Developmental Psychology
2014, Vol. 50, No. 4, 1262–1281

0012-1649/14/$12.00 DOI: 10.1037/a0035073

Teachers' Perceptions of Students' Mathematics Proficiency May Exacerbate Early Gender Gaps in Achievement

Joseph P. Robinson-Cimpian
and Sarah Theule Lubienski
University of Illinois at Urbana-Champaign

Colleen M. Ganley
Florida State University

Yasemin Copur-Gencturk
University of Houston

A recent wave of research suggests that teachers overrate the performance of girls relative to boys and hold more positive attitudes toward girls' mathematics abilities. However, these prior estimates of teachers' supposed female bias are potentially misleading because these estimates (and teachers themselves) confound achievement with teachers' perceptions of behavior and effort. Using data from the Early Childhood Longitudinal Study, Kindergarten Class of 1998–1999 (ECLS-K), Study 1 demonstrates that teachers actually rate boys' mathematics proficiency higher than that of girls when conditioning on both teachers' ratings of behavior and approaches to learning as well as past and current test scores. In other words, on average girls are only perceived to be as mathematically competent as similarly achieving boys when the girls are also seen as working harder, behaving better, and being more eager to learn. Study 2 uses mediation analysis with an instrumental-variables approach, as well as a matching strategy, to explore the extent to which this conditional underrating of girls may explain the widening gender gap in mathematics in early elementary school. We find robust evidence suggesting that underrating girls' mathematics proficiency accounts for a substantial portion of the development of the mathematics achievement gap between similarly performing and behaving boys and girls in the early grades.

Keywords: mathematics achievement gap, gender, teacher perceptions, instrumental variables, propensity score matching

Supplemental materials: http://dx.doi.org/10.1037/a0035073.supp

Gender gaps in mathematics achievement emerge early. Perhaps the strongest evidence for the development of such gaps is found in nationally representative data collected by the U.S. Department of Education (D.O.E.). In particular, studies using the Early Childhood Longitudinal Study, Kindergarten Class of 1998–1999 (ECLS-K) indicate that although the average achievement of boys and girls is similar in kindergarten, a male advantage of about one quarter of a standard deviation emerges by the spring of third grade (Fryer & Levitt, 2010; Husain & Millimet, 2009; Penner & Paret, 2008; Robinson & Lubienski, 2011). Past research has shown that gender gaps in mathematics are at least in part socially constructed and that parents and teachers are socializing agents in the construction of gender (Beilock, Gunderson, Ramirez, & Levine, 2010; Eccles, 1986; Else-Quest, Hyde, & Linn, 2010; Gunderson, Ramirez, Levine, & Beilock, 2012). In the current study, we examine the role that teachers' perceptions may play in the development of gaps in boys' and girls' mathematics proficiency.

Using the ECLS-K data set, we find that elementary school teachers tend to rate the mathematics skills of boys higher than those of girls who perform and behave similarly. This finding might appear to conflict with recent studies that suggest that teachers rate the mathematics abilities and performance of girls at least as favorably as they do those of boys (Fryer & Levitt, 2010; Lavy, 2008; Madon et al., 1998; Robinson & Lubienski, 2011). However, these prior studies do not account for possible behavioral differences between boys and girls and thus may confound teachers' perceptions of students' achievement with their perceptions of students' classroom behavior. We disentangle these issues and show that once teachers' perceptions of behavior are ac-

This article was published Online First December 2, 2013.

Joseph P. Robinson-Cimpian, Department of Educational Psychology, University of Illinois at Urbana-Champaign; Sarah Theule Lubienski, Department of Curriculum and Instruction, University of Illinois at Urbana-Champaign; Colleen M. Ganley, Department of Psychology and Learning Systems Institute, Florida State University; Yasemin Copur-Gencturk, Department of Curriculum and Instruction, University of Houston.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A080147 and Grant R305B100017 to the University of Illinois at Urbana-Champaign. The opinions expressed are those of the authors and do not necessarily represent views of the institute or the U.S. Department of Education. We thank Andrei Cimpian for helpful comments. We appreciate the opportunity to publish this article with commentaries, and we thank Rebecca Bigler, Noel Card, and Andrew Penner for generously providing those commentaries.

Correspondence concerning this article should be addressed to Joseph P. Robinson-Cimpian, Department of Educational Psychology, University of Illinois at Urbana-Champaign, 210F Education, 1310 South 6th Street, Champaign, IL 61820. E-mail: jpr@illinois.edu

counted for, girls' level of mathematics proficiency is actually underrated[1] by teachers—we refer to this as *conditional underrating* because the underrating of girls is revealed when we statistically condition on (or, in some of our analyses, match on) a host of variables, including teachers' perceptions of their behavior.

Then, we extend this work by examining the extent to which this conditional underrating mediates the relation between gender and growth in math performance. Our results suggest that a substantial portion of the growth in the gender gap across the early grades may be explained by teachers' relative underrating of girls' skills. These results add to the mounting evidence that gender gaps are, at least in part, socially constructed.

Why Focus on Gender Gaps in Math Achievement?

Some research has suggested that mathematics gender gaps are small or nonexistent (Hyde, Lindberg, Linn, Ellis, & Williams, 2008; Lindberg, Hyde, Petersen, & Linn, 2010). However, these studies also note that gender gaps may emerge at higher grade levels (Lindberg et al., 2010), among highly selective samples (Lindberg et al., 2010), or when more difficult items are included in tests (Hyde et al., 2008). When comparing these findings with those of studies using ECLS-K and National Assessment of Educational Progress data (which revealed gender gaps in elementary school), the conflicting evidence on gender achievement disparities may seem puzzling. Perhaps one reason that Hyde and colleagues (2008) found minimal gender differences is that their analyses involved state assessments, which are designed to determine whether or not students have attained state standards but are not necessarily intended to home in on the precise achievement students are capable of—thus, these state assessments may potentially suppress the gender gap. As Lindberg et al. (2010) noted, few elementary assessments contain challenging items that distinguish among students at the top end of the scale, where gender gaps are most prevalent (McGraw, Lubienski, & Strutchens, 2006; Robinson & Lubienski, 2011). In contrast, the ECLS-K math assessment uses item response theory (IRT) and an adaptive-stage design to precisely identify the achievement of students. This new evidence suggests that gender gaps do exist—at least on the content assessed by ECLS-K—and they develop during early elementary school (Fryer & Levitt, 2010; Husain & Millimet, 2009; Penner & Paret, 2008; Robinson & Lubienski, 2011). The question then turns to: Why does this gap develop?

Unlike gaps based on race and socioeconomic status (SES), which are in part attributable to differences in schools attended (Fryer & Levitt, 2004), gender gaps in elementary school are unlikely to be due to boys and girls attending different schools (Long & Conger, 2013), having different teachers within those schools, or demographic differences between boys and girls. Hence, one might suspect that gender gaps would not develop as much as race- or SES-based gaps. However, as noted above, the ECLS-K data indicate that from kindergarten to third grade, mathematics achievement gaps grow by about one quarter of a standard deviation between boys and girls, about the same amount of growth as the Black–White gap (Reardon & Robinson, 2008; see also Fryer & Levitt, 2010). Over the same period of time, the mathematics achievement gap remains relatively unchanged between students whose parents' highest degree is a high school diploma and those whose parents have college degrees (an SES proxy); the Hispanic–White gap actually reduces in size (Reardon & Robinson, 2008). Therefore, one of the greatest growths in inequity during elementary school appears to be between boys and girls in the content area of mathematics.

The gender gap in mathematics is unique for another reason: It does not exist when students enter kindergarten. Race-based and SES-based gaps are prominent in the fall of kindergarten (Reardon & Robinson, 2008). Similarly, the gender gap in reading is present in the fall of kindergarten (favoring girls), but it narrows somewhat during elementary school (Robinson & Lubienski, 2011). Taken together, the literature suggests that something special about the experiences of boys and girls during the elementary school years contributes to the mathematics gap that emerges at that time.

Teacher Ratings of Girls and Boys

In examining potential origins of mathematics gender gaps, previous researchers have studied elementary school teachers' ratings of girls' and boys' mathematics abilities. The evidence, however, is mixed, with some studies concluding that teachers hold more positive views of boys' mathematics abilities and others suggesting an advantage for girls.

Earlier research on teachers' interactions with students identified ways in which *boys* appeared to be advantaged by teachers. Teachers tended to hold higher expectations for their male students, as was illustrated by the provision of more specific, positive feedback (Dweck, Davidson, Nelson, & Enna, 1978; Sadker & Sadker, 1986). Other studies suggested that teachers may believe that boys in general have a gift for math. In a study of 38 first-grade teachers, Fennema, Peterson, Carpenter, and Lubinski (1990) found that the teachers more often named boys as the best math students and attributed boys' success to ability and girls' success to effort. Similarly, Tiedemann (2000) found that elementary school teachers attributed students' failures to a lack of ability more for girls than for boys and attributed failure to a lack of effort more for boys than for girls. Moreover, teachers in Tiedemann's study thought that additional effort would benefit the boys more than the girls, who were already perceived by their teachers to be exerting more effort in order to reach the same achievement level as the boys.

In contrast, several recent national studies using ECLS-K data indicate that teachers rate the mathematics performance of *girls* more favorably than that of boys, even when boys outperform girls on a direct cognitive assessment of mathematics (e.g., Fryer & Levitt, 2010; Robinson & Lubienski, 2011). These results are consistent with a recent study of high school teachers in Israel (Lavy, 2008), as well as Madon and colleagues' (1998) study of 56 middle-school teachers in Michigan. Specifically, Madon and colleagues found that teachers rated their seventh-grade girls' performance and effort as higher than that of boys, but tended to rate their abilities equally. Similarly, there is also much research showing that girls earn higher grades (another form of teacher ratings) than boys in most school subjects, including mathematics and science, across all years of schooling (American Association of

[1] We acknowledge that the pattern of teacher ratings can be interpreted in multiple ways and can only be viewed as an "underrating" of girls when adjusting for girls' behavior and approaches to learning. This issue is discussed in more detail later in the article.

University Women Education Foundation, 2008; Catsambis, 1994; Pomerantz, Altermatt, & Saxon, 2002; Willingham & Cole, 1997). Taken together, this more recent research suggests that teachers tend to rate girls' mathematics performance higher than that of boys (but see Tiedemann, 2000).

"Good Girl" Behavior and Teacher Perceptions

One potential explanation for *why* girls might be perceived as more proficient in mathematics could lie in differences between teachers' perceptions of boys' and girls' behavior and effort. Prior research has revealed that there are substantial gender differences in approaches to learning and classroom behavior. Girls tend to exhibit more on-task behavior and positive approaches to learning in school (Forgasz & Leder, 2001; Ready, LoGerfo, Burkam, & Lee, 2005) and fewer "problem behaviors," such as fighting (Rathbun, West, & Germino-Hausken, 2004). In other words, girls tend to be "good girls," more often staying focused on classroom tasks and engaging in behaviors that please the teacher. Evidence for the apparent overrating of girls combined with evidence of gender differences in classroom behaviors raises the question of whether teachers' perceptions of gender differences in achievement may be influenced by gender differences in behavior.

In fact, this question is at the heart of the study conducted by Kenney-Benson, Pomerantz, Ryan, and Patrick (2006), who examined the relation between grades and gender differences in behavior. Their results showed that girls received higher mathematics grades, and this gender disparity could be largely explained by differences in girls' and boys' reported effort and behavior. More specifically, using a sample of 518 students in Illinois, they found that both fifth- and seventh-grade girls reported fewer disruptive behaviors and used more effective learning strategies than did boys. In addition, girls more often held mastery goals (focusing on learning), whereas boys more often held performance goals (obtaining recognition). After adjusting for disruptive classroom behaviors, learning strategies, and achievement goals, the difference between girls' and boys' grades (a measure of teachers' perceptions of competence) was reduced to marginal significance. It is unclear, however, whether teachers' assessments of students were influenced by students' behaviors or whether boys' and girls' differential behaviors led to actual differences in competence. The authors did find that although students' disruptive behaviors were correlated with both grades and test scores, the correlation with grades was stronger. They speculated that the classroom setting might reward positive behavior, whereas achievement tests may not, thus explaining why girls do better in class but not on tests. It is possible that positive learning behaviors do increase girls' mathematics competence but that teachers also factor these behaviors into their grading. The extent to which teachers intentionally reward student behavior when grading, versus conflating behavior with mathematical proficiency, is an open question.

If teachers conflate positive behavior and approaches to learning with mathematics proficiency, then findings from prior ECLS-K studies indicating that teachers overrate girls (Fryer & Levitt, 2010; Robinson & Lubienski, 2011) may be quite different when student behavior and approaches to learning are considered. In Study 1 in the current article, we addressed this question by examining how teachers view the mathematics proficiency of boys and girls who both behave and achieve similarly. However, concerns about gender biases in teachers' perceptions are arguably moot if nothing is affected by those perceptions—an issue examined in Study 2 of this article.

Effects of Teacher Perceptions

Beginning with Rosenthal and Jacobson's (1966, 1968) classic "Pygmalion study," in which teachers were told that some of their (randomly selected) students were on the verge of great intellectual growth, the idea of "self-fulfilling prophecy" has become widely accepted. In these studies, teachers' perceptions about students (based on random information about the student) were related to students' future performance, indicating that teachers' expectations have an effect on their students' achievement. Hundreds of replication studies have been conducted in several domains, and Rosenthal's (1994) meta-analysis of 464 studies revealed an average effect size of $r = .30$. Although some reviews revealed smaller effects in general ($r = .10–.20$; Jussim & Harber, 2005), and when using observational/naturalistic data in particular (Jussim, Eccles, & Madon, 1996), larger effects have been found for children in first and second grades (Raudenbush, 1984). Similar to some prior studies, we used observational data to explore the relation between teacher perceptions and student achievement. However, the focus of our study is not on the effects of teacher perceptions on achievement per se, but rather our focus is on how teacher perceptions may *mediate* the relation between gender and math achievement gains. The difficulty with studying mediation using observational data lies in ensuring that the estimate of the mediator is unbiased (Bullock, Green, & Ha, 2008, 2010; Bullock & Ha, 2011; Imai, Keele, Tingley, & Yamamoto, 2011; Judd & Kenny, 1981; MacKinnon, 2008; Sobel, 2009), a key principle that is often overlooked in mediation analyses (Bullock et al., 2010; Imai et al., 2011). In Study 2, we first used traditional (but potentially biased) approaches to mediation, and then we estimated mediation models that are less prone to bias.

The Current Research

In this article, we present two studies, each using data from the ECLS-K. In the first study, we examined whether teachers on average rate boys' or girls' mathematics proficiency higher after accounting for students' problem behavior, approaches to learning, past and current test scores, and demographic factors. That is, in Study 1 we asked, "Above and beyond other factors, does a student's gender predict the teacher's rating of that student's mathematics proficiency?" If teachers are *conditionally underrating* boys or girls—that is, rating boys' or girls' math proficiency higher, after statistically accounting for prior achievement and prior and current behaviors—we then need to understand the potential implications of teachers' over- and underestimation of students' abilities (Jussim & Harber, 2005). To address this, in the current study we also examined whether teachers' tendency to rate boys or girls higher may be linked to the widening gender gap in mathematics performance in elementary school. That is, in Study 2 we asked, "To what extent do teachers' differential ratings of mathematics proficiency mediate the relation between gender and gains in mathematics achievement?"

Study 1

Method

Data set. The restricted-use ECLS-K data set provides a unique opportunity to examine the extent to which teachers differentially rate boys and girls and the role of these different ratings in the achievement gap. ECLS-K is a nationally representative data set containing information on 21,240 kindergarteners (many of whom were followed through eighth grade), collected by the U.S. D.O.E. (http://nces.ed.gov/ecls/kindergarten.asp). Data were collected in the fall of kindergarten, then again in the spring of kindergarten and Grades 1, 3, and 5. In each wave of data collection, U.S. D.O.E.-trained employees administered a psychometrically valid mathematics test,[2] and each student's teacher was asked to rate the student's mathematics proficiency. Importantly, teachers and students were never informed of the student's score on the mathematics test.

Variables. The following variables were collected at each wave of data collection.

Mathematics direct cognitive assessment score. At each wave, children completed a mathematics assessment, which was developed by the Educational Testing Service based on developmentally appropriate items from widely used assessments, including the Test of Early Mathematics Ability–Third Edition (Ginsburg & Baroody, 2003), the Woodcock–Johnson III (Woodcock, Mather, & McGrew, 2001), and the Peabody Individual Achievement Test–Revised (Markwardt, 1989). Scores on this assessment were based on IRT. Students only completed some of the test items, but based on their performance on these items, the National Center for Education Statistics (NCES) assigned scores that reflected the number of questions the child would have answered correctly had they completed all of the items on the assessment; these scores are referred to as IRT scale scores. The IRT scale scores were then converted to T scores at each wave. These T scores have a mean of 50 and a standard deviation of 10; we then converted these to z scores with a mean of zero and standard deviation of one. We chose to use T scores (converted to z scores) because these are appropriate when conducting longitudinal subgroup analyses (Tourangeau et al., 2001) and lead to more interpretable regression coefficients.

Teacher academic rating score of mathematics proficiency. At each wave, teachers rated the degree to which a child had acquired and demonstrated particular mathematical skills. They rated students on a scale from 1, indicating that the "child has not yet demonstrated skill, knowledge, or behavior," to 5, showing that the "child demonstrates skill, knowledge, or behavior competently and consistently." Teachers rated students across multiple mathematics domains, including number, measurement, geometry, and statistics. For example, in kindergarten, teachers were asked to evaluate how proficiently the student "Orders a group of objects," "Solves problems involving numbers using concrete objects," "Uses a variety of strategies to solve mathematics problems," and demonstrates four other specific skills. The full set of items in each wave of data collection is contained in the supplemental material (see Tables S1–S4). If the particular skill had not yet been taught in class, teachers were asked to select "not applicable." NCES performed Rasch analyses on the teacher rating scale in an effort to (a) create a measure for modeling growth in the teacher ratings, (b) make the ratings more comparable to the direct cognitive assessment, and (c) estimate values for students whose teachers did not complete some items because those skills had not been taught yet. We standardized these teacher ratings within waves to have a mean of zero and standard deviation of one.

Gender. Student gender data were gathered from parent interviews, school records, and field observations. Note that ECLS-K uses the term *gender* instead of *sex*, and we maintain this language.

Race-ethnicity. The responding parent/guardian was asked to indicate the child's race and was also asked whether the child was Hispanic. ECLS-K then created eight mutually exclusive race-ethnicity categories: White, non-Hispanic; Black or African American, non-Hispanic; Hispanic, race specified; Hispanic, no race specified; Asian; Native Hawaiian or other Pacific Islander; American Indian or Alaskan Native; and more than one race specified, non-Hispanic.

SES. This family/household variable is a continuously valued composite of father/male guardian's education; mother/female guardian's education; father/male guardian's occupational prestige; mother/female guardian's occupational prestige; and household income. ECLS-K reports that the composite components were highly correlated, and thus combining them into a single SES composite is a reasonable strategy. The SES composite has a mean of zero and standard deviation of one.

Age at assessment. This variable is the difference between the date when the student was assessed by the ECLS-K test administrator and the child's date of birth. Date of birth was gathered from parent interviews and from school records and is only available in the restricted-use data set.

Externalizing problem behavior. Teachers completed the Externalizing Problem Behavior scale (Tourangeau et al., 2001) for each participating student. Students were rated on a 4-point scale (1 = *never*, 2 = *sometimes*, 3 = *often*, 4 = *very often*) for the following items: argues, fights, gets angry, acts impulsively, disturbs ongoing activities, and (for Grades 3 and 5 only) talks during quiet study time. ECLS-K combined responses on these items into a single scale. Split-half reliabilities for the Externalizing Problem Behavior scale were between .86 and .90 across kindergarten through Grade 5.

Approaches to learning. Each teacher completed the "approaches to learning" scale for each of her or his students. The 4-point scale (1 = *never*, 2 = *sometimes*, 3 = *often*, 4 = *very often*) contains 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. ECLS-K combined these items into a single scale. Split-half reliabilities for the approaches to learning scale ranged between .89 and .91 from kindergarten to Grade 5.

Participants. We restrict our analyses to students who have valid test scores and teacher ratings in all relevant time periods, as well as valid behavior and learning-approaches data. For any given wave, individuals needed valid data for the current wave and all previous waves. Finally, because our preferred analyses account for observable and unobservable teacher characteristics through the use of

[2] For the most recent ECLS-K psychometric report, see http://nces.ed.gov/pubs2009/2009002.pdf

Table 1
Means and Descriptive Statistics for Key Variables

Variable	All students		Boys			Girls		
	M	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*
Spring kindergarten								
Math test score	0.00	1.00	4,653	0.00	1.04	4,571	−0.01	0.95
Teacher rating of math	0.00	1.00	4,653	−0.06	1.02	4,571	0.06	0.98
Externalizing behaviors	1.66	0.64	4,653	1.80	0.68	4,571	1.52	0.56
Approaches to learning	3.12	0.69	4,653	2.97	0.70	4,571	3.27	0.64
Spring first grade								
Math test score	0.00	1.00	3,335	0.05	1.04	3,323	−0.05	0.96
Teacher rating of math	0.00	1.00	3,335	−0.01	1.03	3,323	0.01	0.97
Externalizing behaviors	1.65	0.64	3,335	1.79	0.69	3,323	1.51	0.56
Approaches to learning	3.04	0.71	3,335	2.89	0.71	3,323	3.20	0.67
Spring third grade								
Math test score	0.00	1.00	1,909	0.14	1.01	2,010	−0.13	0.97
Teacher rating of math	0.00	1.00	1,909	0.04	1.01	2,010	−0.04	0.99
Externalizing behaviors	1.68	0.60	1,909	1.81	0.63	2,010	1.56	0.54
Approaches to learning	3.08	0.66	1,909	2.92	0.67	2,010	3.23	0.62
Spring fifth grade								
Math test score	0.00	1.00	534	0.21	0.97	565	−0.20	0.99
Teacher rating of math	0.00	1.00	534	0.09	1.04	565	−0.08	0.96
Externalizing behaviors	1.60	0.52	534	1.71	0.57	565	1.49	0.44
Approaches to learning	3.12	0.63	534	2.96	0.64	565	3.27	0.58

teacher fixed effects (i.e., indicator variables for each teacher), we restrict our analyses to classrooms in which at least one boy and one girl were sampled. The final analytic data set in the spring of kindergarten contains 9,223 students and 1,575 teachers; the sample in the spring of first grade contains 6,658 students and 1,439 teachers; the sample in the spring of third grade includes 3,919 students and 923 teachers; and the sample in the spring of fifth grade includes 1,099 students and 329 teachers.[3] There was no evidence of gender-based differential attrition between grades.

All analyses use the NCES-provided sampling weights to ensure the results are nationally representative. Further, all analyses adjust the standard errors for clustering of students within classrooms and account for heteroskedasticity using cluster and robust options in Stata.

Analytic approach. In this study, we investigate whether teachers rate girls differently than boys in mathematics, after conditioning on a set of covariates. For a naïve estimate, we first explore this question by regressing current teacher ratings on gender. This is achieved through a basic weighted least squares regression. Then we sequentially add covariates to the model to statistically account for differences between boys and girls in terms of age, race, SES, prior achievement, prior and current externalizing problem behavior, prior and current approaches to learning, and current achievement. After adding these covariates, we then see whether teachers rate girls or boys higher, conditional on these factors. The standard errors for these and all subsequent analyses have been appropriately adjusted to account for clustering of students within classrooms.

Results and Discussion

Table 1 displays descriptive statistics for the key variables by grade level and gender. As can be seen, there is no gender difference in math test performance in the spring of kindergarten, but a small gender gap appears in first grade, and this gap grows larger by third and fifth grades. Correlations between all key variables are displayed in Table 2.

When looking at raw differences (the first vertical bar in each period in Figure 1), teachers appear to rate the mathematics skills of girls higher than those of boys in the spring of kindergarten, but not in later periods; in fact, teachers rate boys' mathematics skills higher in third and fifth grades, which mirrors patterns in students' test scores. After we condition on demographic factors and prior mathematics achievement (the third bar in each period), teachers no longer rate boys higher in Grades 3 and 5.[4] These results appear to be gender neutral, as teachers rate boys and girls with equal prior math achievement similarly in Grades 1–5. Still, once we also account for behavior, approaches to learning, prior teachers' ratings of mathematics proficiency, and current mathematics achievement, teachers rate girls' mathematics proficiencies lower than those of boys in each grade. Across the periods, the average amount of this underrating was about 0.1 standard deviations. In other words, teachers rate boys higher than girls after accounting for achievement and their perceptions of students' classroom behaviors.

How to interpret this "conditional underrating" of girls is a bit of a conundrum: When we condition on prior achievement,

[3] The main focus of the article was on mathematics achievement; however, for comparison purposes, we estimate similar models for reading achievement. The sample sizes for reading achievement are as follows: in the spring of kindergarten, 10,885 students and 1,810 teachers; in the spring of first grade, 8,279 students and 1,709 teachers; in the spring of third grade, 4,946 students and 1,122 teachers; and in the spring of fifth grade, 3,795 students and 841 teachers.

[4] When we also condition on current achievement at this step, we get the same results. However, we chose to not include current achievement at this point because it is impossible to tease apart the causal ordering of current achievement and current teacher ratings, and thus it is best to not include this variable until all other variables have been taken into account.

Table 2
Correlations Among Key Variables

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1. Female	—																			
2. Fall K Math	−.01	—																		
3. Spr K Math	−.01	.80	—																	
4. Gr. 1 Math	−.10	.69	.73	—																
5. Gr. 3 Math	−.17	.66	.70	.74	—															
6. Gr. 5 Math	−.21	.64	.67	.71	.85	—														
7. Fall K Tchr Rtng	.01	.48	.42	.37	.34	.33	—													
8. Spr K Tchr Rtng	.09	.51	.53	.43	.42	.40	.60	—												
9. Gr. 1 Tchr Rtng	.00	.49	.53	.49	.51	.49	.34	.48	—											
10. Gr. 3 Tchr Rtng	−.07	.46	.48	.48	.54	.54	.27	.35	.38	—										
11. Gr. 5 Tchr Rtng	−.08	.47	.47	.47	.55	.60	.25	.36	.44	.44	—									
12. Fall K Ext.	−.17	−.19	−.19	−.11	−.11	−.13	−.16	−.22	−.12	−.20	−.16	—								
13. Spr K Ext.	−.22	−.10	−.11	−.06	−.08	−.11	−.09	−.17	−.07	−.13	−.14	.74	—							
14. Gr. 1 Ext.	−.21	−.12	−.15	−.11	−.16	−.15	−.05	−.13	−.21	−.12	−.18	.46	.53	—						
15. Gr. 3 Ext.	−.18	−.07	−.04	−.11	−.11	−.10	−.06	−.10	−.11	−.13	−.16	.42	.49	.52	—					
16. Gr. 5 Ext.	−.21	−.08	−.07	−.07	−.06	−.10	−.05	−.10	−.08	−.10	−.17	.41	.44	.46	.49	—				
17. Fall K AtL	.22	.44	.40	.31	.28	.30	.49	.51	.32	.30	.27	−.54	−.39	−.25	−.23	−.22	—			
18. Spr K AtL	.25	.41	.43	.33	.35	.35	.40	.61	.38	.33	.34	−.48	−.52	−.32	−.30	−.24	.72	—		
19. Gr. 1 AtL	.19	.30	.32	.32	.29	.29	.22	.34	.53	.29	.33	−.31	−.32	−.51	−.36	−.32	.38	.45	—	
20. Gr. 3 AtL	.20	.31	.31	.31	.36	.34	.17	.31	.33	.46	.36	−.38	−.38	−.38	−.54	−.36	.37	.46	.53	—
21. Gr. 5 AtL	.25	.24	.24	.25	.28	.31	.18	.29	.33	.30	.43	−.35	−.32	−.39	−.42	−.55	.34	.38	.45	.53

Note. $N = 1{,}099$; K = Kindergarten; Spr = Spring; Gr. = Grade; Tchr Rtng = Teacher Rating of Math; Ext. = Externalizing problem behaviors; AtL = Approaches to Learning.
$r \geq .06$, $p < .05$; $r \geq .08$, $p < .01$.

gender is *not* a significant predictor of teacher ratings. One might assume this outcome is equitable because we see no gender gap in teacher ratings when boys and girls are equated in terms of prior achievement. Extending this logic, if teacher ratings of math proficiency are truly not gender-biased, there should be no difference between boys' and girls' ratings when boys and girls are equated in terms of even *more* variables (e.g., perceptions of behavior and approaches to learning). That is, the more similar boys and girls are in terms of various factors (e.g., prior performance, behavior, approaches to learning, race, SES, age), the *smaller* the absolute differences in teacher perceptions of boys and girls should become. However, contrary to this logic, this is precisely the point when gender differences *emerge*. This suggests that teachers' ratings of math proficiency are influenced by student gender per se, above and beyond factors such as achievement and behavior. By the standard for equity just described, finding this *lack* of gender neutrality suggests gender inequity in teachers' conditional ratings.

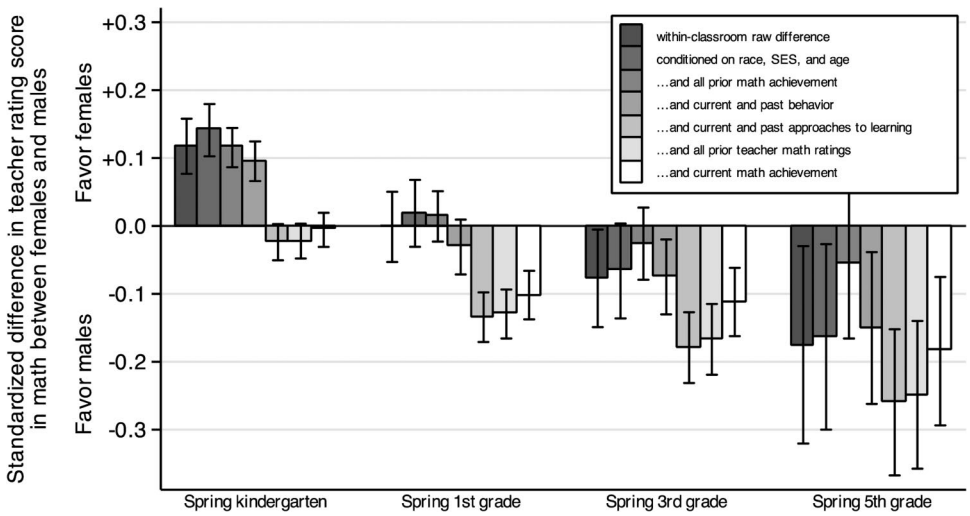


Figure 1. Gender differences in teacher ratings of mathematics proficiency, by wave and model specification. 95% confidence intervals are also presented; if the 95% confidence interval does not cross the 0 line, then the difference is statistically significant at the 5% level ($p < .05$). *Within classroom* refers to these models having teacher fixed effects. SES = socioeconomic status.

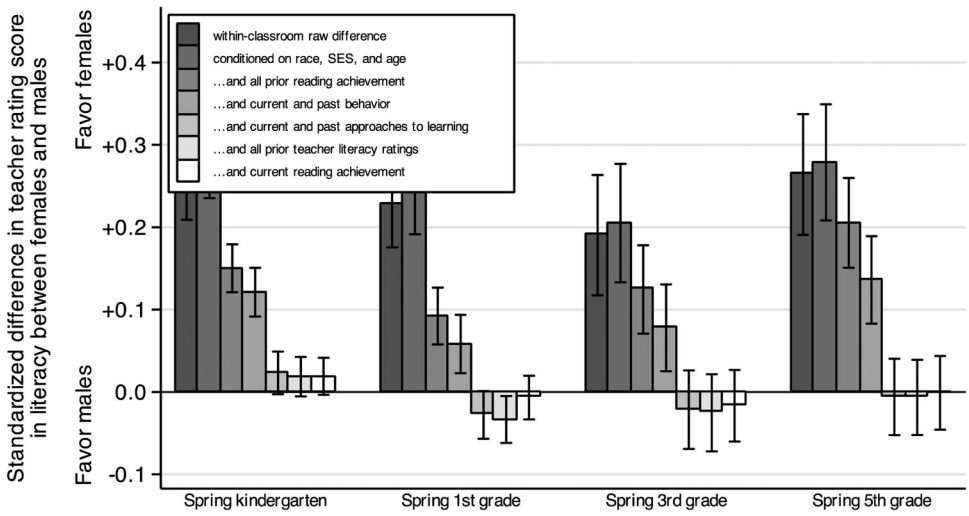


Figure 2. Gender differences in teacher ratings of literacy/reading proficiency, by wave and model specification. 95% confidence intervals are also presented; if the 95% confidence interval does not cross the 0 line, then the difference is statistically significant at the 5% level ($p < .05$). *Within classroom* refers to these models having teacher fixed effects. SES = socioeconomic status.

One explanation for the pattern could be that teachers have an inflated sense of girls' behavior and therefore teachers perceive girls as working harder than they really are. Alternatively, the opposite might be true—that is, teachers may hold girls to a higher standard and thus rate their behavior at least as harshly as boys. Regardless, the results indicate that teachers do tend to conflate their ratings of students' behavior with their ratings of students' mathematics proficiency, and on average, teachers must perceive girls as working harder than similarly achieving boys in order to rate them equally. As is shown in Study 2, teachers' tendency to "conditionally underrate" girls appears to be detrimental to girls' mathematics learning.

It is worth noting that we also explored whether teachers rate racial and ethnic minorities differently when they perform and behave similarly; unlike our findings for gender, there was no consistent evidence suggesting differences in teacher ratings of similar students of different races-ethnicities.[5] We also found that when rating similar boys and girls in reading/literacy, teachers do *not* exhibit gender-biased rating differentials (see Figure 2). Thus, this conditional underrating appears to be specific to girls in the content area of mathematics, suggesting that student gender influences teachers' ratings of the mathematics performance of otherwise similarly performing and behaving boys and girls.

It is also important to point out that over 98% of kindergarten teachers in the data set are women. The corresponding percentages for Grades 1, 3, and 5 are 98%, 94%, and 82%, respectively. Interestingly, when exploring whether this underrating phenomenon varies by teacher gender, we found that, if anything, *female* teachers are more likely than are male teachers to underrate the mathematics performance of girls in Grades 3 ($p = .021$) and 5 ($p = .138$).[6] Although the number of male teachers is large enough to detect some evidence of teacher-gender differences in rating behavior in Study 1, the instrumental-variables approach of Study 2 would require a much larger sample of male teachers to estimate precise mediation estimates by teacher gender. Thus, we proceed with estimating the average mediation estimate of teacher perceptions on the gender gap growth rather than estimating mediation by teacher gender. Yet, because female teachers constitute the vast majority of teachers, the results reported in this article primarily reflect the tendency of *female* teachers to conditionally underrate girls' mathematics proficiency.

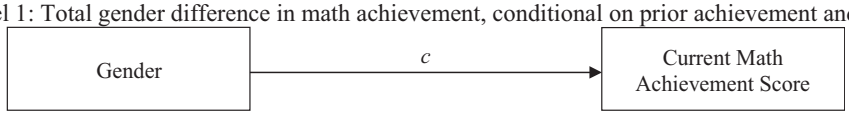
Study 2

After demonstrating in Study 1 that teachers tend to rate the mathematics proficiency of girls lower than that of boys after accounting for differences in achievement and their perceptions of students' behavior, we now turn to the question of whether this tendency to conditionally underrate girls mediates the relation between gender and gains in mathematics achievement. There are several analytic difficulties to be overcome when examining this issue, and the various methods used to address these difficulties have their trade-offs. Here, we present a series of models, ranging

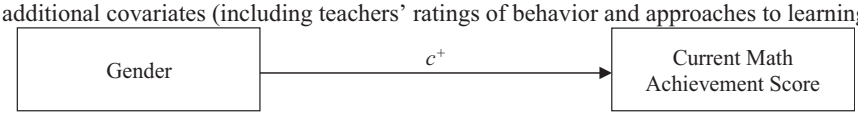
[5] One may also be interested in the tendency for teachers to rate boys and girls differently across race groups (i.e., Gender × Race interactions) and SES levels (i.e., Gender × SES interactions), and the three-way interactions of gender, race, and SES. As a series of supplemental analyses, we explored these interactions, finding that about 5% of them were statistically significant (a percentage that could occur by chance). Across the various waves of analysis, there were no consistent patterns among that 5% (e.g., an interaction that was significant in first grade was not significant in any other grades). Although no consistent evidence of interaction exists in our study, researchers may want to pursue interaction effects with larger samples of minority students.

[6] This conclusion was based on differences between male and female teachers in their tendency to rate girls lower than similar boys. Specifically, these p values are based on the heteroskedastic-robust value cross-level interactions between teacher gender and student gender in hierarchical linear model regressions of teacher rating on teacher gender, student gender, their interaction, race, age, prior and current behavior and learning approaches, and prior teacher ratings and mathematics test scores.

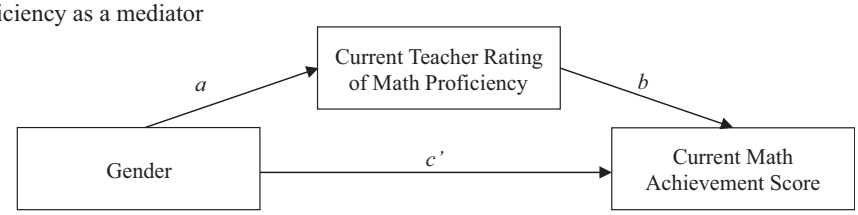
Model 1: Total gender difference in math achievement, conditional on prior achievement and age



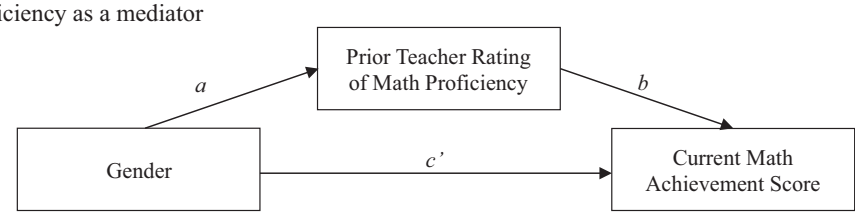
Model 2: Total gender difference in math achievement, conditional on prior achievement and age plus additional covariates (including teachers' ratings of behavior and approaches to learning)



Model 3: Mediation model (building off of Model 2) with current teacher rating of math proficiency as a mediator



Model 4: Mediation model (building off of Model 2) with prior teacher rating of math proficiency as a mediator



Model 5: Mediation model (building off of Model 2) with instrumental variables predicting the mediator, predicted current teacher rating of math proficiency

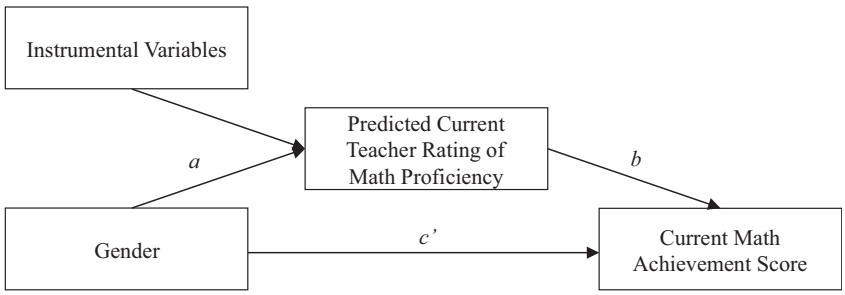
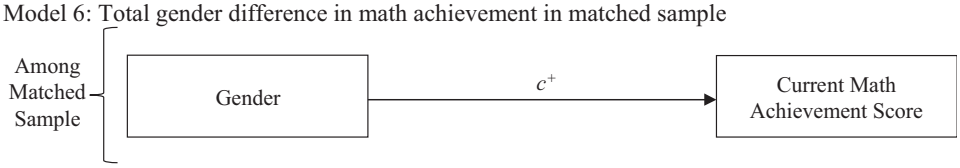


Figure 3. Models of gender differences with and without mediation. For a list of additional covariates included in the models, see Tables 4 and 5. Model 1: Total gender difference in math achievement, conditional on prior achievement and age. Model 2: Total gender difference in math achievement, conditional on prior achievement and age plus additional covariates (including teachers' ratings of behavior and approaches to learning). Model 3: Mediation model (building off of Model 2) with current teacher rating of math proficiency as a mediator. Model 4: Mediation model (building off of Model 2) with prior teacher rating of math proficiency as a mediator. Model 5: Mediation model (building off of Model 2) with instrumental variables predicting the mediator, predicted current teacher rating of math proficiency. Model 6: Total gender difference in math achievement in matched sample. Model 7: Mediation model with instrumental variables predicting the mediator, current teacher ratings of math proficiency in matched sample.

Figure 3 continues

Model 6: Total gender difference in math achievement in matched sample



Model 7: Mediation model with instrumental variables predicting the mediator, current teacher ratings of math proficiency in matched sample

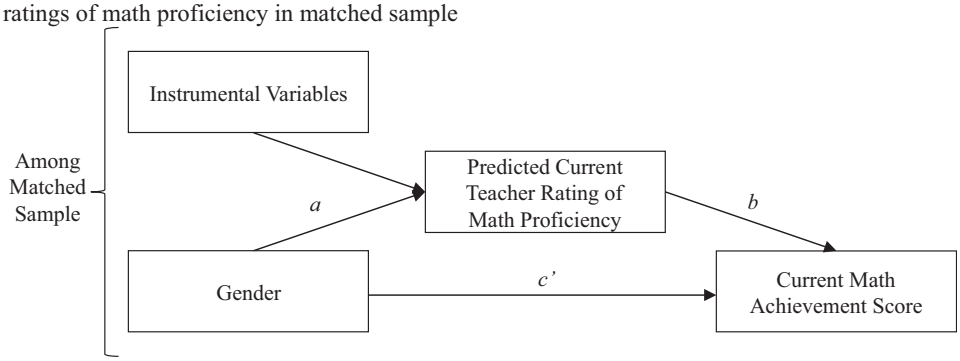


Figure 3 (*continued*).

from the more familiar and straightforward to those that are more novel and complex. We ultimately show that all such models point toward the same conclusion—that teacher perceptions do contribute to gender disparities in mathematics achievement.

Method

Just as in Study 1, we approach this question using the ECLS-K data set. We begin by using regression models to estimate current achievement as a function of gender, age, and prior achievement—these models provide us with an estimate of the gender gap growth within each period (e.g., from the spring of kindergarten to the spring of first grade), adjusting only for age at assessment (see Model 1 in Figure 3). Then, we add a series of covariates to the regression models, to condition on race-ethnicity, SES, prior and current teacher ratings of behavior, and prior and current teacher ratings of approaches to learning (Model 2). Thus, the estimate of the gender gap growth in Model 2 is the conditional growth in the gender gap *before accounting for mediation through teacher ratings of students' mathematics proficiency* (represented by path c^+ in Model 2). We then examine the extent to which teacher perceptions mediate the relation between achievement gains and gender.

In the next three subsections, we present three different approaches to mediation and discuss the merits of each. First, we discuss what we term the *conventional approach to mediation*, which would use the current teacher rating (or perhaps, prior teacher rating) as the mediating variable. As we discuss below, such an approach is limited and may yield biased estimates. Second, we discuss *mediation incorporating instrumental variables*, as a way to address the shortcomings of the conventional mediation approach. In that subsection, we also discuss what instrumental variables are and how we use them in this study. In the third and final subsection, we present an improvement to the instrumental variables approach, in which we match boys and girls in terms of prior achievement and behavior variables and then perform an *instrumental variables-based mediation analysis on the matched sample*. The initial matching stage allows us to rely on less statistical adjustment in the mediation analysis with instrumental variables, which is desirable because it makes fewer modeling assumptions.

Conventional approach to mediation. We begin with the most obvious approach, which is to perform a standard mediation analysis using the *current* teacher rating as the mediator, and then examine the remaining growth in the gender gap that is unexplained by the model (Model 3). The problem with this approach is that the teacher ratings and achievement scores were collected by ECLS-K at the same time, and thus the estimate of the mediation effect will exhibit simultaneity bias because it is unclear which occurred first (the teacher rating or the achievement gains). With a biased estimate of the mediator, we will also obtain a biased estimate of the gender gap growth that is left over after the biased mediator is accounted for (Bullock et al., 2008, 2010; Bullock & Ha, 2011; Imai et al., 2011; Judd & Kenny, 1981; MacKinnon, 2008; Sobel, 2009). Another approach to conventional mediation uses the *prior* teacher rating (instead of the current teacher rating) as a mediator of the gender gap growth because the prior period ratings obviously occurred before the current period gains, thus alleviating concerns about simultaneity bias (Model 4). However, prior teacher ratings are unlikely to explain much of the gender gap growth in the *current* period because, by definition, the prior teachers are not the direct influence on students—rather, the current teachers are. Thus, by using prior teacher ratings as the mediator, the growth in the gender gap that is due to teacher ratings (including current ratings) will be underestimated because prior ratings are a poor substitute for current teacher ratings.

In sum, the advantage of the approach in Model 3 is that it includes current teacher ratings as the mediator, and these ratings are most likely to influence current gender gap growth; however,

	Study 2A N = 6,658 for first grade N = 3,919 for third grade	Study 2B N = 9,363 for first grade N = 5,733 for third grade
Conventional approach to mediation	Current teacher rating of math proficiency as mediator (Model 3)	Current teacher rating of math proficiency as mediator (Model 3)
	Prior teacher rating of math proficiency as mediator (Model 4)	Prior teacher rating of math proficiency as mediator (Model 4)
Mediation incorporating instrumental variables (IV)	*All prior teacher ratings of math proficiency* used as IVs to predict the mediator, current teacher rating of math proficiency (Model 5)	*Most recent prior year teacher ratings of proficiency and behavior* used as IVs to predict the mediator, current teacher rating of math proficiency (Model 5)
Propensity score matching, followed by mediation incorporating IVs (PSM+IV)	Boys and girls are matched before mediation incorporating IVs (Model 7)	Boys and girls are matched before mediation incorporating IVs (Model 7)

Figure 4. Mediation model descriptions for Study 2A and Study 2B. All models are estimated with and without teacher fixed effects, and are estimated for first grade and third grade.

the disadvantage is that the estimates are susceptible to simultaneity bias. The advantage of Model 4 is that it circumvents concerns of simultaneity bias by using prior teacher ratings as the mediator; however, the disadvantage is that it does not include the current teacher ratings, and will thus likely not capture the full extent of mediation due to teacher ratings.

Mediation incorporating instrumental variables (IVs). All together, the problems encountered in these two initial mediation approaches (Models 3 and 4) suggest we need a source of variation in *current* teacher ratings that was determined *prior* to and independent of the current period of test score growth—once such a source is identified, we can obtain an unbiased estimate of the mediator (current teacher perceptions) on current growth in the gender gap (Bullock et al., 2008, 2010). By using an IV approach to *predict current ratings on the basis of predetermined IVs* (Model 5), we can obtain unbiased estimates. IVs are sets of variables that can be used to carve out as-if random variation in the predictor of interest (i.e., current teacher ratings). To be valid, IVs must meet several criteria: (a) they must be correlated with the predictor of interest and (b) they must not have a direct effect on the outcome of interest, although they may have an indirect effect on the outcome through the predictor of interest (Imbens & Angrist, 1994; for an introduction to IV, see Murnane & Willett, 2011). In our study, we demonstrate that these two criteria are plausibly satisfied, and thus the IV approach to mediation is valid.

Two different sets of IVs are used in this study's two components, which we refer to as Study 2A and Study 2B. Figure 4 provides an overview of Study 2A and Study 2B, and how the various mediation models fit within the context of the substudies. We begin by describing our first set of IVs (used in Study 2A): teachers' ratings of a student's mathematics proficiency from all time points *prior* to the current period. Although prior teachers affect student achievement while the student is in their classrooms, past research suggests little or no lasting direct effect of prior teachers' ratings on the student's current achievement after accounting for prior achievement and other factors (Jussim & Harber, 2005; we also find no evidence of lasting direct effects, as we discuss later). However, we argue—and will demonstrate later—that there is an *indirect* effect: That is, prior teachers may influence the current teachers' perceptions (e.g., through notes in a student's cumulative file, conversations between teachers), and, in turn, current teacher perceptions may affect student mathematics performance. Because the prior teacher ratings (i.e., the IVs) affect current teacher perceptions but otherwise have no direct effect on current student gains, they can serve as IVs and provide us with an unbiased mediator in a real-world context (see also Imai et al., 2011, who discuss a related IV approach to mediation analysis). Because there is no direct effect of prior teacher ratings on outcomes and because these prior ratings occur before the onset of the current teacher's in-class experiences with the student, this IV approach circumvents typical pitfalls of using observational data (e.g., reverse causality/simultaneity bias, selection bias).[7] The IV mediation model (Model 5) can be decomposed into two steps: First, the current teacher ratings are predicted by all the covariates included in Model 2 (prior achievement, race, SES, age, teachers' ratings of behavior and approaches to learning) and the IVs. Second, the *predicted* value of the current teacher rating from this analysis is used as the mediator of the relation between gender and growth in math performance.

Study 2B differs from Study 2A in two main respects. First, in contrast with our use of *all* prior teachers' ratings of mathematics proficiency as the IVs in Study 2A, in Study 2B we use only the

[7] Conveniently, this approach taps into the effects of teacher perceptions that are influenced by prior teacher perceptions (net of other variables in the model), leaving behind differences in teacher perceptions that may be driven from other sources that may be susceptible to selection bias. Because the IV estimate taps into the "compliance" of current ratings with prior ratings, it is sometimes referred to as the *complier average treatment effect* or the *local average treatment effect* (Imbens & Angrist, 1994).

Table 3
Evidence of Instrument Validity, by Strategy (IV or PSM + IV), Study, Wave, and Use of Fixed Effects

	Instrumental-variables (IVs) approach							
	Study 2A				Study 2B			
	Spr. first grade (N = 6,658)		Spr. third grade (N = 3,919)		Spr. first grade (N = 9,363)		Spr. third grade (N = 5,733)	
Variable	IV	IV-FE	IV	IV-FE	IV	IV-FE	IV	IV-FE
(a) Strength F (do the IVs predict current ratings?)	44.24	40.98	15.57	9.90	60.78	53.17	50.68	40.17
p value on F	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
(b) Hansen's J (is there a *direct* effect of the IVs on student's achievement gains?)	1.76	1.11	2.75	1.56(!)	0.39	0.14	0.07	0.00
p value on J	.19	.29	.25	.46	.53	.71	.80	1.00
	Propensity score matching + Instrumental-variables (PSM + IV) approach							
	Study 2A				Study 2B			
	Spr. first grade (N = 6,658)		Spr. third grade (N = 3,919)		Spr. first grade (N = 9,362)		Spr. third grade (N = 5,731	
	IV	IV-FE	IV	IV-FE	IV	IV-FE	IV	IV-FE
(a) Strength F (do the IVs predict current ratings?)	42.32	43.19	14.86	7.39	56.93	55.42	38.50	28.22
p value on F	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
(b) Hansen's J (is there a *direct* effect of the IVs on student's achievement gains?)	0.91	0.22	2.26	2.19(!)	1.25	0.04	0.09	0.56
p value on J	.34	.64	.32	.33	.26	.84	.77	.46

Note. IV-FE = IV models with teacher fixed effects. These models correspond to Models 5 and 6 in Figures 3 and 4. IV validity statistics demonstrate (a) whether the IVs strongly predict the treatment [Strength F] (ideal: reject null that they do not predict) and (b) whether there is evidence of a direct effect of the IVs on student achievement gains [Hansen's J] (ideal: fail to reject null that there is no direct effect). (!) indicates the third-grade IV-FE model includes spring of kindergarten teacher ratings in the outcome model as well, but we include an interaction between the prior ratings as an added instrument (similar results obtained without this interaction). Importantly, note that all models pass both IV validity tests.

most recent prior teacher's ratings of mathematics proficiency. This approach does not require a long history of prior teacher ratings (only the most recent), which has the advantage of allowing for an expanded sample of students with nonmissing data. Second, in Study 2B we treat the most recent prior teacher's rating of externalizing problem behavior as an additional IV (instead of a standard covariate, as it was used in Study 2A). We argue that the prior teacher's perceptions of the child's behavior in this second set of IVs is intuitively valid because this perception of behavior (net of other variables, including the current teacher's perceptions of the child's behavior) likely captures bias in excess of bias reflected in ratings of proficiency. For example, there could be bias reflecting a "reputation" a student may have for misbehaving, which is unrelated to academics but that nevertheless influences teachers' general perceptions about the student. In addition, we demonstrate next that both sets of IVs—the set in Study 2A and the new set in Study 2B—pass standard empirical tests assessing the validity of IVs. Thus, the analyses we discuss provide plausibly unbiased estimates of mediation in a real-world setting.

As noted previously, in order for a set of IVs to be valid in this study, two conditions have to hold: (a) the IVs must affect the current teacher's rating of the student's mathematics proficiency and (b) the IVs must have no direct effect on mathematics achievement gains. If these conditions hold, then two empirical results should be obtained in our case: First, the IVs should predict current teacher ratings, above and beyond the other variables in the model. This is assessed by an F test on the joint significance of the IVs in the first-stage equation, the one predicting the current teacher ratings (Stock & Yogo, 2005). Second, there should be no significant direct effect of the IVs on the mathematics test gains. This is assessed by a post hoc test of whether the IVs (by definition, excluded from the second-stage estimation model) have unique power in predicting mathematics gains in the second-stage equation and, therefore, should have been included in the model as covariates and not as IVs (Hansen's J test; Hansen, 1982). Table 3 suggests that, in each wave of analysis and across the overall and within-classroom specifications, both sets of IVs satisfy both criteria. That is, for each of the eight IV analyses performed, the IVs (a) predict current teacher ratings (ps $< .0001$ across all models) and (b) have no significant effect on student achievement gains other than through their effect on the current teacher's perceptions ($.19 < p$s < 1.0). All of these tests support the validity of our analyses.

Propensity score matching, followed by mediation incorporating instrumental variables (PSM + IV). Finally, we perform an analysis similar to the IV analyses discussed above, but with one difference: Before performing the mediation analysis, we first *match* boys and girls in terms of age, race-ethnicity, family SES, prior achievement, and teachers' perceptions of behavior and approaches to learning (Model 7 in Figure 3; note that Model 6 is

the baseline model using the matched sample before mediation for the PSM + IV analysis, just as Model 2 was the baseline model for the IV analysis). That is, we use PSM (Rosenbaum & Rubin, 1984) to identify boys and girls with nearly identical values of behavior and prior performance, among other variables. For details on the matching process, see the Appendix. Using the matched sample, we then see whether teachers rate the math abilities of boys or girls higher.[8] Then, just as in the IV analyses described above, we perform our mediation analysis using the IVs to predict current teacher ratings.

This approach to mediation (i.e., PSM, followed by IV) has clear advantages over the approaches discussed above, especially the conventional approach to mediation. In addition to the benefits of the IV approach over the conventional approach to mediation (e.g., addressing concerns of simultaneity bias), the PSM + IV approach has the added benefit of beginning the mediation analysis with a sample of boys and girls who have nearly identical means and variances on each of the covariates included as "control variables." Matching first is particularly beneficial when the groups being compared (e.g., boys and girls) differ substantially on the covariates included in the regression model. Specifically, models that rely on covariate adjustment (as Models 1–5 do) can yield biased estimates when the mean values of the groups' propensity scores differ by more than 0.5 standard deviations (Rubin, 2001). In our case, boys' and girls' propensity score mean values differ by more than 0.68 standard deviations in all analyses (see Tables S17–S20 in the supplemental material).[9] Thus, if the assumptions of the regression model are not satisfied (e.g., if the functional form of the regression is not correctly specified), then the large covariate differences between boys and girls can lead to biased estimates. However, with PSM, rather than relying on *statistical* adjustment of the large differences between boys and girls on variables such as behavior and approaches to learning, we first identify boys and girls with nearly equivalent values and use them in our mediation analysis.[10] In addition, this matching approach makes the analysis more transparent in two ways: (a) simplifying the analyses so that one can see the underrating of girls when matched to nearly identical boys and (b) reducing the reliance on statistical adjustment in the IV mediation, and thereby reducing a dimension of the conceptual complexity of the IV mediation. Along with the IV approach, we use the PSM + IV approach in Studies 2A and 2B. Table 3 provides evidence suggesting that the use of IVs in the PSM + IV approach is valid in all eight PSM + IV models.

Participants. For analyses using the first set of IVs (i.e., all prior teacher ratings of math proficiency), the ECLS-K participants are identical to those used in Study 1: there were 6,658 students in the first-grade analysis (i.e., gender gap growth from the spring of kindergarten to the spring of first grade) and 3,919 students in the third-grade analysis (i.e., gender gap growth from the spring of first grade to the spring of third grade). As noted above, the second set of IVs does not require a lengthy history of teacher ratings, thus allowing us to expand the sample to students who had missing data and could not be included in the IV analyses in Study 2A. The final sample for Study 2B is 9,362 students in the kindergarten to first-grade analysis, and 5,733 students in the first- to third-grade analysis.[11]

Results and Discussion

Table 4 provides the results for Study 2A. For parsimony, we focus our discussion on the within-classroom estimates (i.e., models *with* teacher fixed effects), which are more methodologically rigorous; note, though, that the results of models without fixed effects are similar to those with fixed effects. We begin by showing the growth in the gender gap when only accounting for prior achievement and age at times of assessment. For example, Model 1 shows that girls' achievement in first grade is, on average, 0.070 standard deviations lower than boys, conditional on prior achievement and age, and when comparing boys and girls within the same classroom. In other words, on average, girls lose ground to boys by about 0.070 standard deviations between the spring of kindergarten and first grade. Between the spring of first and third grades, they lose more ground to boys, about 0.162 standard deviations. Model 2 shows that, on average, girls lose about 0.153 and 0.236 standard deviations in comparison to boys over these periods, respectively, when conditioned on multiple variables (e.g., age, race-ethnicity, prior and current behavior and approaches to learning, and prior mathematics achievement scores) but *not* including teacher ratings.

The mediation results begin with Model 3, which used *current* teacher ratings as the mediator of the gender gap. Model 4 used *prior* teacher ratings as the mediator. As previously mentioned, these conventional approaches may provide biased estimates and understate the true extent of mediation, and thus we focus on Model 5, which uses an IV approach to predict current teacher ratings based on prior teacher ratings. After accounting for the predicted current teacher rating and prior achievement, girls score 0.080 standard deviations lower than boys. Compared with the model not accounting for mediation (i.e., Model 2), Model 5 suggests that the predicted teacher ratings mediate almost half (48%) of the relation between gender and achievement gains between the spring of kindergarten and first grade.

[8] Recall that in Study 1, we examined whether teachers rated boys' or girls' math ability higher after *conditioning* on a host of covariates. Here, we are performing an analogous procedure, but instead of conditioning on a host of variables, we *match* on them. The differences in teachers' perceptions of boys' and girls' math abilities from this matching approach were very similar to those from the conditioning approach in Study 1. To see this, look at the coeffieicnt estimate on female from the IV-FE (fixed effects) Stage 1 columns in Tables S21–S22 in the supplemental material.

[9] The propensity score can be viewed as a composite measure of the covariates included in the model. As can be seen in Tables S17–S20, the variables with the greatest differences between boys and girls are externalizing problem behavior (> 0.37 *SD*s across all externalizing behavior variables and waves) and approaches to learning (> 0.40 *SD*s across all approaches to learning variables and waves).

[10] One drawback here could be that we may not find suitable matches for some students, and thus our sample may become nonrepresentative. Fortunately, we were able to identify suitable matches (i.e., matches within 0.25 *SD*s of the propensity score; Rosenbaum & Rubin, 1985) for all but four students across all waves and studies (Studies 2A and 2B). Hence, our sample remains representative.

[11] Although we demonstrated in Study 1 that teachers underrate girls relative to boys in fifth grade as well, we cannot explore how teacher perceptions affect growth in the gender gap from third to fifth grade because the sets of instruments have low predictive power. That is, the first-stage *F* statistics on the instruments are considerably below the acceptable thresholds established by Stock and Yogo (2005).

Table 4
Study 2A: Female–Male Standardized Differences in Mathematics Achievement, Conditional on Prior-Period Achievement

	Covariate-adjusted models					Matching-based models	
	Raw difference	Conditional difference among boys and girls, *before* accounting for teacher ratings	Conditional difference among boys and girls, after accounting for *current* teacher ratings	Conditional difference among boys and girls, after accounting for *prior* teacher ratings	Conditional difference among boys and girls, after accounting for *predicted current* teacher ratings	Conditional difference among *matched* boys and girls, *before* accounting for teacher ratings	Conditional difference among *matched* boys and girls, *after* accounting for *predicted current* teacher ratings
Test period	WLS [1]	WLS baseline [2]	WLS [3]	WLS [4]	IV [5]	PSM baseline [6]	PSM + IV [7]
Spring first grade (N = 6,658)							
Without teacher fixed effects	−0.066 (0.018)	−0.134 (0.018)	−0.114 (0.018)	−0.133 (0.018)	−0.078 (0.022)	−0.156 (0.020)	−0.101 (0.023)
Percent reduction		baseline	15%	<1%	42%	baseline	35%
With teacher fixed effects	−0.070 (0.020)	−0.153 (0.020)	−0.119 (0.020)	−0.150 (0.020)	−0.080 (0.023)	−0.156 (0.021)	−0.091 (0.021)
Percent reduction		baseline	22%	2%	48%	baseline	42%
Spring third grade (N = 3,919)							
Without teacher fixed effects	−0.158 (0.025)	−0.241 (0.025)	−0.206 (0.025)	−0.229 (0.024)	−0.058 (0.044)	−0.237 (0.027)	−0.092 (0.048)
Percent reduction		baseline	15%	5%	76%	baseline	61%
With teacher fixed effects	−0.162 (0.026)	−0.236 (0.026)	−0.191 (0.025)	−0.225 (0.026)	−0.036 (0.052)	−0.242 (0.027)	−0.045 (0.059)
Percent reduction		baseline	19%	5%	85%	baseline	81%
Regression model also includes							
Last test score	X	X	X	X	X	X	X
Age (in days) at current & prior assessments	X	X	X	X	X	X	X
All prior test scores		X	X	X	X	X	X
Prior & current behavior		X	X	X	X	X	X
Prior & current approaches to learning		X	X	X	X	X	X
Race indicators and SES		X	X	X	X	X	X
Current teacher ratings			X				
Immediately prior teacher ratings				X			
Predicted current teacher ratings (using IV)					X		X

Note. WLS = weighted least squares; IV = instrumental variables; PSM = propensity score matching; SES = socioeconomic status. Numbers in square brackets are model numbers. Heteroskedastic-robust standard errors clustered at the teacher level appear in parentheses to the right of the estimates. X indicates these covariates are included in the model. For all regression coefficients (not just the "female" coefficient) and model statistics, see Tables S9–S12 and S21–S22 in the supplemental materials.

Next, we turn to the period from spring of first grade to third grade. Compared with the model not accounting for mediation (i.e., Model 2), Model 5 suggests that the predicted teacher ratings mediate almost 85% of the relation between gender and achievement gains between the spring of first grade and third grade (i.e., from −0.236 to −0.036).

Despite differing in terms of samples, covariates, and IVs, we find that the patterns from Study 2B (see Table 5) are remarkably similar to those from Study 2A (see Table 4). This is true for not only the first-grade patterns but also the third-grade patterns. For example, comparing the estimates with and without mediation using IV (i.e., Model 2 vs. Model 5), the estimated growth in the gender gap from kindergarten to first grade reduced 46% (from −0.155 to −0.083) in Study 2B, similar to the reduction of 48% (from −0.153 to −0.080) in Study 2A. The first- to third-grade gender gap growth estimate reduced by 66% (from −0.259 to −0.087) in Study 2B, which is a somewhat smaller reduction than the 85% (from −0.236 to −0.036) in Study 2A. Note, too, that the portion mediated in each of these models is statistically significant, as determined by mediation analyses with bootstrapped 95% confidence intervals (MacKinnon, 2008; see Figures 5 and S1). Thus, the IV-based mediation analyses in Studies 2A and 2B all suggest that teachers' perception that boys are more proficient in mathematics than equally behaving and achieving girls mediates a sizable portion of the growth in the early gender gap in mathematics.

The results from mediation analyses using PSM and IVs (PSM + IV) were remarkably similar to the results of the IV analyses just discussed; this is true across the grades examined, whether the models used teacher fixed effects, and whether we used the IVs from Study 2A or Study 2B. That is, whether we performed the IV mediation using covariate adjustment only (i.e., comparing Models 2 and 5) or using matching first (i.e., comparing Models 6 and 7), the results suggest that roughly the same amount of the relation between math achievement gains and gender is mediated by teacher perceptions. For example, in Study 2A with teacher fixed effects (see Table 4), using the IV approach the estimated growth in the gender gap reduced by 48% (from −0.153

Table 5
Study 2B: Female–Male Standardized Differences in Mathematics Achievement, Conditional on Prior-Period Achievement

	Covariate-adjusted models					Matching-based models	
Test period	Raw difference WLS [1]	Conditional difference among boys and girls, *before* accounting for teacher ratings WLS baseline [2]	Conditional difference among boys and girls, after accounting for *current* teacher ratings WLS [3]	Conditional difference among boys and girls, after accounting for *prior* teacher ratings WLS [4]	Conditional difference among boys and girls, after accounting for *predicted current* teacher ratings IV [5]	Conditional difference among *matched* boys and girls, *before* accounting for teacher ratings PSM baseline [6]	Conditional difference among *matched* boys and girls, *after* accounting for *predicted current* teacher ratings PSM + IV [7]
Spring first grade (N = 9,363)							
Without teacher fixed effects	−0.066 (0.015)	−0.133 (0.015)	−0.111 (0.015)	−0.132 (0.015)	−0.074 (0.019)	−0.144 (0.017)	−0.080 (0.022)
Percent reduction		baseline	17%	<1%	44%	baseline	44%
With teacher fixed effects	−0.070 (0.017)	−0.155 (0.017)	−0.117 (0.016)	−0.153 (0.016)	−0.083 (0.021)	−0.152 (0.018)	−0.078 (0.021)
Percent reduction		baseline	24%	<1%	46%	baseline	49%
Spring third grade (N = 5,733)							
Without teacher fixed effects	−0.173 (0.020)	−0.259 (0.020)	−0.229 (0.020)	−0.244 (0.020)	−0.124 (0.030)	−0.246 (0.020)	−0.127 (0.032)
Percent reduction		baseline	12%	6%	52%	baseline	48%
With teacher fixed effects	−0.183 (0.022)	−0.259 (0.020)	−0.215 (0.020)	−0.246 (0.020)	−0.087 (0.033)	−0.265 (0.021)	−0.090 (0.039)
Percent reduction		baseline	17%	5%	66%	baseline	66%
Regression model also includes							
Last test score	X	X	X	X	X	X	X
Age (in days) at current and prior assessments	X	X	X	X	X	X	X
All prior test scores		X	X	X	X	X	X
Current behavior		X	X	X	X	X	X
Immediately prior & current approaches to learning		X	X	X	X	X	X
Race indicators and SES		X	X	X	X	X	X
Current teacher ratings			X				
Immediately prior teacher ratings				X			
Predicted current teacher ratings (using IV)					X		X

Note. WLS = weighted least squares; IV = instrumental variables; PSM = propensity score matching; SES = socioeconomic status. Numbers in square brackets are model numbers. Heteroskedastic-robust standard errors clustered at the teacher level appear in parentheses below estimates. X indicates these covariates are included in the model. For all regression coefficients (not just the "female" coefficient) and model statistics, see Tables S13–S16 and S23–S24 in the supplemental materials.

in Model 2 to −0.080 in Model 5), whereas the same gap growth using the PSM + IV approach was estimated to reduce by 42% (from −0.156 in Model 6 to −0.091 in Model 7). Between first and third grade, the Study 2A PSM + IV mediation analyses with teacher fixed effects suggested that 81% of the relation between gender and achievement gains was mediated by predicted teacher ratings. In Study 2B (see Table 5), the PSM + IV models estimated that 49% and 66% of the relations between gender and achievement gains in the kindergarten-to-first-grade and first-grade-to-third-grade periods, respectively, were explained by predicted teacher ratings. Tables 4 and 5 provide the mediation analyses for all PSM + IV analyses (with and without teacher fixed effects), and Figure 5 illustrates the PSM + IV mediation analyses for the models with teacher fixed effects—we believe the models in Figure 5 present the best estimations of mediation in this article because they reduce bias through matching, IVs, and fixed effects. Note that all mediated paths are statistically significant.

General Discussion

Contrary to recent suggestions that teachers hold more positive assessments of girls' mathematics performance, we find in a nationally representative sample that teachers rate the mathematics skills of girls *lower* than those of boys, after accounting for differences in achievement and teachers' ratings of behavior. That is, after accounting for mathematics achievement histories, behavior, approaches to learning, race, age, SES, and even looking at boys and girls in the same classrooms (in the teacher fixed effects models), girls' skills are rated to be about one tenth of a standard deviation lower than those of their boy classmates. This pattern is consistent throughout elementary school.

Put another way, the results suggest that teachers rate girls on par with similarly achieving boys only if they perceive those girls as working harder and behaving better than those boys. It is important to note that there is no evidence of a similar ratings disadvantage for Black or Hispanic students (relative to White students in the same classroom) and that there is no evidence

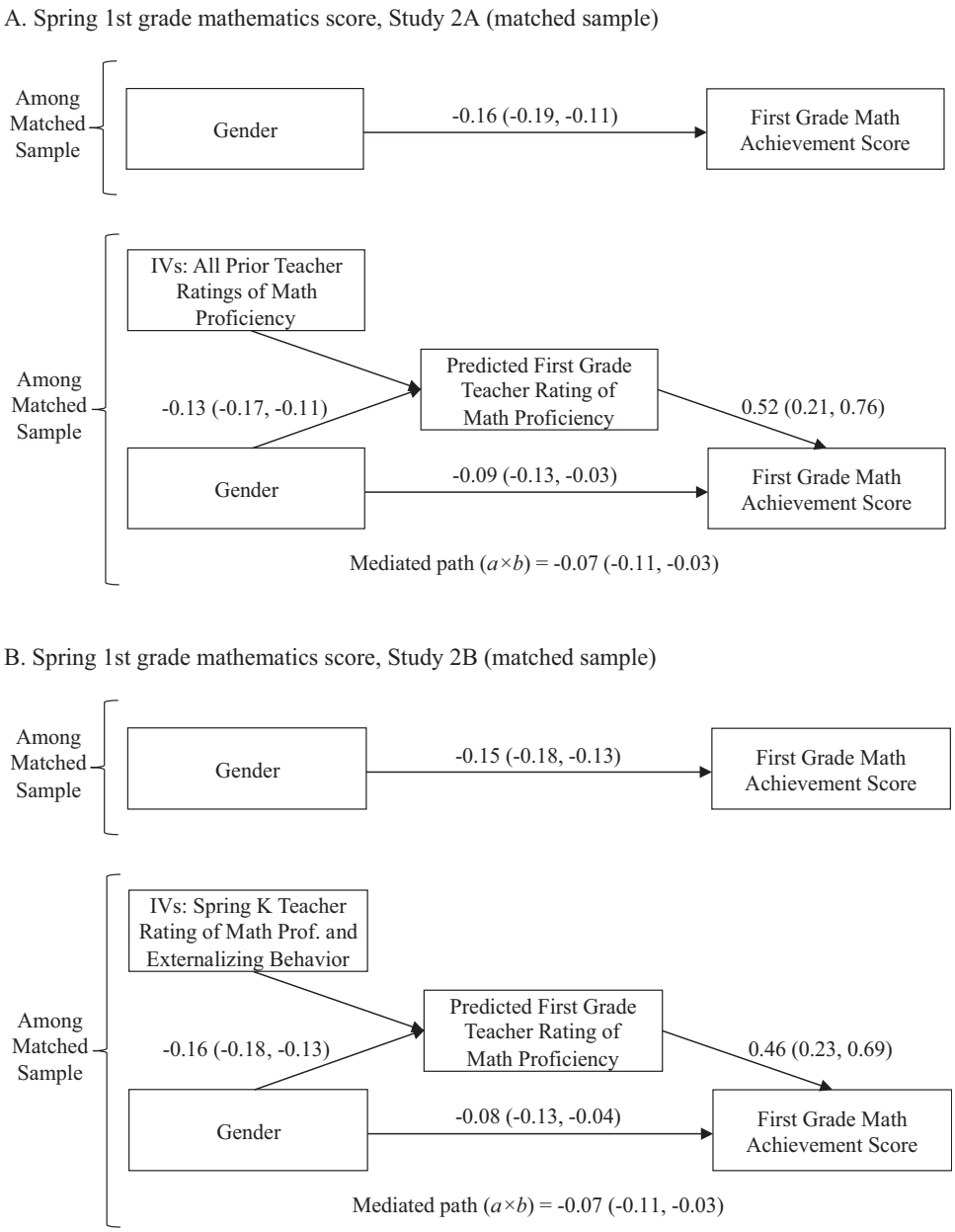


Figure 5. Results of mediation analyses (Model 7, compared with Model 6) for Study 2A and Study 2B with fixed effects (with the matched samples). The numbers in parentheses are 95% bootstrapped confidence intervals based on 1,000 bootstrapped replications (MacKinnon, 2008). See Tables 4 and 5 for additional covariates included in the models. A. Spring first-grade mathematics score, Study 2A (matched sample). B. Spring first-grade mathematics score, Study 2B (matched sample). C. Spring third-grade mathematics score, Study 2A (matched sample). D. Spring third-grade mathematics score, Study 2B (matched sample). IV = instrumental variable; Prof. = proficiency; Spring K = Spring kindergarten.

Figure 5 continues

that girls are rated lower or higher in *reading* after accounting for these factors. This indicates that the teacher underrating phenomenon is unique to girls and mathematics performance.

This study makes an important, and methodologically rigorous, contribution to the literature on the implications of teachers' perceptions. The consistency of the findings across eight different estimation strategies (Studies 2A and 2B, the IV and PSM + IV approaches within each substudy, each estimated with and without teacher fixed effects), as well as the consistent evidence supporting the validity of both IV approaches (see Table 3), provides strong

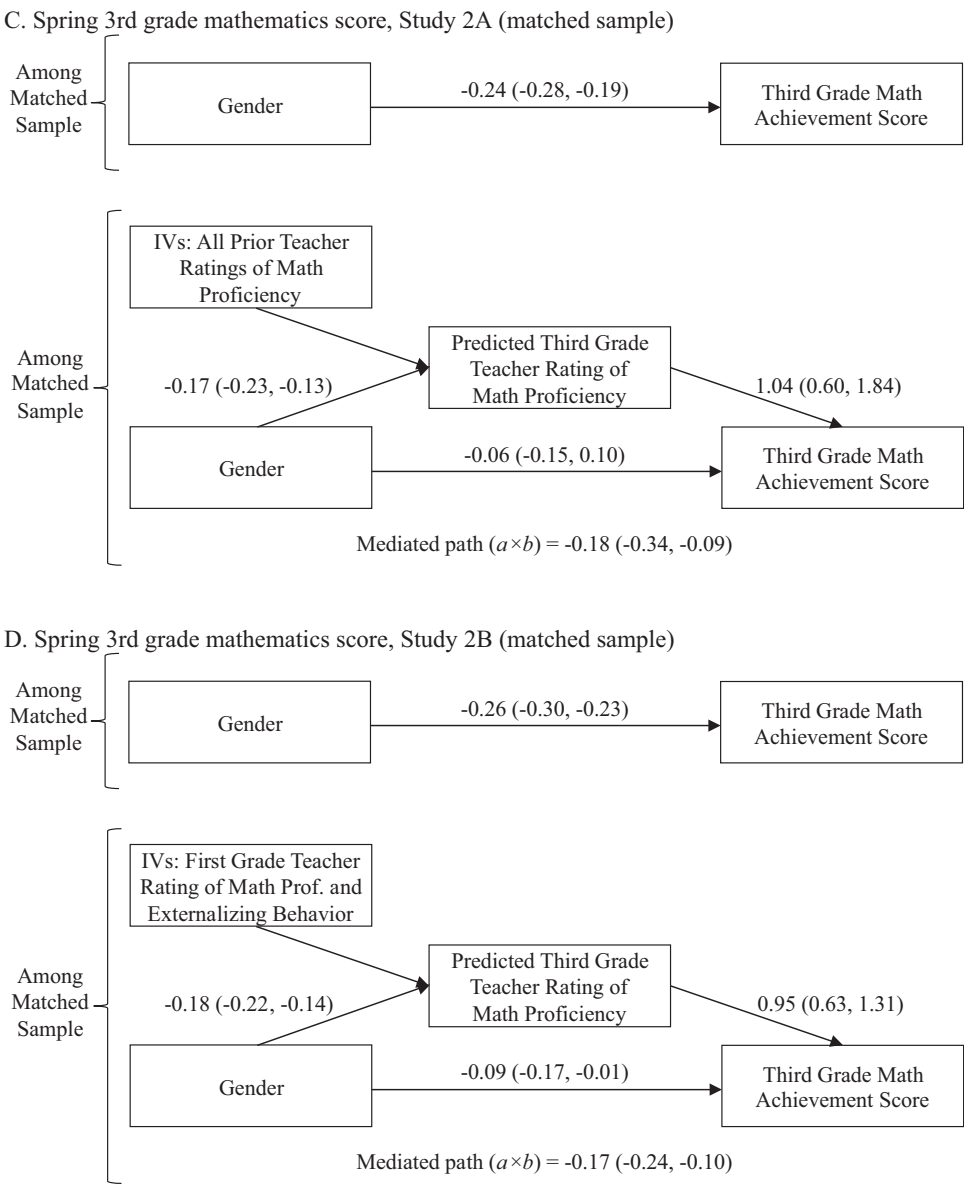


Figure 5 (*continued*).

evidence that gender-biased perceptions mediate much of the early development of the gender gap.

Potential Limitations

Our studies have several limitations worth noting. First, although Study 2 suggests that teachers' perceptions likely affect student learning, our analyses cannot speak to the precise mechanisms through which these effects operate. We speculate about these mechanisms later in the article.

Second, the analyses in Study 2 are not based on experimental designs. Thus, teacher expectations were not randomly assigned. Although we have argued that the use of prior teacher ratings provide a source of exogenous variation and thus yield less biased estimates than conventional mediation approaches, and although the consistency of findings across the different approaches taken and across the different waves of data in Study 2 bolsters confidence in our conclusions, we cannot be certain that all bias has been purged from the estimates.

Third, because perceptions were not randomly assigned to teachers, it is possible that part of what we are terming *teachers' gender-biased perceptions* is attributable to peer, parent, and societal gender-biased perceptions. However, although these other (i.e., nonteacher) factors may contribute to the gender gap, it is doubtful that they explain the majority of the specific estimates of "teacher perceptions" reported here. The teacher rating of student proficiency is not simply an attitudinal measure (e.g., agreement with "boys are naturally better at math"), but rather is a detailed set of ratings of individual students' proficiency in specific mathe-

matics domains (see Tables S1–S4 in the supplemental material). Nevertheless, because the ECLS-K data set does not contain data on the gendered beliefs of these students' parents or peers, we cannot empirically examine their associations with the teachers' ratings of student mathematics proficiency.

Finally, the conclusions we draw from this research pertain to teachers' perceptions of girls' effort, rather than girls' actual effort. Some smaller scale research suggests that teachers might overestimate the degree to which girls actually work harder than boys (e.g., Madon et al., 1998). Hence, it is possible that teachers' perceptions of girls' additional effort identified in this study are somewhat inaccurate. However, evidence from ECLS-K student surveys indicates that girls and boys themselves report differential behavior (e.g., the exhibition of "problem behaviors") that is aligned with ECLS-K teacher reports (Rathbun et al., 2004). Thus, although girls may indeed be working harder than boys (and therefore, it is not just a perception that they are working harder), we only have a measure of teachers' perceptions, and that is why we have used terms such as *perceptions of differences* throughout the current article.

Given these limitations, we are cautious in concluding that growth in gender gaps in mathematics achievement would necessarily be reduced by 35%–85% if teachers judged similarly achieving and behaving boys and girls as mathematically similar. Exact percentages aside, this study *does* suggest that teachers' perceptions mediate the growing gender gap in mathematics achievement during early elementary school.

Teacher Perceptions and Gender: Unanswered Questions

Despite the limitations noted above, using large-scale, nationally representative data in this study, we found that teachers tend to rate the mathematics proficiency of girls lower than that of boys after accounting for achievement and behavior and that this tendency appears to partially account for the growing mathematics gender gap that favors boys during elementary school. These findings add to the mounting evidence of ways in which gender gaps are socially constructed (Beilock et al., 2010; Else-Quest et al., 2010; Halpern et al., 2007) and point toward teachers' perceptions of boys and girls as one important factor influencing gender gaps in early grades. However, our study raises, but does not answer, two additional questions of importance. First, it is unclear why teacher ratings of students' mathematics proficiency are skewed by gender, and second, it is unclear how those skewed teacher perceptions actually influence boys' and girls' achievement. Given the limitations of our data, we can only speculate about potential explanations, suggesting directions for future research.

Potential reasons for gender-biased pattern in teacher ratings. Given the importance of teachers' perceptions of the boys and girls in their mathematics classes, we need to understand more about how those perceptions are shaped. Specifically, why do teachers have to perceive girls as working harder in order to rate them as equally mathematically proficient to boys? In addition to biases potentially stemming from general societal stereotypes about who is good at mathematics, elementary teachers of mathematics may have two additional sources of bias that merit further exploration: teachers' own experiences as learners of mathematics and teachers' experiences with students in their classrooms.

First, the vast majority of elementary teachers are women and may be prone to mathematics anxiety (Beilock et al., 2010). In our analysis, we found that the gender-biased pattern in teacher ratings was stronger for female teachers than for male teachers. This pattern raises the question of whether female teachers' own mathematical insecurities might shape their perceptions of the abilities of the girls and boys in their mathematics classrooms.

Second, perhaps teachers' observations of the students in their classrooms have skewed their perceptions of boys' and girls' mathematics abilities. As noted previously, some research on student motivation suggests that boys are more likely to hold performance goals, whereas girls more often pursue mastery goals (Kenney-Benson et al., 2006). Hence, boys focus more on publicly exhibiting their knowledge (or "showing off") in the mathematics classroom, and are more likely than girls to volunteer answers and get teacher feedback (Li, 1999; Sadker & Sadker, 1986). In addition, as early as first grade, boys have been found to use more novel approaches to solving problems (Fennema, Carpenter, Jacobs, Franke, & Levi, 1998; see also Tiedemann, 2000). Thus, these types of behaviors may lead teachers to perceive boys as "mathematically smarter" than girls. Furthermore, we cannot ignore the fact that girls appear to work harder than boys, on average, despite the fact that boys actually make larger math gains during elementary school (Forgasz & Leder, 2001; Ready et al., 2005). Hence, on one hand, it is not surprising that teachers would assume that boys can achieve more with less effort—a belief that might underlie the patterns found here. However, our results suggest that teachers go beyond what is warranted in assuming that the achievement of boys is higher than girls with similar levels of effort, even when their achievement does not differ.

Mechanisms through which teacher perceptions influence boys' and girls' achievement. Just as this study cannot determine the specific causes of teachers' differential perceptions of boys' and girls' abilities, the data do not lend themselves to exploring the precise mechanisms through which these perceptions operate. However, prior research suggests a number of ways in which teacher expectations are conveyed to students and might ultimately impact student performance.

Some research suggests that teachers' feedback to students may be the mechanism through which expectations are communicated. For example, Dweck and colleagues (1978) found that boys were more likely than girls to receive positive feedback that highlighted the intellectual quality of their work, whereas girls were more often praised for nonintellectual aspects. They linked this pattern to girls' greater tendency to attribute failure to lack of ability. Similarly, Sadker and Sadker (1986) found that teachers gave more attention to boys than girls and provided more specific feedback to boys. Given that self-efficacy, which promotes academic success, is positively related to frequent and immediate feedback provided to students (Pajares, 2002), and given that girls' self-efficacy is particularly likely to be influenced by others' assessments of them (Usher & Pajares, 2006a, 2006b), it may well be that feedback is a potential mechanism through which perceptions affect achievement gaps.

Finally, as mentioned previously, Beilock et al.'s (2010) study suggests that female teachers' own mathematics anxiety might influence their students. Specifically, they concluded that first- and second-grade female teachers' mathematics anxiety exerts a negative influence on girls' beliefs about who is good at mathematics and ultimately on girls' mathematics achievement. Hence, even if

female teachers strive to treat boys and girls equitably, their own perceptions of themselves as female learners who struggle to understand mathematics might be subtly conveyed to students during the course of instruction.

Conclusion

The results of these studies indicate that teachers rate the mathematics proficiency of girls as equal to that of similarly achieving boys only if they perceive the girls as working harder and being better behaved than the boys. This pattern is unique to girls and mathematics, as it did not occur in reading or with other underserved groups (e.g., Black and Hispanic students) in mathematics. This tendency for teachers to underrate girls' mathematics performance relative to boys who perform and behave similarly appears to substantially mediate the development of gender achievement gaps in elementary school. Raising awareness of—and hopefully, reducing—the tendency for teachers to rate boys higher than girls who perform and behave similarly may thus help close the gender achievement gap in mathematics.

Additional insights could be gained from future research that delves further into the ways in which teachers' perceptions impact their interactions with girls and boys, and how, specifically, those interactions affect girls' and boys' mathematics self-concepts and learning. Accounting for the effects of other social influences (e.g., parents, peers, media; Eccles, 1986; Frome & Eccles, 1998; Herbert & Stipek, 2005) may further explain the growing achievement gap in elementary school and, in combination with the current study and related work, can lead to interventions that promote gender equity among even our youngest mathematicians.

References

American Association of University Women Education Foundation. (2008). *Where the girls are: The facts about gender equity in education*. Washington, DC. Retrieved from http://www.aauw.org/files/2013/02/Where-the-Girls-Are-The-Facts-About-Gender-Equity-in-Education.pdf

Beilock, S. L., Gunderson, E. A., Ramirez, G., & Levine, S. C. (2010). Female teachers' math anxiety affects girls' math achievement. *Proceedings of the National Academy of Sciences, USA, 107,* 1860–1863. doi:10.1073/pnas.0910967107

Bullock, J. G., Green, D. P., & Ha, S. E. (2008). Experimental approaches to mediation: A new guide for assessing causal pathways. Retrieved from http://isps.research.yale.edu/conferences/isps40/downloads/BullockGreenHa.pdf

Bullock, J. G., Green, D. P., & Ha, S. E. (2010). Yes, but what's the mechanism? (Don't expect an easy answer). *Journal of Personality and Social Psychology, 98,* 550–558. doi:10.1037/a0018933

Bullock, J. G., & Ha, S. E. (2011). Mediation analysis is harder than it looks. In J. N. Druckman, D. P. Green, J. H. Kuklinski, & A. Lupia (Eds.), *Cambridge handbook of experimental political science* (pp. 508–521) New York, NY: Cambridge University Press. doi:10.1017/CBO9780511921452.035

Catsambis, S. (1994). The path to math: Gender and racial-ethnic differences in mathematics participation from middle school to high school. *Sociology of Education, 67,* 199–215. doi:10.2307/2112791

Dweck, C. S., Davidson, W., Nelson, S., & Enna, B. (1978). Sex differences in learned helplessness: II. The contingencies of evaluative feedback in the classroom; III. An experimental analysis. *Developmental Psychology, 14,* 268–276. doi:10.1037/0012-1649.14.3.268

Eccles, J. S. (1986). Gender-roles and women's achievement. *Educational Researcher, 15,* 15–19. doi:10.3102/0013189X015006015

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin, 136,* 103–127. doi:10.1037/a0018053

Fennema, E., Carpenter, T. P., Jacobs, V. R., Franke, M. L., & Levi, L. W. (1998). A longitudinal study of gender differences in young children's mathematical thinking. *Educational Researcher, 27,* 6–11. doi:10.2307/1176733

Fennema, E., Peterson, P. L., Carpenter, T. P., & Lubinski, C. A. (1990). Teachers' attribution and beliefs about girls, boys, and mathematics. *Educational Studies in Mathematics, 21,* 55–69. doi:10.1007/BF00311015

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). Mahwah, NJ: Erlbaum.

Frome, P. M., & Eccles, J. S. (1998). Parents' influence on children's achievement-related perceptions. *Journal of Personality and Social Psychology, 74,* 435–452. doi:10.1037/0022-3514.74.2.435

Fryer, R. G., & Levitt, S. D. (2004). Understanding the Black-White test score gap in the first two years of school. *The Review of Economics and Statistics, 86,* 447–464. doi:10.1162/003465304323031049

Fryer, R. G., & Levitt, S. D. (2010). An empirical analysis of the gender gap in mathematics. *American Economic Journal: Applied Economics, 2,* 210–240. doi:10.1257/app.2.2.210

Ginsburg, H. P., & Baroody, A. J. (2003). *Test of early mathematics ability* (3rd ed.). Austin, TX: PRO-ED.

Gunderson, E. A., Ramirez, G., Levine, S. C., & Beilock, S. L. (2012). The role of parents and teachers in the development of gender-related math attitudes. *Sex Roles, 66,* 153–166. doi:10.1007/s11199-011-9996-2

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernsbacher, M. A. (2007). The science of sex differences in science and mathematics. *Psychological Science in the Public Interest, 8,* 1–51. doi:10.1111/j.1529-1006.2007.00032.x

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica, 50,* 1029–1054. doi:10.2307/1912775

Herbert, J., & Stipek, D. (2005). The emergence of gender differences in children's perceptions of their academic competence. *Journal of Applied Developmental Psychology, 26,* 276–295. doi:10.1016/j.appdev.2005.02.007

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis, 15,* 199–236. doi:10.1093/pan/mpl013

Husain, M., & Millimet, D. L. (2009). The mythical 'boy crisis'? *Economics of Education Review, 28,* 38–48. doi:10.1016/j.econedurev.2007.11.002

Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B., & Williams, C. C. (2008). Gender similarities characterize mathematics performance. *Science, 321,* 494–495. doi:10.1126/science.1160364

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review, 105,* 765–789. doi:10.1017/S0003055411000414

Imbens, G. W., & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica, 62,* 467–475. doi:10.2307/2951620

Judd, C. M., & Kenny, D. A. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review, 5,* 602–619. doi:10.1177/0193841X8100500502

Jussim, L., Eccles, J., & Madon, S. (1996). Social perception, social stereotypes, and teacher expectations: Accuracy and the quest for the powerful self-fulfilling prophecy. *Advances in Experimental Social Psychology, 28,* 281–388. doi:10.1016/S0065-2601(08)60240-3

Jussim, L., & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controver-

sies. *Personality and Social Psychology Review, 9,* 131–155. doi:10.1207/s15327957pspr0902_3

Kenney-Benson, G. A., Pomerantz, E. M., Ryan, A. M., & Patrick, H. (2006). Sex differences in math performance: The role of children's approaches to schoolwork. *Developmental Psychology, 42,* 11–26. doi:10.1037/0012-1649.42.1.11

Lavy, V. (2008). Do gender stereotypes reduce girls' and boys' human capital outcomes? Evidence from a natural experiment. *Journal of Public Economics, 91,* 2083–2105. doi:10.1016/j.jpubeco.2008.02.009

Li, Q. (1999). Teachers' beliefs and gender differences in mathematics: A review. *Educational Research, 41,* 63–76. doi:10.1080/0013188990410106

Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin, 136,* 1123–1135. doi:10.1037/a0021276

Long, M. C., & Conger, D. (2013). Gender sorting across K–12 schools in the United States. *American Journal of Education, 119,* 349–372. doi:10.1086/669853

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. Mahwah, NJ: Erlbaum.

Madon, S., Jussim, L., Keiper, S., Eccles, J., Smith, A., & Palumbo, P. (1998). The accuracy and power of sex, social class and ethnic stereotypes: A naturalistic study in person perception. *Personality and Social Psychology Bulletin, 24,* 1304–1318. doi:10.1177/01461672982412005

Markwardt, F. C. (1989). *Peabody Individual Achievement Test–Revised: PIAT-R*. Circle Pines, MN: American Guidance Service.

McGraw, R., Lubienski, S. T., & Strutchens, M. E. (2006). A closer look at gender in NAEP mathematics achievement and affect data: Intersections with achievement, race/ethnicity, and socioeconomic status. *Journal for Research in Mathematics Education, 37,* 129–150.

Murnane, R. J., & Willett, J. B. (2011). *Methods matter: Improving causal inference in educational and social science research*. New York, NY: Oxford University Press.

Pajares, F. (2002). Gender and perceived self-efficacy in self-regulated learning. *Theory Into Practice, 41,* 116–125. doi:10.1207/s15430421tip4102_8

Penner, A. M., & Paret, M. (2008). Gender differences in mathematics achievement: Exploring the early grades and the extremes. *Social Science Research, 37,* 239–253. doi:10.1016/j.ssresearch.2007.06.012

Pomerantz, E. M., Altermatt, E. R., & Saxon, J. L. (2002). Making the grade but feeling distressed: Gender differences in academic performance. *Journal of Educational Psychology, 94,* 396–404. doi:10.1037/0022-0663.94.2.396

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. Retrieved from nces.ed.gov/pubs2004/2004007_1.pdfRetrieved from nces.ed.gov/pubs2004/2004007_1.pdf

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,* 85–97. doi:10.1037/0022-0663.76.1.85

Ready, D. D., LoGerfo, L. F., Burkam, D. T., & Lee, V. E. (2005). Explaining girls' advantage in kindergarten literacy learning: Do classroom behaviors make a difference? *Elementary School Journal, 106,* 21–38. doi:10.1086/496905

Reardon, S. F., & Robinson, J. P. (2008). Patterns and trends in racial/ethnic and socioeconomic academic achievement gaps. In H. F. Ladd & E. B. Fiske (Eds.), *Handbook of research in education finance and policy* (pp. 499–518). New York, NY: Routledge. doi:10.1162/edfp.2008.3.1.149

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,* 268–302. doi:10.3102/0002831210372249

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79,* 516–524. doi:10.1080/01621459.1984.10478078

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39,* 33–38. doi:10.2307/2683903

Rosenthal, R. (1994). Interpersonal expectancy effects: A 30-year perspective. *Current Directions in Psychological Science, 3,* 176–179. doi:10.1111/1467-8721.ep10770698

Rosenthal, R., & Jacobson, L. (1966). Teachers' expectancies: Determinants of pupils' IQ gains. *Psychological Reports, 19,* 115–118. doi:10.2466/pr0.1966.19.1.115

Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom: Teacher expectations and student intellectual development*. New York, NY: Holt. doi:10.1007/BF02322211

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services & Outcomes Research Methodology, 2,* 169–188. doi:10.1023/A:1020363010465

Sadker, M., & Sadker, D. (1986). Sexism in the classroom: From grade school to graduate school. *Phi Delta Kappan, 68,* 512–515.

Sobel, M. (2009). Causal inference in randomized and non-randomized studies: The definition, identification, and estimation of causal parameters. In R. Millsap & A. Maydeu-Olivares (Eds.), *The Sage handbook of quantitative methods in psychology* (pp. 3–22). Thousand Oaks, CA: Sage. doi:10.4135/9780857020994.n1

Stock, J. H., & Yogo, M. (2005). Testing for weak instruments in linear IV regression. In D. W. K. Andrews & J. H. Stock (Eds.), *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg* (pp. 80–108). Cambridge, England: Cambridge University Press. doi:10.1017/CBO9780511614491.006

Tiedemann, J. (2000). Gender-related beliefs of teachers in elementary school mathematics. *Educational Studies in Mathematics, 41,* 191–207. doi:10.1023/A:1003953801526

Tourangeau, K., Burke, J., Lê, T., Wan, S., Weant, M., Brown, E., . . . Walston, J. (2001). *ECLS-K Base-Year Public-Use Data Files and Electronic Codebook: User's manual* (NCES 2001–029). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Usher, E. L., & Pajares, F. (2006a). Inviting confidence in school: Invitations as a critical source of the academic self-efficacy beliefs of entering middle school students. *Journal of Invitational Theory and Practice, 12,* 7–16.

Usher, E. L., & Pajares, F. (2006b). Sources of academic and self-regulatory efficacy beliefs of entering middle school students. *Contemporary Educational Psychology, 31,* 125–141. doi:10.1016/j.cedpsych.2005.03.002

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment*. Mahwah, NJ: Erlbaum.

Woodcock, R. W., Mather, N., & McGrew, K. S. (2001). *Woodcock–Johnson III Tests of Cognitive Abilities*. Itasca, IL: Riverside.

Appendix

Propensity Score Matching Process

Propensity score matching proceeded in a series of steps. First, using logistic regression, we predicted each student's gender as a function of covariates such as race, socioeconomic status, behavior, approaches to learning, and prior achievement (and teacher indicators, in the models using teacher fixed effects), weighted by the sampling weight. Second, we obtained each student's predicted propensity score using the coefficient estimates from the logistic regression. Third, we matched. That is, for each girl, we found the boy with the closest propensity score; and similarly, for each boy, we found the girl with the closest propensity score. To match, we used a nearest neighbor (with replacement) strategy. And to help ensure the matches are of good quality, we used a caliper of 0.25 standard deviations of the pooled within-gender propensity score in the unmatched sample, meaning that we only retained matches that differ by less than 0.25 standard deviations of the propensity score (Rosenbaum & Rubin, 1984). Fourth, among the matched sample, we checked for balance, which here entailed three types of checks (see Rubin, 2001): (a) assessing the standardized difference between boys and girls on each covariate, with the goal being small standardized differences after matching; (b) comparing the variance of each covariate for boys and for girls, with the goal being a covariance ratio close to 1 (smaller than .5 or greater than 2 suggest poor balance; Rubin, 2001); and (c) checking for statistically significant differences between boys and girls on covariates included in the matching, with the goal being few to no significant differences after matching. For each of our eight matched samples (i.e., Studies 2A and 2B, each with and without teacher fixed effects, and each for Grades 1 and 3), we obtained excellent balance (see Tables S17–S20 in the supplemental material). Finally, with the matched sample, we performed the mediation analysis. To adjust for any remaining (albeit small) covariate differences between boys and girls in the matched samples, we included all covariates in the final estimation stage, just as we do in Model 5 (Ho, Imai, King, & Stuart, 2007).

Received February 4, 2012
Revision received January 28, 2013
Accepted April 10, 2013 ■