Compensating Discrimination: Behavioral Evidence from Danish School Registers*

Julian Schuessler† Kim Mannemar Sønderskov†

July 20, 2023

Abstract

We suggest that discriminatory practices may vary significantly across decision-makers, which allows for deeper insights into the mechanisms behind discrimination. We study this in the context of biased grading in schools. We develop a theory of teacher biases driven by heuristic beliefs stemming from concrete classroom experiences. Because teachers may also care about grade equality, such a mechanism can lead to either inequality-reinforcing or compensating biases in grading. Based on large-scale administrative data on Danish students, we find strong evidence for highly heterogeneous teacher biases—up to 45% of teachers exhibit a bias that is of the opposite sign as the average bias. Furthermore, there is a robust and substantively large compensation effect. Teachers that experienced a visible demographic group (defined by gender or migration background) academically under-performing relative to a reference group show more positive bias towards that group than teachers where the same group over-performed. We find little evidence for alternative explanations of bias. To fully grasp discrimination, we must go beyond averages and consider the wide variety of biases shaped by individual experiences.

*The authors would like to thank Peter Thisted Dinesen, Simon Calmar Andersen, David Birke, Michela Carlana, Niklas Harder, Jakob Majlund Holm, Macartan Humphreys, Søren Netra, Clara Neupert-Wentz, Adam Peresman, Jakob Schlockermann, as well as participants at the Aarhus Political Behaviour and Institutions section, APSA 2022 annual meeting, CEPDISC 2022, Danish Data Science 2022, DPSA 2022 annual meeting, Wissenschaftszentrum Berlin, workshops at the Center for Experimental-Philosophical Study of Discrimination, the Pufendorf Institute for Advanced Studies, and the Aarhus-Copenhagen Political Behaviour Retreat for valuable feedback. Adam Redman Congleton provided excellent research assistance.

†Department of Political Science and Centre for the Experimental-Philosophical Study of Discrimination (CEPDISC), Aarhus University.

1 Introduction

Discrimination remains at the center of both public debate and scientific investigation. In recent years, an increasingly sophisticated literature has added to our understanding in terms of distinctions between taste-based, statistical, and structural discrimination (Small and Pager 2020), intersectional effects (Dahl and Krog 2018), and the existence of both negative and positive discrimination of the same demographic group, depending on further attributes (Schaeffer, Höhne and Teney 2016; Quadlin 2018). Discrimination is thus increasingly seen as a heterogeneous phenomenon that evades simple theorizing. In this paper, we make the case that there is important heterogeneity in discrimination not only across potential targets but also across decision-makers. At the same time, this heterogeneity opens up possibilities to learn about the deeper causes and mechanisms behind discriminatory behavior. Furthermore, although discrimination is usually inequality-reinforcing, it may also be compensating in specific cases. We develop and test these arguments in the context of biases in school teachers' grading.

A large literature analyzes the sizable gaps in grades across socio-demographic markers such as gender and migration background (e.g., Coenen and Van Klaveren 2016; Andersen and Reimer 2019; Wenz and Hoenig 2020). One explanation put forward is teacher bias (Lavy and Megalokonomou 2019; Lavy 2008; Terrier 2020; Di Liberto, Casula and Pau 2021; Gibbons and Chevalier 2008). However, researchers have noted that determining whether teachers are biased is methodologically challenging (DiPrete and Jennings 2012, p.2, Buchmann, DiPrete and McDaniel 2008, p. 160). Furthermore, it is often unclear why teachers would be biased in one or the other direction. The latter question is especially pressing given the highly varying findings on teachers' biases towards students of different ethnic backgrounds (Alesina et al. 2018; Botelho, Madeira and Rangel 2015; Burgess and Greaves 2013; Gibbons and Chevalier 2008).

We investigate teacher bias using large-scale administrative data on student grades and rich background variables from Denmark. These data allow us to investigate teacher bias in an unusually comprehensive and detailed manner. In order to do so, we differentiate various theoretical mechanisms that incorporate teachers' preferences, beliefs, and professional experiences. Going beyond simple notions of uniform taste-based or statistical discrimination (Small and Pager 2020), we suggest that individual teachers form (potentially distorted) beliefs about groups of students based on their

direct experience with these groups—and that these beliefs drive biases. This means that the magnitude and even the direction of biases may vary across teachers.

Furthermore, given the primacy of academic performance in most school systems and teachers' everyday professional life, we argue that the content of these beliefs centers on groups' perceived academic abilities. However, we argue that beliefs about groups do not necessarily predict the direction in which the discrimination of these groups occurs. When teachers believe a certain group of students to be academically weaker than other groups, does this lead them to *reinforce* such inequalities by—consciously or implicitly—biasing their grading against such groups, as standard accounts suggest? Or, do teachers rather *compensate* for such pre-existing perceived inequalities and exhibit positive bias? We argue that in societies with strong equality norms, such as the Danish context, the latter effect is possible.

Our data allow us to shed light on these questions. Danish 9$^{\text{th}}$ grade students receive grades on written performance both from their teacher as well as from standardized tests. Qua regulation, these two grades are supposed to measure the same abilities. We define teacher bias formally as a causal effect of student characteristics (Lundberg, Johnson and Stewart 2021) and use the difference between the two different grades in Math to estimate grading biases as a function of the migration background and gender of a student, averaging across all teachers as well as for each teacher individually. Our estimates of grading biases amount to around 10% of a standard deviation and indicate an average bias favoring girls (replicating numerous results from other countries, e.g., Terrier 2020) and students with a migration background. However, we point out various problems that may lead to biases in these estimates.

Going beyond simple averages, we document sizable heterogeneity in bias across teachers—in terms of size, but more interestingly also in direction. We estimate that between 30% and 45% of teachers have a bias in the opposite direction as the average would suggest. That is, a sizable minority of teachers favor boys or native Danes.[1] Accordingly, our results suggest that in this context heterogeneous effects are an important phenomenon that cannot be ignored.

What explains the large variation in biases? Following our theory, we argue that

[1]Throughout the paper, we use the term "migrant students" to refer to both first- and second-generation immigrant students. We use the term "native Danes" for students that are third- or higher-generation migrants. Virtually all humans are descendants of migrants (Timmermann and Friedrich 2016), and the concept of "nativity" should therefore be understood as a socially constructed shortcut. We rely on a binary conception of gender, as only a binary variable is available in our data.

teachers observe group-specific academic performance and use this to (partially) update their beliefs, which then influence their biases. In our data, both the relative academic performance of students with a migration background as well as the relative academic performance of girls versus boys vary widely across teachers, which allows us to evaluate these predictions. In essence, we compare the estimated bias of a teacher who has taught a relatively high-performing set of students from a certain demographic to a teacher who has taught relatively lower-performing students from the same demographic group.

We find that the worse a teacher's students with a migration background perform (relative to native Danes), the more positive the teacher's bias towards that group becomes. The same pattern holds for the relative performance of girls versus boys. This is in line with a compensation mechanism.

We find no evidence for alternative explanations of the variation in teacher bias. The share of migrant students a teacher has taught does not moderate bias in the manner predicted by contact theory (Elwert, Keller and Kotsadam N.d.). That is, contact alone is not sufficient to move bias; rather, it appears to depend on the specific information about academic performance transmitted during the contact. Teacher gender and migration background also do not correlate with teacher bias, nor do they appear to moderate the effect of contact or academic performance.

Further results bolster our proposed mechanism. When differentiating first- and second-generation migrants, the relative performance of one group mostly moderates the bias towards that group, but not biases towards the other group. When analyzing specific migrant student groups defined by regional origin or using regional origin fixed effects, we obtain similar results. Furthermore, the relative performance of girls does not moderate bias towards migrants and vice versa, alleviating some concerns about selection bias. When controlling for relative performance on the school level, teacher-level moderation estimates remain essentially unchanged, suggesting a mechanism centered on personal teacher experience that does not spill over onto other teachers. We also explain why a general compensation mechanism—that is, teachers generally compensating weak students irrespective of their group membership—cannot explain our results.

Finally, our estimates of the compensation effect are large. For example, our results imply that teachers who did not observe any academic performance differences between first-generation migrant students and native Danes do not show any bias towards mi-

grant students. The 95% confidence interval for the bias is $[-0.03, 0.01]$ standard deviations. However, a teacher with native Danes outperforming migrant students by one standard deviation is estimated to have a pro-migrant bias of about 0.08 (close to the average bias), while a teacher with migrant students outperforming native Danes by the same magnitude is estimated to biased in the opposite direction by about 0.11 standard deviations. In this sense, the observed performance differences of student groups assigned to a teacher appear to be a major cause of teacher biases.

2 Defining teacher bias

We suggest a theory-based and operationalizable definition of teacher bias that is independent of a specific estimation algorithm (Lundberg, Johnson and Stewart 2021). The extant literature on teacher bias does not offer such a definition, although researchers seem to be aware of numerous problems when *estimating* bias. For example, DiPrete and Jennings (2012) note that prior findings in the literature that were claimed to show gender bias of teachers

> are equally consistent with the contrary hypothesis that parents and teachers accurately observe gender differences in behavior, which affect both learning itself and the production of materials (like homework, reports, and presentations) that factor into the academic evaluation process. (p. 2)

We interpret this as saying that students' group characteristics may affect learning outcomes and thereby, indirectly, grades awarded to them. The left causal graph (Pearl 2009; Morgan and Winship 2015) in Figure 1 reflects this possibility: At any given point in time, student features X (gender, migration background, etc.) affect student ability A and thereby teacher grade T. This alone would create disparities in teacher grades across X. But since teachers are generally supposed to award grades based on (perceived) student abilities, it seems questionable to call such behavior "bias". Instead, it is some sort of direct effect net of abilities—indicated by the arrow from X to T in the right graph—that more intuitively would correspond to bias or discrimination on the part of the teacher. Pearl (2001) was the first to make such an explicit connection between discrimination and direct causal effects in causal graphs. This was recently taken up by Lundberg, Johnson and Stewart (2021) in various sociological examples as well

as in the computer science literature on algorithmic fairness (Zhang and Bareinboim 2018), among others.

Note that this argument relies on a normative statement about what teachers are (not) supposed to factor in when grading—i.e., what the ability variable A consists of. If one were to argue that student demographics are an appropriate source of information for the grading teacher (as "statistical" accounts of discrimination, to be discussed later, suggest), then a direct effect of such demographics would not necessarily constitute bias. Note additionally that the relationship between X and A may reflect other forms of bias than grading bias. That is, prior to the grading decision in question, teachers may have treated groups differently net of their prior abilities, affecting what abilities A students acquired in the first place. Indeed, Carlana (2019) provides evidence that female students with math teachers that exhibit stereotypes against girls, as measured by an implicit association test, obtain lower standardized test scores.[2] Therefore, it is important to keep in mind that studies—like ours—that analyze grading bias can at most yield evidence for "one-shot" discrimination at a single point in time. This may be especially relevant insofar as early interactions between students and teachers matter more than interactions in later grades, as argued by Alexander, Entwisle and Thompson (1987).

We return to challenges in actually estimating such biases from data further below including the role of U and E. In the next section, we introduce our theoretical model of teacher bias that explains why biases would vary across teachers.

[2]Accordingly, the edge $X \to A$ in both graphs in Figure 1 could be thought of as consisting of a part mediated by prior abilities A_{t-1} as well as a prior direct effect. However, since prior abilities are usually not observed, omitting such mediators is irrelevant to our formal analysis.

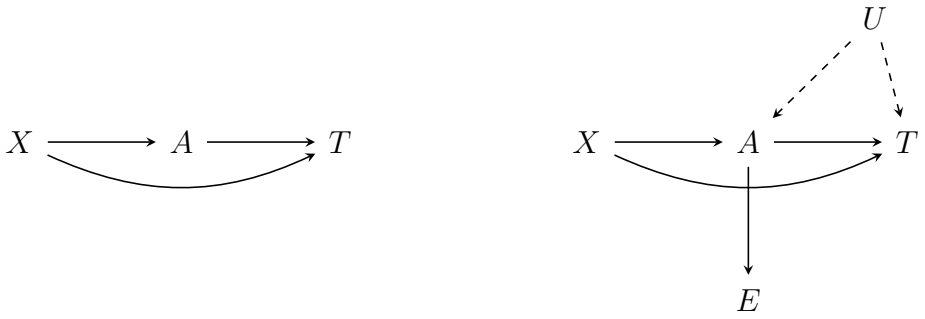

Figure 1: Causal graphs for identification analysis. Left: A basic unconfounded mediation model. Right: Confounded model with a proxy variable for the mediator. X denotes the group membership of a student in terms of gender, migration background, or their interaction. A denotes abilities that are not measured in the data but are observed by the teacher. E is an external grade from a standardized test. T is the grade awarded by the teacher. U are confounders, which may not be measured, that impact students' abilities and also the grading by the teacher.

3 Why are teachers biased?

Extensive research in the social sciences delves into the subject of teacher bias and implicitly adopts a conceptualization of bias akin to the model presented in the previous section (Lavy 2008; Lavy and Sand 2018; Lavy and Megalokonomou 2019; Botelho, Madeira and Rangel 2015; Terrier 2020; Di Liberto, Casula and Pau 2021; Gibbons and Chevalier 2008). Two findings stand out in this literature. Firstly, evidence from numerous countries over recent decades consistently suggests that teachers exhibit a bias in grading that favors female students. Secondly, the assessment of biases against migrant students presents a diverse picture, with both negative and positive estimates. This leads to our base hypothesis:

> *Hypothesis 1 (average biases)*: Teachers bias their grading as a function of student gender and migration background.

Our argument zooms onto the variability in teacher biases. Even in instances where the findings seem consistent, as is the case with gender bias, these results may often obscure significant heterogeneity and deeper underlying mechanisms. The prevailing research tends to focus on average biases, specifically those averaged across all teachers. However, if there is a variation among teachers, then these averages fail to accurately capture many students' lived experiences (Brand and Thomas 2013; Kline and Walters

2021). Additionally, a more profound understanding of the causes of biases can only be achieved by taking into account the diversity among teachers.

However, conventional models of discrimination and bias that focus on potential discriminators' decision-making are often implicitly uniform; similarly, structural accounts of discrimination due to the law or organizational behavior leave little room for heterogeneity (Pager and Shepherd 2008; Small and Pager 2020). It may well be that one reason for this lack of theorizing is the lack of available data. Observational studies of discrimination, especially in the labor market, do usually not include information on decision-makers (e.g., employers), while audit experiments rarely send more than one application to the same decision-maker. This prohibits analyzing the variation in biases across them. A recent exception is an experimental study by Kline and Walters (2021), who sent out multiple artificial CVs to the same employer to estimate variation in discrimination.

3.1 Theorizing Variation in Teacher Bias

Why would we expect teachers' biases to vary? The familiar individual-centered frameworks of "taste" and "statistical" discrimination can serve as a useful starting point (Small and Pager 2020). According to a simple taste-based model, each teacher would have a fixed preference for one student group over the other, which could then influence their grading decisions. As economists and sociologists have noted, this is not a very satisfactory explanation of behavior as it borders on a tautology (Arrow 1998; Charles and Guryan 2011): Biases occur because teachers "like" to be biased in a particular way. However, already the early model by Becker included decision-makers with varying tastes and therefore heterogeneity across discriminators, although again this is rarely reflected in empirical research (Becker 1957).

A "statistical" account of discrimination, on the other hand, assumes that the teacher faces limited information when assessing a student. Even though the teacher and student usually interact repeatedly and in some depth, it appears plausible that the teacher still operates under limited information about the subject-specific skills of the student. This is because both the curriculum content to be assessed as well as the student's skills evolve over the school year. Therefore, at any given point in time, there is residual uncertainty about the true abilities of a student.

Accordingly, teachers could use observable characteristics of students, such as migration background or gender, to inform their grading. Arguably, a standard statisti-

cal model of discrimination would predict reinforcing (Matthew) effects (Burgess and Greaves 2013): If, for example, teachers observe boys generally outperform girls in Math, they would generally bias boys' grades upwards and girls' grades downwards. Intuitively, a boy with an unusually low visible performance would be thought of to have suffered from bad luck or other idiosyncrasies, given boys' general relatively strong performance; therefore, the "optimal" statistical assessment would be higher than the prima facie performance.

However, we argue that such reinforcing mechanisms are not the only way a "statistical" mechanism can play out, especially once one leaves narrow notions of teachers' rationality. Traditional accounts of statistical discrimination rest on a strong notion of rationality. Following such a model, we would think of teachers as only caring about giving grades that reflect true abilities and at the same time accurately observing the relationship between group characteristics and academic performance to improve (statistically update) their assessments. Again, this mechanism is often thought to be uniform across decision-makers. All in all, this picture is unrealistic. A first step towards realism would allow for both taste–possibly unconsciously in the form of an implicit bias (Greenwald, McGhee and Schwartz 1998)– and statistical discrimination components in the teacher's behavior. Second, sociological and psychological evidence question the accuracy of decision-makers beliefs about the relationship between visible demographic markers and individual traits (e.g., Pager and Karafin 2009).

The last point becomes especially relevant once one considers the literature on context effect in schools (Crosnoe 2009; Legewie and DiPrete 2012). This literature suggests important (if varying) effects of school- and classroom composition on students' achievement. There is no similar literature that centers on teachers; however, it stands to reason that teachers as a "class" are similarly influenced by their work environment (Weeden and Grusky 2005). Specifically, availability and representativeness heuristics may lead teachers to overweight (relative to a rational decision-maker) their immediate experience with highly visible demographic subgroups (Bordalo et al. 2016). This opens up the possibility that teachers' beliefs about the performance of student groups vary significantly, and, accordingly, their biases do so as well.

Hypothesis 2 (variation in biases): Grading biases vary across teachers.

Taken together, our account assumes that teacher bias is driven, consciously or unconsciously, by both taste as well as beliefs about the academic performance of

visible demographic subgroups of students. These tastes and beliefs may be variable, and therefore teacher bias may be variable. Teachers' beliefs may be stereotypical and driven by heuristics, and depend to a large extent on their immediate experience with specific student groups.

3.2 The Compensation Effect

These mechanisms do not necessarily lead to predictions about the direction of biases. We have already suggested that a standard statistical discrimination account would lead to inequality-reinforcing biases, that is, a Matthew effect.

> *Hypothesis 3 (Matthew effect)*: Teachers are negatively biased towards groups they believe to perform worse academically and positively biased towards groups they believe to perform better.

However, the opposite effect—a compensation effect—is also possible. This is because teachers have preferences over the overall grade distribution. Specifically, they may prefer equal distributions of grades to less equal distributions. Indeed, van de Werfhorst (2020), based on survey data, reports that school teachers are among the most left-wing professions in Europe, and such attitudes usually go hand in hand with stronger preferences for equality. Furthermore, in our case, Denmark as a Scandinavian country exhibits even stronger and more homogeneous egalitarian norms than most other European countries (Bendixsen, Bringslid and Vike 2018).

If teachers want to equalize educational outcomes, this may interact with decision-making based on heuristical information. Rather than reinforcing pre-existing inequalities, it can lead teachers to compensate those groups which they perceive to perform poorly academically. We call this the *compensation effect*. The central aim of this paper is to investigate whether the compensation effect exists and in what way it materializes.

> *Hypothesis 4 (compensation effect)*: Teachers are positively biased towards groups they believe to perform worse academically and negatively biased towards groups they believe to perform better.

To our knowledge, such a mechanism of a belief-based discrimination mechanism leading to less group inequality has been rarely discussed in the literature. The phenomenon is distinct from the mechanism discussed by Schaeffer, Höhne and Teney

(2016), who show that in a purely statistical model of discrimination, the sign of the discrimination effect may switch depending on further attributes (in their case, education) of the person subject to discrimination. This could but need not, lead to average positive discrimination of the group with lower baseline levels of the outcome (Schaeffer, Höhne, and Teney find no average residual discrimination). In our case, however, the compensation effect always materializes and stems from additional preferences of the discriminating decision-maker that go beyond just making a decision (finding a grade) for a specific subject. Because the teacher wants to reduce overall grade inequality, the teacher pays attention to the typical performance of different groups in her environment and uses this information to reduce inequality, not (only) to assess an individual student.

3.3 Other explanations of variation in bias

Based on our account and the previous literature, we can also generate related but distinct predictions about the causes of teacher bias. Our theoretical mechanisms center on the concrete information about academic performance transmitted to a teacher in her work environment. This presupposes contact with the relevant groups (e.g., migrant students). If a teacher is never exposed to a certain group of students, her beliefs cannot change as a consequence of the exposure. Now, on a more foundational level, it is conceivable that contact alone may be sufficient to impact teachers' biases, akin to mechanisms discussed in the literature on contact theory (Allport 1954; Pettigrew and Tropp 2006; Elwert, Keller and Kotsadam N.d.). The classic contact hypothesis states that contact with out-groups increases sympathy toward the out-group. This argument is therefore conditional on the group membership of the teacher. A test of this would therefore involve, for example, comparing the effect of more exposure to migrant students on teacher bias between migrant teachers and non-migrant teachers.

> *Hypothesis 5 (contact)*: Teachers are more positively biased towards demographic out-groups the more contact they have with the out-group.

Finally, an adjacent and simplified hypothesis is that migration background and other teacher demographics like age and gender impact bias regardless of contact. These factors can impact bias through socialization processes on the bias-taste of a teacher. Prior research has investigated the role of gender, with varying results (Andersen and

Reimer 2019; Coenen and Van Klaveren 2016). Our large administrative data allow us to investigate all of these conjectures.

> *Hypothesis 6 (teacher demographics)*: Teacher demographics impact on biases regardless of contact.

4 Research design and data

Our empirical estimation strategy involves two steps: 1) Estimating teacher bias, on average and separately for each teacher 2) Relating teacher- (and also school-) level bias estimates to teacher and school characteristics. We implement these in one regression framework; however, it is useful to separate them for identification purposes.

4.1 Estimating Average Biases

We have suggested defining teacher bias as a direct causal effect of student attributes on teacher grades, fixing student ability. Note that this definition is also relevant to distinguish group-specific from general Matthew effects or compensation. For example, if teachers were generally awarding weaker (in terms of abilities A) students higher grades and stronger students lower grades than would otherwise be justified, this would lead to possible indirect effects of X on T (insofar as there are effects of demographics on abilities). However, the direct effect, by definition, factors in any teacher behavior based on A. If we find direct effects of X that cannot be explained based on A, then this is on top of any other effect due to the way teachers translate abilities into grades.[3]

In contrast to the standard estimation approach for direct effects, many papers in the teacher bias literature use the fact that teacher grades and standardized test grades are on the same scale and regress the deviation of the teacher grade T from a test grade E on student features X (e.g., Lavy 2008, Terrier 2020):

$$T - E = \alpha + \beta X + \epsilon.$$

Accordingly, there is no explicit regression control for ability. It is therefore not immediately clear whether such an approach estimates a direct effect. A perhaps more

[3]Note that we cannot estimate such general Matthew or compensation effects – that is, the causal effect of A on T. This is because we observed only a proxy of A, the test grade E, and additionally, A and T are possibly confounded, as indicated in the right graph in Figure 1.

intuitive approach would be to control for E in the regression $T = \alpha + \beta X + \delta E + \epsilon$ (e.g., Burgess and Greaves 2013).

To understand this, we turn to the right graph in Figure 1. We denote the test grade by E, as in many applications (including ours), it is often *externally* (co-)graded by a teacher who is not responsible for awarding the teacher grade T. We suggest that the standardized test grade does not perfectly measure ability A but instead is influenced by additional noise factors. The arrow from A to E indicates that A causally affects E. In the psychometric literature on educational testing, such "formative modeling" is a common conceptualization of standardized tests (Markus and Borsboom 2013, 117). The test grade E therefore proxies for A. This creates a first problem for estimating teacher bias. For estimating teacher bias, we would want to statistically control for A; however, we only have the imperfect proxy E available.

A second problem indicated by the right graph in Figure 1 is potentially unobserved confounders U that impact ability and teacher grade. Among other things, these may include teacher competence or "added value"—which impacts A—as well as teacher strictness, which impacts T. We would expect competence and strictness to be correlated due to deeper unobserved variables that impact teacher features (e.g., their training and work experience). Therefore, these variables would constitute confounders of the $A \rightarrow T$ relationship. Additionally, it may be that parents' involvement and educational investments impact A as well as on T insofar as teachers bias their teaching depending on how parents behave. Furthermore, there may be relevant variables that vary on the class and school level (e.g., sociodemographic composition). Such unobserved confounders of mediator and outcome make identification of direct causal effects impossible without stronger assumptions (Knox, Lowe and Mummolo 2020; Lundberg, Johnson and Stewart 2021; Zhang and Bareinboim 2018).

In Appendix B, we discuss a set of stronger assumptions that would allow researchers to estimate teacher bias. Among other things, one would need to assume that the test E measures abilities A in the same way as the teacher grade T. This seems unlikely to be satisfied exactly, and therefore average bias estimates may well suffer from systematic statistical errors.[4]

[4]In Appendix B, we also show that an estimator that adjusts for E via regression control Burgess and Greaves (2013) appears to be strongly biased even under the stronger assumption.

4.2 Estimating the Effects of Teacher Characteristics

However, our main aim is not to estimate teacher bias, but rather to explain it. That is, we relate estimates of teacher bias to teacher- or school-level variables, such as the relative performance of student groups assigned to a teacher. For estimating such causal effects, one main concern is to make sure that there are no unobserved confounders of the teacher and school characteristic and the teacher bias. If our estimates of teacher bias are systematically distorted, this does not introduce a problem, as long as the causes of the distortion are not related via back-door paths (Pearl 2009) to the independent variable (the teacher-level variable).[5] We expand on this in Appendix B, where we show analytically and by means of simulation how we can explain teacher biases even in the face of issues like differential measurement error in the test grade E.

Our primary concern is that certain types of teachers with varying biases systematically sort into specific schools (e.g., urban versus rural schools) and that this is correlated with the number and types of migrant students teachers are exposed to, thereby opening back-door paths between teacher-level variables derived from students' characteristics and teacher bias. However, since this selection happens on the school-, not teacher- or class-level, we use school fixed effects throughout. We thereby approximate a comparison that only happens across teachers within the same school. Furthermore, we will explore such possible biases through various placebo tests, which we discuss in more detail further below.

In terms of estimation, we examine the role of teacher- and school-level characteristics as moderating variables in interactive regression models of the form

$$T_i - E_i = \alpha + \beta X_i + \gamma Z_j + \delta X_i Z_j + C\lambda + \epsilon_i$$

For clarity, we have added subscripts i for students and j for teachers/schools. Here, X_i are students gender and migration background and Z_j are teacher- and school-level variables. Accordingly, estimates of δ will tell us how teacher- and school-level variables moderate the effect of students' demographic characteristics, which is equivalent to moderation of teacher- and school-level biases. C is a matrix of additional control variables, including teacher demographics and school fixed effects.

For testing for the presence of the compensation effect, Z_j will be the average

[5]This is because in a potential regression of teacher bias TB on teacher characteristics TC, $TB = \alpha + \beta TC + \epsilon$, a constant bias in estimates of TB simply moves estimates of the intercept α, but does not affect estimates of β.

performance differences between a group and a reference group (e.g., migrant and native students) for a teacher j, measured on the external test E. The reference group is the same as the reference group for the bias effect estimated by β. Teachers compensate when the interaction effect δ is negative. For example, this would indicate that a teacher whose migrant students outperform native students shows more anti-migrant bias than a teacher where migrant students perform less strongly academically than native ones.

This regression setup allows us to implement all main analyses in one common framework. The only thing that varies between regressions are which subset of the data we use, the composition of the independent variables, as well as the choice of the level on which we cluster standard errors. Since we use full population data, estimation uncertainty comes only from variation in the independent variables. We follow the guidance set out in Abadie et al. (2022) and cluster standard errors on the highest level on which the central independent variables of interest vary. If we focus only on student-level variables, we do not cluster standard errors. If we examine interactions between student demographics and teacher/school characteristics, we cluster on the teacher/school-level. Finally, for descriptive inferences on the distribution of teacher biases, we will also employ multilevel models.

4.3 Data source and sample

We base our analysis on population-wide register data on 9$^{\text{th}}$ grade students in Danish public and private schools for the years 2002–2019.[6] There are grades available for 1,190,234 students in this period. This excludes students in special needs classes as well as those who took the standardized exam at a different institution than the one where they attended school. Due to missing data (some private schools only started to implement the standardized test during the period of study) and duplicate entries, there are between 1,029,539 and 1,039,908 observations in our main analyses.

Furthermore, we make use of data that matches students to teachers. This match is available for the years 2014–2019 for students that attend public schools (about 2/3 of all students, N around 225,000; N of teachers is between 3,085 and 5,341). We use

[6]Data are also available for the years 2020 and 2021. However, in 2020, no standardized tests were performed due to the COVID-19 pandemic. In 2021, tests were performed, but, to counteract the learning difficulties of students during the pandemic, whenever the test grade was below the grade assigned by the teacher, the teacher grade was stored as the official test grade. Accordingly, our estimation strategy is not possible for these two years, which therefore are excluded from the analysis.

these data to investigate teacher-level heterogeneity and mechanisms.

4.4 Outcome variables

To measure teacher bias in grading we use two sets of grades given to students at the end of 9^{th} grade, which is the final year of lower secondary education for many students in the Danish school system. Students are given two grades on written performance in several subjects. One is given by the teacher, while the second is based on a nationally standardized test. By law, these grades are supposed to measure the same skill and only that skill. This sets clear normative standards and simplifies the discussion of the definition of teacher bias: Even if teachers did not grade directly based on student demographics, but instead based on behavior that is caused by these demographics, this would constitute a deviation from the legal framework. We focus on Math, as here the case for the test grade E measuring student abilities A well is strongest.

A further strength of our design is that students receive a separate grade from the teacher on oral performance. This is a special feature of the Danish schooling system; prior papers using data from other countries did not face such a situation. The written grade is therefore empirically less likely to reflect classroom behavior compared to situations in most other countries, as teachers are specifically instructed to separate written from oral performance.

Over the period of study, the stakes involved in the test were raised. Before 2015, the test had no direct formal consequence for students' educational evaluation. Beginning in 2015, passing the test in math was required for admission into vocational training. Furthermore, from 2019 onward, test results became part of the requirement for admission into high school.

Throughout, we standardize E, T, and $T - E$, so that effects can be interpreted as changes in the outcome in terms of standard deviations. Since we are interested in how much of the teacher grade T is explained by bias, the bias measure $T - E$ is standardized by using the (larger) standard deviation of the teacher grade. This allows for a more realistic assessment of effect sizes. On the original grading scale, the standard deviations of T and E are comparable (between 2.87 and 3.13).

4.5 Independent variables: Student demographics

We investigate two sources of teacher bias:

1. Migration background. This is operationalized as separate indicators for first- and second-generation migrants, as well as for a reference group that includes native Danes. Furthermore, we also investigate differences between different regions of origin (reference group: Denmark), regardless of migration generation.

2. The gender of a student. This is operationalized as a binary variable; 0 = male, 1 = female.

First-generation ("1G") migrants are defined in the data by Statistics Denmark as students that were born outside of Denmark and where no parent was born in Denmark and has Danish citizenship. Second-generation ("2G") migrants are students that were born in Denmark, but where no parent was born in Denmark and has Danish citizenship. The reference group ("native Danes", see fn. 1) contains all other students. The operationalization of gender reflects that there are only two genders in the Danish administrative registries.

4.6 Moderators

Our main moderators of interest are teacher-level performance differences of student groups. These are used to differentiate between Hypotheses 3 and 4. For exploring Hypothesis 5 (contact), we measure the share of students from demographic groups a teacher has taught. We also use triple interactions of student demographics, group shares, and teacher gender or migration background to test Hypothesis 5. Teacher gender and migration background are also used to assess Hypothesis 6.

5 Results

We start by estimating average biases. Figure 2 shows gender as well as migrant/non-migrant differences in the teacher grade, the test grade, as well as our bias measure (teacher grade minus test grade) for the whole sample of students across all years. Table A1 in the Appendix provides the full results behind the figure.

Starting on the left, the figure shows that there are no average gender differences in the teacher grades (the point estimate is virtually zero and insignificant). However, female students perform somewhat weaker than males on the external test. The difference amounts to 0.12 standard deviations (or about 0.3 grades on the original grading

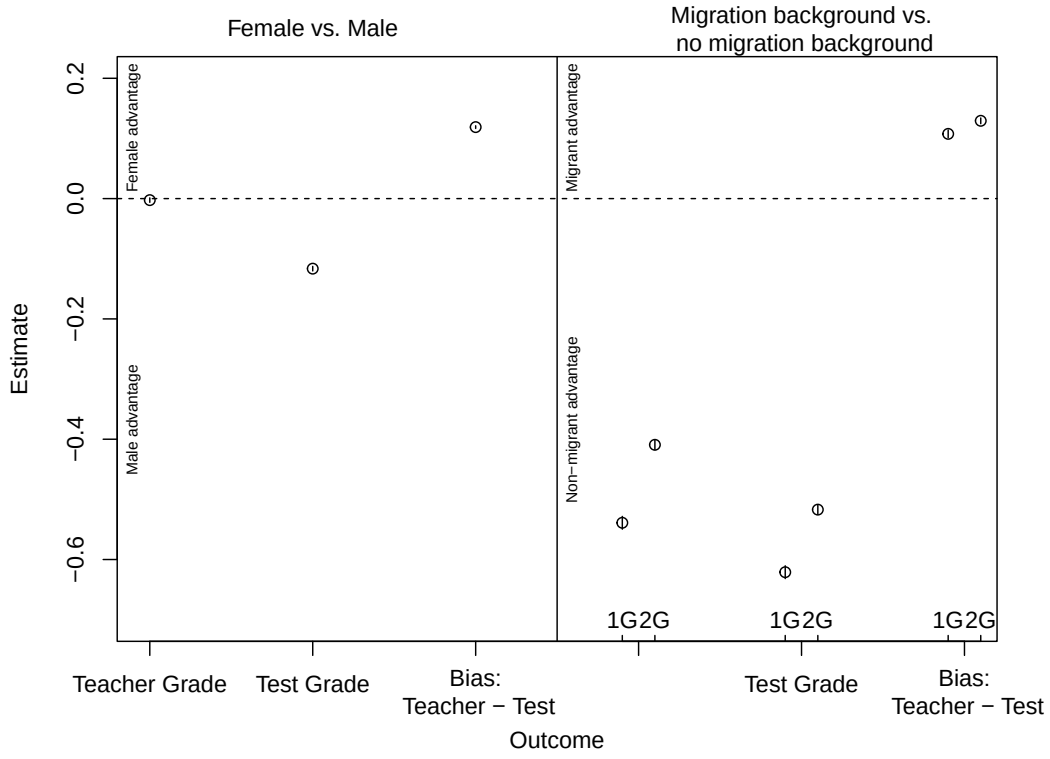

Figure 2: Point estimates and 95% confidence intervals (hardly visible) from student-level regressions. N = 1,038,400 students, pooled over the years from 2002-2019. The x-axis depicts the regression outcome. The first three estimates refer to gender effects, the last six estimates correspond to the effects of migration background. "1G" indicates a comparison of first-generation migrants with non-migrants, "2G" a comparison of second-generation migrants with non-migrants.

scale). This implies that the estimate of average teacher bias is positive and hence that female students seem to be advantaged relative to male students. The bias estimate is 0.12 standard deviations. Accordingly, the zero average difference in teacher grades appears to be made up of negative differences in ability (as proxied by the difference on the test) being outweighed by positively biased teacher grading.

Turning to the role of migration background, we see large differences in both teacher and test grades going into the same direction. First-generation migrant students receive about 0.57 standard deviations lower grades from their teachers than those without a migration background, while second-generation migrant students receive grades lower by about 0.42 standard deviations. For the test grades, the picture is very similar, with average differences of -0.62 (first-generation) and -0.52 (second-generation) standard deviations. Given that the average test grades are lower than the average teacher grades for both first and second generation immigrants compared to natives, the estimates of bias suggests biases in favor of students with a migration background: estimates are 0.11 and 0.13 standard deviations for first- and second-generation students, respectively.[7]

Our estimated effect sizes for the pro-girl bias are very similar to the results in Lavy (2008) (with data from Israel) and Terrier (2020) (with data from France), and somewhat smaller than the estimates based on Italian data in Di Liberto, Casula and Pau (2021). Our results on the role of migration background add to a highly diverse set of findings (Alesina et al. 2018; Botelho, Madeira and Rangel 2015; Burgess and Greaves 2013; Gibbons and Chevalier 2008).

We emphasize again, however, that the estimates of average biases are likely to be systematically distorted due to the methodological problems discussed before. Given their rather small size, it is even imaginable that the true average biases are in the opposite direction of what we find here. Therefore, in the remainder of the paper, we focus on the variability in teacher biases and their deeper causes.

5.1 How do biases vary across teachers?

As a first step of our investigation of the causes of teacher biases, we provide evidence that the biases do indeed vary significantly across teachers. The data for this analysis are based on students and teachers from public schools only (which educate roughly

[7]Further analyses, reported in Table A1 in the Appendix show no evidence of intersectionality; interacting gender and migration background yields very small and insignificant coefficients.

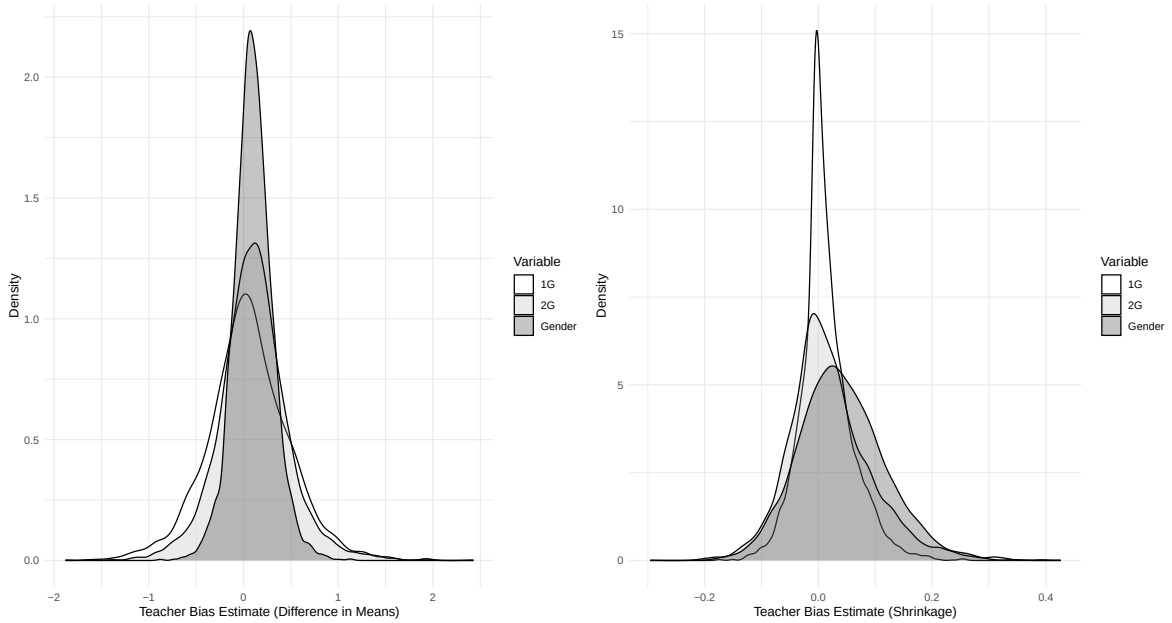

Figure 3: Distributions of teacher-specific bias estimates based. Left: Difference-in-means estimator. Right: Multilevel shrinkage estimator. N = 5,322 (gender); 3,085 (1G); 3,804 (2G) teachers, pooled over the years from 2014–2019.

2/3 of all Danish students in this period), and only for the period 2014–2019 due to data limitations; this yields around 5,300 teachers and the grades of around 225,000 students in total.

We start repeating the analysis from the previous section on the reduced sample and with teacher-fixed effects. Model 8 in Table A1 in the Appendix shows that the results are robust to the inclusion of teacher-fixed effects. Point estimates barely change after their inclusion, but, importantly, the teacher-fixed effects explain a sizable portion (16%) of the variance in the outcome.

Next, we compute the bias outcome $T - E$ for each student of the teacher and aggregate these to each of the demographic groups. For gender bias, there are very few missing values; however, once we turn to computing outcomes for students with a migration background, many teachers drop out of the analysis because they never got teach students with a 1G or 2G migration background. Here, we have data on 3,085 and 3,804 teachers, respectively.

We then use two estimation strategies to estimate heterogeneity in teacher biases. First, we estimate teacher-specific bias by using a simple difference-in-means estimator on the teacher level, e.g., comparing girls and boys. This approach will overestimate the

variance in teacher-specific bias, because each teacher only teaches a limited number of students (the mean number of students per teacher is about 43). The limited sample sizes per teacher introduce sampling variance in each teacher-specific bias estimate that biases the estimated variance upwards. The resulting density plots of bias estimates are in the left part of Figure 3.

Using this approach, we find that about 30% of teachers have a bias against female students (that is, a bias of the opposite sign as the average). About 36% of teachers are estimated to have absolute bias larger than 0.20 standard deviations, and about 5% to have an absolute bias larger than 0.50 standard deviations. Recall that the estimate of average bias amounted to 0.12 standard deviations. Turning to bias based on migration background, recall that the average bias was also positive and amounted to 0.11 (1G) and 0.13 (2G) standard deviations. We estimate that about 45% (1G) and 37% (2G) of teachers show negative bias, i.e. bias against students with a migration background. 59% (1G) and 53% (2G) are estimated to have an absolute bias larger than 0.20; for absolute biases larger than 0.50, the numbers are 23% and 16%. Overall, this indicates large heterogeneity in biases across teachers.[8]

Our second approach uses a shrinkage approach to mitigate the upward bias in variance estimates. We implement an Empirical Bayes estimator using a linear multilevel model where the dependent variable is the bias outcome $T - E$ and the independent variable is the gender or migration background indicator, which is specified to be randomly varying across teachers. Based on this model, we predict teacher-specific random effects, which are estimates of teacher-specific bias. These are shrunk towards the average mean in a data-dependent fashion, which therefore introduces bias (Gelman and Hill 2006). The resulting density plots of bias estimates are in the right part of Figure 3.Using this approach, we still find that about 30% of teachers have a bias against female students. However, the estimated spread of the distribution is reduced. About 23% of teachers are estimated to have absolute bias larger than 0.10 standard deviations, and about 9% to have an absolute bias larger than 0.15 standard deviations. Turning to migration background biases, estimates of the share of teachers with negative biases are again large and comparable to the first approach (43% for both 1G and 2G biases). The spread the distribution is significantly narrowed, especially for 1G bias, reflecting the fact many teachers only teach a few of migrant students (1G: 5% absolute bias larger than 0.10, 1% larger than 0.15; 2G: 16% and 6%).

[8]Histograms of teacher-specific bias estimates are in Appendix **??**.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Female x	-0.10^{***}	-0.11^{***}	-0.10^{***}	-0.10^{***}	-0.11^{***}	-0.11^{***}
Female Performance	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
1G x 1G Performance	-0.12^{***}	-0.13^{***}	-0.12^{***}	-0.11^{***}	-0.09^{***}	-0.13^{***}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)
2G x 2G Performance	-0.12^{***}	-0.11^{***}	-0.12^{***}	-0.11^{***}	-0.10^{***}	-0.13^{***}
	(0.01)	(0.02)	(0.01)	(0.01)	(0.02)	(0.01)
Teacher demographics	Yes	Yes	Yes	Yes	Yes	Yes
School FE	Yes	No	Yes	Yes	Yes	Yes
School-level variables	No	Yes	No	No	No	No
Interaction with share	No	No	Yes	No	No	No
Region FE	No	No	No	Yes	No	No
Only MENA migrants	No	No	No	No	Yes	No
Only non-MENA migrants	No	No	No	No	No	Yes
Num. obs.	125157	125121	125157	125109	112326	115389

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 1: Moderation of Bias through Teacher Experience

Neither of these two estimation approaches is perfect: The first overestimates variance due to sampling noise, the second one underestimates it due to built-in shrinkage bias. However, especially the fact that a sizable share of teachers is robustly estimated to have bias of opposite sign as the average bias points towards considerable heterogeneity that we aim to explain.

5.2 What explains teacher bias?

We now analyze the deeper mechanisms behind teacher bias. Our argument suggests that teachers biases are influenced by the academic performance of visible demographic groups that they are assigned to teach. Therefore, we look at how the effect of student characteristics (the teacher bias) is moderated by the observed academic performance of the relevant group among students assigned to a teacher.

Table 1 presents the main regression estimates. In each model, we interact the female, 1G and 2G migrant dummies with the respective performance of those students on the teacher-level. Throughout, we control for all base terms (student gender, migration background, and teacher-level performance variables) as well as for teacher demographics (age, gender, and migration background). In most models, we also in-

clude school fixed effects. For ease of presentation, we only show the central coefficients of interest. Full results are in Table A2 in the Appendix.

Model 1 in Table 1 shows that for all demographic groups, the interaction effect is negative and amounts to between -0.10 and -0.12 standard deviations. That is, the stronger the relative academic performance of the group to which a particular student belongs, the more negative the bias becomes. Conversely, the weaker the student group is, the more positive the bias is. Accordingly, we see clear evidence for a compensation effect across demographic groups. Teachers discriminate, but the discrimination reduces group inequality.

The effect sizes are considerable. Based on this model, teachers whose 1G migrant students performed as well as native Students are approximately unbiased (95% confidence interval $[-0.03, 0.01]$ standard deviations). However, a teacher with native Danes outperforming migrant students by one standard deviation is estimated to have a pro-migrant bias of about 0.08 (i.e., about 0.10 standard deviations higher than for the teacher with equal-performing groups), while a teacher with migrant students outperforming native Danes by the same magnitude is estimated to biased in the opposite direction by about 0.11 standard deviations. Such a standardized difference of 0.10 is about three times as large as the moderation of teacher bias by implicit association test measures reported in Alesina et al. (2018, p. 14).

For both biases towards females and 2G migrant students, the estimates for teachers where observed group differences are zero are positive and significant (0.09 and 0.06, respectively). Therefore, one could infer that teachers show no taste-bias against 1G migrant students, but positive taste-bias towards female and 2G migrant students. However, the bias of such teachers is still a possible mix of taste-motives and beliefs built through observation outside the classroom. Therefore, we do not think it is appropriate to make such an inference; rather, we should consider comparisons across teachers with different classroom experiences.

Models 2 to 6 in Table 1 show the robustness of this main result in various specifications. Throughout, estimates barely change. Model 2 controls for group performance differences on the school-level and their interaction with student demographics (this model does therefore not include school fixed effects). Model 3 controls for contact, that is, it adds control for the teacher-level shares of 1G/2G migrant students and their interaction with student 1G/2G dummies. (we present more detailed results on contact later). Model 4 includes fixed effects for the regional origin of migrant stu-

dents (irrespective of generation of migrant descend). Model 5 is fitted on the subset of native Danes and migrant students with a MENA (Middle East and North Africa) background; Model 6 is fitted on the subset of native Danes and non-MENA migrants.[9]

The stability of estimates in the last three (regional origin) models is especially interesting, as it suggests that the compensating mechanism occurs even within otherwise homogeneous migrant groups. This is relevant because teachers may also be thought of as holding migrant-group-specific stereotypes about relative performances that lead to bias, irrespective of teachers' classroom experiences. Indeed, when we compute the bias and relative performance measures on the level of the regional origin, we see a very strong negative correlation consistent with a compensation mechanism that superficially suggests group-specific stereotypes (see plot A1 in the Appendix). It appears as if teachers compensate based on (partially empirically grounded) general stereotypes about specific ethnic groups. However, the results here suggest that such group-level relationships are simply aggregates of teacher-level phenomena due to classroom experiences that hold even within ethnic groups (although some simple group-stereotyping may still occur).

The results in Table 1 are cross-sectional in the sense that the teacher-level performance variables are averages of test grades of all students a teacher has taught in the period of study. For example, the bias towards a student who was awarded a graded by the teacher in 2014 is explained by the observed performance of all students of that teacher, including those who were taught after 2014. In Table A3 in the Appendix, we replicate all models in Table 1 with performance-variables computed only based on students a teacher had observed before a given year. This reduces the number of observations significantly, as many teachers in the period of study only taught one class. However, point estimates and significance levels are qualitatively and quantitatively very similar.

5.3 Placebo Tests

In Table 2, we present further analyses that support the validity of these results. In this table, we summarize the results of a series of regressions; full results are in Tables A4 to A6 in the Appendix. Rows indicate student-level indicators, while columns indicate teacher- or school-level variables.

[9]Subsetting on more fine-grained definitions of migrant origin is not possible because of the small sample sizes involved.

The first three rows examine the role of teacher-level performance variables. The diagonal elements correspond to the estimates from Model 1 in Table 1 and show the compensation effect. Off-diagonal elements correspond to placebo tests. For example, the regression coefficient in the second row, first column indicates that the 1G bias is not moderated by the relative performance of girls compared to boys assigned to a teacher, as we would expect. Throughout, one can see that allmost all of these placebo tests pass (i.e., point estimates are very small and statistically insignificant). The only exception is in row three, column two, indicating that the relative performance of 1G migrant student reinforces biases against 2G migrant students. That there is some spillover from experiences with one migrant group to behavior towards another migrant group is consistent with our account, especially because these groups overlap somewhat in terms of regional origin. However, the point estimate is quite small (about 0.04 standard deviations). Therefore, overall, it appears that it is predominantly only the performance of a specific group that moderates bias towards that group.

Furthermore, the results in this table show that our regression strategy does not suffer from "regression to the mean" effects. If this were the case, we would generally expect negative (and possibly large) interaction effects; however, many interaction effect estimates are positive (if insignificant). We expand on this issue in **??**.

Rows four to six in Table 2 mirror the teacher-level analysis on the school level, while also controlling for moderation by teacher-level variables.[10] The diagonal coefficients are small and statistically insignificant, indicating that, when controlling for teacher-level performance, the performance of demographic groups in the school at large does not moderate teachers' biases. Within our theoretical framework, this suggests that teachers' beliefs are mostly informed by their personal classroom experience, not by broader trends on the school-level.

However, the coefficients in rows four and five, column one, indicate that school-level female performance significantly moderates migrant biases, with relatively large point estimates. There is little reason to expect such a causal effect, and it therefore seems likely that this is rather due to the assignment of teachers with anti-migrant biases to schools where girls tend to outperform boys. It is not clear to us why this particular form of unobserved confounding would occur. But, overall, this suggests that inferences about school-level moderation are potentially prone to statistical bias.

[10] Coefficients of teacher-level variables while controlling for school-level interactions are in Table 1, Model 2, and in the Appendix.

Therefore, while we remain confident about the role of teacher-level variables, the Null effects for the school-level moderation may also be a result of unobserved confounding.

Finally, rows seven to nine show estimates of the school-level moderation when not controlling for teacher-level variables. The diagonal entries suggest a compensation effect when it comes to 1G and 2G biases, with similar magnitudes to teacher-level moderation, but an insignificant estimate for gender biases. Yet, the placebo tests in rows eight and nine, column one, fail in the same way as in rows five and six. Therefore, we cannot reliably tell whether school-level moderation is simply aggregating teacher-level effects or whether there are effects beyond that.

5.4 Testing Further Hypotheses

Our final empirical results relate to the role of teachers' demographic attributes as well as mere contact with student groups. These are presented in Table 3. The number of observations is now markedly higher than in most previous models that focused on the performance variables. This is because the performance variables are only available for teachers that actually got to teach a demographic group (specifically, 1G or 2G students). However, if teachers did not teach any migrant students, that simply means that the contact variables (the share of students) is zero.

Model 1 in Table 3 shows that there is no affinity-effect: Teachers with a migration background do not show more (or less) positive bias toward migrant students. Similarly, teacher and student having the same gender does also not correlate with teacher bias.

Model 2 investigates whether teacher contact with more or less students of a demographic group moderates bias. The findings here are also insignificant. To test Hypothesis 5 (contact), model 3 investigates further whether contact moderates bias depending on teacher attributes. We find a statistically significant positive effect, indicating that teachers with a 1G migration background are more positively biased towards 1G students, the more students they teach in their classes. This is inconsistent with Hypothesis 5, which predicts that it is the out-group, not the in-group, that becomes more positively biased as a result of contact. In addition, the point estimate is implausibly large (more than 10 standard deviations). This is possibly due to outliers owing to the extreme right skew in the 1G share variable. Model 4 therefore removes observations in the top 2.5% quantile of the 1G share variable. The result is not robust to this minor adjustment, suggesting it was driven by just a few observations. The corresponding interaction estimate for 2G students and teachers is presented in model

Only teacher-level variables			
	Female Performance Teacher	1G Performance Teacher	2G Performance Teacher
Female	-0.10^{***}	0.00	0.01
	(0.01)	(0.00)	(0.01)
1G	0.01	-0.13^{***}	-0.01
	(0.02)	(0.01)	(0.01)
2G	0.01	0.04^{***}	-0.12^{***}
	(0.02)	(0.01)	(0.01)
School-level variables, controlling for teacher-level variables			
	Female Performance School	1G Performance School	2G Performance School
Female	0.05	0.03	-0.01
	(0.03)	(0.01)	(0.01)
1G	-0.27^{**}	0.05	0.01
	(0.09)	(0.03)	(0.03)
2G	-0.15	0.02	0.01
	(0.08)	(0.03)	(0.03)
Only shool-level variables			
	Female Performance School	1G Performance School	2G Performance School
Female	-0.06^{*}	0.01	0.01
	(0.03)	(0.01)	(0.01)
1G	-0.24^{**}	-0.08^{**}	0.01
	(0.09)	(0.03)	(0.02)
2G	-0.14	0.06^{**}	-0.11^{***}
	(0.08)	(0.03)	(0.02)

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 2: Main interaction estimates (diagonal) as well placebo-tests (off-diagonal) from various regression models. Standard errors in parentheses. Full results in tables A4–A6 in the Appendix.

5 and is insignificant. We note though that our power to detect interaction effects here is lower than in analyses that focus on relative performance measures, as the "shares" are less variable across teachers.

Finally, models 6 and 7 show that teacher demographics also do not moderate the compensation effect. The interaction estimates are very close to zero and insignificant. In sum, we do not find any evidence for teacher demographics or mere contact playing a role in explaining teacher biases.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Female x Female Teacher	−0.01		−0.01	−0.01	−0.01	−0.00	−0.01
	(0.01)		(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
1G x 1G Teacher	−0.06		−0.68**	−0.65	−0.05	−0.03	0.11
	(0.11)		(0.22)	(0.37)	(0.21)	(0.08)	(0.12)
2G x 2G Teacher	−0.10		−0.18	0.10	−0.00	−0.05	−0.08
	(0.07)		(0.17)	(0.14)	(0.14)	(0.08)	(0.07)
Female x Female Share		0.03	0.02	0.10	0.02		
		(0.07)	(0.20)	(0.21)	(0.20)		
1G x 1G Share		0.26	−10.41***	−9.97			
		(0.17)	(2.87)	(6.34)			
2G x 2G Share		0.08			−0.07		
		(0.04)			(0.44)		
Female x Female Teacher x Female Share			0.00	−0.07	0.00		
			(0.13)	(0.14)	(0.13)		
1G x 1G Teacher x 1G Share			10.77***	10.20			
			(2.87)	(6.34)			
2G x 2G Teacher x 2G Share					−0.06		
					(0.38)		
Female x Female Teacher x Female Performance						−0.01	−0.02
						(0.02)	(0.02)
1G x 1G Teacher x 1G Performance						0.21	
						(0.11)	
2G x 2G Teacher x 2G Performance							0.04
							(0.11)
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Num. obs.	223996	223996	223996	218326	223996	150202	174026

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$

Table 3: Role of Teacher Demographics, Group Contact, and Performance

6 Conclusion

In this paper, we have developed and tested a theory of the causes of heterogeneous discrimination across decision-makers. We suggested that biases are driven by beliefs partially influenced by teachers' concrete classroom experiences with the varying relative performance of visible demographic groups. Furthermore, we suggested that as teachers may not only care about accuracy when giving grades, but also about the overall distribution of grades across groups, influences of pre-existing differences in groups' academic ability may lead to reinforcing or compensating biases. We found evidence for large heterogeneity in teacher biases and compensating effects. Both the heterogeneity and the size of the compensating bias are substantial and robust to various specifications and tests. We found very little evidence for alternative explanations of bias.

The substantial size of the interaction effects we find, their robustness, and their specificity suggest that bias due to unobserved confounding is less of a danger than in many other observational studies. Still, there may be concerns. Scholars could search for specific sources of exogenous variation in student groups' performances to complement our results.

Furthermore, our data only cover students in 9^{th} grade. While this is the important final year of Danish school students' primary education, teacher bias may occur before and after this. As discussed, previous biases could be reflected in the abilities that are measured by the standardized tests in our data. As such, our analysis can only paint a partial picture of the trajectory of discrimination throughout students' educational career.

Investigating heterogeneity in discrimination across decision-makers is a fruitful avenue for future research. It requires sufficient data for each decision-maker, appropriate designs to identify the bias and ideally theories, measurements and designs for investigating the deeper causes of such bias. The only related study we are aware of is Kline and Walters (2021), who, based on job audit experiment data with four to eight observations per decision-maker, also find evidence for substantial heterogeneity in biases.

The compensation mechanism that we uncovered entails that in micro-contexts (on the teacher-level) inequalities in academic abilities cause offsetting biases of individual decision makers. In research on discrimination, such individual behavior appears to be

an unusual finding. We suggested that in our case, this phenomenon stems from general preferences for equal grades that teachers exhibit, but we were unable to provide direct evidence for this mechanism. It is imaginable that compensation mechanisms are at play in other contexts as well. In decision-making at the organization or firm-level, this would entail, for example, that organizations with imbalances in demographic composition exhibit biases that offset the imbalances. We are not aware of studies providing evidence in this direction. Indeed, the recent meta-study of experimental studies of gender discrimination in job applications by Galos and Coppock (2023) suggests the opposite, inequality-reinforcing pattern.

The compensation mechanism aggregates to the macro-level, as shown in the Appendix; student groups who perform worse academically profit from more positive teacher bias. On this level, the relationship appears to be equivalent to results from an affirmative action policy. But, its roots are in individual decisions, not regulations or mandates.

Our results should motivate researchers to consider the possibility that discrimination may not be uniform, neither across targets nor across decision-makers. We hope it also prompts more elaborate theories of discrimination that combine taste, statistical and heuristic, as well as additional preference mechanisms, as suggested in our theoretical argument.

References

Abadie, Alberto, Susan Athey, Guido W. Imbens and Jeffrey M. Wooldridge. 2022. "When should you adjust standard errors for clustering?" *The Quarterly Journal of Economics* 138(1):1–35. Publisher: Oxford Academic.

Alesina, Alberto, Michela Carlana, Eliana La Ferrara and Paolo Pinotti. 2018. Revealing stereotypes: Evidence from immigrants in schools. Technical report National Bureau of Economic Research.

Alexander, Karl L., Doris R. Entwisle and Maxine S. Thompson. 1987. "School performance, status relations, and the structure of sentiment: Bringing the teacher back in." *American sociological review* pp. 665–682. Publisher: JSTOR.

Allport, Gordon Willard. 1954. "The nature of prejudice.". Publisher: Addison-wesley Reading, MA.

Andersen, Ida Gran and David Reimer. 2019. "Same-gender teacher assignment, instructional strategies, and student achievement: New evidence on the mechanisms generating same-gender teacher effects." *Research in Social Stratification and Mobility* 62:100406.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S0276562417302093*

Arrow, Kenneth J. 1998. "What has economics to say about racial discrimination?" *Journal of economic perspectives* 12(2):91–100. Publisher: American Economic Association.

Becker, Gary Stanley. 1957. *The economics of discrimination: an economic view of racial discrimination*. Chicago: The University of Chicago Press.

Bendixsen, Synnøve, Mary Bente Bringslid and Halvard Vike. 2018. "Introduction: Egalitarianism in a Scandinavian context." *Egalitarianism in Scandinavia: Historical and contemporary perspectives* pp. 1–44.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli and Andrei Shleifer. 2016. "Stereotypes*." *The Quarterly Journal of Economics* 131(4):1753–1794.
URL: *https://doi.org/10.1093/qje/qjw029*

Botelho, Fernando, Ricardo A. Madeira and Marcos A. Rangel. 2015. "Racial Discrimination in Grading: Evidence from Brazil." *American Economic Journal: Applied Economics* 7(4):37–52.
URL: *https://pubs.aeaweb.org/doi/10.1257/app.20140352*

Brand, Jennie E. and Juli Simon Thomas. 2013. "Causal effect heterogeneity." *Handbook of causal analysis for social research* pp. 189–213. Publisher: Springer.

Buchmann, Claudia, Thomas A. DiPrete and Anne McDaniel. 2008. "Gender Inequalities in Education." *Annual Review of Sociology* 34(1):319–337.
URL: *https://www.annualreviews.org/doi/10.1146/annurev.soc.34.040507.134719*

Burgess, Simon and Ellen Greaves. 2013. "Test Scores, Subjective Assessment, and Stereotyping of Ethnic Minorities." *Journal of Labor Economics* 31(3):535–576.
URL: *https://www.journals.uchicago.edu/doi/10.1086/669340*

Carlana, Michela. 2019. "Implicit stereotypes: Evidence from teachers' gender bias." *The Quarterly Journal of Economics* 134(3):1163–1224. Publisher: Oxford University Press.

Charles, Kerwin Kofi and Jonathan Guryan. 2011. "Studying discrimination: Fundamental challenges and recent progress." *Annu. Rev. Econ.* 3(1):479–511. Publisher: Annual Reviews.

Coenen, Johan and Chris Van Klaveren. 2016. "Better test scores with a same-gender teacher?" *European Sociological Review* 32(3):452–464. Publisher: Oxford University Press.

Crosnoe, Robert. 2009. "Low-Income Students and the Socioeconomic Composition of Public High Schools." *American Sociological Review* 74(5):709–730. Publisher: SAGE Publications Inc.
URL: *https://doi.org/10.1177/000312240907400502*

Dahl, Malte and Niels Krog. 2018. "Experimental evidence of discrimination in the labour market: intersections between ethnicity, gender, and socio-economic status." *European Sociological Review* 34(4):402–417. Publisher: Oxford University Press.

Di Liberto, Adriana, Laura Casula and Sara Pau. 2021. "Grading practices, gender bias and educational outcomes: evidence from Italy." *Education Economics* pp. 1–28. Publisher: Taylor & Francis.

DiPrete, Thomas A. and Jennifer L. Jennings. 2012. "Social and behavioral skills and the gender gap in early educational achievement." *Social Science Research* 41(1):1–15. Publisher: Elsevier.

Elwert, Felix, Tamás Keller and Andreas Kotsadam. N.d. "Rearranging the Desk Chairs: A Large Randomized Field Experiment on the Effects of Close Contact on Interethnic Relations." *American Journal of Sociology*. Forthcoming.

Galos, Diana Roxana and Alexander Coppock. 2023. "Gender composition predicts gender bias: A meta-reanalysis of hiring discrimination audit experiments." *Science Advances* 9(18):eade7979. Publisher: American Association for the Advancement of Science.
URL: *https://www.science.org/doi/full/10.1126/sciadv.ade7979*

Gelman, Andrew and Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Gibbons, Stephen and Arnaud Chevalier. 2008. "Assessment and age 16+ education participation." *Research Papers in Education* 23(2):113–123. Publisher: Routledge _eprint: https://doi.org/10.1080/02671520802048638.
URL: *https://doi.org/10.1080/02671520802048638*

Greenwald, Anthony G., Debbie E. McGhee and Jordan LK Schwartz. 1998. "Measuring individual differences in implicit cognition: the implicit association test." *Journal of personality and social psychology* 74(6):1464. Publisher: American Psychological Association.

Kline, Patrick and Christopher Walters. 2021. "Reasonable Doubt: Experimental Detection of Job-Level Employment Discrimination." *Econometrica* 89(2):765–792. Publisher: Wiley Online Library.

Knox, Dean, Will Lowe and Jonathan Mummolo. 2020. "Administrative records mask racially biased policing." *American Political Science Review* 114(3):619–637. Publisher: Cambridge University Press.

Lavy, Victor. 2008. "Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment." *Journal of Public Economics* 92(10-11):2083–2105.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S0047272708000418*

Lavy, Victor and Edith Sand. 2018. "On the origins of gender gaps in human capital: Short- and long-term consequences of teachers' biases." *Journal of Public Economics* 167:263–279.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S0047272718301750*

Lavy, Victor and Rigissa Megalokonomou. 2019. Persistency in Teachers' Grading Bias and Effects on Longer-Term Outcomes: University Admissions Exams and Choice of Field of Study. Technical Report w26021 National Bureau of Economic Research Cambridge, MA: .
URL: *http://www.nber.org/papers/w26021.pdf*

Legewie, Joscha and Thomas A. DiPrete. 2012. "School Context and the Gender Gap in Educational Achievement." *American Sociological Review* 77(3):463–485. **URL:** *http://journals.sagepub.com/doi/10.1177/0003122412440802*

Lundberg, Ian, Rebecca Johnson and Brandon M Stewart. 2021. "What is your estimand? Defining the target quantity connects statistical evidence to theory." *American Sociological Review* 86(3):532–565. Publisher: SAGE Publications Sage CA: Los Angeles, CA.

Markus, Keith A. and Denny Borsboom. 2013. *Frontiers of test validity theory: Measurement, causation, and meaning*. Routledge.

Morgan, Stephen L and Christopher Winship. 2015. *Counterfactuals and causal inference*. Cambridge University Press.

Pager, Devah and Diana Karafin. 2009. "Bayesian bigot? Statistical discrimination, stereotypes, and employer decision making." *The Annals of the American Academy of Political and Social Science* 621(1):70–93. Publisher: Sage Publications Sage CA: Los Angeles, CA.

Pager, Devah and Hana Shepherd. 2008. "The sociology of discrimination: Racial discrimination in employment, housing, credit, and consumer markets." *Annu. Rev. Sociol* 34:181–209. Publisher: Annual Reviews.

Pearl, Judea. 2001. Direct and Indirect Effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, ed. Jack Breese and Daphne Koller. pp. 411–420.

Pearl, Judea. 2009. *Causality*. 2nd ed. Cambridge: Cambridge University Press.

Pettigrew, Thomas F. and Linda R. Tropp. 2006. "A meta-analytic test of intergroup contact theory." *Journal of personality and social psychology* 90(5):751. Publisher: American Psychological Association.

Quadlin, Natasha. 2018. "The mark of a woman's record: Gender and academic performance in hiring." *American Sociological Review* 83(2):331–360. Publisher: SAGE Publications Sage CA: Los Angeles, CA.

Schaeffer, Merlin, Jutta Höhne and Céline Teney. 2016. "Income advantages of poorly qualified immigrant minorities: Why school dropouts of Turkish origin earn more in Germany." *European Sociological Review* 32(1):93–107. Publisher: Oxford University Press.

Small, Mario L and Devah Pager. 2020. "Sociological perspectives on racial discrimination." *Journal of Economic Perspectives* 34(2):49–67.

Terrier, Camille. 2020. "Boys lag behind: How teachers' gender biases affect student achievement." *Economics of Education Review* 77:101981.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S0272775718307714*

Timmermann, Axel and Tobias Friedrich. 2016. "Late Pleistocene climate drivers of early human migration." *Nature* 538(7623):92–95. Number: 7623 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/nature19365*

van de Werfhorst, Herman G. 2020. "Are universities left-wing bastions? The political orientation of professors, professionals, and managers in Europe." *The British Journal of Sociology* 71(1):47–73.

Weeden, Kim A and David B Grusky. 2005. "The case for a new class map." *American Journal of Sociology* 111(1):141–212.

Wenz, Sebastian E. and Kerstin Hoenig. 2020. "Ethnic and social class discrimination in education: Experimental evidence from Germany." *Research in Social Stratification and Mobility* 65:100461. Publisher: Elsevier.

Zhang, Junzhe and Elias Bareinboim. 2018. Fairness in decision-making—the causal explanation formula. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Online Appendix

"Teachers' Ethnic and Gender Biases: Behavioral Evidence from Danish Registry Data"

Table of Contents

A Further Results A2
A.1 Average Biases A2
A.2 Heterogeneity in Teacher Bias A2
A.3 Aggregation to Region-Level A9
B Identification Analysis A9

A Further Results

A.1 Average Biases

Table A1: Average Teacher Biases

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Outcome:	Teacher Grade	External Grade	T - E	T - E	T - E	T - E	T - E	T - E
Female	−0.00	−0.12***	0.12***	0.12***	0.16***	0.10***	0.10***	0.10***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
1G	−0.54***	−0.62***	0.11***	0.10***	0.15***	0.07***	0.07***	0.05***
	(0.01)	(0.01)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)
2G	−0.41***	−0.52***	0.13***	0.13***	0.14***	0.13***	0.13***	0.11***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)
Female x 1G				0.01				
				(0.01)				
Female x 2G				−0.01				
				(0.00)				
Year x Female					−0.01***			
					(0.00)			
Year x 1G					−0.01***			
					(0.00)			
Year x 1G					−0.01***			
					(0.00)			
School FE	No	No	No	No	No	No	No	Yes
Num. obs.	1038402	1038402	1029539	1029539	1038400	225406	354619	225406

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

A.2 Heterogeneity in Teacher Bias

Table A2: Moderation of Bias through Teacher Experience

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Female	0.09***	0.09***	0.09***	0.09***	0.09***	0.09***
	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
Female Performance	0.06***	0.06***	0.06***	0.06***	0.05***	0.06***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
1G	−0.01	0.02	−0.05**	−0.07***	0.01	−0.02
	(0.01)	(0.02)	(0.01)	(0.01)	(0.02)	(0.01)
2G	0.06***	0.10***	0.05***		0.07***	0.06***

	(0.01)	(0.02)	(0.01)		(0.01)	(0.01)
1G Performance	0.01	0.00	0.01	0.01	0.01	0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
2G Performance	0.00	0.01	0.00	0.00	0.00	0.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Female Teacher	-0.03^{**}	-0.03^{**}	-0.03^{**}	-0.03^{**}	-0.03^{**}	-0.03^{***}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
1G Teacher	-0.13^{*}	-0.13^{*}	-0.13^{*}	-0.13^{*}	-0.13^{*}	-0.11
	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
2G Teacher	-0.07	-0.06	-0.07	-0.07	-0.06	-0.05
	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.06)
Age Teacher	0.00^{*}	0.00	0.00^{*}	0.00^{*}	0.00^{*}	0.00^{*}
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Female x Female Performance	-0.10^{***}	-0.11^{***}	-0.10^{***}	-0.10^{***}	-0.11^{***}	-0.11^{***}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
1G x 1G Performance	-0.12^{***}	-0.13^{***}	-0.12^{***}	-0.11^{***}	-0.09^{***}	-0.13^{***}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)
2G x 1G Performance	0.03^{***}	0.04^{***}	0.03^{***}	0.03^{***}	0.03^{**}	0.04^{***}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
1G x 2G Performance	-0.01	-0.02	-0.01	-0.01	-0.04	-0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)
2G x 2G Performance	-0.12^{***}	-0.11^{***}	-0.12^{***}	-0.11^{***}	-0.10^{***}	-0.13^{***}
	(0.01)	(0.02)	(0.01)	(0.01)	(0.02)	(0.01)
(Intercept)		0.18^{**}				
		(0.06)				
Female Performance		-0.16^{**}				
School		(0.05)				
1G Performance		0.03				
School		(0.02)				
2G Performance		-0.05^{**}				
School		(0.02)				
Female x Female Performance		0.03				
School		(0.04)				
1G x 1G Performance		0.05				
School		(0.03)				
2G x 1G Performance		0.02				
School		(0.03)				
1G x 2G Performance		0.02				
School		(0.03)				
2G x 2G Performance		0.00				
School		(0.03)				
Female x Female Share			0.18			
			(0.09)			

1G x 1G Share			0.40*			
			(0.19)			
2G x 1G Share			0.25			
			(0.17)			
1G x 2G Share			0.07			
			(0.05)			
2G x 2G Share			0.01			
			(0.04)			
Teacher demographics	Yes	Yes	Yes	Yes	Yes	Yes
School FE	Yes	No	Yes	Yes	Yes	Yes
School-level variables	No	Yes	No	No	No	No
Control for Interaction with Share	No	No	Yes	No	No	No
Region FE	No	No	No	Yes	No	No
Only MENA migrants	No	No	No	No	Yes	No
Only non-MENA migrants	No	No	No	No	No	Yes
Num. obs.	125157	125121	125157	125109	112326	115389

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table A3: Moderation of Bias through Teacher Experience: Panel measures of Teacher-Level Variables

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Female x	-0.10^{***}	-0.10^{***}	-0.10^{***}	-0.10^{***}	-0.11^{***}	-0.11^{***}
Female Performance	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
1G x 1G Performance	-0.12^{***}	-0.12^{***}	-0.12^{***}	-0.11^{***}	-0.08^{***}	-0.13^{***}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)
2G x 2G Performance	-0.11^{***}	-0.10^{***}	-0.11^{***}	-0.10^{***}	-0.09^{***}	-0.12^{***}
	(0.01)	(0.02)	(0.01)	(0.01)	(0.02)	(0.01)
Teacher demographics	Yes	Yes	Yes	Yes	Yes	Yes
School FE	Yes	No	Yes	Yes	Yes	Yes
School-level variables	No	Yes	No	No	No	No
Interaction with Share	No	No	Yes	No	No	No
Region FE	No	No	No	Yes	No	No
Only MENA migrants	No	No	No	No	Yes	No
Only non-MENA migrants	No	No	No	No	No	Yes
Num. obs.	100423	100387	100423	100380	88751	91498

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table A4: Moderation of Bias through Teacher Experience: Placebo Tests 1

	Model 1	Model 2	Model 3
Female	0.09^{***}	0.10^{***}	0.10^{***}
	(0.00)	(0.00)	(0.00)
Female Performance	0.05^{***}		
	(0.01)		
1G	0.05^{***}	-0.01	-0.01
	(0.01)	(0.01)	(0.01)
2G	0.11^{***}	0.06^{***}	0.06^{***}
	(0.01)	(0.01)	(0.01)
Female teacher	-0.02^{***}	-0.03^{**}	-0.03^{**}
	(0.01)	(0.01)	(0.01)
1G Teacher	-0.07	-0.13^{*}	-0.13^{*}
	(0.04)	(0.06)	(0.06)
2G Teacher	-0.04	-0.07	-0.07
	(0.04)	(0.07)	(0.07)
Age Teacher	0.00	0.00^{*}	0.00^{*}
	(0.00)	(0.00)	(0.00)
Female x Female Performance	-0.12^{***}		
	(0.01)		
1G x Female Performance	0.01		
	(0.02)		
2G x Female Performance	0.01		
	(0.02)		
1G Performance		0.01	0.01
		(0.01)	(0.01)
2G Performance		0.00	-0.00
		(0.01)	(0.01)
Female x 1G Performance		0.00	
		(0.00)	
1G x 1G Performance		-0.13^{***}	-0.13^{***}
		(0.01)	(0.01)
2G x 1G Performance		0.04^{***}	0.04^{***}
		(0.01)	(0.01)
1G x 2G Performance		-0.01	-0.01
		(0.01)	(0.01)
2G x 2G Performance		-0.12^{***}	-0.12^{***}
		(0.01)	(0.01)
Female x 2G Performance			0.01
			(0.01)
School FE	Yes	Yes	Yes
Num. obs.	223903	125157	125157

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table A5: Moderation of Bias through Teacher Experience: Placebo Tests 2

	Model 1	Model 2	Model 3
(Intercept)	0.19^{***}	0.19^{**}	0.20^{**}
	(0.04)	(0.06)	(0.06)
Female	0.10^{***}	0.11^{***}	0.10^{***}
	(0.00)	(0.01)	(0.01)
Female Performance	0.05^{***}		
	(0.01)		
1G	0.03^{*}	0.03	0.03
	(0.01)	(0.02)	(0.02)
2G	0.10^{***}	0.11^{***}	0.11^{***}
	(0.01)	(0.02)	(0.02)
Female Performance	-0.09^{*}		
School	(0.05)		
Female Teacher	-0.03^{***}	-0.03^{**}	-0.03^{**}
	(0.01)	(0.01)	(0.01)
1G Teacher	-0.11^{**}	-0.13^{*}	-0.13^{*}
	(0.03)	(0.06)	(0.06)
2G Teacher	-0.07	-0.06	-0.06
	(0.04)	(0.07)	(0.07)
Age Teacher	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)
Female x Female Performance	-0.12^{***}		
	(0.01)		
1G x Female Performance	0.02		
	(0.02)		
2G x Female Performance	0.01		
	(0.02)		
Female x Female Performance	0.05		
	(0.03)		
1G x Female Performance	-0.27^{**}		
School	(0.09)		
2G x Female Performance	-0.15		
School	(0.08)		
1G Performance		0.00	0.00
		(0.01)	(0.01)
2G Performance		0.01	0.00
		(0.01)	(0.01)
1G Performance		0.01	0.03
School		(0.02)	(0.02)
2G Performance		-0.05^{**}	-0.04^{*}

School		(0.02)	(0.02)
Female x 1G Performance		−0.00	
		(0.00)	
1G x 1G Performance		−0.13***	−0.13***
		(0.01)	(0.01)
2G x 1G Performance		0.04***	0.04***
		(0.01)	(0.01)
Female x 2G Performance			0.01
			(0.01)
1G x 2G Performance		−0.02	−0.02
		(0.01)	(0.01)
2G x 2G Performance		−0.11***	−0.11***
		(0.02)	(0.02)
Female x 1G Performance		0.03	
School		(0.01)	
1G x 1G Performance		0.05	0.05
School		(0.03)	(0.03)
2G x 1G Performance		0.02	0.02
School		(0.03)	(0.03)
Female x 2G Performance			−0.01
School			(0.01)
1G x 2G Performance		0.02	0.02
School		(0.03)	(0.03)
2G x 2G Performance		0.01	0.01
School		(0.03)	(0.03)
School FE	Yes	Yes	Yes
Num. obs.	223903	125121	125121

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table A6: Moderation of Bias through Teacher Experience: Placebo Tests 3

	Model 1	Model 2	Model 3
(Intercept)	0.19^{***}	0.19^{***}	0.20^{***}
	(0.04)	(0.04)	(0.04)
Female	0.09^{***}	0.11^{***}	0.10^{***}
	(0.00)	(0.01)	(0.00)
Female Performance	-0.04		
	(0.04)		
1G	0.03^{*}	0.02	0.02
	(0.01)	(0.02)	(0.02)
2G	0.11^{***}	0.11^{***}	0.11^{***}
	(0.01)	(0.02)	(0.02)
Female Teacher	-0.03^{***}	-0.03^{***}	-0.03^{***}
	(0.01)	(0.01)	(0.01)
1G Teacher	-0.11^{**}	-0.12^{***}	-0.12^{***}
	(0.03)	(0.04)	(0.04)
2G Teacher	-0.07	-0.08	-0.08
	(0.04)	(0.04)	(0.04)
Age Teacher	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)
Female x Female Performance	-0.06^{*}		
School	(0.03)		
1G x Female Performance	-0.24^{**}		
School	(0.09)		
2G x Female Performance	-0.14		
School	(0.08)		
1G Performance		0.02	0.02
School		(0.01)	(0.01)
2G Performance		-0.03^{**}	-0.04^{***}
School		(0.01)	(0.01)
Female x 1G Performance		0.01	
School		(0.01)	
1G x 1G Performance		-0.08^{**}	-0.08^{**}
School		(0.03)	(0.03)
2G x 1G Performance		0.06^{*}	0.06^{*}
School		(0.03)	(0.03)
1G x 2G Performance		0.01	0.01
School		(0.02)	(0.02)
2G x 2G Performance		-0.11^{***}	-0.11^{***}
School		(0.02)	(0.02)
Female x 2G Performance			0.01
School			(0.01)
School FE	Yes	Yes	Yes
Num. obs.	223996	221590	221590

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

A.3 Aggregation to Region-Level

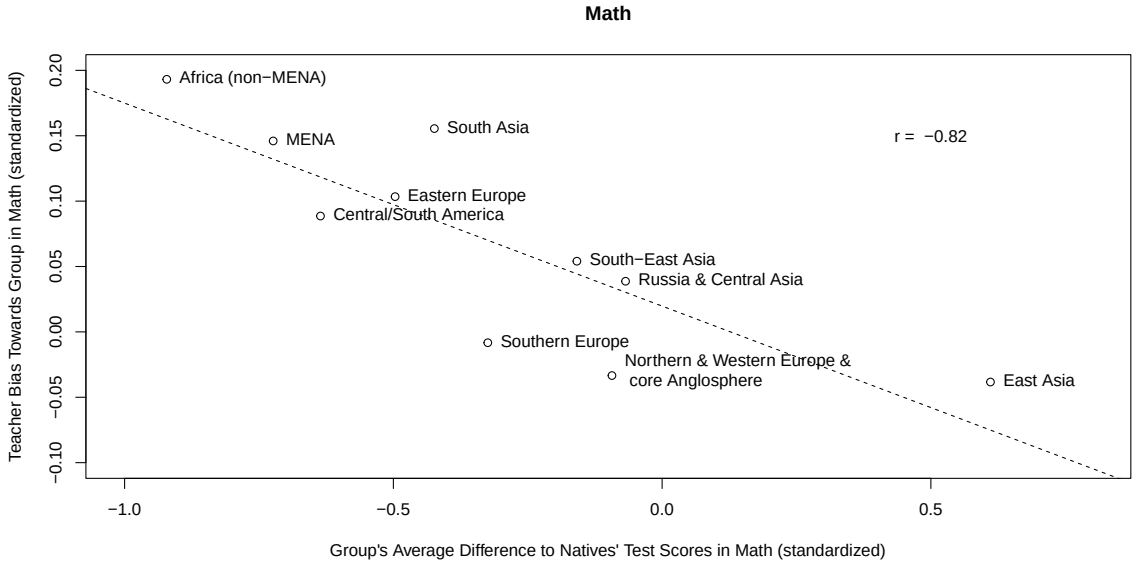

Figure A1: Results (point estimates) differentiating among the region of origin for students with a migration background. Computed as averages on the region-level. Teacher bias towards students of a specific regional background is depicted on the y-axis while groups' average difference in test scores compared to non-migrants is on the x-axis.

B Identification Analysis

We first show how one could estimate average biases, and then discuss assumptions under which causal interactions between student demographics and teacher-/school-characteristics can be estimated. It turns out that the assumptions needed for these two aims are different, and that they appear more plausible for the interaction case. We illustrate these points by means of simulations.

Consider a linear version of the DAG in Figure 1:

$$A = \alpha_A + \beta_1 X + U + \epsilon_A, \tag{1}$$

$$T = \alpha_T + \beta_2 X + \beta_3 A + U + \epsilon_T, \tag{2}$$

$$E = \alpha_E + \beta_4 A + \epsilon_E. \tag{3}$$

β_2 is the effect of interest–the direct effec ot student demographics on teacher grades, controlling for abilities. Subtracting E from T and simplifying yields

$$T - E = \alpha_{T-E} + \beta_2 X + (\beta_3 - \beta_4)A + (U + \epsilon_T - \epsilon_E). \quad (4)$$

If we assume $\beta_3 = \beta_4$— that is, the causal effect of ability on teacher grades is the same as the causal effect of ability on test grades—, A drops out from the equation. The DAG further implies that X is independent from the composite error term made up of U, ϵ_T and ϵ_E, so that the β_2 can be estimated by simple linear regression.

The requirement that $\beta_3 = \beta_4$ seems very strong. Furthermore, we need to assume that X does not directly affect E; that is, there is no differential measurement error. However, this is likely if for example girls perform worse on standardized tests due to lower risk aversion compared to boys (Niederle and Vesterlund 2010) or if migrants perform on worse on Math tests due to language difficulties (which are not supposed to be measured by Math tests).

However, by running regressions of $T - E$ on X for each teacher j separately, we obtain (biased) estimates $\widehat{TB}_j$ of the bias of that teacher, TB_j. We then have

$$\widehat{TB}_j = \gamma_j + TB_j, \quad (5)$$

where γ_j measures teacher-level systematic and random deviations of the estimate from the true bias. When we are interested in how teacher-level characteristics Z_j affect this teacher bias, we can estimate the model

$$\widehat{TB}_j = \delta Z_j + \gamma_j + \epsilon_j, \quad (6)$$

which is possible as long as teacher-characteristics are independent from other determinants of teacher bias (γ_j, ϵ_j). One way we test this is using the placebo tests described in the main text.

To underscore these points, we now present the results from simulations consistent with the discussed data-generating processes. We consider two scenarios:

1. A best-case scenario illustrating how average biases could be estimated in principle

2. A more realistic scenario where average bias cannot be estimated due to differences in how grades relate to A as well as differential measurement error, but nonetheless effects of teacher characteristics can be estimated.

We give R code for our simulations that can be directly used to replicate our analyses. The code for the first scenario looks as follows:

```
set.seed(39823)

N <- 100000
Nteacher <- 1000
```

```
df_1 <- data.frame(adj = NA, diff = NA, inter = NA)

for(rep in 1:1000){

  X <- rbinom(n = N, size = 1, prob = 0.5)

  U <- rnorm(n = N)

  A <- 0.5*X + U

  E <- 0.9*A + rnorm(n = N, sd = 0.25)

  Z <- runif(n = Nteacher, min = -0.1, max = 0.3)
  Z <- rep(Z, each = N/Nteacher)

  T <- X*Z + 0.9*A + 2*U

  df_1[rep, "adj"] <- coef(lm(T ~ X + E))["X"]
  df_1[rep, "diff"] <- coef(lm(T - E ~ X))["X"]
  df_1[rep, "inter"] <- coef(lm(T - E ~ X*Z))["X:Z"]

}
```

Here, we create 1000 samples of 100,000 students and 1,000 teachers. Each teacher teaches 100 students. The variable interpretations are consistent with our notation. In this particular setup, ability A affects test grade E and teacher grade T the same ($\beta_2 = \beta_4 = 0.9$). Z varies uniformly on the teacher-level between -0.1 and 0.3 such that its mean is 0.1. Therefore, the average bias (the average direct effect of X on T) is also 0.1. The interaction between X and Z, the causal effect of the teacher characteristics on biases, is 1.

We evaluate three estimators: First, a regression of T on X, controlling for E, which estimates average biases (the "adjusted estimator"). Second, a regression of $T - E$ on X (our estimator of average biases) (the "difference estimator"). Third, a regression of $T - E$ on the interaction of student demographics X and teacher characteristics Z (the "interaction estimator"), which is what we use to estimate the effect of teacher characteristics on biases.

Density plots of the sampling distributions of the estimators are depicted in Figure A2. Across the samples, the adjusted estimator is heavily biased. Its mean is about -0.8, whereas the true effect is 0.1. This means that it is very distorted both in terms of its sign as well as in terms of the magnitude of the bias.

The difference estimator, on the other hand, appears to be exactly unbiased (mean ≈ 0.1). The same holds for the interaction estimator, whose mean is also approximately equal to the true interaction effect. This confirms our analysis. Note that the

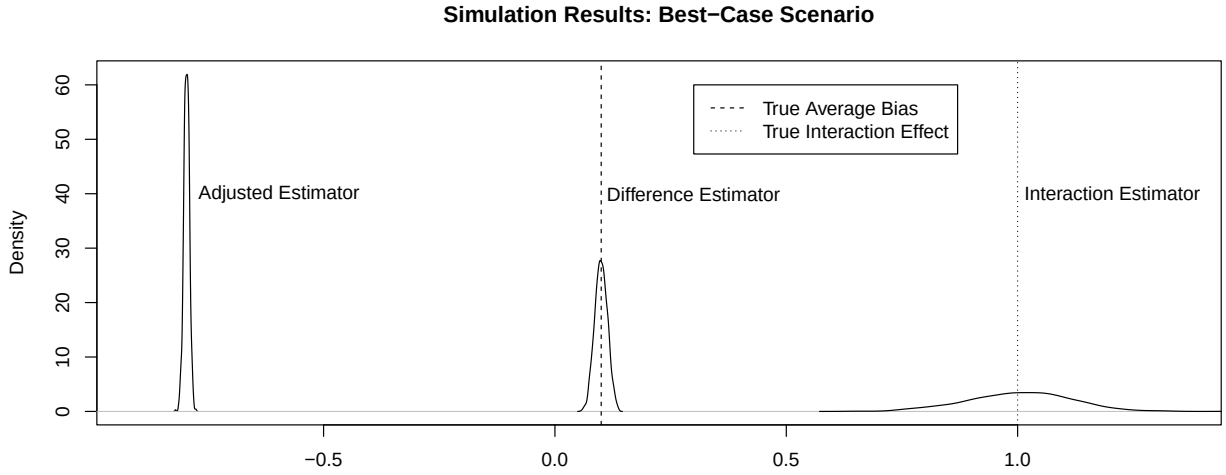

Figure A2: Results from simulations: Best-case scenario. Sampling distributions of three estimators: The adjusted estimator, the difference estimator (both aiming to estimate average biases), and the interaction estimator.

simulations clearly show that the interaction estimator is more variable than the other estimators. This is because for the first two estimators, the independent variable varies on the student-level, whereas for the interaction estimator, it varies on the teacher-level. However, this is simply reflected in larger standard errors in our empirical analyses.

Figure A3 shows the results from simulations where we change the generation of the test variable E as follows:

```
E <- -0.2*X + 0.7*A + rnorm(n = N, sd = 0.25)
```

Accordingly, student demographics affect test results directly, and the test measures ability in a less precise manner than the teacher grade. This can be deemed realistic, and we therefore call this the "realistic scenario". All other parameters stay the same. Therefore, the average bias is still 0.1, while the interaction effect is 1.

As a result of these changes, both the adjusted and the difference estimator of average biases are systematically distorted. The mean of the adjusted estimator is now about 0.19, while the mean of the difference estimator is now about 0.40. Both clearly overestimate true average biases. However, consistent with our argumentation, the interaction estimator still appears to be approximately unbiased.

Taken together, this underlines how we may be able to evaluate the causal effects of teacher characteristics on biases even when we cannot estimate those biases.

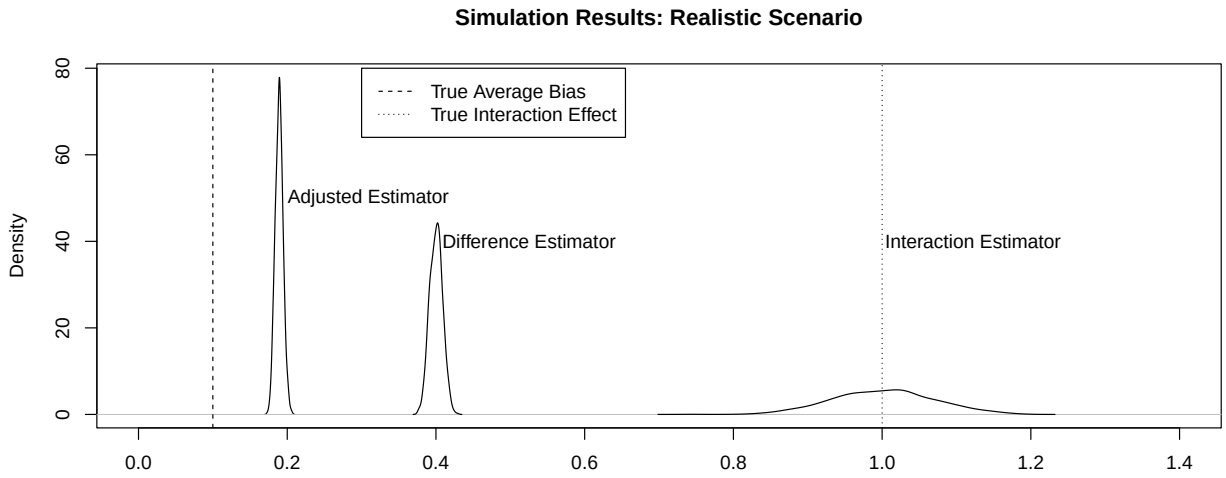

Figure A3: Results from simulations: Realistic scenario. Sampling distributions of three estimators: The adjusted estimator, the difference estimator (both aiming to estimate average biases), and the interaction estimator.

References

Niederle, Muriel and Lise Vesterlund. 2010. "Explaining the gender gap in math test scores: The role of competition." *Journal of economic perspectives* 24(2):129–144. Publisher: American Economic Association.