.048	6,764
Australia	0.010	0.010	28,061	Canada	0.061	0.024	23,802
Morocco	0.010	0.007	60,700	Taiwan	0.059	0.015	25,150
Malta	0.009	0.017	3,324	Lebanon	0.049	0.045	14,560
Romania	0.009	0.014	30,112	Chile	0.041	0.042	18,790
Algeria	0.009	0.017	6,664	Finland	0.039	0.026	14,591
Macedonia	0.009	0.019	16,012	Lithuania	0.038	0.074	16,282
South Africa	0.008	0.005	55,534	Indonesia	0.033	0.023	20,262
Singapore	0.008	0.007	42,232	Singapore	0.032	0.022	26,598
Poland	0.008	0.023	2,656	Scotland	0.031	0.093	1,872
Indonesia	0.007	0.011	27,214	Australia	0.029	0.023	22,970
Denmark	0.007	0.018	7,148	Switzerland	0.028	0.073	3,138
Slovak Rep.	0.007	0.019	8,296	Colombia	0.028	0.060	3,218
Palestine	0.006	0.045	710	UAE	0.024	0.061	5,584
Finland	0.006	0.011	16,163	Sweden	0.021	0.031	17,036
England	0.004	0.010	15,483	Thailand	0.018	0.021	18,064
Netherlands	0.004	0.017	11,303	Philippines	0.017	0.036	8,460
Iceland	0.004	0.044	1,806	Tunisia	0.009	0.053	13,562
Chile	0.004	0.012	20,734	Syria	0.007	0.076	3,428
Portugal	0.003	0.019	8,760	South Africa	0.003	0.018	41,928
Armenia	0.003	0.028	13,820	Estonia	0.001	0.089	4,356
Czech Rep.	0.003	0.019	12,248	Morocco	-0.001	0.025	37,376
Thailand	0.003	0.009	23,156	Armenia	-0.004	0.069	8,346
Oman	0.003	0.020	3,842	Botswana	-0.005	0.031	13,018
Sweden	0.003	0.011	24,166	Germany	-0.006	0.163	654
Botswana	0.002	0.010	18,698	Georgia	-0.006	0.085	5,442
Lithuania	0.001	0.021	20,982	Iran	-0.012	0.109	320
Canada	0.000	0.007	32,356	Kazakhstan	-0.020	0.056	10,554
Colombia	0.000	0.016	9,476	Qatar	-0.025	0.074	3,370
Norway	-0.001	0.013	11,430	Czech Rep.	-0.025	0.088	4,738
Tunisia	-0.002	0.011	20,010	Moldova	-0.040	0.058	9,294
Spain	-0.004	0.016	9,298	Ghana	-0.040	0.049	4,164
Switzerland	-0.005	0.036	3,250	Serbia	-0.045	0.086	6,070
Georgia	-0.009	0.025	8,188	Latvia	-0.051	0.087	5,400
Cyprus	-0.011	0.012	27,888	Bahrain	-0.060	0.130	2,094
Latvia	-0.013	0.037	5,608	Belgium	-0.071	0.092	2,332
France	-0.015	0.021	10,060	Jordan	-0.245	0.183	586
Ukraine	-0.017	0.035	7,354				
Bulgaria	-0.018	0.033	9,140				
Italy	-0.020	0.041	554				
Philippines	-0.021	0.014	11,094				
Jordan	-0.023	0.058	668				
Belgium	-0.026	0.025	2,806				
Kuwait	-0.031	0.085	1,024				
Mongolia	-0.040	0.024	2,204				

Note: This table shows estimated role model effects from regressions of standardized test scores (Panel a) or standardized job preferences (Panel b) on a $FemaleStudent_i \times FemaleTeacher_j$ interaction term, student fixed

effects, teacher fixed effects, as well as other control variables from our preferred specification (see Section 5) for the different country-subsamples indicated on the left of each panel. Panel (a) shows 79 point estimates and Panel (b) shows 71 point estimates. The smaller number of estimated role-model effects on job preferences is due to missing data on job preferences for Algeria, Azerbaijan, Bosnia and Herzegovina, El Salvador, Honduras, Poland, Mongolia, and Yemen. Three countries with less than 150 observations are omitted from the figure (Albania, Northern Ireland and Pakistan).

Table A10: Global Heterogeneity for Role Model Effects in Job Preferences

Panel A Std. Dependent Variable: Job Preferences	GDP per capita	Human Development Index	Gender Equality Index	University Enrolment
Role model effect	0.0509***	0.0505***	0.0512***	0.0389***
	(0.0074)	(0.0075)	(0.0076)	(0.0078)
Role model effect * above median	0.0237**	0.0238**	0.0200*	0.0403***
	(0.0101)	(0.0101)	(0.0104)	(0.0109)
Countries	67	68	67	63
Obs.	745681	749109	730591	678641
Predicted effect for below median group	0.051***	0.051***	0.051***	0.039***
p-value	0.0000	0.0000	0.0000	0.0000
Predicted effect for above median group	0.075***	0.074***	0.071***	0.079***
p-value	0.0000	0.0000	0.0000	0.0000

Panel B Std. Dependent Variable: Job Preferences	Bank Account Ownership	Fertility	Science Score M-F Gap	Math Score M-F Gap
Role model effect	0.0487***	0.0701***	0.0435***	0.0371***
	(0.0079)	(0.0071)	(0.0067)	(0.0064)
Role model effect * above median	0.0251**	-0.0142	0.0425***	0.0568***
	(0.0102)	(0.0101)	(0.0097)	(0.0098)
Countries	67	68	71	71
Obs.	747523	749109	776713	776713
Predicted effect for below median group	0.049***	0.07***	0.044***	0.037***
p-value	0.0000	0.0000	0.0000	0.0000
Predicted effect for above median group	0.074***	0.056***	0.086***	0.094***
p-value	0.0000	0.0000	0.0000	0.0000

Note: This table shows estimated role model effects from regressions of standardized job preference on an above-median dummy variable of the characteristic shown in the column title, a $FemaleStudent_i \times FemaleTeacher_j$ interaction term, and an interaction of those two variables. Additional controls include student fixed effects, teacher fixed effects, as well as other control variables from our preferred specification (see Section 5). Standard errors clustered at the school level are in parentheses. ***, **, and * mark estimates statistically different from zero at the 10%, 5%, and 1% significance levels.

Appendix B

Additional Information about the Meta-Analysis

B1: Data Collection

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

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

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

Next, we removed duplicates (keeping the latest version) within and across the five data sources, thus narrowing our dataset to 4,150 studies. After that, we dropped all studies pre-

registered on Social Science Registry that matched our keywords but failed to include test scores or grades among their primary outcomes. We further filtered these results from SSR on study status, keeping only projects classified as complete and offering available results. After these pre-processing steps, we manually screened the title, abstract, and where necessary, the introduction of the remaining 1,838 studies and excluded those that did not match all pre-registered inclusion criteria. We then performed full-text assessments of 174 articles to identify point estimates. Next, we removed all studies that did not allow us to calculate *standardized* role model effects and standard errors (e.g., if they did not report the standard deviation of the outcome). This left us with 24 studies reporting at least one same-sex role model effect.

To avoid overlooking studies that did not use our keyword combinations, we identified studies that had more than 50 citations. For these highly cited papers, we collected the top ten papers that cited these seminal studies using the "cited by" functionality on Google Scholar. Through this process, we thereby identified 130 additional potential role model studies. Of those 130 studies, none reported a same-sex role model effect. That left our final sample of 24 studies. Figure B1 summarizes the data collection using the PRIMSA flow chart.

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

the difference between the female teacher effect for girls and the female teacher effect for boys. Recovering the standard error for this difference is impossible without making further assumptions. However, by assuming a zero covariance between both estimates, we recover the standard error of the difference as the square root of the sum of squared standard errors.

Furthermore, for each estimate we recorded the following information:

- Study ID
- Citation (APA)
- Abstract
- Link to publication (DOI or PDF)
- Citation count as of Nov. 25 (same as indicator for 100+ citations)
- Main outcome (test Score or grade)
- Number of observations
- Effect size
- Standard error as reported
- Effect size in std. dev
- Standard error in std. dev
- Subject
- Estimation method
- Country
- Level of education
- Identification of main analysis
- Identifying variation in the main specification
- First year of measurement
- Last year of measurement
- Coefficient specification type (the coefficient's type of interaction)
- Subsample of students
- Single subject
- Most controlled estimate
- Heterogeneous effect (same as subsample of students)
- Heterogeneity type (if the coefficient is subsample. E.g., gender, single vs multiple teachers, native vs foreign students…)
- Included in appendix
- Model/table (the exact table/column location of the estimate)
- Fixed effects
- Controls
- Comments

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

Consistency check: After the conclusion of the data collection, two predoctoral researchers not involved in the initial coding randomly selected 5 studies and replicated the data collection. Any ambiguities identified through this process were resolved in discussions with a co-author on this project. We recorded whether a study was checked for consistency, whether inconsistencies were found, and how they were resolved. Out of 106 replicated estimates, 2 inconsistent estimates and 3 inconsistent standard errors were found, yielding an error rate of 4.71%. These false values were corrected in the base dataset.

In addition to replicating the data collection for 5 studies, all estimates in the remaining 19 studies were cross-checked by a different research assistant. Any ambiguities identified through this process were resolved in discussions with a co-author on this project. This yielded an error rate of 7.65% and false values were corrected in the base dataset.

Figure B1: Data Collection Flowchart

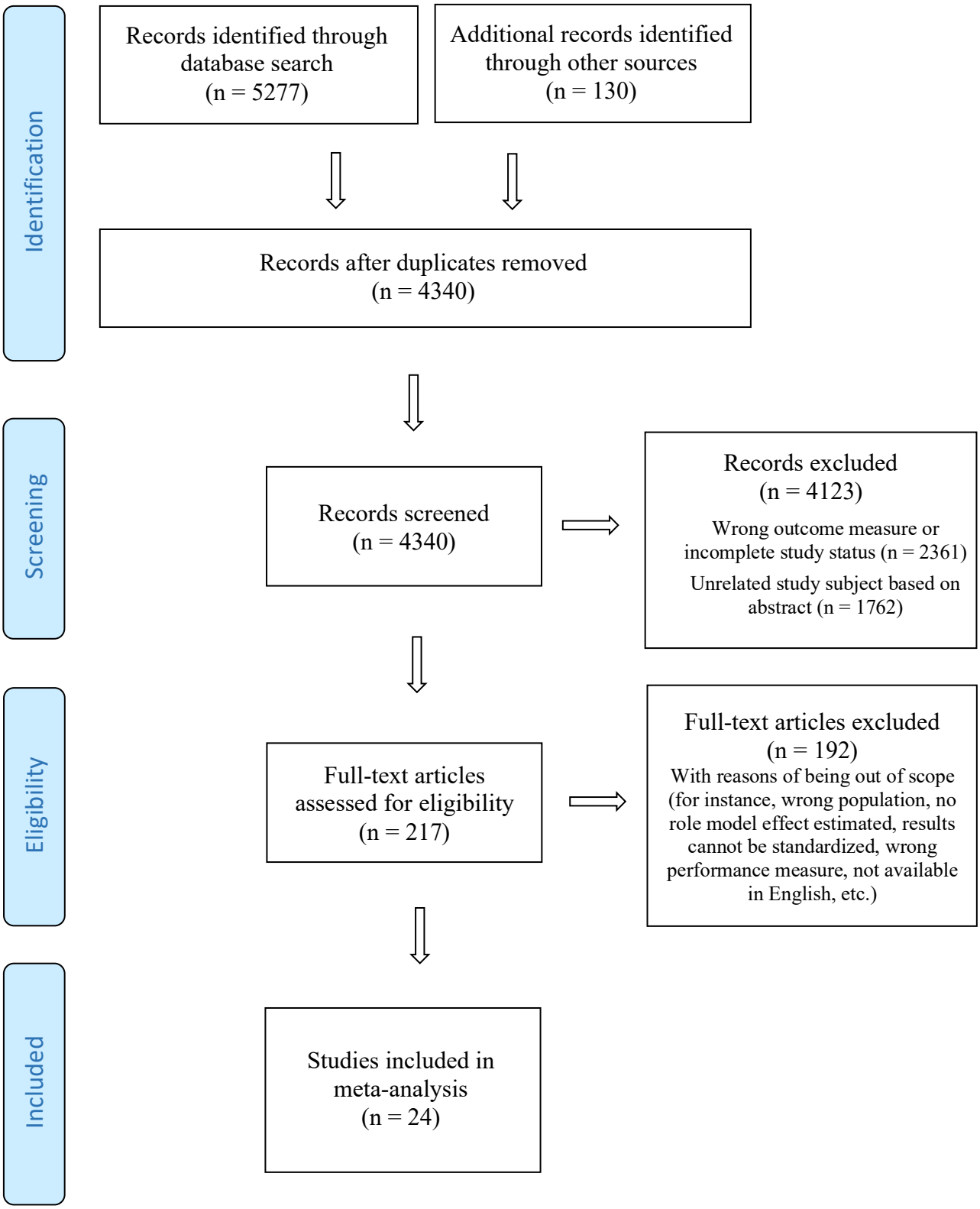


Table B1: List of Included Studies

1	Ammermüller and Dolton (2006)
2	Antecol et al. (2015)
3	Bhattacharya et al. (2020)
4	Buddin and Zamarro (2008)
5	Carrington et al. (2008)
6	Coenen and Klaveren (2016)
7	Dee (2007)
8	Eble and Hu (2020)
9	Escardíbul and Mora (2013)
10	Evans (1992)
11	Gong et al. (2018)
12	Hermann and Diallo (2017)
13	Holmlund and Sund (2008)
14	Hwang and Fitzpatrick (2021)
15	Lee et al. (2019)
16	Lim and Meer (2017)
17	Lim and Meer (2020)
18	Lindahl (2007)
19	Mulji (2016)
20	Muralidharan and Seth (2016)
21	Neugebauer et al. (2011)
22	Rakshit and Sahoo (2020)
23	Xu and Li (2018)
24	Xu (2020)

Appendix B2: Meta-Analysis Using "Most Controlled" Estimates

In this Appendix we conduct our meta-analysis with a more restrictive sample of estimates. We now focus on the "most controlled" estimates within each study, defined as those from model specifications using the largest amount of control covariates and narrowest within-group variation. From these estimates we additionally exclude "first difference" estimates, defined as effects of 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 from our 24 selected studies.

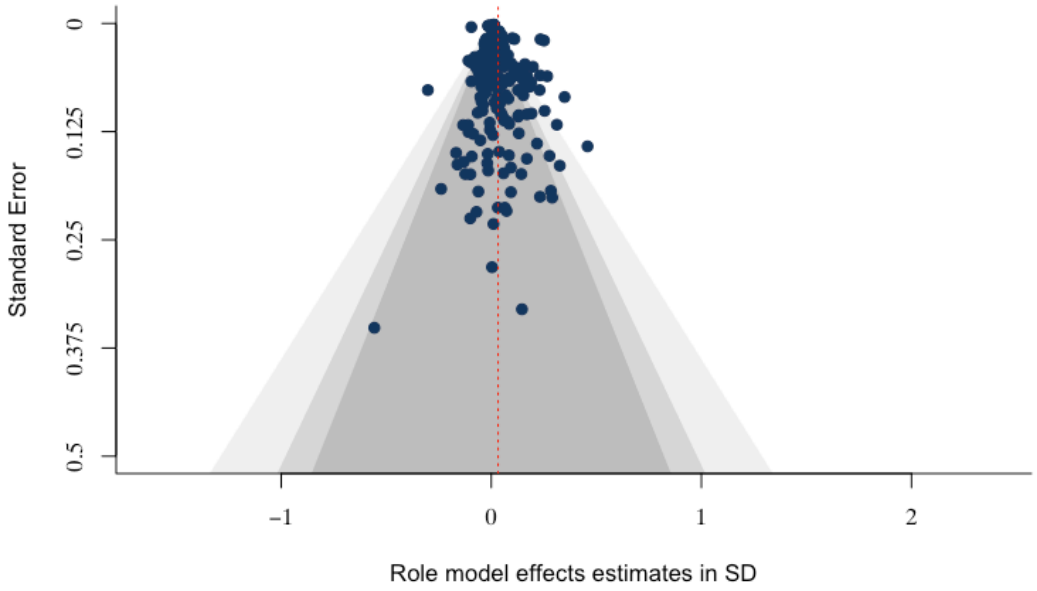
Table B2.1: Role Model Effect Estimates Corrected for Publication Bias, using the "Most Controlled" Set of Estimates

Estimation method	Significance threshold for selection	Average effect	Standard error	95% Confidence Interval		Standard deviation of effect
Three-level REML	-	0.031	(0.014)	0.005	0.060	0.06
Trim and Fill	-	0.007	(0.004)	-0.002	0.015	0.057
PET-PEESE	-	0.006	(0.001)	0.005	0.007	0.000
Limit-Meta	-	0.031	(0.161)	-0.285	0.347	0.060
3-Parameter Selection	10%	0.024	(0.005)	0.015	0.034	0.037
3-Parameter Selection	5%	0.033	(0.006)	0.022	0.044	0.041
3-Parameter Selection	1%	0.035	(0.006)	0.024	0.047	0.042
Andrews & Kasy (t)	10%	0.007	(0.002)	0.003	0.011	0.007
Andrews & Kasy (t)	10%, 5%	0.007	(0.001)	0.005	0.009	0.007
Andrews & Kasy (t)	10%, 5%, 1%	0.008	(0.001)	0.006	0.010	0.007
Andrews & Kasy (N)	10%	-0.017	(0.014)	-0.044	0.010	0.056
Andrews & Kasy (N)	10%, 5%	-0.011	(0.015)	-0.040	0.018	0.065
Andrews & Kasy (N)	10%, 5%, 1%	-0.012	(0.019)	-0.049	0.025	0.086

Note: As benchmark, 3-level REML shows the estimated role model effect without correcting for publications bias (see Section 2.2). 3-level REML: Inverse variance method used for pooling estimates. Restricted maximum likelihood estimator of the standard deviation of the effect size and its confidence interval. Knapp-Hartung adjustment for the uncertainty in the between-study heterogeneity applied to the standard error of the effect size. Trim and Fill: Inverse variance method used for pooling estimates. Restricted maximum likelihood estimator of the standard deviation of the effect size. Q-profile method used for the confidence interval of the standard deviation of the effect size. Knapp-Hartung adjustment for the uncertainty in the between-

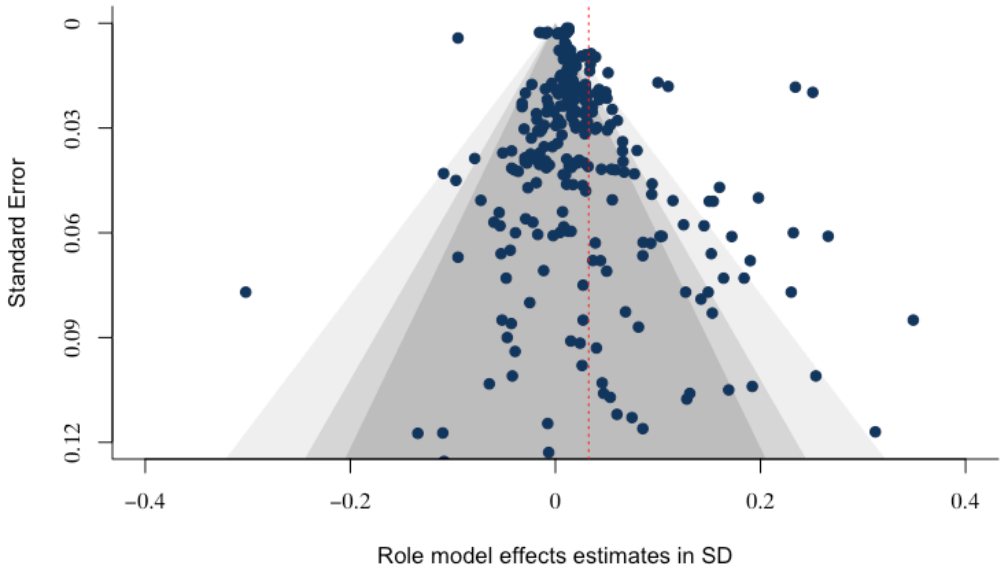
study heterogeneity applied to the standard error of the effect size. PET-PEESE: We use estimates from the precision-effect test (PET) model rather than from the precision-effect estimate with standard error (PEESE) model because the one-sided *t*-test of intercept for the PET model does not reject the null hypothesis at the 5% level. Estimates weighted by their inverse variance. We use the *rma.uni()* function in *R* for implementing this method. Limit-Meta: Uses 3-level REML (see above) as input. 3-Parameter Selection: We use 0.05, 0.025, and 0.01 as jumps in the publication probability function. Restricted maximum likelihood estimator of the standard deviation of the effect size, the standard deviation of the effect size and their confidence intervals. Andrews and Kasy: We use 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, at the 0.05 and 0.025, and at the 0.05, 0.025 and 0.01 significance levels for both positive and negative effects. We allow the probability of publications bias to be asymmetric. We produce estimate using Kasy's App: https://maxkasy.github.io/home/metastudy. Other correction methods: Andrews and Kasy's (2019) 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).

Figure B2.1: Funnel Plot of "Most Controlled" Role Model Effect Estimates



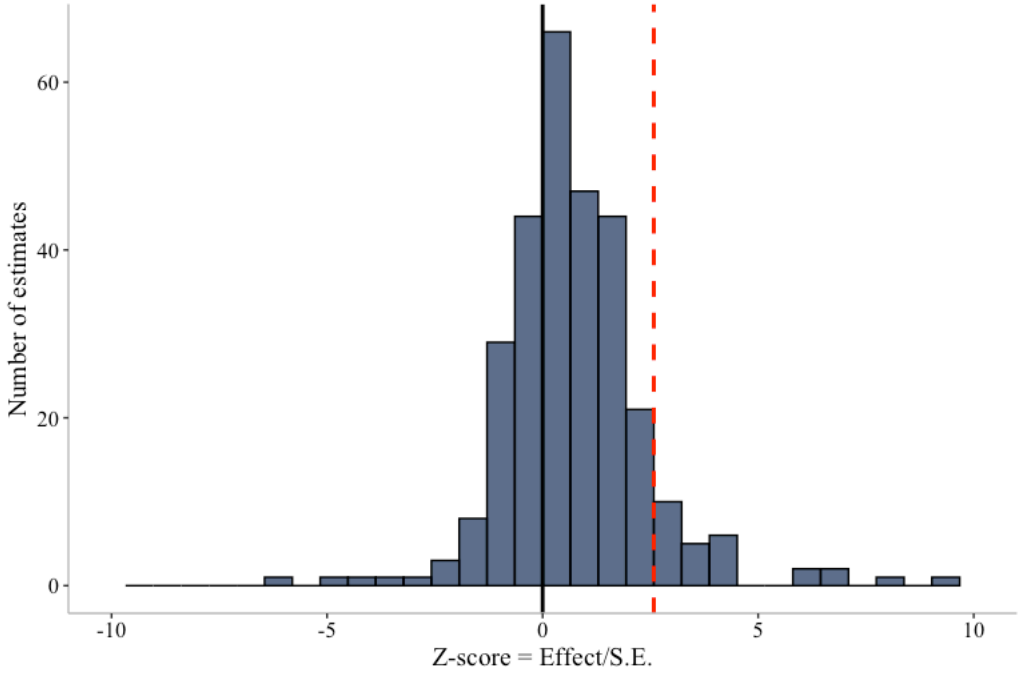
Note: This figure shows a scatterplot of 296 most-controlled role model effects estimates from 24 studies on the x-axis, with their standard error on the y-axis. For this figure we exclude 1 outlying role model estimate of size 2.07 SD with a standard error of 5.42 SD. The grey shaded areas mark the traditional thresholds for statistical significance at the 10%, 5% and 1% level. The vertical dotted line marks our estimated average role model effect of 0.032 SD.

Figure B2.2: Zoom into Funnel Plot of "Most Controlled" Role Model Effect Estimates



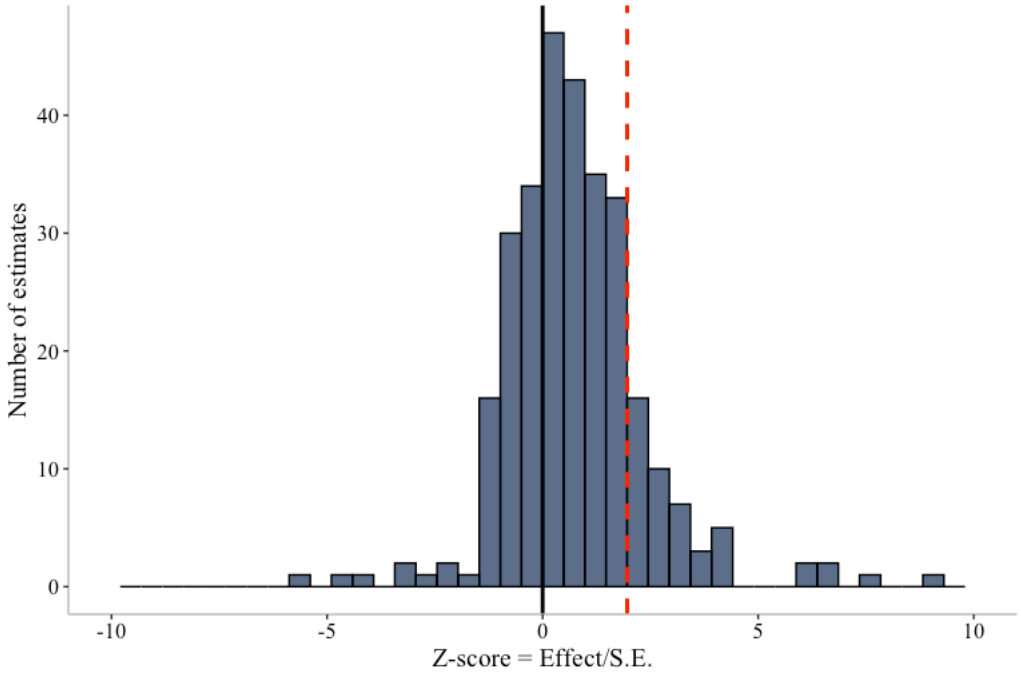
Note: This figure shows a zoomed-in subsection of the scatterplot in Figure B2.4 where most estimates are present. The grey shaded areas mark the traditional thresholds for statistical significance at the 10%, 5% and 1% level. The vertical dotted line marks our estimated average role model effect of 0.032 SD.

Figure B2.3: *Z*-scores Distribution of "Most Controlled" Role Model Effect Estimates, with 99% Two-sided Critical Value Marked



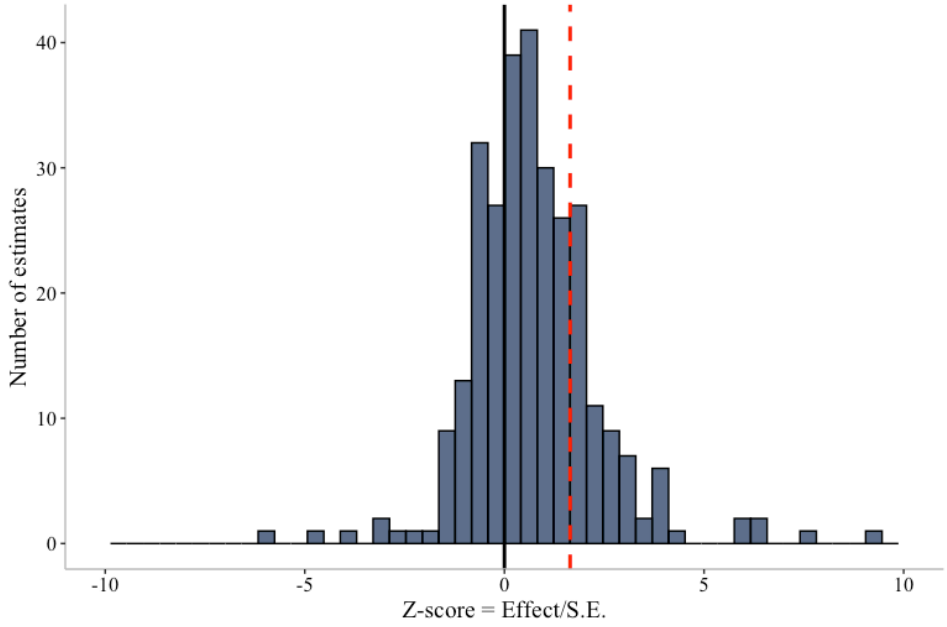
Note: This figure shows *z*-scores of 297 most-controlled role model effects estimates from 24 studies. The vertical dashed line marks 2.58, the two-sided test critical value of 99% for the normal distribution. The histogram uses a bin width of 0.645.

Figure B2.4: *Z*-scores Distribution of "Most Controlled" Role Model Effect Estimates, with 95% Two-sided Critical Value Marked



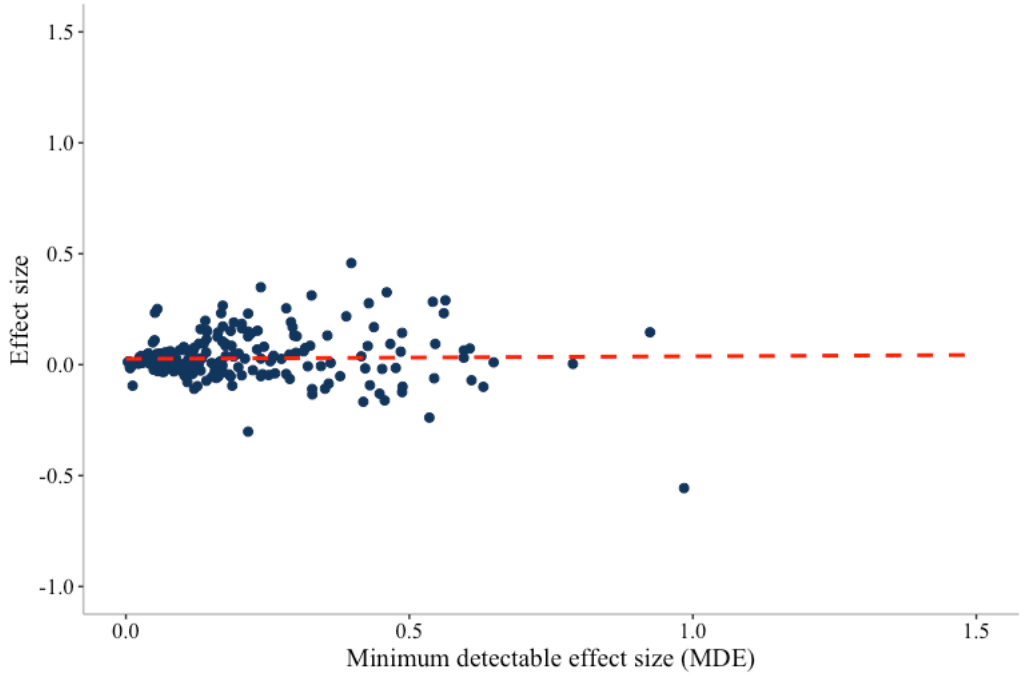
Note: This figure shows *z*-scores of 297 most-controlled role model effects estimates from 24 studies. The vertical dashed line marks 1.96, the two-sided test critical value of 95% for the normal distribution. The histogram uses a bin width of 0.49.

Figure B2.5: *Z*-scores Distribution of "Most Controlled" Role Model Effect Estimates, with 90% Two-sided Critical Value Marked



Note: This figure shows *z*-scores of 297 most-controlled role model effects estimates from 24 studies. The vertical dashed line marks 1.64, the two-sided test critical value of 90% for the normal distribution. The histogram uses a bin width of 0.41.

Figure B2.6: Minimum Detectable Effect Size (MDE) Plot of "Most Controlled" Role Model Estimates



Note: This figure shows the relationship between role model effect estimates (y-axis) and their corresponding ex-post MDE size (x-axis), calculated by multiplying the standard error by 2.8 (and thus assuming 80% power). Each dot represents one different role model effect estimate. To increase readability, this figure excludes one outlying

role model estimate of size 2.07 SD with an MDE of 15.19 SD. The dashed line shows the linear fit from regressing all role model estimates on MDEs. The slope estimate from this regression is 0.132, with a standard error of 0.005 clustered at the study level. Excluding the outlying estimate yields a slope of 0.117 and standard error of 0.084.

Table B2.2: Meta-Regression of Role Model "Most Controlled" Estimates

			95% CI	
	Coef.	Std.err.		
Intercept	0.077	(0.090)	-0.101	0.255
Variation (base = Experimental)				
Quasi-experimental	0.010	(0.057)	-0.103	0.122
Continent (base = Africa)				
Asia	-0.051	(0.054)	-0.157	0.055
Europe	-0.072	(0.053)	-0.175	0.032
North America	-0.112**	(0.054)	-0.218	-0.006
School level (base = Secondary)				
Elementary	-0.027	(0.034)	-0.093	0.040
Both	-0.012	(0.035)	-0.081	0.056
Outcome (base = Grades)				
Test scores	0.024	(0.046)	-0.066	0.113
Test for significance of moderators (p-value)	0.001			
Test for residual heterogeneity (p-value):	<0.0001			
Variance components (τ)				
Between studies	0.0039			
Within studies	0.0003			

Note: Three-level meta-regression of role model effect estimates on grades and test scores, estimated using the *meta* package in *R*. The first level contains studies, and the second level contains estimates within each study. All moderators are coded at the study level. We only consider the "most controlled" estimates within each study, and additionally exclude "first difference" estimates. Our resulting subset of estimates includes 297 estimates from our 24 selected studies. Standard errors in parentheses. ***, **, and * mark estimates statistically different from zero at the 1%, 5%, and 10% significance levels.

B3: Pre-registration and deviations

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

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

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

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

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

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

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

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

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

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

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

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

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

1. *Type of main outcome (Test score or grade)*
2. *Number of observations for main results*
3. *Record one main effect, as identified by authors. For this effect, record:*
    a. *Effect size as reported*
    b. *Standard error as reported*
    c. *Effect size in standard deviations of the outcome*
    d. *Standard error in standard deviations of the outcome*
    e. *Subject (e.g., math)*

   f. *Country where the study takes place*
   g. *Level of education (e.g., grade 8)*
4. *Identification of main analysis (e.g. experiment, natural experiment, observational)*
5. *Identifying variation in the main specification (e.g., between students, within schools, within classrooms)*
6. *Data first year of measurement*
7. *Data last year of measurement*
8. *Indicator for 100+ citations.*

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

Things we did differently: Instead of coding only or a few main estimates, we coded *all* role model estimates from each relevant study's main text and appendix. We decided to expand the data collection to all estimates to be more thorough. In addition to the information listed above we also recorded the following information: Citation count as of Nov. 25 (same as indicator for 100+ citations), Main outcome (Test Score or grade), Number of observations, Effect size, Standard error as reported, Effect size in std. dev, Subject, Country, Level of education, Identification of main analysis, Identifying variation in the main specification, First year of measurement, Last year of measurement.

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

Instead of adding an indicator for 100+ citations, we added an indicator for a study having 50+ citations. We used this indicator to identify overlooked studies (see below). We lowered the bar to 50+ studies because we wanted to increase our chances of finding overlooked studies.

Despite our pre-registration's statement that an explanation for each excluded study would be recorded, such reasons were given succinctly or not at all in cases when the study was clearly not related to role model effects or failed most of our inclusion criteria. The 50+ indicator is based on the Google Scholar citation count on January 25, 2022.

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

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

Pre-registration of main effect recording:

We will report:

- *Meta regression results using all studies in our spreadsheet. We will estimate this model using a random-effect (RE) meta-regression. We will use the DerSimonian and Lair (1986) method to estimate the weights unless this becomes analytically impracticable;*

else will use more standard restricted maximum likelihood methods. This estimate will be produced using the meta regress, random(dlaird) functionality in the Stata meta-analysis command suite.

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

Pre-registration of publication bias correction:

We will report:

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

Things we did differently: In addition to the analysis in our pre-analysis plan we also implemented the ten other publication bias correction methods shown in Figure 3. We deviated from the pre-analysis plan because we found that results were quite sensitive to the exact correction method used.

Appendix C

Additional information about PIRLS and TIMSS

C.1 Sampling

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

C.2 Plausible Values

The test answers for each student are transformed into estimates of a student's subject-specific ability. For each student, the IEA calculates five *plausible values* per subject. These are different estimates of the student's latent subject-specific ability based on their answers. Each of the five sets of plausible values is standardized by setting the unweighted mean of all countries that participated in TIMSS 1995 to 500 points and setting their standard deviation to 100. To enable measurement of trends over time, achievement data from later TIMSS assessments (e.g., TIMSS 2011) were transformed to these same metrics. This was done by concurrently scaling the data from each successive assessment with the data from the previous assessment—a process known as concurrent calibration—and applying linear transformations

to place the results from each successive assessment on the same scale as the results from the previous assessment (see TIMSS 2019 Technical Report, Chapter 11, page 558). To simplify our analysis, we use the average of all five plausible values for each student as our main outcome variable. For simplicity, and following other studies who have worked with this data, we use the term "students' test score" to refer to the average of these five values. Previous work using TIMSS data shows that regression analysis results are generally robust to this simplification (e.g., Bietenbeck and Collins, 2020).

C.3 Construction of Base Dataset

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

In total, we excluded 19 out of 731 country-study-grade-wave combinations because of survey implementation issues. We excluded two country-grade-wave cases with empty student or teacher background files, such that student or teacher sex cannot be recovered; this issue occurred in Bulgaria and in South Africa in grade 8 in wave 1995. We also excluded 17 country-grade-wave combinations in which students could not be linked to their teachers and classroom. Those issues took place in the first wave of TIMSS grade 8 in 1995 and were due to miscoding in some schools of the key variable linking students to teachers. Those survey implementation issues are documented in the 1995 user guide as "implementation issues," and

lead to duplicate student observations with multiple test scores because student identifiers and student-teacher linking codes are miscoded in the Student-Teacher Linkage files (AST* and BST* files). We analyzed those files for all countries, grades, and waves. We confirm the implementation issues reported for 12 country-grade-wave cases in the 1995 documentation, and we exclude entirely country-waves from our analyses for which issues affected 97% to–98% of student observations in grade 8 in those countries (Belgium Flanders, Cyprus, Czech Republic, Hungary, Iran, Israel, Latvia, Lithuania, New Zealand, Romania, Slovak Republic, and Slovenia). In addition, we also excluded five country-grade-wave cases from our analyses for which we found evidence of similar implementation issues affecting more than 10% of student observations (Canada grades 4 and 8 affecting 59% of observations, Germany grade 8 affecting 66% of observations, England grade 8 affecting 18% of observations, Belgium Flanders grade 8 2003 affecting 16% of observations).

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

In the base dataset, job preference has a mean of 2.55 and a standard deviation of 1.02, subject enjoyment as a mean of 3.11 and a standard deviation of 0.91, and subject confidence has a mean of 3.12 and a standard deviation of 0.82. We use those means and standard deviations to standardize these three variables for our analysis (see Section 5).

Appendix D

Additional Role Model Effects Estimates on Subject Enjoyment and Subject Confidence

Table D1: Global Heterogeneity for Role Model Effects on Enjoyment

Panel A Std. Dependent Variable: Enjoyment	GDP per capita	Human Development Index	Gender Equality Index	University enrolment
Role model effect	0.0617***	0.0562***	0.1040***	0.0561***
	(0.0060)	(0.0060)	(0.0060)	(0.0061)
Role model effect * above median	0.0455***	0.0530***	-0.0390***	0.0527***
	(0.0083)	(0.0083)	(0.0084)	(0.0088)
Countries	77	78	77	71
Obs.	1033698	1038624	1005134	926103
Control	0.062***	0.056***	0.104***	0.056***
p-value control	<0.0001	<0.0001	<0.0001	<0.0001
above median GDP	0.107***	0.109***	0.065***	0.109***
p-value	0.0000	0.0000	0.0000	0.0000

Panel B Std. Dependent Variable: Enjoyment	Account ownership	Fertility	Science score M-F gap	Math score M-F gap
Role model effect	0.0710***	0.0940***	0.0601***	0.0585***
	(0.0064)	(0.0060)	(0.0059)	(0.0054)
Role model effect * above median	0.0257***	-0.0150*	0.0506***	0.0611***
	(0.0084)	(0.0083)	(0.0081)	(0.0080)
Countries	77	78	82	82
Obs.	1036896	1038624	1088056	1088056
Control	0.071***	0.094***	0.06***	0.058***
p-value control	<0.0001	<0.0001	<0.0001	<0.0001
above median GDP	0.097***	0.079***	0.111***	0.12***
p-value	<0.0001	<0.0001	<0.0001	<0.0001

Note: This table shows estimated role model effects from regressions of standardized subject enjoyment on an above-median dummy variable of the characteristic shown in the column title, a $FemaleStudent_i \times FemaleTeacher_j$ interaction term, and an interaction of those two variables. Additional controls include student fixed effects, teacher fixed effects, as well as other control variables from our preferred specification (see Section 5). Standard errors clustered at the school level are in parentheses. ***, **, and * mark estimates statistically different from zero at the 10%, 5%, and 1% significance level.

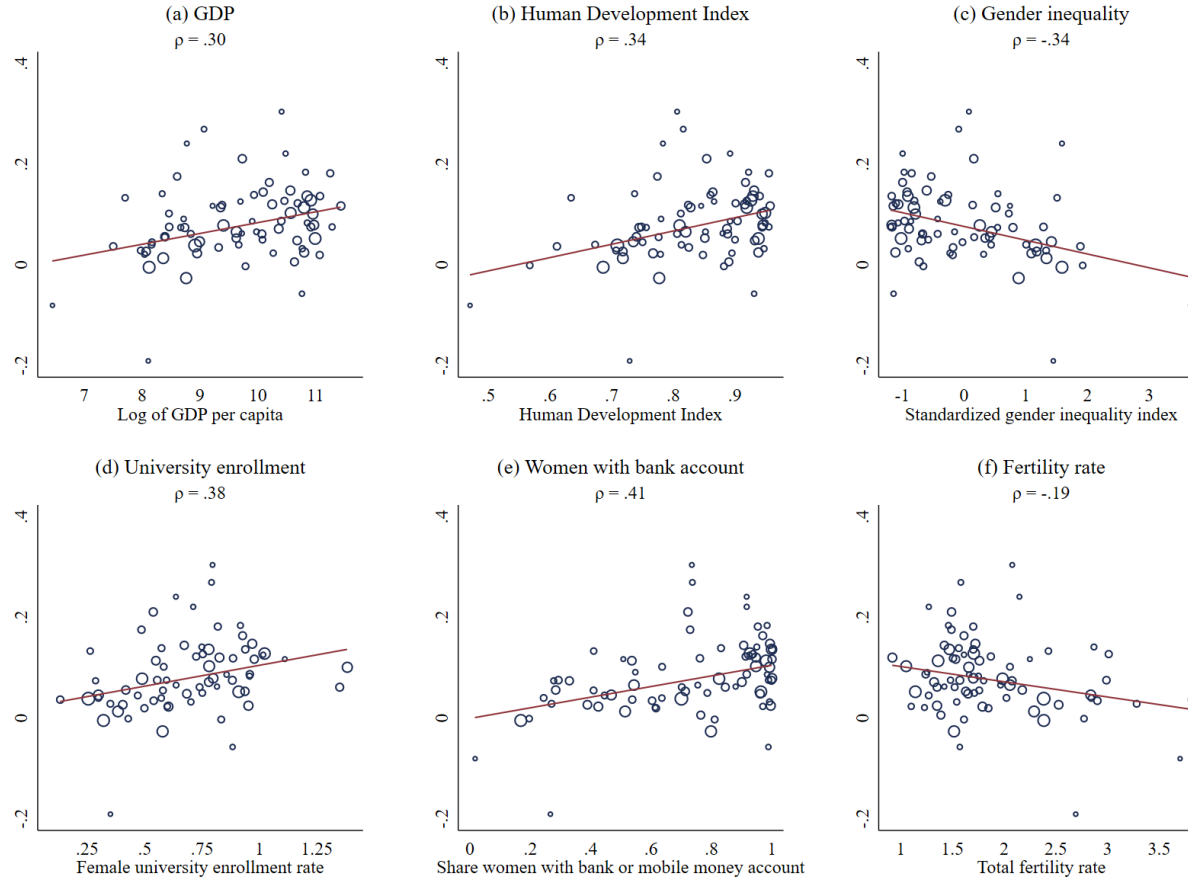
Table D2: Global Heterogeneity for Role Model Effects on Subject Confidence

Panel A Std. Dependent Variable: Confidence	GDP per capita	Human Development Index	Gender Equality Index	University enrollment
Role model effect	0.0301***	0.0275***	0.0573***	0.0234***
	(0.0062)	(0.0062)	(0.0055)	(0.0062)
Role model effect * above median	0.0288***	0.0332***	-0.0255***	0.0388***
	(0.0081)	(0.0081)	(0.0082)	(0.0085)
Countries	77	78	77	71
Obs.	1040154	1045318	1011700	932473
Control	0.03***	0.028***	0.057***	0.023***
p-value control	<0.0001	<0.0001	<0.0001	0.000183
above median GDP	0.059***	0.061***	0.032***	0.062***
p-value	0.0000	0.0000	0.0000	0.0000

Panel B Std. Dependent Variable: Confidence	Account ownership	Fertility	Science score M-F gap	Math score M-F gap
Role model effect	0.0343***	0.0549***	0.0314***	0.0354***
	(0.0065)	(0.0056)	(0.0060)	(0.0055)
Role model effect * above median	0.0211**	-0.0163**	0.0336***	0.0301***
	(0.0083)	(0.0080)	(0.0078)	(0.0077)
Countries	77	78	82	82
Obs.	1043570	1045318	1098172	1098172
Control	0.034***	0.055***	0.031***	0.035***
p-value control	<0.0001	<0.0001	<0.0001	<0.0001
above median GDP	0.055***	0.039***	0.065***	0.066***
p-value	0.0000	0.0000	0.0000	0.0000

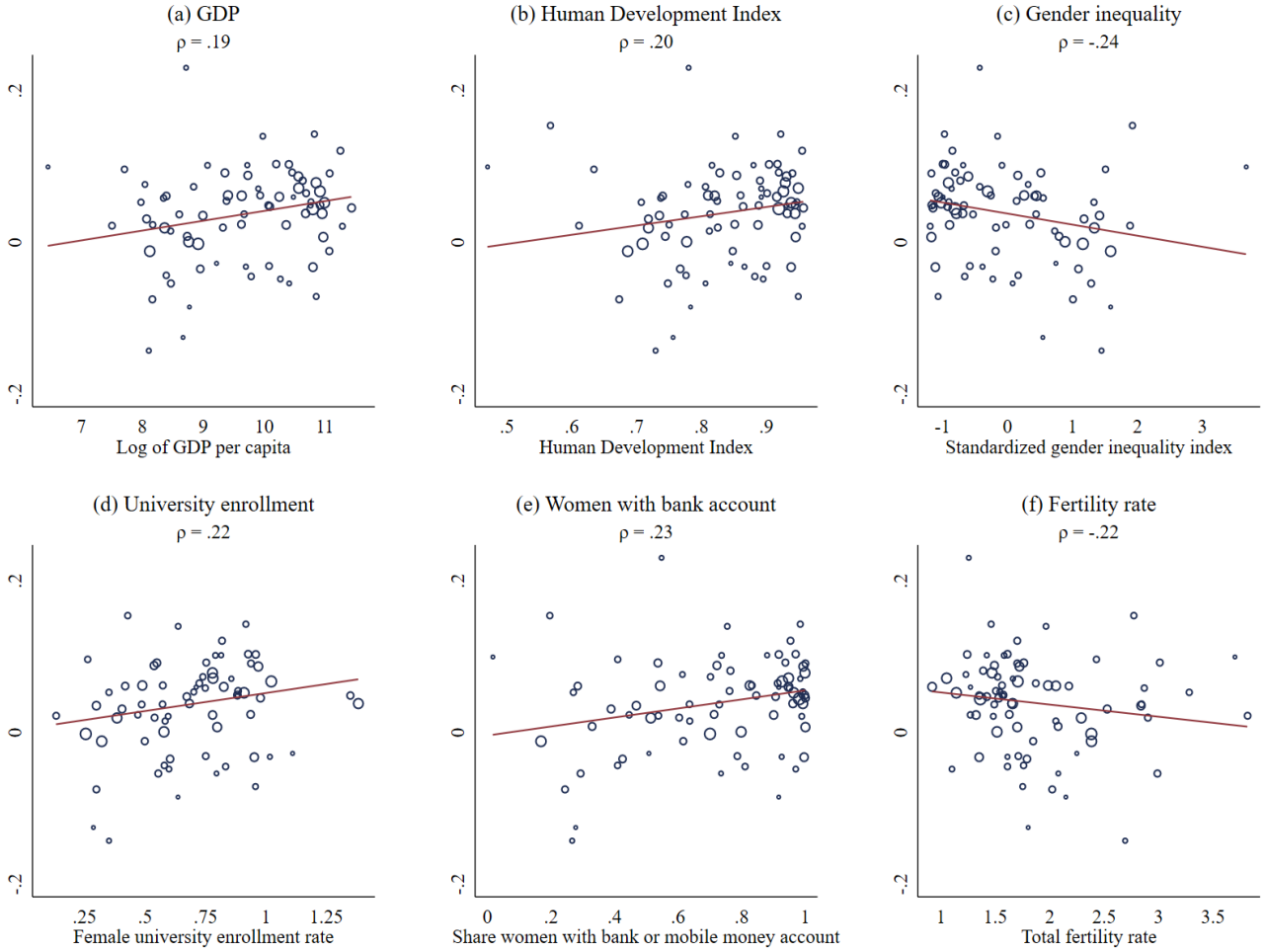
Note: This table shows estimated role model effects from regressions of standardized subject confidence on an above-median dummy variable of the characteristic shown in the column title, a $FemaleStudent_i \times FemaleTeacher_j$ interaction term, and an interaction of those two variables. Additional controls include student fixed effects, teacher fixed effects, as well as other control variables from our preferred specification (see Section 5). Standard errors clustered at the school level are in parentheses. ***, **, and * mark estimates statistically different from zero at the 10%, 5%, and 1% significance level.

Figure D1: Role Model Effects on Subject Enjoyment and Country-level Correlates



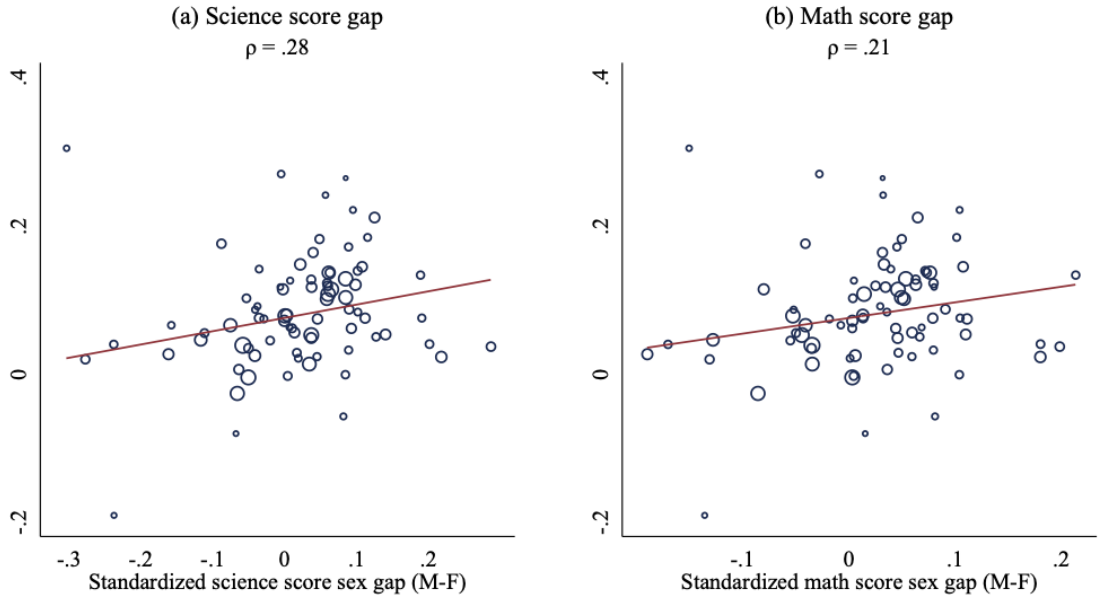
Note: These panels show the relationship between the estimated role model effects on standardized subject enjoyment shown in Figure A6 and different country level characteristics. ρ shows the Pearson's correlation coefficient between the two variables, the line shows a fitted least squares regression line. The characteristic shown in Panel (a) is log GDP per capita from 2019 which is taken from the World Bank World Development Indicators 2019. This characteristic is not available for Palestine, Scotland, Syria, and Taiwan. The characteristic shown in Panel (b) is the Human Development Index computed by the UN. This characteristic is not available for Palestine, Scotland, and Taiwan. The characteristic shown in Panel (c) is the standardized Gender Inequality Index (GII) from the Human Development Report 2020 published by the UN. The GII is calculated using this formula: GII = $\sqrt[3]{Health * Empowerment * LFPR}$ 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 the female population with at least some secondary education, and SE_M is the male population with at least some secondary education. *LFPR* is computed as the mean of male and female labor force participation rates: *LFPR* = $\frac{LFPR_F+LFPR_M}{2}$. The GII is missing for Hong Kong, Palestine, Scotland, and Taiwan. The characteristic shown in Panel (d) is the female university enrollment rate 2016/17. The female university enrollment rate is computed as the ratio of total female enrollment in tertiary education, regardless of age, to the female population of the age group that officially corresponds to tertiary education. This rate can hence be larger than 1, for example, if the number of overage women in tertiary education is large. The data are taken from the Gender Data Portal of the World Bank. This characteristic is available for all countries except for Japan, Lebanon, Palestine, Scotland, Taiwan, Turkey, Ukraine, and the United Arab Emirates. The characteristic in Panel (e) is the share of women of the female population aged 15+ who owned a bank or mobile money account in 2017. Data taken from the Gender Data Portal of the World Bank. This characteristic is not available for Iceland, Palestine, Scotland, and Taiwan. The characteristic shown in Panel (f) is the total fertility rate in 2019. The data are taken from the Gender Data Portal of the World Bank. This characteristic is not available for Palestine, Scotland, and Taiwan.

Figure D2: Role Model Effects on Subject Confidence and Country-level Correlates



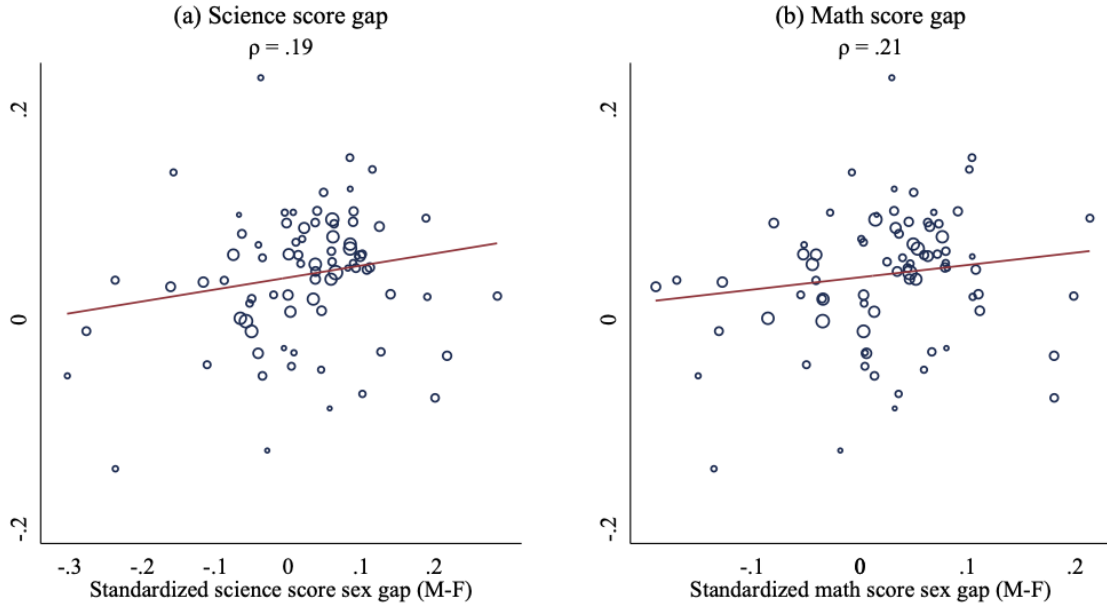
Note: These panels show the relationship between the estimated role model effects on standardized subject confidence shown in Figure A6 and different country level characteristics. ρ shows the Pearson's correlation coefficient between the two variables, the line shows a fitted least squares regression line. The characteristic shown in Panel (a) is log GDP per capita from 2019 which is taken from the World Bank World Development Indicators 2019. This characteristic is not available for Palestine, Scotland, Syria, and Taiwan. The characteristic shown in Panel (b) is the Human Development Index computed by the UN. This characteristic is not available for Palestine, Scotland, and Taiwan. The characteristic shown in Panel (c) is the standardized Gender Inequality Index (GII) from the Human Development Report 2020 published by the UN. The GII is calculated using this formula: GII = $\sqrt[3]{Health * Empowerment * LFPR}$ 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 the female population with at least some secondary education, and SE_M is the male population with at least some secondary education. *LFPR* is computed as the mean of male and female labor force participation rates: $LFPR = \frac{LFPR_F+LFPR_M}{2}$. The GII is missing for Hong Kong, Palestine, Scotland, and Taiwan. The characteristic shown in Panel (d) is the female university enrollment rate 2016/17. The female university enrollment rate is computed as the ratio of total female enrollment in tertiary education, regardless of age, to the female population of the age group that officially corresponds to tertiary education. This rate can hence be larger than 1, for example, if the number of overage women in tertiary education is large. The data are taken from the Gender Data Portal of the World Bank. This characteristic is available for all countries except for Japan, Lebanon, Palestine, Scotland, Taiwan, Turkey, Ukraine, and the United Arab Emirates. The characteristic in Panel (e) is the share of women of the female population aged 15+ who owned a bank or mobile money account in 2017. Data taken from the Gender Data Portal of the World Bank. This characteristic is not available for Iceland, Palestine, Scotland, and Taiwan. The characteristic shown in Panel (f) is the total fertility rate in 2019. The data are taken from the Gender Data Portal of the World Bank. This characteristic is not available for Palestine, Scotland, and Taiwan.

Figure D3: Role Model Effects on Subject Enjoyment and Test Score Gaps between Boys and Girls



Note: This figure shows the bivariate relationships between the estimated role model effects on standardized subject enjoyment shown in Figure A6 (on the y-axes) and the standardized sex gap (M-F) in science (Panel a) or math (Panel b) (on the x-axes). 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 79 countries for which we have role model effects on subject enjoyment.

Figure D4: Role Model Effects on Subject Confidence and Test Score Gaps between Boys and Girls



Note: This figure shows the bivariate relationships between the estimated role model effects on standardized subject confidence shown in Figure A6 (on the y-axes) and the standardized sex gap (M-F) in science (Panel a) or math (Panel b) (on the x-axes). 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 79 countries for which we have role model effects on subject confidence.