Abstract Title Page

Not included in page count.

Title: The effects of teachers' gender-stereotypical expectations on the development of the math gender gap

Author(s): Joseph P. Robinson, Sarah T. Lubienski, & Yasemin Copur (University of Illinois at Urbana-Champaign)

Abstract Body

Background / Context:

Scholars have identified mathematics gender gaps favoring males as early as kindergarten or first grade, particularly at the top of the achievement distribution (Penner & Paret, 2008; Rathbun, West & Germino-Hausken, 2004; Robinson & Lubienski, 2011). These relatively small achievement disparities precede larger differences in students' career choices. For example, men recently earned 82% of engineering bachelor's degrees, while women earned only 18% (Dey & Hill, 2007). Women's under-representation in math-related careers both limits the pool of talented people contributing to those fields and leaves disproportionate numbers of women in lower-paying occupations.

In examining the possible origins of these early math gender gaps, previous researchers looked inside mathematics classrooms and found that teachers tended to hold higher expectations of their male students and to view mathematics as a male domain (Li, 1999). Yet, in contrast to this previous work, recent, large-scale studies suggest that teachers actually rate the performance of girls more favorably than the performance of males (e.g., Fryer & Levitt, 2010; Robinson & Lubienski, 2011). Given gender disparities in mathematics-related careers, the new findings seem to be promising news if teachers' positive assessments help level the playing field for future generations of women in STEM careers.

However, these initial estimates of teachers' female bias may be misleading, confounding achievement with behavior and learning approaches. Indeed, prior research has revealed that girls tend to exhibit more on-task behavior and positive approaches to learning behavior in schools (Forgasz & Leder, 2001; Ready, LoGerfo, Lee & Burkam, 2005). Hence, teachers might conflate "good girl" behavior with mathematics proficiency. This study untangles these issues, examining whether teachers in a national sample rate boys' math proficiency higher than that of girls when boys and girls behave similarly, have similar approaches to learning, and have the same past and current test scores. This study also examines whether teachers' tendency to rate boys or girls higher is causally linked to the widening gender gap in mathematics in early elementary school.

Purpose / Objective / Research Question / Focus of Study:

In prior research, mathematics achievement gaps favoring males were found to widen during early elementary school; however, teachers tended to rate girls' mathematics proficiency higher than that of boys with similar mathematics test scores (Robinson & Lubienski, 2011). This research builds upon this prior work by examining the following two research questions:

Study 1: Do teachers still rate the mathematics proficiency of girls higher when boys and girls are equated in terms of demographics, prior achievement, behavior, and teacher-reported approaches to learning?

Study 2: If teachers do have a tendency to rate observationally-similar boys and girls differently, do these differential ratings have an effect on the development of the mathematics gender gap in elementary school?

We used a large-scale, nationally-representative, longitudinal dataset to address these two questions. Study 1 revealed that teachers actually rate *boys* higher than observationally similar girls. In other words, teachers' overrating of girls in our prior study was probably due to girls' tendency to behave in particularly appropriate ways. Study 2 showed the damaging effects of teachers' stereotyped beliefs on girls' math achievement, echoing previous studies on teacher expectancy. However, unlike the experimental manipulations in the expectancy literature, the treatment is not simply being told that a student is “gifted/about-to-bloom” (which may or may not be believable to a teacher familiar with that student) but rather the teacher’s *stated belief* of the student’s math proficiency. Thus, we in fact expect larger effects of teacher expectancy than demonstrated in prior research (see Raudenbush, 1984, for a meta-analysis of earlier teacher-expectancy studies).

Setting:

This research uses the Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), which is nationally representative of the kindergarten class of 1998-99 when the NCES-provided sampling weights are used.

Population / Participants / Subjects:

The ECLS-K dataset contains information on 21,240 kindergarteners (many of whom were followed through 8th grade), collected by the U.S. Department of Education. To account for the complex sampling design of the original data collection and ensure the results are nationally representative, all analyses use the appropriate child sampling weights, following the recommendation of the National Center for Education Statistics (NCES; U.S. Department of Education, NCES, 2001).

We restrict our analyses to students who have valid achievement measures (both direct cognitive assessments and teacher ratings of student proficiency) in the current period (i.e., the period from which the dependent variable comes) and in all prior periods (as these will serve as covariates), as well as valid behavior and learning-approaches data. Finally, since our preferred analyses will account for observable and unobservable teacher characteristics through the use of teacher fixed effects, we restrict our analyses to classrooms where at least one boy and one girl were sampled. The final analytic dataset in the spring of first grade (i.e., the first period of our outcome models) contains 6,658 students.

Intervention / Program / Practice:

Given that we are analyzing a large-scale dataset, there is no typical “intervention” involved. However, in Study 2, the “intervention” can be thought of as the degree to which a teacher rates a student’s content area proficiency higher than would be predicted by prior achievement, behavior, learning approaches, age, race, SES, gender, and fixed characteristics of the teacher (and thus the school). The intervention is plausibly exogenous because we instrument current teacher ratings on prior teachers’ ratings, conditional on the above covariates (see below for more details).

Research Design:

Study 1 uses OLS regression with teacher fixed effects (i.e., a vector of indicator variables for each teacher in the dataset) to predict whether—and to what degree—teachers rate girls’ math proficiency differently than the proficiency of observationally similar boys. In addition to the teacher fixed effects, we also adjust the standard errors to account for the clustering of students within classrooms.

Study 2 uses prior teacher ratings as instrumental variables for the current teacher rating to estimate the effects of teacher ratings on the development of the gender gap. These models will also use teacher fixed effects, condition on a series of demographic, behavioral, and learning-approaches variables, as well as adjust the standard errors for clustering.

In order for a set of instrumental variables to be valid, it must display the following two features: (1) it must predict the variable it is instrumenting (here, current teacher ratings) above and beyond the covariates and fixed effects already included in the model and (2) it cannot be related to the outcomes (that is, current achievement) other than through its direct effect on the variable it is instrumenting (again, conditional on other covariates and fixed effects in the model). In our study, both of these conditions hold, which we discuss further below.

Data Collection and Analysis:

Study 1. Our goal is to investigate if teachers rate girls differently than observationally-similar boys in math and reading. [We focus on the math results in the main text of this abstract, but we present the reading results for Study 1 and 2 in Appendix B.] For a naïve estimate, we first explore this question by regressing current teacher ratings on gender. Then we add covariates to the model to statistically condition out differences between boys and girls in terms of age, race, SES, prior achievement, behavior (i.e., teacher reports of how often the student externalizes problem behavior), and teacher-reported approaches to learning. The “approaches to learning” scale contained the following items: shows eagerness to learn new things, works independently, keeps belongings organized, easily adapts to changes in routine, persists in completing tasks, pays attention well, and (for grades 3 and 5 only) follows classroom rules. After adding these covariates, we then see if teachers rate girls or boys higher, conditional on these factors.

Study 2. After seeing which gender group tends to be rated higher than observationally similar students in the other gender group, our goal in study 2 is to test if this gender-based overrating has an effect on the development of the gender gap in elementary school.

Our theory—reflected below in Equation (1)—posits that achievement (Y) for student i in the current period t is a function of the student’s achievement history (i.e., achievement in all prior n periods: $Y_{it-1}, Y_{it-2}, \dots, Y_{it-n}$), current teacher ratings (TR_{it}), whether the student is female, and a host of other factors collectively termed **X** (this includes the student’s race, age at each assessment, family SES, the student’s current and past behavior, and the student’s current and past approaches to learning new material). Student achievement is likely also affected by the teacher ($\mathbf{T_j}$) a child has and the school ($\mathbf{S_k}$) the students attends. Since our interest here focuses on the gender achievement gap, and since boys and girls should be roughly equally distributed across teachers and schools (at least in elementary school, see Long & Conger, 2011), $\mathbf{T_j}$ and $\mathbf{S_k}$ may not play as integral a role in our conclusions as they would if we were studying something such as the effect of class size; thus, we place $\mathbf{T_j}$ and $\mathbf{S_k}$ in brackets to signify their *general* importance to achievement, but likely trivial role in this study. Other influences on achievement are assumed to vary randomly, indicated by the error term ε_{it}. Of particular interest to the current study are (1) the relationship between achievement and teacher ratings (i.e., $g(TR_i)$) and (2) how the gender gap is affected by teacher ratings (i.e., how the $female$ relationship to conditional Y_{it} changes between models with and without $g(TR)$).

$$Y_{it} = f(Y_{it-1}, Y_{it-2}, \dots, Y_{it-n}) + g(TR_i) + female_i + \mathbf{X_i}[+\mathbf{T_j} + \mathbf{S_k}] + \varepsilon_{it} \qquad (1)$$

Ordinary least squares (OLS) regression. The standard approach to this question might involve a regression such as the following:

$$Y_{it} = (Y_{it-1}, Y_{it-2}, \dots, Y_{it-n})'\boldsymbol{\beta} + \delta(TR_i) + \lambda female_i + \mathbf{X_i} + \varepsilon_{it} \qquad (2)$$

where current achievement (Y_{it}) is predicted by past achievement, current teacher ratings (TR_{it}), gender, and other factors (**X**) described above. The assumption in this model is that the conditional mean of the error term is zero: i.e., $E\left[\varepsilon_{it}\middle|\hat{Y}_{it}\right] = 0$. Consider, however, the very plausible scenario where a teacher rates a student's current proficiency highly because she observed the student making achievement gains. In this case of reverse causality, the estimated coefficient on teacher rating is not a pure effect of the teacher rating, but is at least partially due to the gains affecting the rating.

An instrumental variables (IV) approach. To eliminate the possibility of reverse causality or of systematically underestimating the effect of teacher expectations, we propose an instrumental variables (IV) approach. [IV estimates are common in economics and in quasi-experimental education research for inferring causality from observational data; see Murnane & Willett, 2010, for a discussion.] The instruments of the lagged teacher ratings from all *prior* n periods ($\{TR_{it-1}, TR_{it-2}, \dots TR_{it-n}\}$) are used to predict the current teacher rating (TR_{it}). Since the prior ratings occur temporally prior to the current period gains and current teacher ratings, the concerns of reverse causality and underestimation are eliminated. For example, current teacher ratings can be affected by a host of factors, including the gains the teacher observes in her class. However, the IV approach does not use the *actual* current teacher rating. Instead, this approach uses the a teacher rating score that is *predicted* based on prior teacher ratings, thus ensuring that the estimates of the effect of teacher ratings on student achievement gains are not biased by the current teacher's observation of the student's ability or progress.

Interpreting IV analyses as causal estimates hinges on the validity of the instruments. To be valid, conditional on all other covariates, the instruments (prior teacher ratings) must predict current teacher ratings (which they do) and must not be correlated with the error term in the outcome model (this assumption is supported by Hansen's *J* statistics, which are possible here because the model is overidentified). Thus, there is considerable evidence that our instruments are plausibly exogenous and that the estimates on teacher expectations can be treated as causal effect estimates. In other words, conditional on the covariates in the model, prior teacher ratings are correlated with current teacher ratings, but beyond that correlation, the prior ratings do not predict students' learning while in the current teacher's classroom.

The IV approach can be thought of as a two-stage approach, where the first stage (Equation 3) predicts current teacher ratings on the basis of prior teacher ratings and other variables. In the second stage (Equation 4), we use the predicted (not actual) values of the current teacher ratings to predict current achievement, conditional on all the variables included in stage one (except the prior teacher ratings, which should not affect current outcomes other than through their effect on current ratings, which is already included in the stage-two model).

$$TR_{it} = (Y_{it-1}, \dots, Y_{it-n})'\boldsymbol{\beta} + \delta(TR_{it-1}, \dots, TR_{it-n})'\boldsymbol{\eta} + \mu female_i + \mathbf{X}_\mathbf{i}'\boldsymbol{\theta} + \mathbf{T}_\mathbf{j} + \nu_{it} \quad (3)$$

$$Y_{it} = (Y_{it-1}, Y_{it-2}, \dots, Y_{it-n})'\boldsymbol{\gamma} + \delta\left(\widehat{TR_{\imath}}\right) + \lambda female_i + \mathbf{X}_\mathbf{i}'\boldsymbol{\psi} + \mathbf{T}_\mathbf{j} + \varepsilon_{it} \quad (4)$$

We compared the estimates of λ from Equation 4 with the estimates from an OLS model similar to Equation 4 without the term $\delta\left(\widehat{TR_{\imath}}\right)$ to examine how accounting for teacher expectation effects alters the estimate of the female coefficient.

Findings / Results:

In Study 1, we show that in each period examined, teachers rated girls' math skills lower than those of observationally similar boys. Across the periods, the average amount of underrating was just over 0.1 SDs. [See Figure 1] [It is worth noting that we also explored whether teachers rate observationally-similar racial and ethnic minorities differently; unlike our findings for gender, there was no consistent evidence suggesting differences in teacher ratings of similar students of different races/ethnicities. We also found that females are not overrated in reading (see Figures 3 and 4). Thus, this underrating phenomenon is specific to females in the content area of mathematics.]

In Study 2, Table 1 shows that on average girls lose about 0.137 SDs in comparison to boys between kindergarten and first grade when we do not account for the effects of teacher ratings [see Model 4]. When we account for the effects of teacher ratings [see Model 8], girls lose only 0.080 SDs over the same period—a 42% reduction in girls' losses. Similarly, our models suggest girls' losses between first and third grade would be reduced by 74% if they were not underrated, and their third-to-fifth grade losses would be reduced by 57% (in the non-fixed effects model)[2,3].

Conclusions:

As Robinson and Lubienski (2011) demonstrated, the math gender gap develops early—in the first few years of formal schooling, growing from nonexistent in the fall of kindergarten to a male advantage of about 0.25 standard deviations by third grade. Study 1 demonstrates that teachers rate the math skills of girls lower than those of observationally similar boys. That is, conditioning on math achievement histories, behavior, approaches to learning, race, age, SES, and even looking at boys and girls with the same teachers, girls' skills are rated to be more than one-tenth of a standard deviation lower than boys. This pattern is consistent throughout elementary school. Lamentably, even when conditioning on *current* math achievement, girls are still rated lower (as shown in Figure 1). There is no evidence of similar ratings disadvantage for black or Hispanic students; and there is no evidence that girls are rated higher in reading. Thus, this teacher underrating phenomenon is unique to girls and math performance.

Study 2 demonstrates that girls lose ground in math to boys in every period examined (from the spring of kindergarten through fifth grade), consistent with recent studies (Fryer & Levitt, 2010; Robinson & Lubienski, 2011). However, when we account for the effects of teachers' expectancies, we find that girls lose far less ground. Our analyses tested the instruments used (i.e., we tested if prior teacher ratings were correlated with conditional achievement gains in a way other than through teacher ratings), and we found no evidence to suggest they were invalid. Overall, the results suggest if teachers did not believe that boys had higher math proficiency than similar girls, then girls would lose about 40-75% less ground in math achievement in each period examined. Raising awareness of—and hopefully, reducing—the tendency for teachers to rate males higher in math may thus go a long way to close the gender achievement gap in math.

[2] We prefer the instrumental variables model without teacher fixed effects for this one period because the instrument strength is very low in the fixed effects model, which could lead to biased estimates. For all other periods examined, we prefer the fixed effects models, which have more than sufficient instrument strength (see Stock & Yogo, 2005).

[3] Note that ECLS-K only surveyed teachers at grades K, 1, 3, and 5; thus, grades 2 and 4 are missing teacher ratings of students. Hence, we would likely have stronger predictors of the outcome in Equation 3 if these grade-2 and -4 data were collected, which could have helped improve the measure of instrument strength here and facilitated interpreting the teacher fixed effects models in grade 5.

Appendices

Not included in page count.

Appendix A. References

Dey, J. G. & Hill, C. (2007). *Beyond the pay gap*. American Association of University Women Educational Foundation.

Forgasz, H., & Leder, G. (2001). 'A+ for Girls, B for Boys': Changing perspectives on gender equity and mathematics. In B. Atweh, H. Forgasz, B. Nebres (Eds.), *Sociocultural Research on Mathematics Education: An International Perspective* (pp. 347-366). New Jersey: Lawrence Erlbaum Associates.

Fryer, R. G., & Levitt, S. D. (2010). An empirical analysis of the gender gap in mathematics. *American Economic Journal: Applied Economics, 2*(2), 210-240.

Li, Q. (1999). Teachers' beliefs and gender differences in mathematics: A review. *Educational Research, 41*(1), 63-76.

Long, M. C., & Conger, D. (2011). *Gender sorting across public high schools and its possible effects*. Paper presented at the annual conference of the Association for Education Finance and Policy, Seattle, WA.

Murnane, R. J., & Willett, J. B. (2010). *Methods matter: Improving causal inference in educational and social science research*. New York: Oxford University Press.

Penner, A. M., & Paret, M. (2008). Gender differences in mathematics achievement: Exploring the early grades and the extremes. *Social Science Research, 37*(1), 239–253.

Rathbun, A. H., West, J., & Germino-Hausken, E. (2004). *From kindergarten through third grade: Children's beginning school experiences* (NCES 2004-007). Washington, DC: National Center for Education Statistics.

Raudenbush, S. W. (1984). Magnitude of teacher expectancy effects on pupil IQ as a function of the credibility of expectancy induction: A synthesis of findings from 18 experiments. *Journal of Educational Psychology, 76*(1), 85-97.

Ready, D. D., LoGerfo, L. F., Lee, V. E., & Burkam, D. T. (2005). Explaining girls' advantage in kindergarten literacy learning: Do classroom behaviors make a difference? *Elementary School Journal, 106*(1), 21-38.

Robinson, J. P., & Lubienski, S. T. (2011). The development of gender achievement gaps in mathematics and reading during elementary and middle school: Examining direct cognitive assessments and teacher ratings. *American Educational Research Journal, 48*(2), 268–302.

Stock, J. H., & Yogo, M. (2005). Testing for weak instruments in linear IV regression. In D. W. K. Andrews and J. H. Stock (Eds.), *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg*, pp. 80-108. Cambridge: Cambridge University Press.

Appendix B. Tables and Figures

Figure 1.

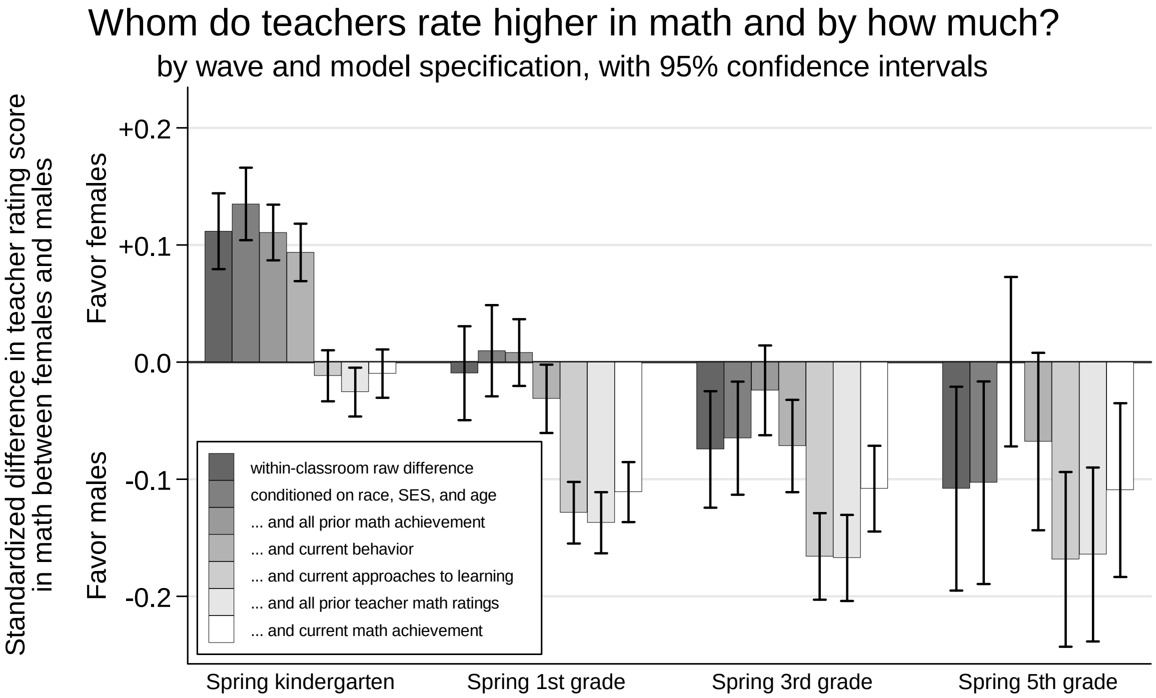

Figure 2.

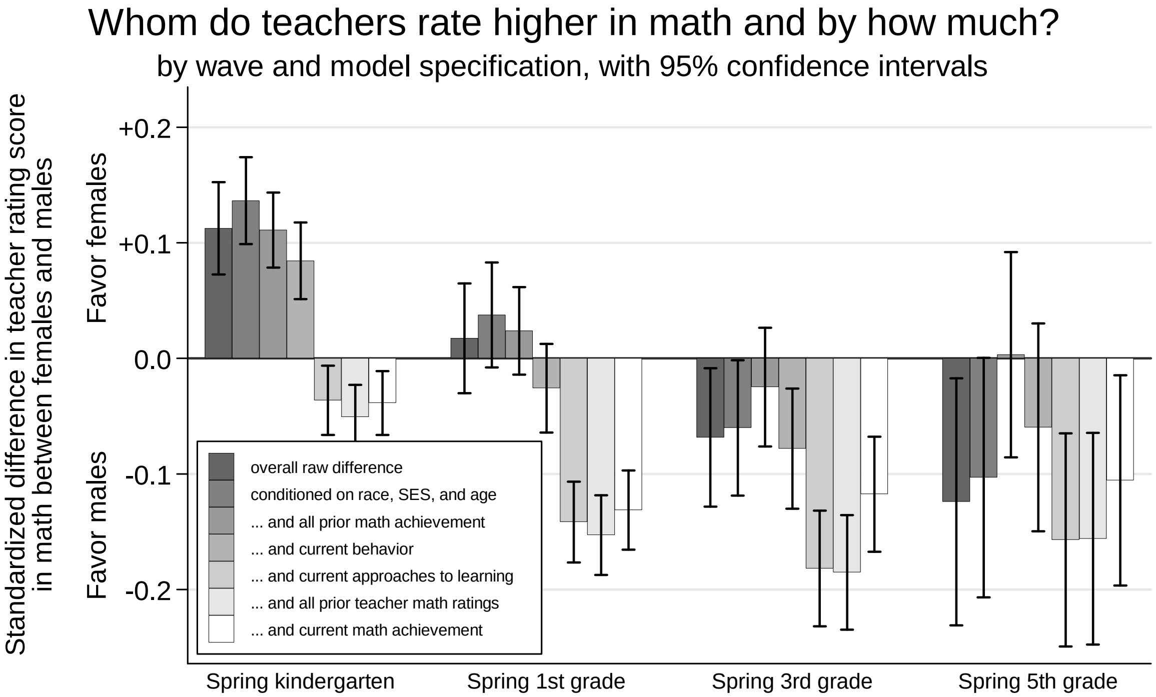

Figure 3.

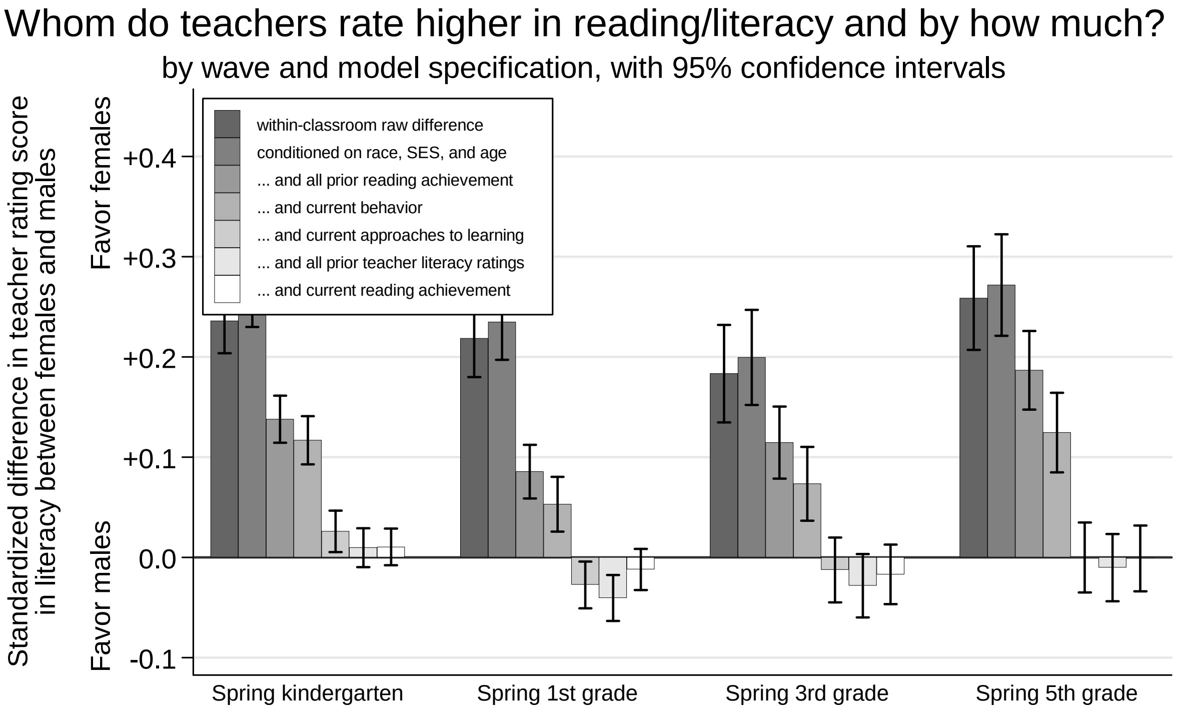

Figure 4.

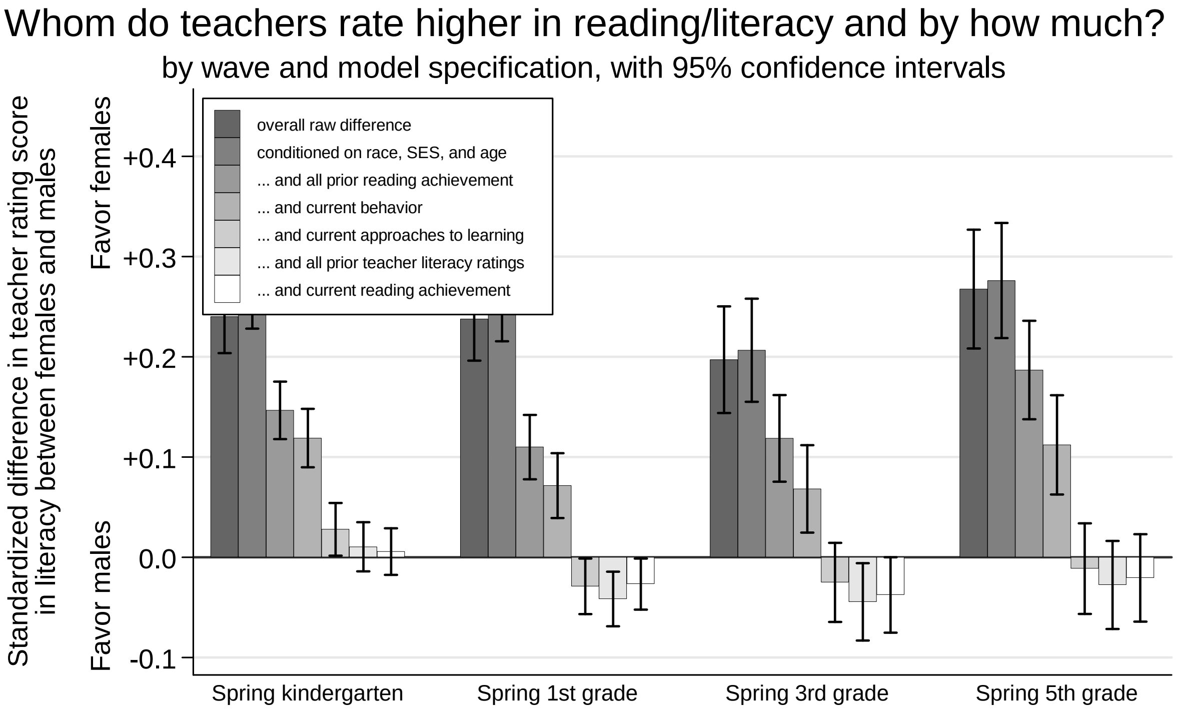

Table 1. Male-female differences in math achievement, conditional on prior-period achievement.

	Raw difference		Covariate-adjusted		Current teacher rating treated as exogenous		Prior teacher ratings used as IVs for current teacher rating	
	OLS	FE	OLS	FE	OLS	FE	IV	IV-FE
Test period	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
Spring 1st grade (N=6,658)	-0.064	-0.062	-0.126	-0.137	-0.108	-0.107	-0.077	-0.080
	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.018)	(0.019)
Spring 3rd grade (N=3,919)	-0.149	-0.157	-0.228	-0.208	-0.201	-0.170	-0.082	-0.054
	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.035)	(0.036)
Spring 5th grade (N=1,099)	-0.080	-0.103	-0.169	-0.188	-0.143	-0.142	-0.072	0.020
	(0.030)	(0.028)	(0.035)	(0.033)	(0.034)	(0.033)	(0.040)	(0.072)
Model includes								
Last test score	X	X	X	X	X	X	X	X
Age at current and prior assessments	X	X	X	X	X	X	X	X
All prior test scores			X	X	X	X	X	X
Prior & current behavior dummies			X	X	X	X	X	X
Prior & current ATL dummies			X	X	X	X	X	X
Race dummies and SES			X	X	X	X	X	X
Teacher fixed effects		X		X		X		X
Current teacher ratings					X	X	IV	IV

Note: We highlighted the models we are comparing in the same color. For example, without teacher fixed effects, comparing observationally similar boys and girls without accounting for teacher rating effects yields the estimates in column 3, which we would then compare to the estimates in column 7 (i.e., compare the columns in yellow). When using teacher fixed effects, we compare the estimates in column 4 to those in column 8 (i.e., compare the columns in orange).

Table 2. Male-female differences in reading achievement, conditional on prior-period achievement.

	Raw difference		Covariate-adjusted		Current teacher rating treated as exogenous		Prior teacher ratings used as IVs for current teacher rating	
	OLS	FE	OLS	FE	OLS	FE	IV	IV-FE
Test period	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
Spring 1st grade (N=8,279)	0.012	0.003	-0.035	-0.054	-0.025	-0.041	-0.021	-0.031
	(0.012)	(0.012)	(0.012)	(0.011)	(0.011)	(0.010)	(0.012)	(0.011)
Spring 3rd grade (N=4,946)	0.036	0.043	-0.023	-0.024	-0.015	-0.019	0.001	-0.012
	(0.018)	(0.018)	(0.018)	(0.018)	(0.017)	(0.017)	(0.019)	(0.018)
Spring 5th grade (N=3,795)	-0.017	-0.006	-0.051	-0.036	-0.047	-0.032	-0.029	-0.025
	(0.017)	(0.018)	(0.018)	(0.019)	(0.018)	(0.019)	(0.020)	(0.019)
Model includes								
Last test score	X	X	X	X	X	X	X	X
Age at current and prior assessments	X	X	X	X	X	X	X	X
All prior test scores			X	X	X	X	X	X
Prior & current behavior dummies			X	X	X	X	X	X
Prior & current ATL dummies			X	X	X	X	X	X
Race dummies and SES			X	X	X	X	X	X
Teacher fixed effects		X		X		X		X
Current teacher ratings					X	X	IV	IV

Note: We highlighted the models we are comparing in the same color. For example, without teacher fixed effects, comparing observationally similar boys and girls without accounting for teacher rating effects yields the estimates in column 3, which we would then compare to the estimates in column 7 (i.e., compare the columns in yellow). When using teacher fixed effects, we compare the estimates in column 4 to those in column 8 (i.e., compare the columns in orange).