International Journal of
STEM Education

RESEARCH Open Access

# Teachers' race and gender biases and the moderating effects of their beliefs and dispositions

Yasemin Copur-Gencturk[1*], Ian Thacker[2] and Joseph R. Cimpian[3]

## Abstract

**Background** Women and people of color continue to be underrepresented in many STEM fields and careers. Many studies have linked societal biases against the mathematical abilities of women and people of color to this underrepresentation, as well as to earlier measures of mathematical confidence and performance. Recent studies have shown that teachers may unintentionally have biases that reflect those in broader society. Yet, many studies on teachers' reports of students' abilities use data in the field—not experimental data—and thus often cannot say if the findings reflect bias or actual differences. The few experimental studies conducted suggest bias against the abilities of girls and students of color, but the prior work has limitations, which we seek to address (e.g., local samples, no exploration of moderators, no preregistration).

**Methods** In this preregistered experiment of 458 teachers across the U.S., we randomly assigned gender- and race-specific names to solutions to math problems, then asked teachers to rate the correctness of the solution, as well as the student's math ability and effort. Teachers also completed scales reflecting their own beliefs and dispositions, which we then assessed how those beliefs/dispositions moderated their biases. We used multilevel modeling to account for the nested data structure.

**Results** Consistent with our preregistered hypotheses, when the solution was not fully correct, findings suggest teachers thought boys had higher ability, even though the same teachers did not report differences in the correctness of the solution or perceived effort. Moreover, teachers who reported that gender disparities no longer exist in society were particularly likely to underestimate girls' abilities. Although findings revealed no evidence of racial bias on average, teachers' math anxiety moderated their ability judgments of students from different races, albeit with only marginal significance; teachers with high math anxiety tended to assume that White students had higher math ability than students of color.

**Conclusions** The present research identifies teachers' beliefs and dispositions that moderate their gender and racial biases. This experimental evidence sheds new light on why even low-performing boys consistently report higher math confidence and pursue STEM—namely, their teachers believe they have higher mathematical ability.

**Keywords** Bias, Equity, Gender, Race

## Introduction

The underrepresentation of women and students of color in mathematically intense STEM (science, technology, engineering, and math) majors remains a persistent challenge in the United States (Cimpian et al., 2020; Warikoo et al., 2016). For example, despite nearly equivalent

*Correspondence:
Yasemin Copur-Gencturk
copurgen@usc.edu
[1] University of Southern California, Los Angeles, USA
[2] University of Texas at San Antonio, San Antonio, USA
[3] New York University, New York, USA

mean mathematics achievement between boys and girls throughout K-12 and bachelor's and Master's mathematics degree completion, there are large gender differences in doctoral degree completion (~27% female; National Science Foundation [NSF], 2020). In other math-intensive disciplines such as computer science, engineering, and physics, large gender disparities exist at all higher-education degree levels and in the labor market (NSF, 2020). Further, despite gradual improvements to racial diversity in the STEM workforce, as of 2021 there were still disproportionately fewer Black and Hispanic recipients of Bachelor's, Master's, and Doctoral degrees as well as fewer workers in STEM fields (National Center for Science and Engineering Statistics [NCSES], 2023). There are many competing hypotheses for why there are gender and racial disparities in STEM, with explanations stemming from fields of economics, psychology, sociology (Thacker et al., 2022), and even biology in the case of gender (Halpern et al., 2007). Yet, evidence is converging on the idea that unconscious and conscious biases held by teachers comprise an important factor that can contribute to gender and race disparities in academic outcomes and perpetuate STEM-specific stereotypes (see Cheryan et al., 2017; Dixon & Rousseau, 2005; Wang & Degol, 2017; Warikoo et al., 2016 for reviews).

Mathematics teachers are particularly well positioned to shape students' STEM academic self-concepts and help them overcome stereotypes; but unfortunately, teachers themselves are embedded in a society in which race and gender stereotypes are pervasive and are not immune to bias. Prior research has documented that teachers' biases are revealed in several situations. Teachers' perceptions of student mathematical ability (e.g., Copur-Gencturk et al., 2020) and attributions of students' performance (e.g., Espinoza et al., 2014; Reyna, 2008; Wang & Hall, 2018) differ by student race and gender; teachers' grading of student work differ by students' gender (Lavy & Sand, 2015); and teachers' recommendations for special education and gifted programs differ by students' race (Copur-Gencturk et al., 2022; Donovan & Cross, 2002; Elhoweris et al., 2005; Morgan, 2020; U.S. Department of Education, 2021). Further, teachers in the U.S. hold implicit and explicit racial biases that are similar to the general public (Starck et al., 2020).

Although prior work has found empirical evidence suggesting that teachers may have biases (e.g., Chin et al., 2020; Copur-Gencturk et al., 2022; Dennesen et al., 2022; Starck et al., 2020), fewer studies have explored the factors moderating these biases. Beliefs about ability, stable personal dispositions, and personal experiences are theorized to contribute to certain biases (Graham, 2017; Graham & Williams, 2009). When teachers evaluate student work, sometimes they attribute students' successes and failures to effort or ability in ways that reflect stereotypes (Espinoza et al., 2014; Fennema et al., 1990; Reyna, 2008; Tiedemann, 2000, 2002; Wang & Hall, 2018), and such biased assessments are thought to be influenced by social norms, as well as teachers' beliefs about ability, various stable personality dispositions, and situational and personal experiences (Graham & Williams, 2009). While some studies have investigated relationships between teacher bias and teacher belief factors such as self-efficacy and explicitly biased attitudes and beliefs (see Denessen et al., 2022, for a review), fewer look specifically at the roles of STEM-specific beliefs, dispositions, and experiences in shaping biases that are relevant to classroom contexts. Further, most STEM-specific studies investigating teacher bias were conducted with teachers outside of the United States or were conducted with teachers who had similar characteristics (e.g., those attending the same professional development program), which might have limited the extent to which the observed bias patterns could be generalizable to a national sample of U.S. teachers.

Thus, the purpose of this study was to investigate the biases and potential moderators of such biases among a sample of teachers across the United States (i.e., a national sample) so that the findings would not be bound to certain locations and would be more generalizable to U.S. teachers. In particular, building on our prior work (Copur-Gencturk et al., 2020), we investigated the extent to which teachers' evaluations of students' work and their attributions of students' performance to ability and effort differed by gender and race. We also explored the extent to which teachers' beliefs and dispositions moderated their biases. Furthermore, we preregistered our planned analyses and hypotheses in advance of data collection (see https://aspredicted.org/GNH_RXF for all preregistration information[1]), a practice encouraged to promote transparency in research (Reich, 2021).

## Prior work on teacher bias

### Explicit and implicit bias

Researchers distinguish between implicit and explicit forms of bias (e.g., Greenwald & Krieger, 2006). Explicit biases are discriminatory attitudes and behaviors that people are consciously aware of, are under the control of the individual, and are often captured using self-reported surveys of attitudes towards social groups and beliefs about the existence of inequality in society (e.g., Henry & Sears, 2002; Swim et al., 1995). Because explicit bias is intentional by definition, self-reports can be readily

[1] Note that the full preregistration refers to an additional study that goes beyond the scope of this paper.

altered as beliefs change or to portray the individual in a positive light and make them appear less prejudiced. In general, people are often reluctant to report explicit discriminatory beliefs and attitudes on explicit measures of bias, which tend to be poor predictors of educational disparities compared with more automatic, implicit forms of bias. For example, Nosek and Smyth (2011) found only weak relations between explicit gender bias and implicit math-male stereotypes among adults who volunteered to complete the Implicit Associations Test (IAT) at their publicly available website.

In contrast to explicit bias, implicit biases make use of quick, effortless, and automatic cognitive processing of the observed environment (Bargh, 1994; Greenwald & Banaji, 1995; Kahneman, 2011; Strack & Deutsch, 2004) and tend to arise in ambiguous situations where social information, such as information about race, ethnicity, or gender, might implicitly signal missing information, such as education levels (e.g., Aigner & Cain, 1977; Arrow, 1973; Bertrand & Duflo, 2017; Bertrand et al., 2005; Phelps, 1972). Implicit gender biases associating men with STEM are held across the world, and countries in which the general public holds stronger associations between boys and science also tend to have larger gender disparities in eighth-grade mathematics and science achievement (Nosek et al., 2009). Similarly, people across the United States hold negative anti-Black implicit racial attitudes, and much like the general public, in-service teachers in the U.S. hold implicit racial biases at similar levels (Starck et al., 2020) and exhibit behaviors consistent with ubiquitous implicit race and gender stereotypes (e.g., Carlana, 2019; Copur-Gencturk et al., 2020), as we discuss in greater detail in the next section.

## Teacher bias

From a birds-eye view, the evidence on the impacts of teacher bias appears to be mixed. A recent systematic review of the literature on teacher biases investigated the extent of teachers' implicit attitudes and stereotypes and their relationships with teacher factors and student outcomes (Dennesen et al., 2022). The authors reviewed 49 total studies and tallied significant relationships between teacher factors, student outcomes, and teacher implicit biases across multiple domains of implicit bias (e.g., implicit stereotypes and attitudes regarding special education needs, religion, physical appearance, giftedness, race, and gender). Findings across studies were mixed. With regard to teaching and student outcomes, the authors identified 17 relevant studies, and of them 6 (35%) found significant associations between teaching/student outcomes and implicit bias. Further, 11 of the 17 included explicit measures, of which 3 (27.3%) found significant relationships between outcomes and explicit measures. However, it should be noted that this systemic review included teaching and learning outcomes across a wide range of student outcomes (e.g., motivation, level of physical activity, self-concept, acceptance into honors societies) and teaching outcomes (e.g., grading errors, perceptions of student facial expressions, judgment of math skills, responses to misbehaviors). Further, of the 49 studies reviewed, only three were specific to STEM classrooms (i.e., De Kraker Pauw et al., 2016; Nürnberger et al., 2016; Thomas, 2017). Yet, when assessing the evidence pertaining specifically to STEM classrooms, the evidence appears to be more conclusive.

## Teacher bias in the STEM classroom

Teachers can hold biases that are specific to STEM-disciplines. For example, there are widespread stereotypes that White students  and men have greater natural ability in mathematics and other math-intensive disciplines. Post-secondary practitioners in math-intensive STEM disciplines (mathematics, physics, computer science, and engineering) view their fields as requiring higher levels of innate ability than other disciplines (Leslie et al., 2015). Furthermore, because women and Black individuals are often stereotyped as having lower levels of innate ability than men or White individuals (Kirkcaldy et al., 2007; Lecklider, 2013), and are stereotyped as having lower ability in STEM-disciplines (Copur-Gencturk et al., 2021; McGee & Martin, 2011; Rogers, 2020), they may feel pressure to avoid STEM fields that require "brilliance". Such STEM-specific beliefs about ability held by postsecondary instructors were found to be significantly correlated with women's and Black students' underrepresentation in Ph.D. attainment (Leslie et al., 2015). Robinson-Cimpian and colleagues (2014) also found that teachers in grades K, 1, 3, and 5 also demonstrate gender-biased beliefs about mathematical proficiency. The authors analyzed data from Early Childhood Longitudinal Study, Kindergarten Class of 1998–1999 (ECLS-K) revealing that teachers only perceived the mathematical proficiency of boys to be equal to that of girls when the girls were *also* viewed as working harder and behaving better. These teacher biases were specific to mathematics and were not found in teachers' perceptions of student reading proficiency. Moreover, these patterns of bias were replicated with a new ECLS-K cohort collected 12 years later (Cimpian et al., 2016).

Some studies have also found that K-12 teachers have STEM-specific biases associating boys with STEM, and these biases have consequences for students. Of the small number of studies investigating K-12 STEM teacher bias and correlates, experimental evidence suggests that teachers spanning from elementary to secondary school hold implicit gender biases that are linked with their tracking decisions and evaluations of student work. For example, studies have linked teachers' implicit boy–science associations with gender stereotypical mathematics tracking decisions. Nürnberger (2016) used the implicit associations test (IAT) to measure German pre-service elementary (grades 1–4) STEM teachers' male–mathematics associations finding that implicit gender stereotypes predicted teachers' gender-biased school career recommendations in math/science. Similarly, Carlana (2019) found implicit male–science stereotypes among Italian in-service 8th-grade teachers that were associated with gender-biased mathematics tracking decisions and gender disparities in their students' mathematics achievement. Thomas (2017) found that in-service physical science teachers of grades 6–8 in Austria also hold implicit boy–science biases, and that such biases positively predicted their male students' and negatively predicted female students' self-concept and intrinsic value. Additional lab-based research used IAT methods to assess male–STEM career, male–science aptitude, and male–learning styles associations among in-service secondary and postsecondary teachers in the Netherlands, finding that male teachers with a STEM background had stronger male–science aptitude associations compared with other groups (De Kraker Pauw et al., 2016). However, of these studies, all were conducted in countries outside of the U.S., focused exclusively on gender bias, and only one (Nürnberger et al., 2016) included explicit measures, which measured overtly discriminatory beliefs (e.g., "Boys are often talented for doing math.") rather than more modern, subtle discriminatory beliefs (e.g., "Discrimination against women is no longer a problem in the United States"; Swim et al., 1995). Further, these studies measured implicit bias in situations that do not resemble real teaching situations (e.g., using IAT methods).

Fewer studies have investigated racial biases among K-12 STEM educators. Kumar and colleagues (2015) used IAT methods and explicit measures to show that White middle-school teachers in the U.S. with implicit preferences for White versus Arab American and Chaldean American people were less likely to report promoting mutual respect among students and less likely to engage in culturally adaptive practices for resolving racial conflict in the classroom. A composite measure of teachers' explicit negative beliefs about minority and poor students was found to predict performance-focused classroom practices. As with the research on STEM teachers' gender bias, this study measured overt rather than covert discriminatory racial beliefs (e.g., "Generally, minority students are not as interested in school and schoolwork as White students") and employed implicit measures that do not parallel real teaching situations.

Other scholars have focused on detecting STEM instructor race- and gender-biases as they might occur in the classroom. As one of the pioneering works in this area, Copur-Gencturk and colleagues (2020), conducted an experimental study with K-12 mathematics teachers in a southern U.S. state who were participating in professional development programs. In that study, teachers were asked to evaluate the same set of student mathematics work to which a gender- and ethnicity-linked first name had been randomly assigned. Student names varied by three ethnicities (Black, Hispanic, White) and two genders (male, female), and solutions were either correct, incorrect, or partially correct. The authors found that teachers did not show bias in their grading (i.e., their evaluations of the correctness of student solutions), but showed biases against the mathematical ability of students. When the student solution was incorrect, teachers tended to infer that the student had greater mathematical ability when a male name appeared on the work compared with a female name. However, when the student solution was partially correct, teachers of color gave higher ability ratings to White male students compared with female students of color. The authors concluded that teachers of color showed greater racial bias in their ability judgments and provided a number of possible explanations for this finding, including that teachers might have been aware of the true purpose of the study and that White teachers in particular might have been more careful to disguise their biases. The authors also posited that ability bias found among teachers of color might have been due to internalized racism stemming from their own experiences of discrimination in mathematics. However, the authors did not collect evidence to substantiate these hypotheses, for example, by assessing teachers' suspicions or teachers experiences of discrimination. Further, because the teachers in this particular study came from a southern state and attended in-person professional development that was led by a team of teacher educators who devoted time to building rapport with teachers, it was not clear whether these findings would hold true for teachers teaching in other parts of the United States or in online contexts with fewer opportunities to build researcher–participant rapport.

A related strand of research has also shown that teachers attribute students' performance to different factors, depending on the students' race or gender (Graham, 2017; Graham & Williams, 2009; Reyna, 2008). K-12 teachers attribute the low performance of girls to low levels of natural ability and the failure of boys to their lack of effort (Espinoza et al., 2014; Fennema et al., 1990; Tiedemann, 2000, 2002). With regard to teachers' race-based attributions, some research finds that uncritical, overly positive feedback is more often provided to low-performing Black and Hispanic students compared with White students of the same achievement levels, particularly among White teachers with low school-based social support (Harber et al., 2012). Teachers' attributions of low student performance to a malleable factor, such as a lack of knowledge or effort, versus a factor outside their control, such as ability, have different psychological and behavioral consequences. For instance, on the one hand a teacher may attribute the low performance of a student to a lack of effort, which students can readily control and improve, which may elicit concern from the teacher and communicate to the student that they need to try harder. On the other hand, if the teacher attributes low performance to a lack of ability, a cause that is not easily controlled by the student, teachers may experience sympathy for the "helpless" student and lower their expectations, which can communicate to the student that they have low ability (Graham, 1984, 2017). And yet, while some studies have provided experimental evidence that teachers' ability attributions differed by the students' race and gender (e.g., Copur-Gencturk et al., 2020), no experimental studies have investigated whether teachers' effort attributions differed by student race or gender.

In sum, STEM-specific gender and racial biases are held by teachers at all levels of education from elementary to secondary school. Evidence from lab settings using the IAT, in settings where teachers evaluate fictitious students, and when teachers attribute reasons for their own students' performance all indicate that teachers inequitably associate boys with STEM, and attribute their successes to mathematical ability, and such biases are linked to teacher's decision-making and evaluations of student ability and effort. Furthermore, given that teacher bias may be present year-after-year, even potentially small effects of being underestimated by a single teacher has the potential to build up and snowball into larger gender disparities over time, potentially leading to the more apparent gender and racial disparities found in completion rates of math-intensive degrees in higher education (NCSES, 2023; NSF, 2020). Indeed, as children progress through school and are exposed to gender stereotypes by their teachers, they also internalize them, as evidenced by girls developing the perspective that boys are more likely to be "really really smart" starting at age six (Bian et al., 2017), and demonstrating male–math associations starting in second grade (Cvencek et al., 2011). Yet, as noted, of the studies on teacher bias that we have reviewed, most were conducted in countries other than the U.S., measured only gender bias, and were conducted in lab-based settings that do not necessarily capture how biases might operate in authentic teaching situations. Of the experimental studies that measure teacher bias in a more authentic setting (i.e., while evaluating student work; Copur-Gencturk et al., 2020), this prior research did not explore important moderators that could potentially play a role in the persistence of teacher bias. In the next section, we summarize existing research indicating potential moderators of STEM teacher bias.

## Potential moderators of teacher bias

When teachers evaluate student math work, they may implicitly or explicitly consider *why* the student was successful or not, and explain the outcome in terms of student or social characteristics (Jacobson et al., 2022). Teachers' beliefs, dispositions, and personal experiences are thought to explain why their decisions, evaluations, and attributions of students' performance differ by race and gender (Graham, 2017; Graham & Williams, 2009). Attributional theory posits that a number of antecedents can moderate attributions that people make to explain the performance of others, such as the individuals' past personal history, dispositions, beliefs about social norms, and their habits of attributing success to themselves and failures to others (also called "hedonic bias"; Graham & Williams, 2009). For example, a systematic review of 79 empirical studies identified several teacher factors that moderated their attributions for students' successful or unsuccessful performance (Wang & Hall, 2018). This included factors related to teachers' experiences such as teaching experience, level of education, and experience teaching students with learning disabilities, as well as personal dispositions and beliefs about the role of effort in academic success. Findings also confirmed that student characteristics of race, gender, and disability status also predicted teacher attributions to effort and ability. However, despite theory predicting that teacher factors and student factors are important components of attributional bias, little to no experimental studies have looked at relations between the two. As such, we aimed to explore whether teacher's prior experiences, beliefs, and dispositions would moderate their attributional gender and racial biases, as predicted by attributional theory.

Namely, we contend that teachers' dispositions and beliefs about social norms, their personal history of discrimination, and implicit theories of intelligence are all expected to shape whether an individual makes controllable vs. uncontrollable attributions. We now discuss four potential antecedents (beliefs about sexism, experiences of race and gender discrimination, implicit theories of mathematics intelligence, and mathematics anxiety) that might be expected to moderate such biases.

### *Explicit sexism and racism*

Social perceptions are considered antecedents to potentially biased effort or ability evaluations (Graham

& Williams, 2009). Explicit beliefs about sexism and racism are intentional, conscious, discriminatory attitudes and behaviors toward women and people of color. Explicit sexism and racism are typically measured by directly asking people to report their beliefs and attitudes about social groups, using surveys. Survey scales measuring of explicit bias can range from overt measures (e.g., agreement that "Women are generally not as smart as men") to more subtle measures such as the Modern Sexism and Modern Racism Scales which assess beliefs less directly, for example, by asking questions about the existence of gender discrimination in U.S. society (Swim et al., 1995); a measure of explicit sexism that has been used with teachers in educational contexts in the U.S. (e.g., Degner et al., 2019; Storage et al., 2020). Such explicit beliefs are considered to be subtle but explicit measures of biases with stronger beliefs that men and women are treated equally indicating higher levels of modern sexism.

Generally speaking, teachers report low levels of explicitly negative racial attitudes (Starck et al., 2020) and low levels of agreement with gender-specific stereotypes about mathematics (Carlana, 2019; Copur-Gencturk et al., 2021; Nürnberger et al., 2016). However, among the minority of teachers who do report biases, those explicit biases have been shown to be associated with other sets of beliefs and other forms of biases. For example, teachers tend to overwhelmingly disagree that "boys tend to be smarter than girls at math", but the minority of teachers who agree with such statements also hold essentialist beliefs that social categories are natural entities (Nürnberger et al., 2016) and believe that innate ability or "brilliance" is required for success in mathematics (Copur-Gencturk et al., 2021).

Explicit sexism and racism have also been linked to implicit forms of bias, though the results are mixed. In a systematic review of the literature, Denessen et al., (2022) found that, of the 23 studies that assessed both teachers' implicit and explicit bias, ten (43.5%) reported non-significant associations between them, eight (34.5%) reported mixed results wherein multiple explicit measures were included revealing some significant and some non-significant associations with implicit measures, and five (22%) reported significant positive relationships. However, it should be noted that many of these studies were not conducted in STEM-specific settings, and the few that were found positive but non-significant associations between explicit and implicit bias measures using the IAT (De Kraker Pauw et al., 2016; Nürnberger et al., 2016; Thomas, 2017; also see Carlana, 2019). However, few if any research studies have explored whether explicit biases might relate to how teachers make biased effort and ability ratings of students work. In this study, we explored whether teachers' modern sexist beliefs were related to gender and racial biases.

### *Personal experiences of race and gender discrimination*

Teachers' personal experiences of gender and racial discrimination in the mathematics classroom may be potential moderators of their biased ability and effort judgments. Based on prior research and theory, experiences of discrimination might sway teachers' perceptions of students in one of two ways. On the one hand, research on internalized oppression theorizes that teachers' prior experiences of discrimination might lead teachers to accept and perpetuate negative stereotypes that are imposed upon them by observers (e.g., Fredrickson & Roberts, 1997; Jost et al., 2004). For example, oppressed groups can internalize social stereotypes and a sense of inferiority in order to justify the existing social order and tend to be internalized to a greater extent among those who are targeted by the bias (Jost et al., 2004). For example, girls and women who are viewed by observers as an object and judged based on physical appearance internalize the outsider's view of themselves, and self-objectify by viewing themselves and other women as objects (Fredrickson & Roberts, 1997; Grogan, 2021). On the other hand, explicit awareness of one's own experiences of discrimination may be negatively related with bias, given that such awareness of discrimination is considered to be an indicator of the absence of bias (Henry & Sears, 2002; Starck et al., 2020; Swim et al., 1995). For example, Black mathematics and engineering students at a primarily White institution reported in interviews that, of the many motivations driving them to achieve and maintain their academic success over their academic career, their persistence through experiences of racism and exposure to racial stereotypes was one of the motivators toward high achievement in math and engineering (McGee & Martin, 2011). Similarly, a mathematics teacher who self-reports an experience of discrimination must be aware of that discrimination, and such awareness might represent resilience to stereotypes which could be negatively related with their implicit and explicit bias. However, despite these potential explanations for why relations may exist, few studies have investigated relations between personal experiences of discrimination and bias. As such, we sought to explore whether teachers' experiences of race and gender discrimination were related to teacher bias.

### *Implicit theories of mathematics intelligence*

Unarticulated beliefs about whether intelligence is malleable or fixed (also called growth- vs fixed-mindset; or implicit theories of intelligence) are considered to be important factors that influence how individuals view and persist through challenges and predict attributions people make for their own successes or failures and the successes and failures of others (Boaler, 2013; Weiner, 2005; Yeager & Dweck, 2012). For example, a teacher who believes that mathematical ability is fixed and unchangeable might interpret a student's failure on an exam to their lack of mathematical ability and infer that the student's failure is beyond their control, potentially leading them to perceive the student as hopelessly unskilled, and expend less effort to support this student as a result. Evidence suggests that elementary and secondary mathematics teachers disproportionately attribute natural ability as an explanation for their girl students' low performance and to their boy students' high performance (Espinoza et al., 2014; Fennema et al., 1990; Tiedemann, 2000, 2002). For this reason, we expected that teachers' beliefs that mathematical ability is a fixed entity would predict their gender biases favoring male students and their ability attributions for the success of boys and the failure of girls.

### *Mathematics anxiety*

Personality dispositions, such as trait-level mathematics anxiety, are expected to shape potentially biased ability and effort evaluations (Graham, 2017; Graham & Williams, 2009). Individual differences in the way that people interpret and explain events can shape their causal reasoning processes in the classroom. Trait-level mathematics anxiety, in particular, may be a personal disposition that shapes how teachers interpret failure and the failure of their students.

Mathematics anxiety can be defined as a relatively enduring disposition that is characterized by feelings of fear and anxiety in response to doing mathematics or considering the prospect of doing mathematics driven partly by a fear of failure (e.g., Ramirez et al., 2018b). Mathematics anxiety is linked to how people appraise mathematics experiences and outcomes, with high levels of anxiety being associated with interpreting failure as being an indicator of low ability (Dweck, 1975; Ramirez et al., 2018b; Wilson, 2011). Furthermore, findings show that early elementary female teachers' anxiety levels predicted their female students' beliefs that boys are good at math, which in turn predicted their students' mathematics achievement (Beilock et al., 2010). This suggests that teachers' math anxiety may play a role in the transmission of stereotypes from teacher to student. As such, we hypothesized that teachers' mathematics anxiety would moderate their gender and racial bias.

## Current study

As mentioned, few experimental studies have been conducted to investigate STEM teachers' biases as they may be revealed in actual classroom settings. Our study aimed to replicate prior work (e.g., Copur-Gencturk et al., 2020) by testing the generalizability of their findings with regard to teachers' biases in their evaluations of students' math performance and perceptions of math ability by collecting data from teachers in schools across the United States. Our study goes beyond prior work in two ways: (1) by exploring how teachers' beliefs and dispositions moderate their gender and racial biases and (2) by exploring teachers' gender and racial biases in their perceptions of student effort. Finally, we aimed to show transparency in our approach to the study design and analysis, which were carefully thought out and had clear purposes. This is why before we began the data collection, we preregistered the research questions, hypotheses, and planned analyses (see https://aspredicted.org/GNH_RXF for preregistration). Namely, we aimed to answer the following questions:

1. Are there systematic differences in teachers' evaluations of student performance (i.e., grading of student work), their evaluations of the effort they assume students put into the work, and their estimations of students' mathematical ability that could be explained by the students' gender or race?
2. To what extent are teachers' beliefs and dispositions (mathematics anxiety, beliefs about mathematical intelligence, levels of modern sexist beliefs, perceptions of being underestimated because of race or gender) related to such biases?

Regarding our first research question, given that teachers' ratings for the correctness of students' solutions did not differ by the students' gender or race in prior work (e.g., Copur-Gencturk et al., 2020, 2022), as stated in the preregistration, we did not anticipate finding gender or racial bias in teachers' evaluations of students' written work (H1). However, we did expect to find that teachers would rate the students' ability higher when a White or male student name appeared on work in situations where there was ambiguity in the work (i.e., when the students' work is not completely correct) compared with a non-White (Hispanic/Black) or girl name. This prediction was based on theory positing that people tend to attribute the successes of members of stereotyped groups to effort and non-stereotyped groups to ability (Graham, 2017), and is consistent with previous findings (e.g., Copur-Gencturk et al., 2020; Tiedeman, 2000; 2002). For the same reasons, we also expected that teachers would rate the effort of students' work higher for female students' work.

Regarding our second research question, we hypothesized that (H2a) teachers who had higher levels of mathematics anxiety may draw on their gender and racial biases more often than other teachers when evaluating students' work because mathematics anxiety can impair a person's cognitive ability (e.g., Beilock et al., 2010; Ramirez et al., 2018a). We also hypothesized that (H2b) teachers who believed that mathematical intelligence is fixed and innate would tend to show more gender bias because prior research suggests that teachers who believe that mathematical ability is fixed and innate also believe that boys, but not girls, have this ability (Copur-Gencturk et al., 2021). We had no specific preregistered hypotheses for the modern sexism moderator. We added modern sexism to our model to explore whether teachers' beliefs that gender disparities exist in current society were related to gender bias.

## Methods

### Study context

We partnered with an education marketing company to create a random sample of elementary and middle school teachers from across the U.S. The teachers were contacted via email to inform about study and asked them to participate (see the "Recruitment Details" section in Appendix A of Additional file 1 for details). To ensure the data were collected from teachers who were teaching mathematics at the elementary or middle school level during the data collection period, teachers were asked to answer a set of screening questions. Our records indicate that 58.5% of the data were collected from teachers who were teaching mathematics at the elementary school level, whereas the remaining data were from middle school mathematics teachers.

Table 1 presents the background characteristics of the teachers in the study sample along with a nationwide representative sample of U.S. public elementary and secondary teachers. Our sample is similar to the nationally representative sample of U.S. teachers in terms of age, certification type, and geographic locations of their schools. Yet our sample includes more experienced teachers as well as a greater percentage of female and White teachers.

**Table 1** Background characteristics of teachers in the present sample compared with a nationwide sample

| Teacher or school characteristic | Study sample (%) | Nationwide sample of U.S. public school teachers (%) |
|---|---|---|
| Sex[a] | | |
| Female | 84.9 | 76.5 |
| Male | 13.8 | 23.5 |
| Prefer not to say | 1.3 | N/A |
| Race/ethnicity[a] | | |
| White | 86.5 | 79.3 |
| Black | 5.7 | 6.7 |
| Hispanic | 5.0 | 9.3 |
| Other | 2.8 | 4.6 |
| Age[a] | | |
| Under 30 | 12.0 | 15.0 |
| 30–39 | 34.3 | 27.9 |
| 40–49 | 31.4 | 29.0 |
| 50–59 | 18.8 | 20.7 |
| 60 and over | 3.5 | 7.4 |
| Regular certification[a] | 96.9 | 90.4 |
| Years of teaching experience[a] | | |
| Less than 3 | 1.3 | 9.0 |
| 3–9 | 27.3 | 28.3 |
| 10–20 | 46.5 | 39.9 |
| More than 20 | 24.9 | 22.8 |
| School region[b] | | |
| Midwest | 27.5 | 21.4 |
| Northeast | 16.4 | 19.5 |
| South | 37.3 | 40.3 |
| West | 18.8 | 18.8 |

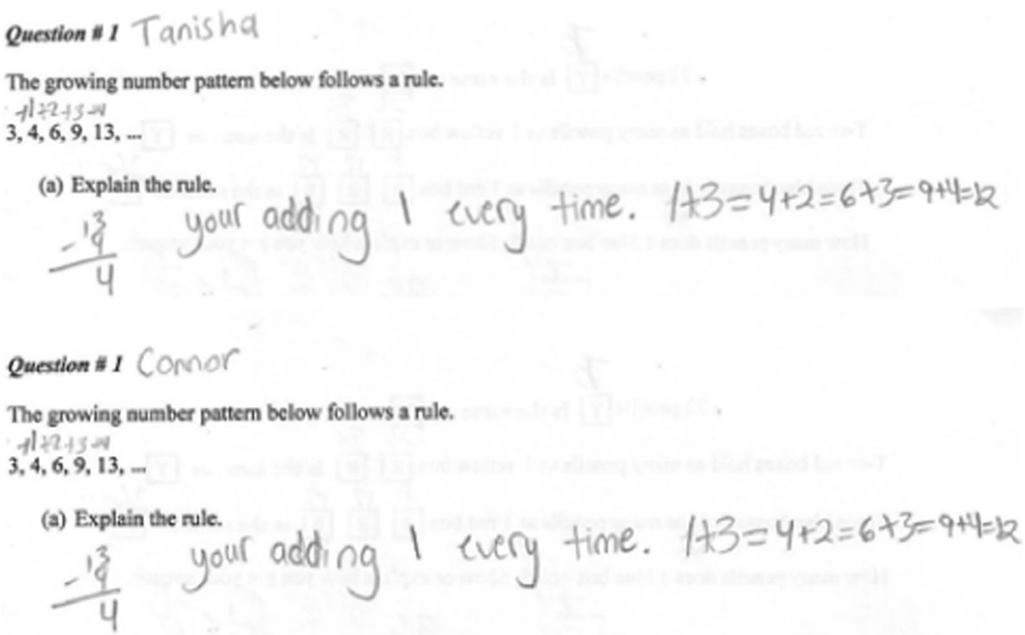
[a] From *Digest of Education Statistics*, https://nces.ed.gov/programs/digest/d20/tables/dt20_209.22.asp

[b] From *Common Core of Data: America's Public Schools*, https://nces.ed.gov/ccd/tables/201920_summary_2.asp

The participating teachers were told a deceptive story similar to the one used by Copur-Gencturk and colleagues (2020), namely, that we had field-tested a set of items to create an assessment for middle school children that could capture their advanced problem-solving abilities and mathematical reasoning skills, which could reveal their mathematical ability. We stated that we needed the teachers' help to identify the problems that will be included in the final assessment. As in the previous study, our aim in using this story was to more accurately capture teachers' potential biases. All the teachers in the study were given the same 18 student responses that were used in the original study (see Copur-Gencturk et al., 2020, for further details). After teachers evaluated the student work, they were asked a set of questions regarding their background.

### Survey development

We used the same 18 student responses that were used in the study by Copur-Gencturk and colleagues (2020). The solutions came from real students who answered three National Assessment of Educational Progress (NAEP) problems. For each problem, two responses were incorrect, two were partially correct, and two were fully correct. Prior to the full study, we assessed whether teachers associated the student names used in the original study with the targeted gender and race. In a separate



***Question # 1*** Tanisha

**The growing number pattern below follows a rule.**

+1 +2 +3 +4

**3, 4, 6, 9, 13, ...**

**(a) Explain the rule.**

$-13$ 9 / 4 your adding 1 every time. 1+3=4+2=6+3=9+4=12

***Question # 1*** Connor

**The growing number pattern below follows a rule.**

+1 +2 +3 +4

**3, 4, 6, 9, 13, ...**

**(a) Explain the rule.**

$-13$ 9 / 4 your adding 1 every time. 1+3=4+2=6+3=9+4=12

**Fig. 1** A partially correct student solution assigned different names

investigation, 139 teachers were presented with 60 names that were selected to differ by race and gender, with the original names among them (for details, see "Names Selection" in Appendix A of Additional file 1). Teachers indicated whether they associated each name with a different race (Black, White, Hispanic), gender (girl, boy), and SES (high, medium, low), and were given the option to state "I don't know" for each. Teachers largely associated the original 18 names of the students with the targeted gender and race. As such, we used the same student names as in the original study. As in the original study, we created 24 different survey forms in which the gender- and race-associated names were randomly assigned to these student solutions, with an equal number of incorrect, partially correct, and correct solutions associated with each gender and race (see Fig. 1).

Although our study design was similar to the one by Copur-Gencturk and colleagues (2020), our survey differed from theirs in two ways. First, we added information to the outcome measures to help teachers make ranking decisions based on common reference points (i.e., letter grades for the correctness of student solutions, and fifth graders as a reference point for the mathematical ability of students). By including the same reference points across teachers, we aimed to reduce the potential for teachers to change the rating scales depending on the group they were evaluating. For instance, teachers who held biases against a certain group of students might lower their expectations of that group or attribute the group's successes to effort and their failures to ability (e.g., Fennema et al., 1990). Without establishing a fixed group (here, all fifth graders) as the reference, teachers might have been more likely to raise or lower expectations in accordance with their own biases. Second, we asked teachers to evaluate each student solution based on the amount of effort they thought the student had put into it because prior research has suggested that teachers attribute effort to girls' successes more than with boys (Espinoza et al., 2014; Fennema et al., 1990; Tiedemann, 2000, 2002). Thus, by comparing teachers' effort ratings for an identical student solution, with the only difference being the assigned names, we would be able to capture teachers' potential biases more accurately.

Given this rationale, we modified the measures created by Copur-Gencturk et al. (2020) to assess teachers' judgments of correctness and mathematical ability, and also created a scale for teachers to judge student effort. Although prior studies have used scales to measure teachers' effort and ability attributions for successes and failures of their lowest and highest achieving students in mathematics (Espinoza et al., 2014; Tiedemann, 2000, 2002), a key difference in our study is that we asked teachers to evaluate *fictitious* students rather than their own students. Given the relative novelty of our measures, we conducted cognitive interviews with seven in-service teachers to ensure that the modified items and language from additional scales were interpreted by teachers in the way we had intended (see "Item Development" in Appendix A of Additional file 1).

**Table 2** Descriptive statistics for the beliefs and experiences of teachers in the study

| | **Sample (*N* = 458)** | | **Sample based on predefined criteria (*N* = 413)** | |
|---|---|---|---|---|
| | **Without weights** | **With weights** | **Without weights** | **With weights** |
| Teacher characteristic | | | | |
| Female | 0.85 (0.36) | 0.86 (0.35) | 0.85 (0.36) | 0.86 (0.34) |
| White | 0.86 (0.34) | 0.86 (0.35) | 0.86 (0.34) | 0.86 (0.35) |
| Black | 0.06 (0.23) | 0.06 (0.24) | 0.05 (0.22) | 0.05 (0.23) |
| Hispanic | 0.05 (0.22) | 0.05 (0.22) | 0.05 (0.22) | 0.05 (0.22) |
| Other race | 0.03 (0.17) | 0.03 (0.18) | 0.03 (0.17) | 0.04 (0.19) |
| Full credential | 0.97 (0.17) | 0.98 (0.15) | 0.98 (0.15) | 0.98 (0.13) |
| Experience (in years) | 14.96 (7.71) | 14.84 (7.73) | 15.22 (7.77) | 15.14 (7.80) |
| Suspicion weight | 0.82 (0.35) | 0.97 (0.12) | 0.81 (0.36) | 0.97 (0.12) |
| Moderators | | | | |
| Modern sexism | 2.33 (0.75) | 2.40 (0.74) | 2.35 (0.76) | 2.41 (0.75) |
| Fixed mindset | 1.68 (0.66) | 1.72 (0.65) | 1.68 (0.66) | 1.71 (0.65) |
| Experiences of sexism | 1.71 (0.92) | 1.68 (0.89) | 1.74 (0.94) | 1.70 (0.91) |
| Experiences of racism | 1.18 (0.46) | 1.18 (0.47) | 1.17 (0.44) | 1.17 (0.44) |
| Math anxiety | 2.03 (0.80) | 2.01 (0.79) | 2.06 (0.80) | 2.04 (0.79) |
| Outcomes | | | | |
| Correctness | 7.46 (4.05) | 7.55 (4.09) | 7.50 (4.04) | 7.60 (4.07) |
| Effort | 5.45 (2.01) | 5.46 (2.03) | 5.46 (2.01) | 5.47 (2.02) |
| Ability | 63.56 (28.03) | 64.07 (28.22) | 63.62 (27.86) | 64.19 (27.94) |

The values shown are means, followed by standard deviations in parentheses. The Other race category included teachers from other races, such as Asian, Native American, or two or more races

### Analytic sample

Of the 458 teachers who completed the survey, we restricted our analysis to the teachers who completed the survey in a reasonable amount of time ($N = 413$). The exclusion criterion we set in advance was to remove the data of teachers whose time spent evaluating students' solutions fell outside the 5th and 95th percentile of the distribution solutions. The rationale behind this decision was, similar to that of Copur-Gencturk and colleagues (2020), that teachers who completed the survey who spent very little time or who took a very long time on the students' work might not have paid attention to the study names or might have forgotten the instructions. Those who were excluded were not statistically different from those included in the analyses in terms of gender, $\chi^2(2) = 1.28$, $p = 0.53$, or race, $\chi^2(3) = 4.79$, $p = 0.19$ (see Table 2 for the descriptive statistics for the full and analytic sample). Additionally, teachers' scores on the fixed mindset scale, $t(456) = 0.16$, $p = 0.87$; Cohen's $d = 0.02$; the modern sexism scale, $t(456) = -1.05$, $p = 0.30$ Cohen's $d = -0.16$, their underestimations of mathematical ability attributable to gender or race [$t(456) = -1.64$, $p = 0.10$; Cohen's $d = -0.26$ for gender and $t(456) = 1.05$, $p = 0.29$; Cohen's $d = 0.17$] were not statistically different between those who met the exclusion criteria and those in the analytic sample. However, teachers who met the exclusion criteria had less teaching experience than those in the analytic sample, $t(456) = -2.17$, $p = 0.03$; Cohen's $d = 0.34$, and had higher scores on their math anxiety, $t(456) = -2.12$, $p = 0.03$; Cohen's $d = -0.33$.

### Suspicion check

Given that some teachers might be prone to social desirability and not present their actual response patterns to the researchers from whom they heard about the study for the first time through an email, we designed precautionary steps to ensure the validity of the data collected. Thus, we included suspicion check items in the survey so that we would be able to identify the teachers who might have become suspicious of the purpose of the study and altered their responses. The two suspicion check items were adapted from a prior work that also involved deception (Blackhart et al., 2012; see the Suspicion Check Items section in Appendix B of Additional file 1). The first item was an open-ended response to the following prompt: "In your own words, what is the present study about?" Participants then responded to the statement, "Did you believe, at any time, that the study had a purpose other than what the researchers had described to you?" with a "Yes" or

**Table 3** Description of suspicion check codes and sample responses

| Category | Description | Sample responses |
|---|---|---|
| Suspicious | Clear indication or evidence in a teacher's response that he or she is suspicious that the study captured his or her implicit bias in evaluating student work | "I guessed that you're looking to see if my responses change based on the perceived gender and/or ethnicity of the students."<br>"It showed names of students that reflect cultural and ethnic backgrounds as well as gender." |
| Not suspicious | No clear indication or evidence in a teacher's response that he or she is suspicious that the study captured his or her implicit bias | "How the thought process of male and female students determines their performance on a mathematics assessment."<br>"Comparing different ethnicities' performance on math assessments. Also, looking at how educators evaluate student work." |

"No" response. Participants who responded "Yes" were also asked to "Please explain" in an open-ended response.

We had identified these teachers before conducting the planned data analysis by coding their responses to all the suspicion check items. We first defined what constituted evidence of suspicion (i.e., a clear indication of the teacher's acknowledgment that the researcher intended to capture gender and racial differences in the teachers' ratings) and then defined the two scoring categories: suspicious and not suspicious (see Table 3). This process was followed by the development of a training document that also included sample responses and the rationale behind our codes (see Additional file 1 for the raters' training materials).

Five raters independently coded the data each teacher produced by answering these items. The first author met with the external raters to clarify any questions they might have about the coding process or the training document that was previously shared with them. The kappa statistic across the five raters was 0.79, indicating that we reliably identified the suspicious teachers. In fact, all five raters agreed that 55 teachers (12.0%) showed evidence that they knew the study was designed to capture their gender or racial biases in their evaluations, and all five raters agreed that 342 teachers (74.7%) did not indicate that they became suspicious of the purpose of the study. Still, to show the transparency of our work, we reported the results regardless of whether teachers showed any suspicion that the study focused on teachers' potentially different evaluations based on students' gender and race (see Additional file 1: Tables S4, S6, and S8).

### Power analysis

To ensure that the study had enough power to capture teachers' biases as well as the role of moderators in these biases, we conducted a power analysis. One of the central aspects of our design involves a cross-level interaction effect for moderation, and it is recognized in the multilevel statistical literature that power to detect such interaction effects hinges on a range of factors at both levels of the study design (Mathieu et al., 2012; Scherbaum & Ferreter, 2009; Snijders & Bosker, 2011). Indeed, Scherbaum and Ferreter (2009) note that "estimates of statistical power of cross-level interactions are much more complex than the computations for simple main effects or variance components … and there is little guidance that can be provided in terms of simple formulas" (p. 363).

Given these limitations and complexities, simulations that mimic the data-generating process are often used to determine power for cross-level interactions. We conducted Monte Carlo simulations because neither traditional (e.g., G*Power) nor modern (e.g., Optimal Design, PowerUp!) power-analysis software packages fully capture the unique design elements of our study or allow us to estimate moderation effects. We begin by specifying the data-generating model, based on prior literature and the minimum effect sizes we seek to have power to detect. For all analyses, we perform two-level random effects analyses and assume a two-tailed hypothesis test with $\alpha = 0.05$. Each power analysis is based on 1000 simulations.

When correctness or effort are the outcomes, there are 18 level-1 items nested within 400 level-2 (e.g., simulated teacher) observations, for a total of 7200 observations. Additionally, the intraclass correlation coefficients (ICCs) are set at 0.07, a value reflecting prior studies in this area. Given these assumptions for correctness and effort, we have 85% and 82% power to detect gender and race effects of 0.07 SDs, respectively.

When ability is the outcome, there are two necessary changes: first, the preregistered hypotheses concern only 12 of the 18 level-1 items, so we reduce the level-1 observations correspondingly. Second, our preregistered analysis follows the approach of Copur-Gencturk et al. (2020)

and adds the teacher's rating of the solution correctness as a covariate in the model. Based on that prior work, the correctness rating is highly predictive of the ability rating (equivalent to a level-1 $R^2$ of 0.72), but teachers do not vary considerably in their average ratings of correctness. As such, adding correctness as a covariate greatly reduces the level-1 error variance, while leaving the level-2 error variance largely unchanged. This results in an ICC of 0.25 for the model when ability is the outcome. Consequently, precision improves for ability under these assumptions, so that we have 86% and 84% power to detect gender and race effects of 0.04 SDs, respectively.

Turning to moderation, we assume a gamma distribution for the moderators, more realistically reflecting that many of the moderators are skewed and not normally distributed: for example, teachers tend to be concentrated on the reported growth-orientation side of the mindsets scale. When correctness or effort are the outcomes, under our assumptions, we have 86% and 81% power to detect a moderation effect of 0.07 SDs on gender and race, respectively. When ability is the outcome, and given the above assumptions, we have 86% and 80% power to detect a moderation effect of 0.04 SDs on gender and race, respectively.

Finally, it is worth noting that one of the preregistered hypotheses concerns a small subgroup—namely, non-White teachers. If we assume that 20% of the teachers are non-White, then for testing the hypothesis that non-White teachers underestimate non-White student ability, the number of level-2 observations decreases to 80. This greatly reduces power for this research question, such that instead of the 84% power to detect a race effect of 0.04 SDs in the overall study, we only have 24% power to detect the same sized effect among this subsample.

Overall, given the stated assumptions, the study is well powered to detect main and moderation effects in the range of 0.04–0.07 SDs, effect sizes that are quite small. The one exception to this is analysis on the subsample of non-White teachers, which is underpowered.

## Measures

### *Correctness*

Teachers were asked to evaluate the mathematical soundness of each student solution by assigning a grade on a 13-point scale, ranging from $F$ to $A+$.

### *Mathematical ability*

After teachers rated the correctness of a given solution, they were asked to estimate the mathematical ability of that student compared with the population of U.S. fifth graders on a 100-point scale, ranging from *very low mathematical ability* to *very high mathematical ability*.

### *Effort*

Teachers were also asked to "evaluate the level of effort evident in the student's response" on an 8-point scale, ranging from "*The student put minimum effort into this response*" to "*The student put maximum effort into this response.*"

### *Student gender and race*

Each student solution was randomly assigned to a female or male name associated with being Black, White, or Hispanic. We created variables for implied gender (boy versus girl) and implied race (White versus Black or Hispanic).

### *Display order*

The order of the student solutions was randomized, and a variable was created indicating the order in which teachers evaluated the student work.

### *Modern sexism*[2]

The modern sexism scale (Swim et al., 1995) consisted of eight items used to measure whether teachers believed that gender disparities exist in society (e.g., "Women often miss out on good jobs due to sexual discrimination"; α = 0.86). Teachers responded on a 5-point agreement scale. We used standardized mean scores in our analysis. Higher values on this scale represented greater agreement that gender discrimination was not present in current society.

### *Fixed mindset*

To measure teachers' perceptions that mathematical intelligence is fixed, we used a scale adapted from Dweck (1999) that consisted of eight items. Teachers reported their beliefs about mathematical intelligence being fixed (e.g., "Your math intelligence is something about you that you can't change very much") on a 5-point agreement scale (α = 0.91). We used standardized mean scores, with higher values indicating that teachers believed mathematical ability was innate/fixed.

### *Experiences of discrimination based on gender or race*

Teachers also reported their own experiences of gender or racial discrimination (adapted from Torres et al., 2010) by stating their experiences of being underestimated in mathematics because of their gender while in Grades K-12 (four items for gender and racial discrimination,

[2] We did not measure teachers' modern racism beliefs because we were concerned that the items on the modern racism scale (McConahey, 1986) were worded in a way that might sound offensive to teachers of color. We were also concerned that these items might signal to teachers the purpose of our study. See the Limitations section for more detail.

respectively α = 0.94 for gender and α = 0.86 for race). Responses ranged from 1 (*never*) to 5 (*very frequently*). The standardized mean scores were used in the analyses. Higher values on this scale represented teachers who felt their mathematical ability was frequently underestimated based on their gender or race.

#### Mathematics anxiety

Teachers also completed a 9-item mathematics anxiety scale (adapted from Ganley et al., 2019) that captured their mathematics-related feelings of panic, worry, and self-consciousness. All statements were rated on a scale from 1 (*Not true of me at all*) to 5 (*Very true of me*; α = 0.93). We used standardized mean scores in our analysis, with higher values on this scale indicating teachers reported a higher level of mathematics anxiety.

#### Suspicion-check weight

We created a weight variable ranging from 0 to 1 for the number of raters who identified a teacher as suspicious. For example, if all five raters identified a teacher as suspicious, that teacher's data were given an analytic weight of 0 (i.e., they were essentially dropped from the analysis because they were deemed suspicious by all raters), and if four out of five raters identified a teacher as suspicious, that teacher's data were given a weight of 0.2 (i.e., some, but not much, weight because most raters thought they were suspicious). If none of the raters identified a teacher as suspicious, that teacher's data were given a weight of 1 (i.e., the full analytic weight). To show transparency in our work, we report the unweighted results (see Additional file 1: Tables S4, S6, and S8).

#### Analytic approach

For each outcome of interest, we examined whether teachers' evaluations of students' correctness, ability, or effort differed by the students' gender and race by conducting 2-level hierarchical linear modeling (HLMRaudenbush & Bryk, 2002; Snijders & Bosker, 2011) in which teachers' ratings of students' solutions ($\text{TchrRating}_{it}$) were predicted by student's gender ($\beta_{1t}$), and race ($\beta_{2t}$) in Level 1 along with the item and item order as fixed effects. In all these models, the intercept was estimated as being random, whereas all the slopes were estimated as having fixed effects. This approach allowed us to take into account the nested structure of the data (i.e., each teacher evaluated a set of student solutions).[3] When we investigated teachers' evaluations of the correctness of students' solutions and students' effort, we combined incorrect, partially correct, and fully correct solutions, given that when we preregistered our planned analysis, we did not assume that teachers' grading and effort ratings would be different depending on the solution levels:

$$\begin{aligned}\text{TchrRating}_{it} = &\beta_{0t} + \beta_{1t}\text{student\_gender}_{it}\\ &+ \beta_{2t}\text{student\_race}_{it}\\ &+ \sum \text{item} + \sum \text{itemorder} + \varepsilon_{it},\end{aligned}$$

$\beta_{0t} = \gamma_{00} + \omega_{ot}, \beta_{kt} = \gamma_{k0}$ for all $k \neq 0$.

While examining the extent to which students' race and gender impacted teachers' ability ratings, we had expected, from drawing on prior literature, that teachers' biases about students' mathematical ability would be revealed when a student's solution was ambiguous. Thus, we investigated whether there was a systematic difference in teachers' ratings of students' ability when students' work was not completely correct (i.e., partially correct or incorrect).[4] Additionally, as we specified in the preregistration, we had planned on adjusting for differences in perceptions of solution correctness to improve the precision of the estimate, consistent with the approach in the original study by Copur-Gencturk et al. (2020). Thus, as shown in the equation below, we also included teacher-centered correctness ratings as a Level-1 predictor ($\beta_{3t}$) and the teacher's mean correctness ratings in Level 2 ($\gamma_{01}$).

$$\begin{aligned}\text{Ability\_rating}_{it} = &\beta_{0t} + \beta_{1t}\text{student\_gender}_{it}\\ &+ \beta_{2t}\text{student\_race}_{it}\\ &+ \beta_{3t}\text{Correctness\_rating}_{it}\\ &+ \sum \text{item} + \sum \text{itemorder} + \varepsilon_{it},\end{aligned}$$

$\beta_{0t} = \gamma_{00} + \gamma_{01}\text{Mean\_correctness\_ratings}_t + \omega_{ot}, \beta_{kt} = \gamma_{k0}$ for all $k \neq 0$.

The second research question aimed to investigate the extent to which teachers' beliefs and dispositions (e.g., mathematics anxiety) were moderating their gender or racial biases. To do so, each of the teacher's belief and disposition indicators was added as a Level-2 variable to the aforementioned equations to predict the average differences in teachers' ratings (i.e., intercept) as well as the slope for the student's gender or race.[5] Specifically, the following model was tested for teachers' beliefs

[3] To increase the precision of our estimates, we added item and item order as fixed effects. We also conducted the reported analyses without the fixed effects and obtained similar results.

[4] To show transparency in our approach, we have also reported the weighted and unweighted results for different levels of student solutions separately for the full and analytic sample in Additional file 1: Tables S1–S3.

[5] We centered the moderators around their means for interpretability of the results.

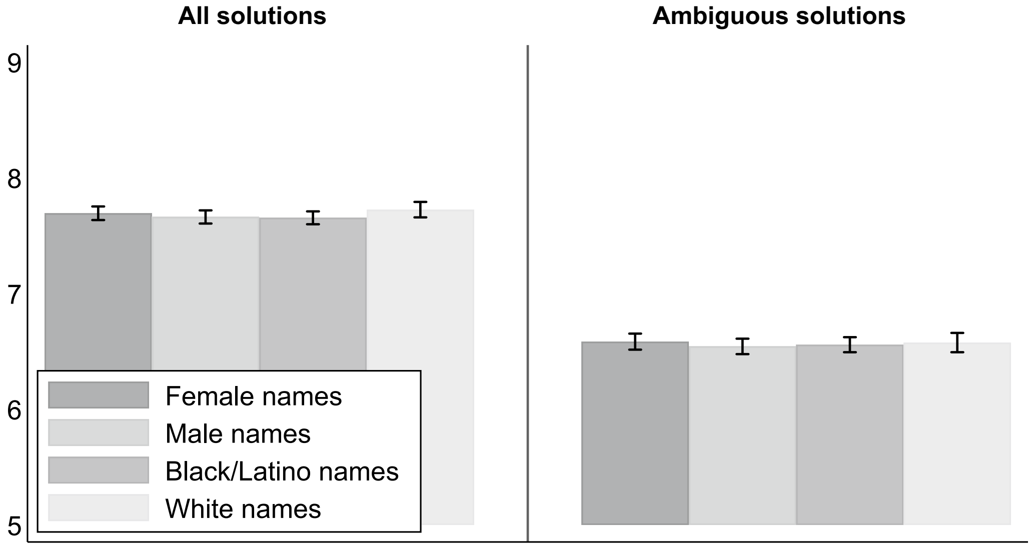


**Fig. 2** Mean correctness score for student solutions by level of correctness, gender, and race of student names. Heteroskedastic-robust standard errors clustered on teachers appear as bars around the mean estimates. Models also include controls for item and item positioning on the questionnaire

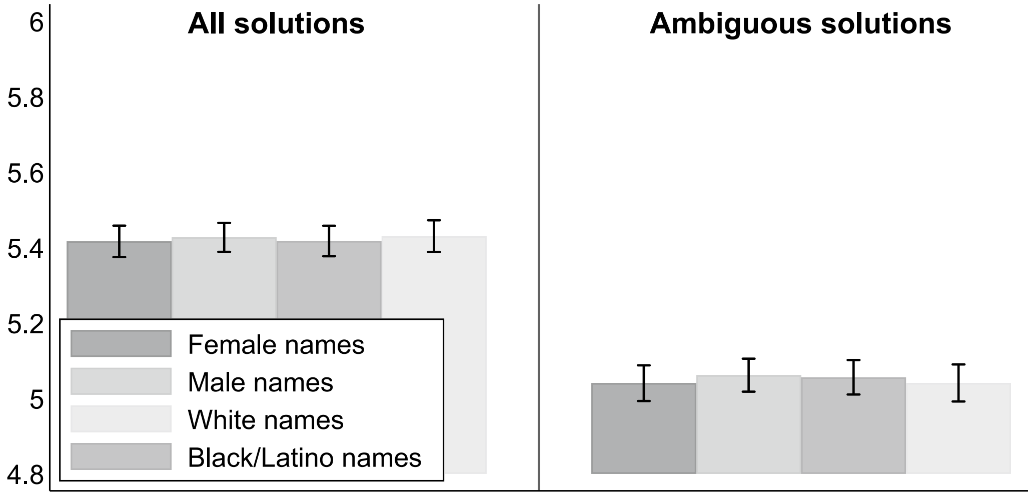


**Fig. 3** Mean effort ratings for student solutions by level of correctness, gender, and race of student names. Heteroskedastic-robust standard errors clustered on teachers appear as bars around the mean estimates. Models also include controls for item and item positioning on the questionnaire

and dispositions and how the gender differentials were moderated:

$$\text{Teacher\_rating}_{it} = \beta_{0t} + \beta_{1t}\text{studen}t\text{\_gender}_{it} + \beta_{2t}\text{student\_race}_{it} + \sum \text{item} + \sum \text{itemorder} + \varepsilon_{it},$$

$$\beta_{0t} = \gamma_{00} + \gamma_{01}T\text{\_moderator}_{t} + \omega_{ot},$$

$$\beta_{1t} = \gamma_{10} + \gamma_{11}T\text{\_moderator}_{t},$$

$\beta_{kt} = \gamma_{k0}$ for all $k \neq \{0, 1\}$.

A similar model was run to examine teachers' racial bias, in that the slope for the student race (the $\beta_{2t}$ coefficient) was predicted by the moderators, rather than the $\beta_{1t}$ having the moderators.

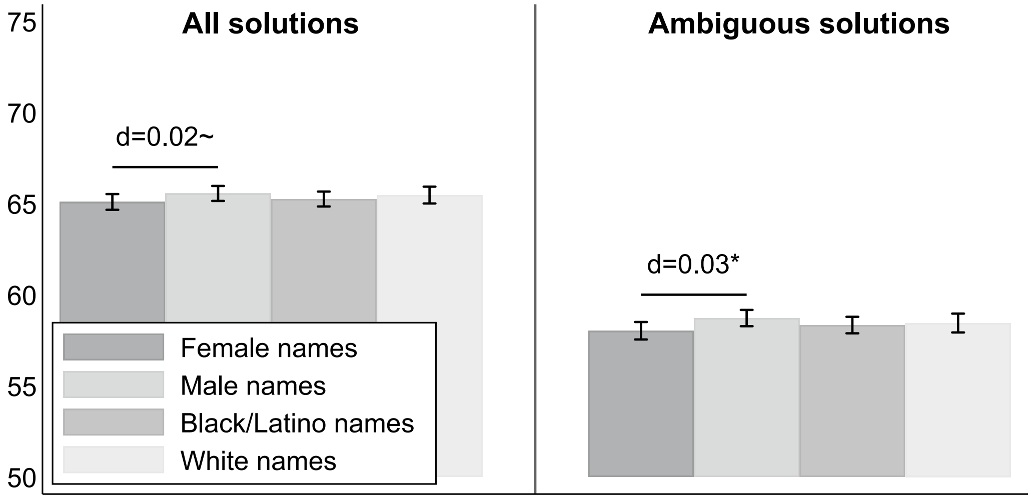


**Fig. 4** Mean ability ratings for ambiguous student solutions by gender and race of student names. Heteroskedastic-robust standard errors clustered on teachers appear as bars around the mean estimates. Models also include controls for correctness ratings, item, and item positioning on the questionnaire

## Results

### Teachers' evaluations of students' work, effort, and ability

Our analysis of teachers' evaluations of the correctness of students' work suggested that teachers did not grade student work differently based on the students' gender and race (see Fig. 2). Similarly, teachers' evaluations of students' effort did not differ by the students' gender or race (Fig. 3). In contrast, teachers' ratings of students' mathematical ability favored boys in ambiguous situations (i.e., when solutions were not completely correct) when their own evaluations of the correctness of the students' work were taken into account (see Fig. 4). Teachers' ratings of male students' ability were, on average, 0.7 points higher than their ratings of girls' ability (effect size of 0.03 *SD*,[6] $p = 0.025$). However, the difference in teachers' ability ratings of boys and girls was not significant when the ratings of suspicious teachers were not weighted (effect size of 0.02 *SD*, $p = 0.144$). Disaggregating the ambiguous solutions into the incorrect and partially correct solutions reveals that the gender gap is larger for the incorrect than the partially correct solutions. This effect is most pronounced for the incorrect solutions, but the difference is only statistically significant in the analytic sample with weights (effect size of 0.04 *SD*, $p = 0.025$).

### The moderating role of teachers' beliefs and dispositions in their evaluations of students' work, effort, and ability

Of the five moderators (teachers' sexist beliefs, fixed mindset about mathematical ability, self-reported underestimations of mathematical ability attributable to gender or race, and mathematics anxiety), we found that none moderated teachers' evaluations of the correctness of student work or the effort they assumed students put into their work (see Tables 4 and 5).

In terms of teachers' biases about students' mathematical ability, however, teachers' sexist beliefs and math anxiety moderated their ability ratings for boys and girls (see Table 6). As shown in Fig. 5, teachers with some of the highest reported beliefs (i.e., above the 75th percentile) that gender discrimination is no longer a problem in society gave higher ability ratings to the same student work when a boy name was assigned; in contrast, among teachers who reported believing that gender discrimination was a bigger problem (i.e., those below the 50th percentile), there was no statistically significant differences in how they rated boys' and girls' math ability. On the other hand, the more math anxiety teachers reported having, the higher they rated girls' math ability; nevertheless, the differences in the ability

[6] We calculated the effect sizes by dividing the coefficient for the gender variable by the standard deviation of the outcome variable (i.e., similar to Cohen's *d*). There are other approaches of calculating effect sizes such as dividing the coefficient by the standard deviation of the level-1 variance, given that gender is a level-1 variable (Hedges, 2007). Copur-Gencturk and colleagues (2020) divided the estimate by a measure of the level-1 variance after adjusting for covariates, which can be interpreted as the standard deviation remaining within an individual after accounting for other factors. Our rationale for calculating the effect size by using the standard deviation of the outcome variable was to make it parallel to the way effect sizes were calculated in the power analysis. By doing so (i.e., dividing by the standard deviation of the outcome variable rather than dividing by the square root of the level-1 error variance), the denominator is larger and the effect size appears smaller. Thus, the estimates presented here are smaller than in Copur-Gencturk et al. (2020) in part due to the effect size calculation. Because we provide all the variance decomposition information (as suggested by Hedges, 2007), readers can easily move between the metrics. For example, if we used the remaining level-1 error variance as the denominator here (from column 2 of Table 6), the effect size of gender bias is 0.06 SD [=0.70/sqrt(145.32)].

**Table 4** HLM results for teachers' evaluations of the correctness of students' solutions and the moderating roles of teachers' beliefs and dispositions

| Predictor | No moderator | | Modern sexism | | Fixed mindset | | Experiences of sexism | | Experiences of racism | | Math anxiety | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Boys | − 0.03 | − 0.03 | | − 0.03 | − 0.03 | − 0.03 | − 0.03 | − 0.03 | − 0.03 | − 0.03 | − 0.03 | − 0.03 |
| | (0.06) | (0.06) | | (0.06) | (0.06) | (0.06) | (0.06) | (0.06) | (0.06) | (0.06) | (0.06) | (0.06) |
| White | 0.07 | 0.07 | | 0.07 | 0.07 | 0.07 | 0.07 | 0.07 | 0.07 | 0.07 | 0.07 | 0.07 |
| | (0.07) | (0.07) | | (0.07) | (0.07) | (0.07) | (0.07) | (0.07) | (0.07) | (0.07) | (0.07) | (0.07) |
| Moderator | | 0.02 | | 0.06 | − 0.02 | − 0.06 | 0.05 | 0.04 | − 0.09 | − 0.10~ | 0.09~ | 0.08 |
| | | (0.06) | | (0.06) | (0.06) | (0.05) | (0.06) | (0.05) | (0.06) | (0.06) | (0.05) | (0.06) |
| Boys* moderator | | 0.10 | | | 0.01 | | − 0.04 | | 0.03 | | − 0.01 | |
| | | (0.06) | | | (0.06) | | (0.06) | | (0.05) | | (0.06) | |
| White* moderator | | | | 0.02 | | 0.12 | | − 0.04 | | 0.09 | | − 0.01 |
| | | | | (0.07) | | (0.08) | | (0.07) | | (0.07) | | (0.07) |
| Intercept | 8.45*** | 8.45*** | | 8.45*** | 8.45*** | 8.45*** | 8.45*** | 8.45*** | 8.45*** | 8.45*** | 8.45*** | 8.45*** |
| | (0.23) | (0.23) | | (0.23) | (0.23) | (0.23) | (0.23) | (0.23) | (0.23) | (0.23) | (0.23) | (0.23) |
| Variance intercept | 0.49 | 0.49 | | 0.49 | 0.49 | 0.49 | 0.49 | 0.49 | 0.49 | 0.49 | 0.49 | 0.49 |
| | (0.07) | (0.07) | | (0.07) | (0.07) | (0.07) | (0.07) | (0.07) | (0.07) | (0.07) | (0.07) | (0.07) |
| Variance residual | 6.27 | 6.26 | | 6.27 | 6.27 | 6.26 | 6.26 | 6.26 | 6.26 | 6.26 | 6.27 | 6.27 |
| | (0.20) | (0.20) | | (0.20) | (0.20) | (0.20) | (0.20) | (0.20) | (0.20) | (0.20) | (0.20) | (0.20) |
| ICC | 0.07 | 0.07 | | 0.07 | 0.07 | 0.07 | 0.07 | 0.07 | 0.07 | 0.07 | 0.07 | 0.07 |

The weighted results are reported here. Values in parentheses indicate robust standard errors clustered at the teacher level. The model also includes the item and item order as fixed effects. Scores for students' correctness could range from 1 (*F*) to 13 (*A*+)

$N = 413$. $^{\sim}p < 0.10$. $^{***}p < 0.001$

**Table 5** HLM results for teachers' evaluations of students' effort and the moderating role of teachers' beliefs and dispositions

| Predictor | No moderator | | Modern sexism | | Fixed mindset | | Experiences of sexism | | Experiences of racism | | Math anxiety | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Boys | 0.01 | 0.01 | | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |
| | (0.03) | (0.03) | | (0.03) | (0.03) | (0.03) | (0.03) | (0.03) | (0.03) | (0.03) | (0.03) | (0.03) |
| White | 0.01 | 0.01 | | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |
| | (0.04) | (0.04) | | (0.04) | (0.04) | (0.04) | (0.04) | (0.04) | (0.04) | (0.04) | (0.04) | (0.04) |
| Moderator | | 0.04 | | 0.04 | 0.03 | 0.01 | − 0.06 | − 0.06~ | − 0.01 | − 0.04 | − 0.01 | − 0.00 |
| | | (0.04) | | (0.04) | (0.04) | (0.04) | (0.04) | (0.04) | (0.04) | (0.04) | (0.04) | (0.04) |
| Boys* moderator | | 0.03 | | | 0.00 | | 0.00 | | − 0.04 | | 0.02 | |
| | | (0.03) | | | (0.03) | | (0.03) | | (0.03) | | (0.03) | |
| White* moderator | | | | 0.05 | | 0.04 | | 0.01 | | 0.02 | | 0.01 |
| | | | | (0.03) | | (0.04) | | (0.04) | | (0.04) | | (0.03) |
| Intercept | 4.19*** | 4.19*** | | 4.19*** | 4.19*** | 4.19*** | 4.19*** | 4.19*** | 4.19*** | 4.19*** | 4.19*** | 4.19*** |
| | (0.12) | (0.12) | | (0.12) | (0.12) | (0.12) | (0.12) | (0.12) | (0.12) | (0.12) | (0.12) | (0.12) |
| Variance intercept | 0.38 | 0.38 | | 0.38 | 0.38 | 0.38 | 0.38 | 0.38 | 0.38 | 0.38 | 0.38 | 0.38 |
| | (0.04) | (0.04) | | (0.04) | (0.04) | (0.04) | (0.04) | (0.04) | (0.04) | (0.04) | (0.04) | (0.04) |
| Variance residual | 1.79 | 1.79 | | 1.79 | 1.79 | 1.79 | 1.79 | 1.79 | 1.79 | 1.79 | 1.79 | 1.79 |
| | (0.05) | (0.05) | | (0.05) | (0.05) | (0.05) | (0.05) | (0.05) | (0.05) | (0.05) | (0.05) | (0.05) |
| ICC | 0.18 | 0.18 | | 0.18 | 0.18 | 0.18 | 0.18 | 0.18 | 0.18 | 0.18 | 0.18 | 0.18 |

The weighted results are reported here. Values in parentheses indicate robust standard errors clustered at the teacher level. The model also includes the item and item order as fixed effects. Scores for effort could range from 1 (*minimum effort*) to 8 (*maximum effort*)

$N = 413$. $^{\sim}p < 0.10$. $^{***}p < 0.001$

**Table 6** HLM results for teachers' ratings of students' mathematical ability and the moderating role of teachers' beliefs and dispositions for solutions that are not completely correct

| Predictor | No moderator | | Modern sexism | | Fixed mindset | | Experiences of sexism | | Experiences of racism | | Math anxiety | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Boys | 0.48 | 0.70* | 0.70* | 0.70* | 0.70* | 0.70* | 0.70* | 0.70* | 0.70* | 0.70* | 0.70* | 0.70* |
| | (0.46) | (0.31) | (0.31) | (0.31) | (0.31) | (0.31) | (0.31) | (0.31) | (0.31) | (0.31) | (0.31) | (0.31) |
| White | 0.21 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 |
| | (0.61) | (0.40) | (0.40) | (0.40) | (0.40) | (0.40) | (0.40) | (0.40) | (0.40) | (0.40) | (0.40) | (0.40) |
| Moderator | | | − 0.25 | 0.09 | 0.55 | 0.39 | − 0.30 | 0.05 | − 0.20 | − 0.21 | 0.47 | − 0.03 |
| | | | (0.47) | (0.44) | (0.53) | (0.50) | (0.46) | (0.45) | (0.50) | (0.47) | (0.48) | (0.47) |
| Boys* moderator | | | 0.81** | | − 0.18 | | 0.21 | | 0.04 | | − 0.57* | |
| | | | (0.31) | | (0.32) | | (0.33) | | (0.37) | | (0.28) | |
| White* moderator | | | | 0.20 | | 0.20 | | − 0.75~ | | 0.09 | | 0.66~ |
| | | | | (0.43) | | (0.42) | | (0.42) | | (0.38) | | (0.37) |
| Teacher-centered correctness | | 5.33*** | 5.32*** | 5.33*** | 5.33*** | 5.33*** | 5.33*** | 5.33*** | 5.33*** | 5.33*** | 5.33*** | 5.33*** |
| | | (0.11) | (0.11) | (0.11) | (0.11) | (0.11) | (0.11) | (0.11) | (0.11) | (0.11) | (0.11) | (0.11) |
| Mean correctness | | 6.58*** | 6.57*** | 6.57*** | 6.57*** | 6.57*** | 6.59*** | 6.59*** | 6.58*** | 6.58*** | 6.56*** | 6.56*** |
| | | (0.37) | (0.38) | (0.38) | (0.37) | (0.37) | (0.38) | (0.38) | (0.38) | (0.38) | (0.38) | (0.38) |
| Intercept | 77.41*** | 60.50*** | 60.52*** | 60.50*** | 60.51*** | 60.51*** | 60.51*** | 60.54*** | 60.51*** | 60.51*** | 60.52*** | 60.50*** |
| | (1.43) | (1.09) | (1.09) | (1.09) | (1.09) | (1.09) | (1.10) | (1.09) | (1.09) | (1.09) | (1.09) | (1.09) |
| Variance intercept | 92.15 | 54.59 | 54.58 | 54.57 | 54.40 | 54.40 | 54.58 | 54.59 | 54.56 | 54.56 | 54.55 | 54.55 |
| | (8.68) | (5.24) | (5.25) | (5.25) | (5.26) | (5.26) | (5.26) | (5.26) | (5.24) | (5.24) | (5.23) | (5.23) |
| Variance residual | 311.25 | 145.32 | 145.14 | 145.31 | 145.31 | 145.30 | 145.30 | 145.18 | 145.32 | 145.32 | 145.23 | 145.21 |
| | (11.20) | (5.17) | (5.16) | (5.17) | (5.17) | (5.17) | (5.17) | (5.14) | (5.17) | (5.17) | (5.16) | (5.16) |
| ICC | 0.23 | 0.27 | 0.27 | 0.27 | 0.27 | 0.27 | 0.27 | 0.27 | 0.27 | 0.27 | 0.27 | 0.27 |

The weighted results are reported here. Values in parentheses indicate robust standard errors clustered at the teacher level. The model also includes the item and item order as fixed effects

$N = 413$. $^{\sim}p < 0.10$. $^{*}p < 0.05$. $^{**}p < 0.01$. $^{***}p < 0.001$

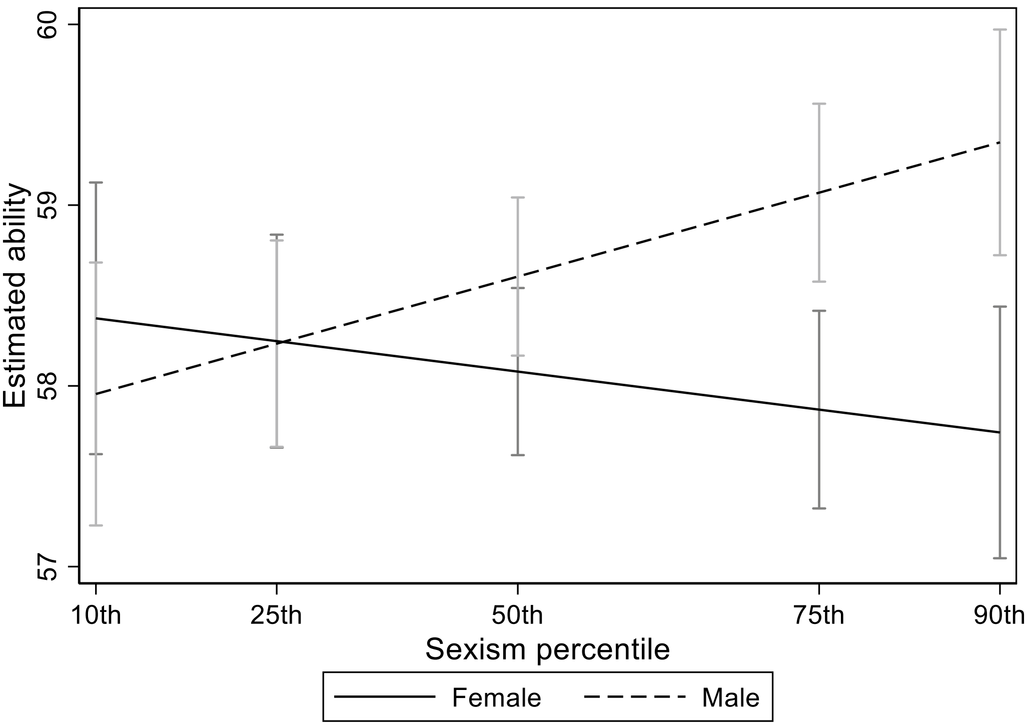


**Fig. 5** Teachers' sexist beliefs as a moderator of their ability ratings of boys and girls for the same student work. The bars indicate ± 1 standard error of the prediction

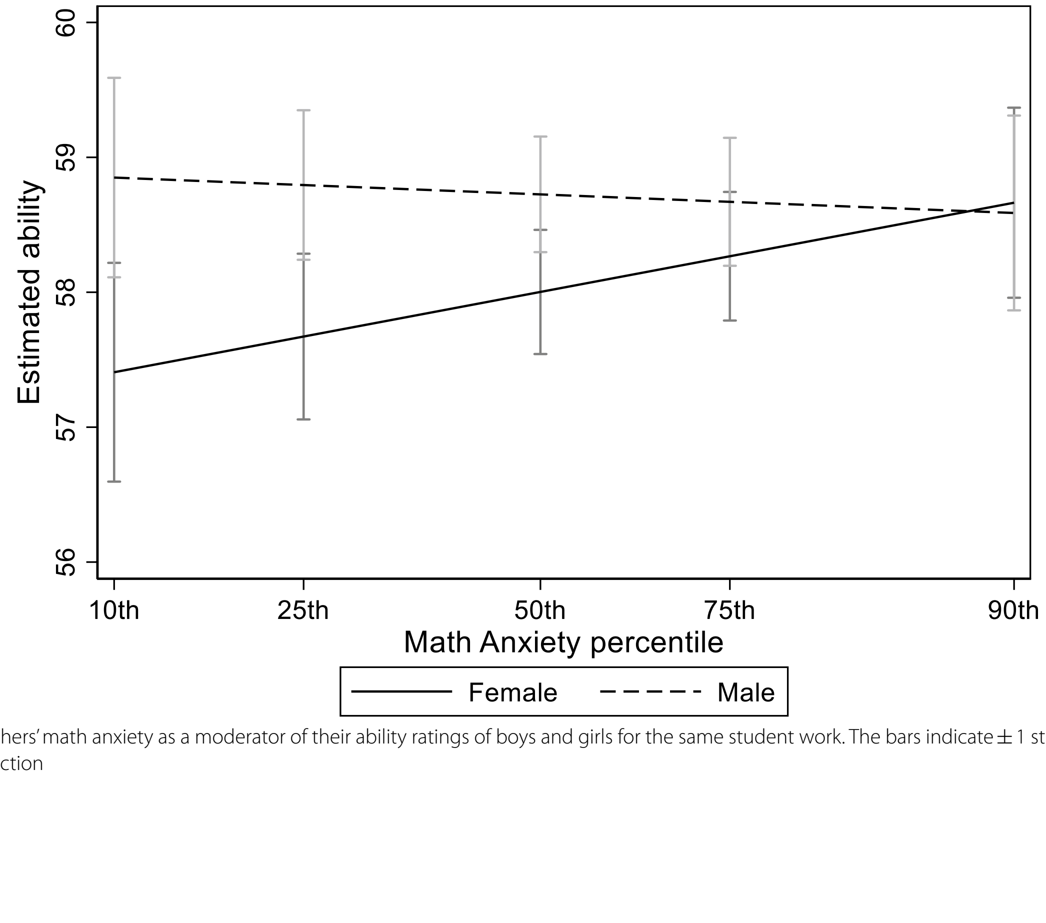


**Fig. 6** Teachers' math anxiety as a moderator of their ability ratings of boys and girls for the same student work. The bars indicate ± 1 standard error of the prediction

**Table 7** Preregistered research questions, hypotheses, and findings

| Research question | Hypotheses | Findings |
|---|---|---|
| RQ1. Are there systematic differences in teachers' evaluations of student performance (i.e., grading of student work), their evaluations of the effort they assume students put into the work, and their estimations of students' mathematical ability that could be explained by the students' gender or race? | H1. We hypothesize that we will not find any systemic differences in teachers' evaluation of students' written work (i.e., their ratings for the correctness of students' solutions) based on the students' gender or race, as found in prior empirical research (Copur-Gencturk et al., 2020). This prediction was based on theory positing that people tend to attribute the successes of members of stereotyped groups' to effort and non-stereotyped groups to ability (Graham, 2017; Graham & Williams, 2009), and is consistent with previous findings (e.g., Copur-Gencturk et al., 2020; Fennema et al., 1990; Tiedeman, 2000, 2002). However, we expect to find that teachers rate the students' ability higher when a White or male student name appears on work in situations where there is ambiguity in the work (i.e., when the students' work is not completely correct) compared with a non-White (Hispanic/Black) or girl name, consistent with previous findings. We also expect that teachers will rate the effort of students' work higher for female students' work | ✓ There were on average no differences in correctness ratings for gender (an effect size of 0.01 *SD* favoring girls, $p=0.59$) or for race (an effect size of 0.02 *SD* favoring White students, $p=0.32$)<br>✓ Teachers gave higher ability ratings to male names for not fully correct solutions (an effect size of 0.03 *SD* favoring boys, $p=0.03$)<br>✕ No race-based differences in teacher ability ratings were detected (underpowered; an effect size of 0 *SD* favoring White students, $p=0.78$)<br>✕ No gender-based differences in teacher evaluations of effort (an effect size of 0.01 *SD* favoring boys, $p=.72$) |
| RQ2. To what extent are teachers' beliefs and dispositions (mathematics anxiety, beliefs about mathematical intelligence, levels of sexist beliefs, perceptions of being underestimated because of race or gender) related to such biases? | [We had no specific preregistered hypotheses for the modern sexism moderator. We added modern sexism to our model to explore whether teachers' beliefs that gender disparities exist in current society were related to gender bias.] | _ Modern sexism moderated teachers' gender-biased ability evaluations for not fully correct solutions (an effect size of 0.03 *SD* favoring boys, $p=0.01$) |
| | H2a. Teachers who have higher levels of mathematics anxiety may draw on their biases more often than other teachers when evaluating students' work because mathematics anxiety can impair a person's cognitive ability (e.g., Beilock et al., 2010; Ramirez, Shaw, & Maloney, 2018a, 2018b) | ✕ Teachers' gender-biased ability evaluations for not fully correct solutions were moderated by math anxiety but in the opposite direction (an effect size of 0.02 *SD* favoring girls, $p=0.04$)<br>✓ Math anxiety moderated teachers' race-biased ability ratings for not fully correct solutions (an effect size of 0.02 *SD* favoring White students, $p=0.07$) |
| | H2b. Teachers who believe that mathematical intelligence is fixed and innate will tend to show more gender bias because prior research suggests that teachers who believe that mathematical ability is fixed and innate also believe that boys, but not girls, have this ability (Copur-Gencturk et al., 2021) | ✕ Fixed mindset did not moderate teachers' gender-biased ability evaluations for not fully correct solutions (an effect size of 0.01 *SD* favoring girls, $p=0.57$) |

The official preregistration can be accessed here using the following link: https://aspredicted.org/GNH_RXF. Also note that the full preregistration refers to an additional study that goes beyond the scope of this paper. The ✓ indicates that the preregistered hypothesis was confirmed; the ✕ indicates that it was not confirmed; and the _ indicates that we did not have a specific directional hypothesis for this question ratings of boys and girls by teachers at specific levels of mathematics anxiety were not significant (see Fig. 6).

In robustness checks, for all outcomes, the role of these moderators in teachers' biases were similar for both the analytic sample and the full sample without weights (see Additional file 1: Tables S4–S9).

## Discussion

In this study, we aimed to understand teachers' biases as they were revealed in math teachers' practices and the moderating role that teachers' beliefs and dispositions played in their biases. In particular, we focused on teachers' potential biases in three areas: their evaluations of the correctness of student work, the effort they assumed the student put into the work, and the student's mathematical ability. Our work aimed to extend prior work, particularly that by Copur-Gencturk and colleagues (2020), by testing the extent to which the observed bias patterns held true in a national sample of teachers as well as how teachers' beliefs and dispositions might help us understand their biases. A summary of our preregistered research questions, hypotheses, and findings can be found in Table 7.

### Limitations

Before we discuss our findings, we wish to acknowledge that we were unable to investigate the potential biases of teachers of color or the factors moderating their biases because we were unable to recruit a sufficient number of teachers of color. Thus, further research is needed to explore the biases among teachers of color. Second, although we aimed to replicate and extend the work of Copur-Gencturk et al. (2020), the differences in study designs should be taken into account when interpreting our results. In particular, we did not have a system of trust in place with the teachers in our study because these teachers heard about the study for the first time from a group of researchers they did not know. In contrast, the teachers in the former study were first informed by a colleague they knew who was a coordinator of the yearlong professional development program they were attending. During the design phase of our study, we had anticipated that some teachers might be suspicious that their bias was being measured; therefore, we included items that would identify suspicious teachers, and we had preregistered our plans for conducting analyses that would take suspicious teachers into account. Our findings confirmed that the results were contingent on teachers' suspicions for the areas in prior work that indicated bias could occur. The results of teachers' ratings of the correctness of student work or the effort they thought the student put into it did not differ substantially regardless of whether the teachers became aware of the nature of the study. In contrast, the results differed for teachers' ability ratings depending on whether the teachers were aware of the true nature of the study. Taken together, our findings suggest that teachers who were suspicious consciously monitored their responses in the areas where they had unconscious biases. We argue that our study draws attention to an important issue that can arise when conducting research in sensitive areas. We further suggest that more research needs to investigate why some teachers are more cautious and alert than others that their biases are being measured and how study findings might be affected by these teachers. Third, using experimental methods to assess teacher bias has distinct tradeoffs. On the one hand, experimentally manipulating gender and race of a students' name enabled us to capture bias situated in a relevant context of evaluating student work rather than a lab setting. On the other hand, such manipulations make it difficult to disentangle whether the biases detected were a result of implicit or explicit processing. Fourth, although we explored the moderating role of teachers' modern sexist beliefs in their unconscious bias, we did not include a measure of modern racism. The existing modern racism scale (McConahay, 1986) has been critiqued as being too conceptually similar to "old-fashioned" racism, as not having been systematically updated in several decades, and as not making use of more recent and subtle language around race, among other criticisms (Morrison et al., 2017). In addition, we were concerned that items on the scale sound offensive and that the inclusion of these items might raise teachers' suspicions about the purpose of the study. For example, one of the items was "Blacks have more influence upon school desegregation plans than they ought to have" (McConahay, 1986, p. 212). Future research on explicit bias might pursue the creation and validation of an updated modern racism scale that uses items that would be more appropriate for studies investigating subtle racist beliefs.

### Biased judgments of student ability

Our results are in alignment with the finding of Copur-Gencturk et al. (2020) that teachers did not grade students' work differently, but did show bias against the mathematical ability of female students in ambiguous situations. The magnitude of the bias we detected against the mathematical ability of girls was similar to that found in the original study for all solution levels (i.e., incorrect, partially correct, and fully correct). These two studies with different populations both suggested that teachers assumed boys had a higher math ability than girls, especially for less correct solutions. We believe such consistent evidence may shed new light on what contributes to males' higher mathematical confidence (e.g., Ganley & Lubienski, 2016), as well as on the persistence of low-performing males in mathematics-intensive college majors (Cimpian et al., 2020). In fact, across international contexts, male students fairly reliably demonstrate much higher confidence in mathematics than would be suggested by their mathematics performance relative to that of females (Else-Quest et al., 2010; Ganley & Lubienski, 2016). It is possible that teachers may be signaling to boys with lower math performance that their ability is high, which may in turn boost their confidence despite their low performance. We would also urge future researchers to examine how students' experiences of interacting with and learning from K-12 teachers who believe that boys—particularly low-achieving boys—have relatively higher mathematical abilities than girls may contribute to the gender gaps in mathematics-intensive majors.

We would also like to point out that, though the ability bias effect sizes that we detected were relatively small, these ability biases were consistently detected across teachers of grade 1–8. Even small signals that boys have greater ability can potentially snowball into a larger cumulative effect when those signals are repeated over several years of schooling, across multiple teachers and on many assignments (Cimpian et al., 2016; Robinson-Cimpian et al., 2014). Future research might investigate the potential cumulative effects of receiving signals that reinforce stereotypes when they are repeated and come from multiple teachers.

Our findings differed from those of Copur-Gencturk and colleagues (2020) in that we did not detect bias against the mathematical ability of students of color. In the former study, more profound racial bias was detected among teachers of color when measured in an in-person professional development setting in a southern U.S. state. While we were unable to recruit a sufficient number of teachers of color for this study to detect the presence of similar biases among teachers of color, we also collected data from a broader group of teachers across the U.S. which may be revealing a smaller magnitude of ability bias as compared with teachers from the original study which was conducted in a "southern state". A recent study conducted with a national sample of about 1000 mathematics teachers also documented racial bias in teachers' evaluations of math ability (albeit marginally significantly; Copur-Gencturk et al., 2022). Their findings suggested that teachers who work in schools serving a higher ratio of Black students seemed to show more racial bias against Black students' mathematical ability. Taking the findings of these studies together, it seems that racial bias against students of color seems to occur more among teachers who work with these students, indicating that teachers working in racially diverse schools are more prone to generalizing stereotypical beliefs about students of color.

We did not find teachers' evaluations of student effort to differ by students' race or gender. Teachers did not seem to attribute the success of girls or students of color to more effort or the failure of boys or White students to a lack of effort. Our findings indicated that teachers attributed the differences in students' performance to ability rather than effort. This result is not what we expected, and is only partly in line with prior research finding that teachers attribute the successes of their top performing boy students to ability, but the successes of girls to effort (Espinoza et al., 2014; Fennema et al., 1990; Tiedemann, 2000). Yet, while there were discrepancies around the findings pertaining to effort attributions, there were also several differences between prior research and our study that might explain the discrepancies. For example, prior studies were conducted with teachers as they evaluated their *own* students with whom they had many experiences to ground their effort judgments, while our study asked teachers to evaluate the effort of *fictitious* students based on a single math solution. Future studies might consider improving upon the experimental study design in such a way that better enables more authentic evaluations of student effort.

Our work contributes to the literature by identifying which sets of teacher beliefs and dispositions moderated their biases. We found that teachers who disagreed that sexism exists in society predicted higher ability ratings for boys than girls for identical solutions that were not fully correct. This result suggests that this bias is stronger among teachers who maintain that gender inequity is not a social issue. Indeed, our findings are in alignment with an experimental study conducted with managers or those with managerial experience in veterinary medicine, a profession in which women have become well represented (Begeny et al., 2020). Specifically, Begeny and colleagues (2020) found that those who thought gender discrimination was not an issue in the profession evaluated men as being more competent than women and advised higher salaries for men. The authors contended that individuals whose profession has a strong representation of women might infer that the field has become more equitable, overlook subtle manifestations of gender bias, and subconsciously perpetuate the enactment of such a bias. This contention could be applied to our findings. Teachers' observations of girls' and boys' similar performance on standardized mathematics tests or the progress toward gender equity perceived to be made in society may lead teachers to overlook subtle displays of gender bias that exist around mathematical ability. These teachers then become the ones who subconsciously contribute to the perpetuation of such biases.

For this reason, we argue that teachers may need targeted professional development to encourage them to recognize the negative consequences of overestimating the progress toward gender equality, including perpetuating gender bias around mathematical ability. Yet, despite the best intentions driving efforts to create anti-bias trainings, existing interventions generally show very small, short-term effects that do not necessarily connect with biased behaviors (Forscher et al., 2019). Schmader et al. (2022) argue that many trainings are ineffective because the current methods are not well aligned with the scientific understanding of how and when biases emerge, and that designers of such trainings should direct more attention to creating multi-pronged interventions that address the multiple pathways in which biased behaviors may manifest. Indeed, one successful example that uses such an approach is an intervention by Devine et al. (2012), which treats stereotype bias as a "bad habit" that can be broken by creating awareness of bias and interrupting it using multiple strategies. Efforts to build on the science of implicit bias and adapt these interventions for teacher trainings are underway (Rimm-Kaufman & Thomans, 2021), and it is only a matter of time before effective anti-bias professional development for teachers will be identified. As such, when effective trainings have been identified, our findings suggest that teachers' who report that gender disparities do not exist might benefit most from gender bias training.

Furthermore, our study underscores the importance of studying more subtle beliefs and perceptions to understand teachers' biases. Although this study provides insights into how teachers' subtle sexist beliefs affect their gender-stereotypical judgments of students' mathematical ability, further research is needed to explore how teachers' modern racist beliefs moderate their biases regarding mathematical ability.

## Conclusions

This study provides evidence that teachers tend to rate the abilities of boys higher than those of girls when their mathematical solutions are not correct. Though the bias is rather small in magnitude, it can have a compounding effect since the evidence suggests teachers underestimate girls anew each successive year of elementary education studied (Cimpian et al., 2016; Robinson-Cimpian et al., 2014). Given that emerging studies have documented the disproportionate share of low-performing males in mathematics-intensive college majors (e.g., Cimpian et al., 2020), the consistent gender bias—even seemingly small—across experimental studies in which teachers boost the ability ratings of boys who provide incorrect solutions points to an earlier source for the increased math confidence among low-performing males: their teachers' unjustified confidence in them. Additionally, our study calls attention to the potentially harmful consequences for female students of teachers' beliefs that gender equity has been accomplished in society. Similarly, more attention should be paid to investigating the negative consequences of racial bias in schools and in society for students of color.

### Abbreviations

| | |
|---|---|
| IAT | Implicit associations test |
| NAEP | National Assessment of Educational Progress |
| SD | Standard deviation |
| SES | Socioeconomic status |
| STEM | Science, technology, engineering, and mathematics |

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40594-023-00420-z.

**Additional file 1.** Supplementary Materials.

### Acknowledgements

Not applicable.

### Author contributions

All authors conceptualized the study and its design. Dr. Thacker collected data. Dr. Copur-Gencturk prepared and analyzed the data. Drs. Copur-Gencturk and Thacker wrote the initial draft. Dr. Cimpian conducted the power analysis and wrote the corresponding section, and he was a major contributor to the interpretation of the data analysis and results. All authors read, revised, and approved the final manuscript.

### Funding

This project is funded by the Herman + Rasiej Mathematics Initiative.

### Availability of data and materials

All survey materials are provided in the additional materials. The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

## Declarations

### Competing interests

The authors declare that they have no competing interests.

Received: 3 November 2022 Accepted: 12 April 2023
Published online: 24 April 2023

### References

Aigner, D. J., & Cain, G. G. (1977). Statistical theories of discrimination in labor markets. *ILR Review, 30*(2), 175–187.

Arrow, K. (1973). The theory of discrimination. In O. Ashenfelter & A. Rees (Eds.), *Discrimination in labor markets* (pp. 3–33). Princeton University Press.

Bargh, J. A. (1994). The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition. Basic processesIn R. S. Wyer & T. K. Srull (Eds.), *Handbook of social cognition* (Vol. 1, pp. 1–40). Lawrence Erlbaum Associates.

Begeny, C. T., Ryan, M. K., Moss-Racusin, C. A., & Ravetz, G. (2020). In some professions, women have become well represented, yet gender bias persists—Perpetuated by those who think it is not happening. *Science Advances, 6*(26), eaba7814.

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, 107*(5), 1860–1863.

Bertrand, M., Chugh, D., & Mullainathan, S. (2005). Implicit discrimination. *American Economic Review, 95*(2), 94–98.

Bertrand, M., & Duflo, E. (2017). Field experiments on discrimination. In A. V. Banerjee & E. Duflo (Eds.), *Handbook of economic field experiments* (Vol. 1, pp. 309–393). North Holland Publishing.

Bian, L., Leslie, S. J., & Cimpian, A. (2017). Gender stereotypes about intellectual ability emerge early and influence children's interests. *Science, 355*(6323), 389–391.

Blackhart, G. C., Brown, K. E., Clark, T., Pierce, D. L., & Shell, K. (2012). Assessing the adequacy of postexperimental inquiries in deception research and the factors that promote participant honesty. *Behavior Research Methods, 44*(1), 24–40.

Boaler, J. (2013). Ability and mathematics: The mindset revolution that is reshaping education. *Forum, 55*(1), 143–152.

Carlana, M. (2019). Implicit stereotypes: Evidence from teachers' gender bias. *The Quarterly Journal of Economics, 134*(3), 1163–1224.

Cheryan, S., Ziegler, S. A., Montoya, A. K., & Jiang, L. (2017). Why are some STEM fields more gender balanced than others? *Psychological Bulletin, 143*(1), 1–35.

Cimpian, J. R., Kim, T. H., & McDermott, Z. T. (2020). Understanding persistent gender gaps in STEM. *Science, 368*(6497), 1317–1319.

Cimpian, J. R., Lubienski, S. T., Timmer, J. D., Makowski, M. B., & Miller, E. K. (2016). Have gender gaps in math closed? Achievement, teacher perceptions, and learning behaviors across two ECLS-K cohorts. *AERA Open, 2*(4), 1–19.

Copur-Gencturk, Y., Cimpian, J. R., Lubienski, S. T., & Thacker, I. (2020). Teachers' bias against the mathematical ability of female, black, and Hispanic Students. *Educational Researcher, 49*(1), 30–43.

Copur-Gencturk, Y., Thacker, I., & Cimpian, J. R. (2022). Teacher bias in the virtual classroom. *Computers & Education, 191*, 104627.

Copur-Gencturk, Y., Thacker, I., & Quinn, D. (2021). K-8 teachers' overall and gender-specific beliefs about mathematical aptitude. *International Journal of Science and Mathematics Education, 19*, 1251–1269.

Cvencek, D., Meltzoff, A. N., & Greenwald, A. G. (2011). Math–gender stereotypes in elementary school children. *Child Development, 82*(3), 766–779.

De Kraker-Pauw, E., van Wesel, F., Verwijmeren, T., Denessen, E., & Krabbendam, L. (2016). Are teacher beliefs gender-related? *Learning and Individual Differences, 51*, 333–340.

Degner, J., Mangels, J., & Zander, L. (2019). Visualizing gendered representations of male and female teachers using a reverse correlation paradigm. *Social Psychology, 50*(4), 233.

Denessen, E., Hornstra, L., van den Bergh, L., & Bijlstra, G. (2022). Implicit measures of teachers' attitudes and stereotypes, and their effects on teacher practice and student outcomes: A review. *Learning and Instruction, 78*, 101437.

Devine, P. G., Forscher, P. S., Austin, A. J., & Cox, W. T. (2012). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of Experimental Social Psychology, 48*(6), 1267–1278.

Dixson, A. D., & Rousseau, C. K. (2005). And we are still not saved: Critical race theory in education ten years later. *Race Ethnicity and Education, 8*(1), 7–27.

Donovan, S., & Cross, C. T. (Eds.). (2002). *Minority students in special and gifted education*. National Academy Press.

Dweck, C. S. (1975). The role of expectations and attributions in the alleviation of learned helplessness. *Journal of Personality and Social Psychology, 31*(4), 674.

Dweck, C. S. (1999). *Self-theories: Their role in motivation, personality, and development*. Psychology Press.

Elhoweris, H., Mutua, K., Alsheikh, N., & Holloway, P. (2005). Effect of children's ethnicity on teachers' referral and recommendation decisions in gifted and talented programs. *Remedial and Special Education, 26*(1), 25–31.

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*(1), 103–127.

Espinoza, P., da Luz, A., Fontes, A. B., & Arms-Chavez, C. J. (2014). Attributional gender bias: Teachers' ability and effort explanations for students' math performance. *Social Psychology of Education, 17*(1), 105–126.

Fennema, E., Peterson, P. L., Carpenter, T. P., & Lubinski, C. A. (1990). Teachers' attributions and beliefs about girls, boys, and mathematics. *Educational Studies in Mathematics, 21*, 55–65.

Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2019). A meta-analysis of procedures to change implicit measures. *Journal of Personality and Social Psychology, 117*(3), 522–559.

Fredrickson, B. L., & Roberts, T. A. (1997). Objectification theory: Toward understanding women's lived experiences and mental health risks. *Psychology of Women Quarterly, 21*(2), 173–206.

Ganley, C. M., & Lubienski, S. T. (2016). Mathematics confidence, interest, and performance: Examining gender patterns and reciprocal relations. *Learning and Individual Differences, 47*, 182–193.

Ganley, C. M., Schoen, R. C., LaVenia, M., & Tazaz, A. M. (2019). The construct validation of the math anxiety scale for teachers. *AERA Open, 5*(1), 2332858419839702.

Graham, S. (1984). Communicating sympathy and anger to Black and White children: The cognitive (attributional) consequences of affective cues. *Journal of Personality and Social Psychology, 47*, 40–54.

Graham, S. (2017). An attributional perspective on motivation in ethnic minority youth. In J. T. Decuir-Gunby & P. A. Schutz (Eds.), *Race and ethnicity in the study of motivation in education*. Routledge.

Graham, S., & Williams, C. (2009). An attributional approach to motivation in school. In K. R. Wentzel & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 11–33). Routledge.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*(1), 4.

Greenwald, A. G., & Krieger, L. H. (2006). Implicit bias: Scientific foundations. *California Law Review, 94*(4), 945–967.

Grogan, S. (2021). *Body image: Understanding body dissatisfaction in men, women, and children*. Routledge.

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), 1–51.

Harber, K., Gorman, J., Gengaro, F., Butisingh, S., Tsang, W., & Ouellette, R. (2012). Stu- dents' race and teachers' social support affect the positive feedback bias in public schools. *Journal of Educational Psychology, 104*(4), 1149–1161.

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics, 32*(4), 341–370.

Henry, P. J., & Sears, D. O. (2002). The symbolic racism 2000 scale. *Political Psychology, 23*(2), 253–283.

Jost, J. T., Banaji, M. R., & Nosek, B. A. (2004). A decade of system justification theory: Accumulated evidence of conscious and unconscious bolstering of the status quo. *Political Psychology, 25*(6), 881–919.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kirkcaldy, B., Noack, P., Furnham, A., & Siefen, G. (2007). Parental estimates of their own and their children's intelligence. *European Psychologist, 12*(3), 173.

Kumar, R., Karabenick, S. A., & Burgoon, J. N. (2015). Teachers' implicit attitudes, explicit beliefs, and the mediating role of respect and cultural responsibility on mastery and performance-focused instructional practices. *Journal of Educational Psychology, 107*(2), 533.

Lavy, V., & Sand, E. (2015). *On the origins of gender human capital gaps: Short and long term consequences of teachers' stereotypical biases* (Working Paper No. w20909). National Bureau of Economic Research.

Lecklider, A. (2013). *Inventing the egghead: The battle over brainpower in American culture*. University of Pennsylvania Press.

Leslie, S. J., Cimpian, A., Meyer, M., & Freeland, E. (2015). Expectations of brilliance underlie gender distributions across academic disciplines. *Science, 347*(6219), 262–265.

Mathieu, J. E., Aguinis, H., Culpepper, S. A., & Chen, G. (2012). Understanding and estimating the power to detect cross-level interaction effects in multilevel modeling. *Journal of Applied Psychology, 97*(5), 951–966.

McConahay, J. B. (1986). Modern racism, ambivalence, and the modern racism scale. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 91–125). Academic Press.

McGee, E. O., & Martin, D. B. (2011). "You would not believe what I have to go through to prove my intellectual value!" Stereotype management among academically successful Black mathematics and engineering students. *American Educational Research Journal, 48*(6), 1347–1389.

McShane, B. B., & Böckenholt, U. (2017). Single-paper meta-analysis: Benefits for study summary, theory testing, and replicability. *Journal of Consumer Research, 43*(6), 1048–1063.

Morgan, H. (2020). Misunderstood and mistreated: Students of color in special education. *Voices of Reform, 3*(2), 71–81. Retrieved from https://www.voicesofreform.com/article/18595-misunderstood-and-mistreated-students-ofcolor-in-special-education

Morrison, T. G., & Kiss, M. (2017). Modern Racism Scale. In V. Zeigler-Hill & T. Shackelford (Eds.), *Encyclopedia of personality and individual differences* (pp. 2951–2953). Springer.

National Science Foundation. (2020). *Science and engineering degrees, by race/ethnicity of recipients: 2008–18*. Arlington, VA. Retrieved from https://ncsesdata.nsf.gov/sere/2018/

National Center for Science and Engineering Statistics. (2023). *Women, minorities, and persons with disabilities in science and engineering 2023*. Arlington, VA. Retrieved from ncses.nsf.gov/pubs/nsf23315

Nosek, B. A., & Smyth, F. L. (2011). Implicit social cognitions predict sex differences in math engagement and achievement. *American Educational Research Journal, 48*(5), 1125–1156.

Nosek, B. A., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, A., Bar-Anan, Y., Bergh, R., Cai, H., Gonsalkorale, K., Kesebir, S., Maliszewski, N., Neto, F., Olli, E., Park, J., Schnabel, K., Shiomura, K., Tulbure, B.T., Wiers, R.W., … & Kesebir, S. (2009). National differences in gender–science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences*, *106*(26), 10593–10597.

Nürnberger, M., Nerb, J., Schmitz, F., Keller, J., & Sütterlin, S. (2016). Implicit gender stereotypes and essentialist beliefs predict preservice teachers' tracking recommendations. *Journal of Experimental Education, 84*(1), 152–174.

Phelps, E. S. (1972). The statistical theory of racism and sexism. *American Economic Review, 62*(4), 659–661.

Ramirez, G., Hooper, S. Y., Kersting, N. B., Ferguson, R., & Yeager, D. (2018a). Teacher math anxiety relates to adolescent students' math achievement. *AERA Open, 4*(1).

Ramirez, G., Shaw, S. T., & Maloney, E. A. (2018b). Math anxiety: Past research, promising interventions, and a new interpretation framework. *Educational Psychologist, 53*(3), 145–164.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.

Reich, J. (2021). Preregistration and registered reports. *Educational Psychologist, 56*(2), 101–109.

Reyna, C. (2008). Ian is intelligent but Leshaun is lazy: Antecedents and consequences of attributional stereotypes in the classroom. *European Journal of Psychology of Education, 23*(4), 439–458.

Rimm-Kaufman, S. E., & Thomas, K. T. (2021). *'I can't seem to connect with my students!': How white, middle class teachers can apply psychology to teach students who are different from them—a Practice Brief for Educators* [Issue Brief]. American Psychological Association Division 15. Retrieved from: https://apadiv15.org/wp-content/uploads/2021/07/Practice-Brief-Rimm-Kaufman-Thomas.pdf

Robinson-Cimpian, J. P., Lubienski, S. T., Ganley, C. M., & Copur-Gencturk, Y. (2014). Teachers' perceptions of students' mathematics proficiency may exacerbate early gender gaps in achievement. *Developmental Psychology, 50*(4), 1262–1281.

Rogers, K. D., Jr. (2020). Centering the M in STEM: A review of Black students' math experiences. *The Negro Education Review, 71*(1–4), 7–52.

Scherbaum, C. A., & Ferreter, J. M. (2009). Estimating statistical power and required sample sizes for organizational research using multilevel modeling. *Organizational Research Methods, 12*(2), 347–367.

Schmader, T., Dennehy, T. C., & Baron, A. S. (2022). Why antibias interventions (need not) fail. *Perspectives on Psychological Science, 17*(5), 1381–1403.

Snijders, T. A., & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage.

Starck, J. G., Riddle, T., Sinclair, S., & Warikoo, N. (2020). Teachers are people too: Examining the racial bias of teachers compared to other American adults. *Educational Researcher, 49*(4), 273–284.

Storage, D., Charlesworth, T. E., Banaji, M. R., & Cimpian, A. (2020). Adults and children implicitly associate brilliance with men more than women. *Journal of Experimental Social Psychology, 90*, 104020.

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review, 8*(3), 220–247.

Swim, J. K., Aikin, K. J., Hall, W. S., & Hunter, B. A. (1995). Sexism and racism: Old-fashioned and modern prejudices. *Journal of Personality and Social Psychology, 68*(2), 199–214.

Thacker, I., Copur-Gencturk, Y., & Cimpian, J. R. (2022). Teacher bias: A discussion with special emphasis on gender and STEM learning. In T. L. Good & M. McCaslin (Eds.), *The Routledge Encyclopedia of Education: Educational Psychology Edition*. Routledge.

Thomas, A. E. (2017). Gender differences in students' physical science motivation: Are teachers' implicit cognitions another piece of the puzzle? *American Educational Research Journal, 54*(1), 35–58.

Tiedemann, J. (2000). Gender-related beliefs of teachers in elementary school mathematics. *Educational Studies in Mathematics, 41*(2), 191–207.

Tiedemann, J. (2002). Teachers' gender stereotypes as determinants of teacher perceptions in elementary school mathematics. *Educational Studies in Mathematics, 50*, 49–62.

Torres, L., Driscoll, M. W., & Burrow, A. L. (2010). Racial microaggressions and psychological functioning among highly achieving African-Americans: A mixed-methods approach. *Journal of Social and Clinical Psychology, 29*(10), 1074–1099.

U.S. Department of Education. (2021). *42nd annual report to Congress on the implementation of the Individuals with Disabilities Act, 2009*. Author.

Wang, H., & Hall, N. C. (2018). A systematic review of teachers' causal attributions: Prevalence, correlates, and consequences. *Frontiers in Psychology, 9*, 2305.

Wang, M.-T., & Degol, J. L. (2017). Gender gap in science, technology, engineering, and mathematics (STEM): Current knowledge, implications for practice, policy, and future directions. *Educational Psychology Review, 29*(1), 119–140.

Warikoo, N., Sinclair, S., Fei, J., & Jacoby-Senghor, D. (2016). Examining racial bias in education: A new approach. *Educational Researcher, 45*(9), 508–514.

Weiner, B. (2005). Motivation from an attribution perspective and the social psychology of perceived competence. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 73–84). Guilford Press.

Wilson, T. (2011). *Redirect: The surprising new science of psychological change.* Penguin UK.

Yeager, D. S., & Dweck, C. S. (2012). Mindsets that promote resilience: When students believe that personal characteristics can be developed. *Educational Psychologist, 47*(4), 302–314.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.