Ethnic and social class discrimination in education: Experimental evidence from Germany

Wenz, Sebastian E.; Hoenig, Kerstin

This is an accepted manuscript of an article published in
Research in Social Stratification and Mobility, 2020, Vol. 65, Article 100461
available online at: https://doi.org/10.1016/j.rssm.2019.100461

Ethnic and social class discrimination in education: Experimental evidence from Germany

Sebastian E. Wenz[a,*], Kerstin Hoenig[b]
[a] *GESIS Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8 50667 Cologne, Germany*
[b] *Leibniz Institute for Educational Trajectories (LIfBi), Wilhelmsplatz 3, 96047 Bamberg, Germany*

ARTICLE INFO

Keywords: Discrimination; Experimental research; Education; Teachers; Ethnic inequality; Social inequality

ABSTRACT

Even though social class is at least as predictive of educational achievement as ethnicity in virtually all developed countries, experimental research on discrimination in education has overwhelmingly focused on the latter. We investigate both ethnic discrimination and social class discrimination by elementary school teachers in Germany. We conceptualize discrimination as causal effects of signals and use directed acyclic graphs (DAGs) to disentangle ethnic from social class discrimination. In our experiment, we asked randomly sampled elementary school teachers who teach immigrants to evaluate an essay written by a fourth-grader. Employing a 2x2x3 factorial design, we varied essay quality, child's gender, and ethnic and socioeconomic background using names as stimuli. We do not find evidence for discrimination in grading. However, our findings for teachers' expectations of children's future performance suggest a discriminatory bias along the lines of both ethnicity and social class. The effect is conditional on essay quality—it only holds true for the better essay. We interpret our findings as evidence for models that highlight situational moderators such as the richness of information and ambiguity—e.g., statistical discrimination—but as evidence against simpler models of ingroup-favoritism or outgroup derogation, e.g., social identity theory or taste discrimination.

1. Introduction

Race, ethnicity, and immigrant background are powerful predictors of educational achievement and attainment in virtually all developed countries, among which Germany shows comparatively large disparities between children with immigrant background—in particular those of Turkish descent—and natives (Dustmann, Frattini, & Lanzara, 2012; Heath, Rothon, & Kilpi, 2008; Marks, 2005, a; OECD, 2016).[1]

One possible cause for these disparities in education is discrimination by teachers. In the US and the UK, the issue of racial and ethnic discrimination in education has been debated for quite some time now. In contrast, in continental Europe, including Germany, a broader discussion that explicitly separates discrimination from other sources of educational inequality has just recently begun. Yet, recent literature reviews agree in concluding that for all these countries we know only little on the question to what extent, if at all, racial or ethnic disparities in educational achievement are due to discrimination and, if so, what mechanisms are at work (Diehl & Fick, 2016; Farkas, 2003; Heath et al., 2008; Kristen, 2006a; Mickelson, 2003).

While it is a well established finding that socioeconomic or social class background is—again: in virtually all developed countries—an even stronger predictor for educational achievement and attainment than ethnicity (Breen & Jonsson, 2005; Marks, 2005b; OECD, 2016), social class discrimination in education has not been studied to the same extent as ethnic discrimination and, thus, we know even less about it.

Therefore, in this article, our main research question is whether we find empirical evidence for or against both ethnic and social class discrimination by elementary school teachers in Germany. To address the limitations of observational studies in identifying discrimination, we turn to an experimental design. As a methodological contribution, we address several limitations of experimental designs that assess discrimination in education. Most importantly, we discuss how to and seek to define, identify, and estimate ethnic discrimination independently from social class discrimination using names as treatments. Additionally, we explicitly theorize discrimination in education and design our experiment in a way that enables indirect testing of mechanisms suggested by different theories of discrimination.

* Corresponding author.
E-mail addresses: sewenz@gmail.com (S.E. Wenz), kerstin.hoenig@lifbi.de (K. Hoenig).

[1] Clearly, race, ethnicity, and immigrant background are not one and the same thing. However, throughout this article, we will use these terms interchangeably as our main concern with previous studies and the solution we propose applies to all three concepts in essentially the same way.

https://doi.org/10.1016/j.rssm.2019.100461
Received 28 February 2019; Received in revised form 20 December 2019; Accepted 20 December 2019

2. Conceptualizing discrimination

Our definitions of ethnic discrimination and social class discrimination are built on both the substantive literature on discrimination as causal effect (Blalock, 1967; Blank, Dabady, & Citro, 2004; Heckman, 1998; Pager & Shepherd, 2008; Quillian, 2006) and the methodological literature on causality (Greiner & Rubin, 2010; Imai, Tingley, & Yamamoto, 2013; Pearl, 2001, 2009; Pearl, Glymour, & Jewell, 2016; Rubin, 1986; VanderWeele & Hernán, 2012; Wang & Sobel, 2013).

We define *ethnic discrimination* as the total causal effect of an ethnic signal sent out by an individual on how this individual is treated by another individual. Analogously, *social class discrimination* refers to the total causal effect of a signal carrying information about social class.

Defining discrimination as *total causal effect* of *signals* avoids two major problems: first, defining discrimination as causal effect of *signals* avoids confusing it with unconditional inequality. Characteristics such as race, ethnicity, or sex can be seen as assigned very early in life—e.g., at conception—and, afterwards, immutable (Greiner & Rubin, 2010; Holland, 1986; Rubin, 1986). When defined as total causal effect of such an immutable characteristic, it can be hard to distinguish discrimination from unconditional inequality (Greiner & Rubin, 2010; Sen & Wasow, 2016; VanderWeele & Hernán, 2012). Panel (a) of Fig. 1 illustrates the problem: Following the rules of the graphical theory of causality (Pearl, 2009), identifying the total causal effect of ethnicity at conception, E_C, on a teacher-driven outcome such as a grade or track recommendation, Y, by means of covariate adjustment, requires conditioning on confounder C_C, social class at conception, but forbids conditioning on M, a mediator such as performance. The resulting definiens of ethnic discrimination would be equivalent to inequality in grades or track recommendations between different ethnic groups conditional on social class but not conditional on possibly remaining differences in performance between these groups. This is not what we think of as discrimination.

Secondly, by defining discrimination as *total* causal effect, we avoid direct effect conceptualizations of discrimination that are often proposed as solution to the first problem (e.g., Imai et al., 2013; Pearl, 2001, 2014; VanderWeele & Hernán, 2012; Wang & Sobel, 2013). In panel (a) of Fig. 1, $E_C \rightarrow Y$ would be the direct effect definition of ethnic discrimination. While they help distinguishing discrimination more clearly from inequality, the meaning of direct effects depends on the number and nature of all corresponding indirect effects. That is, changing M or adding another mediator to the relation between E_C and Y, would change the *definition* of discrimination. This appears to us as undesirable. In contrast, the definition, estimation, and interpretation of total causal effects is straightforward (Gangl, 2010; Pearl, 2001). Panel (b) of Fig. 1 introduces signals of ethnicity, E_S, and social class, C_S, whose total effects on Y constitute ethnic and social class discrimination, respectively.

Also, the definitions proposed here offer a treatment that is perfectly manipulable in theory and practice and, thus, allow to pose well-defined causal questions that can be studied using both observational and experimental designs (Greiner & Rubin, 2010; Holland, 1986; Rubin, 1986).

3. Identifying discrimination

According to the definitions above, identifying ethnic or social class discrimination is equivalent to identifying the total causal effect of an ethnic or social class signal, respectively. Identifying these effects requires assumptions about the causal paths that connect signals with the outcome of interest. Since we are interested in discrimination as it occurs in the population, we also need to ensure that our sample allows for causal inference. We argue that these points hold for both observational and experimental designs but are often ignored in experimental studies.

3.1. Observational studies

In observational studies, ethnic discrimination is usually—often implicitly—conceptualized as a direct effect of the form $E_C \rightarrow Y$ in panel (a) of Fig. 1. In panel (a) of Fig. 1, it is necessary to condition on C_C and M to identify the effect $E_C \rightarrow Y$. Social scientists have long argued that ethnic or racial discrimination might at least partly be a problem of social class discrimination (e.g., Blalock, 1967; Mickelson, 2003; Myrdal, 1944). In Germany, for example, immigrants with Turkish background are overrepresented in the lower classes and have generally worse labor market outcomes than the ethnic majority (Kalter, 2008; Kogan, 2007). Therefore, in most observational studies on ethnic discrimination in education, measures of social class background or proxies such as parental education are controlled for in addition to measures of prior performance or achievement.

Evidence from such studies on Germany is rather mixed: in the first study of its kind for Germany, Kristen (2006b), who does not control for social class, finds only weak evidence for discrimination to the disadvantage of students of Italian background (Kristen, 2006b, footnote 7) and no evidence for discrimination against students of Turkish background in both grades and track recommendations in a sample of six schools from the city of Mannheim.

Using larger samples, others (e.g., Lüdemann & Schwerdt, 2013) find statistically significant negative ethnic residuals in grades or track recommendations controlling for performance and ability measures.

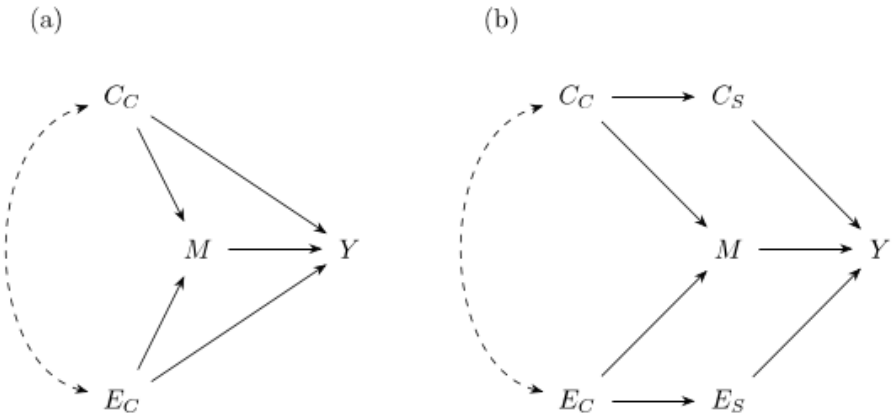

Fig. 1. Stylized directed acyclic graphs (DAGs) visualizing the consequences of different definitions of discrimination in education, where Y is a teacher-driven outcome, e.g., a grade or track recommendation, and M is a mediator, e.g., performance. Panel (a) illustrates problems of conceptualizing discrimination as causal effect of immutable characteristics assigned very early in life—e.g., ethnicity and social class at conception, E_C and C_C, respectively. Panel (b) illustrates the solution by conceptualizing discrimination as causal effects of signals—e.g., signals of ethnicity, E_S, or social class, C_S.

However, these differences largely vanish when social class is controlled for (e.g., Becker & Beck, 2012; Gresch, 2012; Lüdemann & Schwerdt, 2013), or even turn into positive residuals (e.g., Gresch, 2012). In other studies ethnic penalties remain even after proxies for social background (e.g., parental education, Kiss, 2013) are introduced.

The evidence from observational studies with regard to social class is stronger: conditional on measures of competence, performance, or achievement, many studies find that teachers recommend or prefer lower tracks for students from lower class families and grade them worse than students of higher class background (e.g., Bos, Tarelli, Bremerich-Vos, & Schwippert, 2012; Bos, Wendt, Köller, & Selter, 2012; Ditton, 2013; Wendt, Bos, Köller, Schwippert, & Kasper, 2016). However, following the logic visualized in panel (a) of Fig. 1 and the underlying substantive literature cited above, identifying social class discrimination in the sense of $C_C \rightarrow Y$ requires conditioning on M and E_C. Studies that do so—i.e., additionally control for ethnicity, immigrant background, or proxies such as language use at home—find that disadvantages of students from lower class families decrease in size but typically remain statistically significant and are larger than the disadvantages of immigrants (e.g., Becker & Beck, 2012; Ditton, Krüsken, & Schauenberg, 2005; Gresch, 2012; Schneider, 2011). Given these findings, it is remarkable that most of these studies do not investigate social class discrimination explicitly (but cf. Schneider, 2011).

The major limitation of observational studies on discrimination is that they typically provide so-called residual estimates of discrimination only (Oaxaca, 1973). As all other estimates from observational studies, they hinge on the strong assumption that all relevant controls have been included in the model and measured without error (Oaxaca, 1973; Pearl, 2009; Rubin, 1974). With regard to education, residual estimates might either over- or underestimate discrimination due to under- or overcontrolling of key variables, respectively (Holzer & Ludwig, 2003): undercontrolling could arise from imperfect measures of students' actual performance at school. Overcontrolling and, thus, underestimation of discrimination, might happen in the face of biased test scores due to mechanisms such as stereotype threat (Croizet, 2008; Steele & Aronson, 1995).

3.2. Experimental studies

In experimental studies, ethnic discrimination is usually—again: often implicitly—conceptualized as either total effect of an ethnic signal or information, e.g., $E_S \rightarrow Y$ in panel (b) of Fig. 1, or some kind of direct effect of an ethnic signal or information. Identification of the total causal effect, $E_S \rightarrow Y$ in panel (b) of Fig. 1, is seemingly straightforward in experiments: While different strands in the literature use different terminology, all agree that a key advantage of successful randomization is that it, in combination with few other assumptions, blocks all backdoor-paths from E_S to Y, since it renders E_S "parentless" (Pearl et al., 2016, p. 78). Put differently, E_S becomes indepedent of the potenial outcomes and, thus, exogenous (Hernán & Robins, 2020; Imbens, 2004; Pearl, 2009; Pearl, Glymour, & Jewell, 2016; Rubin, 1974, 1978).

However, we argue that experimental studies on ethnic discrimination, when using names or pictures (e.g., Deming, Yuchtman, Abulafi, Goldin, & Katz, 2016; Jacquemet & Yannelis, 2012; Sprietsma, 2013; van Ewijk, 2011; Weichselbaumer, 2016), apply treatments that potentially carry both ethnic *and* social class signals (for this and similar arguments see Bertrand & Mullainathan, 2004; Figlio, 2005; Fryer & Levitt, 2004; Gaddis, 2015, 2017a, 2017b). So, these studies will typically not identify ethnic discrimination according to the definition above, but confound ethnic with social class discrimination (cf. Bertrand & Mullainathan, 2004).

Fig. 2 visualizes the problem: names, N, are determined by both E_C and C_C and affect both E_S and C_S. Hence, the challenge for researchers setting up an experiment to study ethnic discrimination is to construct a name variable whose values send varying ethnic signals but remain constant on their social class signal. The popular approach of using *typical* ethnic *minority* and *typical* ethnic *majority* names is no solution, as a typical majority name usually carries a different class signal than a typical minority name. Thus, since social class is not held constant in the experimental manipulation, it is possible that any discrimination found in these studies is—partly or even completely—the result of class differences, not ethnicity.

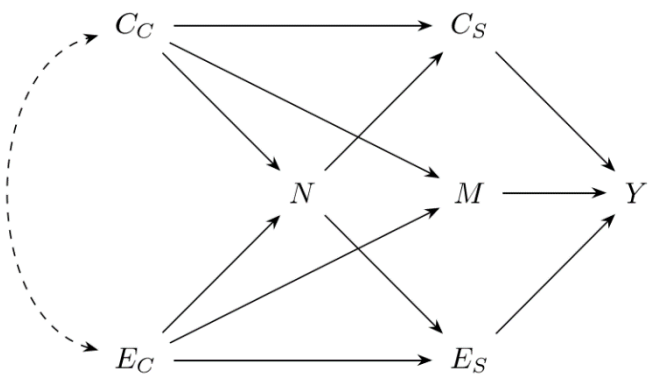

Fig. 2. Stylized directed acyclic graph (DAG) showing the problems of identifying ethnic discrimination, $E_S \rightarrow Y$, and social class discrimination, $C_S \rightarrow Y$, using names, N, as treatments. A randomized assignment of names, N, blocks the backdoor paths through ethnic and social class at conception, E_C and C_C. However, if N carries both ethnic and social class signals, E_S and C_S, the backdoor paths $E_S \longleftarrow N \rightarrow \quad \rightarrow Y$ and $C_S \longleftarrow N \rightarrow E_S \rightarrow Y$ remain open.

While vignette designs that make use of extensive descriptions of students (e.g., Glock, Krolak-Schwerdt, & Pit-ten Cate, 2015; Hanna & Linden, 2012; Schulze & Schiener, 2011) may help to address the problem, they have other limitations of which the most important is that they oftentimes create rather artificial settings. Due to the large amount of information on individual students and because teachers often have to rate several students, these designs make it difficult to provide a reasonable cover story that hides the true purpose of the study. Therefore, in principle, names remain an attractive stimulus in experiments on discrimination. However, to the best of our knowledge, no experimental study in the field of education had made a conscious effort to disentangle ethnic discrimination from social class discrimination using names as stimuli at the time our study was conducted. Tobisch and Dresel (2017) have since published a study with a similar design that replicates our main findings.

A small number of experimental studies has examined the issue of discrimination by teachers in Germany. In her sample of $N = 88$ teachers, Sprietsma (2013) finds that essays with randomly assigned Turkish names receive grades of about 0.1 standard deviation worse than essays with a German name on them. Also, teachers are on average 11 percentage points less likely to recommend the highest track, *Gymnasium*, to a student with a Turkish name compared to a student with a German name; no difference is found for the intermediate *Realschule*. To signal ethnicity, Sprietsma (2013) uses typical ethnic minority and ethnic majority first names; social class is thus not held constant.

Using extensive vignettes, Schulze and Schiener (2011) find that education students at four German universities do not discriminate based on the language students speak at home, but do discriminate based on parental education. While Glock, Krolak-Schwerdt, Klapproth, and Böhmer (2013) also hold constant students' socioeconomic background, they do not report estimates of social class discrimination. In their sample of Luxembourgian teachers, they do find evidence for ethnic discrimination to the disadvantage of students with a Portuguese name in track recommendations, though.

3.3. Sampling and sample of analysis

Unbiased causal inference about discrimination in a population of

interest not only requires identifying discrimination in a given sample. It also depends on sampling strategy and sample of analysis (Berk, 1983; Elwert & Winship, 2014; Heckman, 1979; Hernán, Hernández-Díaz, & Robins, 2004).

Regarding sampling, only few experimental studies in the field of discrimination in education rely on random samples from a clearly defined population (e.g., Sprietsma, 2013; van Ewijk, 2011). Many experimental studies rely on convenience samples, often drawn from populations such as university students in education programs or preservice teachers (e.g., Bonefeld & Dickhäuser, 2018; Glock, Krolak-Schwerdt, & Pit-ten Cate, 2015; Schulze & Schiener, 2011). In their sample, Glock et al. (2015) even mix teachers from one European country with preservice teachers from another. Such approaches reduce external validity and, thus, increase bias in estimates of population parameters.

However, a simple random sample of teachers or schools may not be the sample of interest: for the labor market, Heckman (1998), based on Becker (1971), makes a usually ignored point and argues that due to self-selection of employees into non-discriminating firms, random samples of firms are unlikely to provide unbiased estimates of discrimination as it takes place in the market as a whole (Heckman, 1998, p. 102). This argument also applies to elementary education in Germany, where teachers might differ systematically in their cognitive, affective, and behavioral responses towards ethnic minority students depending on whether they teach them. Such a difference might be due to self-selection of teachers into schools with different student compositions, to a change in attitudes and beliefs as a consequence of teaching minority students, or to the self-selection of ethnic minority and ethnic majority parents into different schools with different teachers. We have not seen this exact point being explicitly addressed in any empirical study on discrimination in education, although van Ewijk (2011) makes a similar point and oversamples schools with a share of at least 25% non-ethnic Dutch students.

4. Explaining discrimination in education

4.1. Opportunities for discrimination in education

Discrimination by teachers may occur in at least three situations: day-to-day classroom interactions (e.g., by calling more on some students than others), the evaluation of a specific performance (e.g., grading a test or assignment), and the evaluation of a student's general potential or ability (e.g., decisions about ability grouping or tracking). In our experiment we take a look at the two latter types and deduce different predictions from different theories of discrimination depending on the situation. This enables us to provide an indirect test of the mechanisms suggested by these theories.

4.2. The transition from elementary school to secondary school in Germany

The transition from elementary school to secondary school in Germany provides us with a test case scenario in which both teachers' grading and expectation formation matter. Even though the 16 German federal states are responsible for their education policy, key features of the education system that are pertinent to our experiment are similar across states. All children start elementary school around the age of six. Usually, a single teacher teaches the main subjects and there is no formal ability grouping or streaming. Students are tracked into different school types after four years of elementary school when they are on average 10 years old.[2] The number and specifics of tracks differ between states, but in all states, the highest track is the *Gymnasium*, which leads to the *Abitur*, the highest secondary degree and entrance ticket to university. In all states, elementary school teachers give official recommendations suggesting the track they believe would be ideal considering the child's potential. The major determinants of these recommendations are grades, but teachers are, legally or factually, asked to consider the child's overall potential.

Because tracking between different school types occurs unusually early in the German education system and because a child's track is largely determined by the teacher's recommendation, discrimination by elementary school teachers can have especially severe consequences for children's educational attainment. However, our theoretical and methodological contributions are of much broader significance, as both the grading of students' performance in a non-anonymous setting and some form of tracking, streaming, or ability grouping, dependent largely on teachers' evaluations and grading, takes place in virtually all education systems. Furthermore, teachers' expectations may turn into self-fulfilling prophecies that can be especially harmful to students from stigmatized groups (Jussim, Eccles, & Madon, 1996; Jussim & Harber, 2005; Lorenz, Gentrup, Kristen, Stanat, & Kogan, 2016).

4.3. Discrimination in grading and expectations: theories and hypotheses

The most obvious and, as we are going to show in this section, theoretically important difference between forming expectations and grading a manifest performance such as a written essay is the amount and reliability of individuating information available to the teacher. While in principle an essay provides the teacher with all information needed to grade it, tracking or grouping decisions are based on expectations that are themselves necessarily based on imperfect information about the latent construct of ability or potential and—especially in situations of explicit between school tracking—a yet unobserved future. As we intend to provide evidence on the mechanisms governing teachers' judgments, in this section we theorize discrimination in grading and expectation formation and deduce hypotheses from different theoretical perspectives.

4.4. In-group favoritism, prejudice, and tastes

People tend to categorize others automatically and immediately on the basis of salient category cues such as race, ethnicity, or sex, distinguishing between ingroups, i.e., those groups they belong to, and outgroups, i.e., those they do not belong to (Allport, 1954; Sumner, 1906). Numerous studies have documented that categorization is a general and basically inevitable process that is typically followed by biased perceptions, evaluations, and behavior in favor of ingroups and ingroup members (Fiske, 1998).

Explanations for these observations include rational actors that simply discriminate because group preferences enter their utility function (Becker, 1971), real or perceived conflicts about limited resources and relative social and economic positions between groups (Blumer, 1958; Campbell, 1965; Sherif, 1966; Sherif, Harvey, White, Hood, & Sherif, 1961), the importance of an individual's social identity in satisfying a need for self-esteem (Tajfel, 1982; Tajfel & Turner, 1986), and mainly cognitive processes as a consequence of categorization. We argue that all these theories predict ingroup favoritism—and, to a lesser extent, outgroup derogation—regardless of the amount and reliability of individuating information available.

Given that most teachers in German elementary schools are of German ethnicity, belong—as university graduates employed in the public sector—to the upper middle class, and evidently hold negative stereotypes and prejudices about students from lower classes and with a Turkish background (Glock & Klapproth, 2017; Wenz, Olczyk, & Lorenz, 2016), we deduce the following hypotheses:

$Hypothesis1_a$: German teachers discriminate in *both* essay grading and expectation formation by virtue of students' ethnic background to the disadvantage of students with a name signaling a Turkish background.

[2] In two states—Berlin and Brandenburg—most children are tracked after six years at the age of 12.

Hypothesis 1_b: German teachers discriminate in *both* essay grading and expectation formation by virtue of students' social class background to the disadvantage of students with a name signaling a lower social class background.

4.5. The role of imperfect information and ambiguity

theoretical contributions that point to situational moderators of the link between categorization and the application of stereotypes or prejudice and, hence, discrimination include statistical discrimination theory (Aigner & Cain, 1977; Arrow, 1973; Phelps, 1972), the continuum model (Fiske, Lin, & Neuberg, 1999; Fiske & Neuberg, 1990), and the theory of aversive racism (Dovidio & Gaertner, 2008; Gaertner & Dovidio, 1986).

Statistical discrimination (e.g., Aigner & Cain, 1977) points to imperfect knowledge as the key reason for why rational decision makers discriminate on the basis of observable group characteristics. Following statistical discrimination theory, teachers should be expected to construct a weighted average of observed individual performance and assumed group ability to estimate a student's individual ability: the lower the reliability of the individual information, the further the estimate is pulled towards the assumed group mean, that is, towards the teacher's stereotype.

The continuum model (Fiske et al., 1999; Fiske & Neuberg, 1990) acknowledges available information as one of "two primary factors" in more or less category or stereotype driven and, hence, more or less discriminatory judgments and behavior. If the target is of minimal interest or relevance for the perceiver in the very moment of categorization, perceivers are motivated to allocate attention to individuating information and move down the continuum from category-based "affect, cognitions, and behavioral tendencies" toward a "piecemeal integration" of individual attributes (Fiske et al., 1999, 233). Of course, this process of recategorization and, eventually, piecemeal integration may only be started if the available information is rich enough and the perceiver has the time and the cognitive capacity to take it into account.

The theory of aversive racism (Dovidio & Gaertner, 2008; Gaertner & Dovidio, 1986) suggests that in modern societies, where negative prejudice and discrimination to the disadvantage of ethnic and social minorities is condemned by a majority of people, many people are motivated to uphold a positive self-image as unprejudiced non-discriminator but at the same time hold negative implicit prejudices and stereotypes about outgroups and outgroup members. These aversive racists are expected to discriminate only in situations that are ambiguous enough to not reveal the discriminatory behavior to themselves and others.

All theories discussed in this section assume that the application of stereotypes and prejudice is dependent on the situation: as for the logic of the situation teachers find themselves in when grading a written essay, sufficiently motivated teachers should *not* show any discriminatory biases, since all relevant information is available (Aigner & Cain, 1977; Fiske et al., 1999) and a discriminatory bias is hard to hide (Gaertner & Dovidio, 1986).

However, when the same teachers need to predict future development of students when recommending tracks, the available information based on an essay might not be perceived as perfectly reliable (Aigner & Cain, 1977), or—put differently—might not be rich, diagnostic, or clear enough (Dovidio & Gaertner, 2008; Fiske et al., 1999; Fiske & Neuberg, 1990; Gaertner & Dovidio, 1986). In this case, teachers should make use of beliefs about group means and stereotypes in general (Aigner & Cain, 1977), may have not enough information to go all the way from a category-based response to a piecemeal-based response (Fiske et al., 1999; Fiske & Neuberg, 1990), and may take the opportunity to hide a judgment based on stereotypes or prejudices behind vague information and the ambiguity of the situation (Gaertner & Dovidio, 1986).

Based on how statistical discrimination theory, the continuum model, and the theory of aversive racism acknowledge the reliability of information and the ambiguity of situations as moderating factors that increase the likelihood of category based judgments, we deduce the following hypotheses:

Hypothesis 2_a: German teachers *do not* discriminate when grading a written essay *but do so* when forming expectations by virtue of students' ethnic background to the disadvantage of students with a name signaling a Turkish background.

Hypothesis 2_b: German teachers *do not* discriminate when grading a written essay *but do so* when forming expectations by virtue of students' social class background to the disadvantage of students with a name signaling a lower social class background.

4.6. Expectations regarding effect heterogeneity

Models of statistical discrimination theory suggest that discrimination may differ along the distribution of observed performance (Aigner & Cain, 1977; Phelps, 1972). Accordingly, if teachers engage in reliability-based statistical discrimination, it is possible that the discriminatory effect not just varies but even changes its direction over the range of observed performance.

Therefore, studies that only assess discrimination at one point along the performance distribution are severely limited: without making further strong assumptions, they neither allow inference about the average level or direction of individual discrimination nor can they say anything about group discrimination—not even which group is on average suffering from it (also see Heckman, 1998; Heckman & Siegelman, 1993; Neumark, 2012, for this argument). Interestingly, both van Ewijk (2011) and Sprietsma (2013) use a set of essays with varying quality but do not investigate the corresponding interaction effect. Bonefeld and Dickhäuser (2018) report a discriminatory bias against immigrants for a subset of teachers when performance was poor but not average.

While investigating them is straightforward, predicting heterogeneous effects across the competence distribution is less so. Clearest guidance in this regard comes from the formal models of statistical discrimination theory. They suggest that discrimination should differ along the distribution of observed performance if teachers attach different reliabilities to performance signals from different groups (Aigner & Cain, 1977; Phelps, 1972). However, deriving concrete hypotheses would require additional assumptions, since we don't know the group-specific reliabilities teachers perceive—i.e., the performance of which group is perceived more accurately (Glock et al., 2015; Kaiser, Südkamp, & Möller, 2016). We also don't know where along the competence distribution exactly teachers see the performance presented in the experiment and how risk-averse teachers really are (cf. Maaz et al., 2008).

With regard to the other two theories that highlight situational moderators such as the richness of information and ambiguity, it might certainly be the case that essays of different quality relate to these mechanisms: the continuum model (Fiske et al., 1999; Fiske & Neuberg, 1990) proposes that category-based affect, cognition, and behavioral responses are the default mode of human cognition. Only if the initial categorization of a person does not seem to fit the data, a recategorization process is started that potentially leads all the way down to a piecemeal integration of the available data and, thus, to an individuating response. It could be, for example, that the better of two essays is so good that the teacher finds it difficult to achieve a fit to the stereotype of a student with Turkish background and, instead of a category-based response, looks very carefully for individuating information and behaves accordingly. At the same time, the teacher might achieve a good fit for this essay and the stereotype of a German student from an upper middle class family. This might result in similar predictions for these students and no discrimination on the basis of ethnic signals that could disadvantage the Turkish student. If at the same time, the bad essay's nature is such that it allows the teacher to proceed with a category-based response in both cases—because the essay is rather

average, not very good, not very bad—there should be discrimination on the basis of the ethnic signal and corresponding stereotypes. Of course, this example also works the other way around—with a very bad and an average essay. In either case, the resulting pattern of this scenario would be an interaction effect of discriminatory responses with essay quality.

Similarly, relying on mechanisms from aversive racism theory we might also predict such an interaction effect. If performance is clearly very good or very bad, aversive racists will not apply stereotypes and prejudices in their judgments and behavior.

5. Research design and data collection

We conducted an online experiment in the spring of 2010 to identify and estimate discrimination in the grading of a specific assignment and in track recommendations. Our test subjects were elementary school teachers from the German federal state of Baden-Wuerttemberg who taught German at the time of data collection. We employed a 2 × 2 × 3 factorial design, in which we varied essay quality, child's gender, and child's social class and ethnic background between essays and teachers. Gender and background were varied by random assignment of names to the essays. Teachers graded one essay only to reduce respondent burden and the likelihood of them guessing the true purpose of the study. Afterwards, teachers were asked to answer a short questionnaire. Documentation of all material used in the study can be found in the online supplementary material at https://osf.io/43exq/.

5.1. Sampling and contact

Our main goal in sampling was to ensure a high external validity of our findings. To this end, we drew a random sample of 720 schools from a list of all elementary schools in the state of Baden-Wuerttemberg and contacted them via e-mail. The recipient—in most cases probably the school's secretary—was asked to forward the e-mail to all teachers at the school who taught German at that time.[3] As an incentive to participate, we had a lottery of three gift certificates for an online book store worth 20 Euros each, as well as the option to receive information about results. 237 teachers participated in the survey.[4] Teachers were told that the goal of the study was to find differences in teachers' evaluations due to different teacher characteristics and experiences.

5.2. Essays

Each teacher was presented with one of two essays of different quality. Essays were obtained from a fourth-grade class from Baden-Wuerttemberg. They were about 200 words long and were based on the assignment to write a story around a given title. The two essays for our experiment were chosen based on results of a pretest in the neighboring state of Bavaria. 27 teachers from randomly sampled Bavarian elementary schools were asked to grade six essays and guess the gender of the author without receiving any information about the author except for age and grade level. Based on the results of the pretest, we selected two essays that were of different quality and comparatively gender-neutral.

5.3. Children's names

To identify and estimate ethnic and class discrimination, we chose one male and one female name each that signal a German upper middle class (Jakob, Sophie), German lower class (Justin, Jacqueline), or Turkish background (Ayse, Murat), respectively. We made sure that the German names we selected are about equally prevalent in the relevant birth cohort and that none of the names are linked to a certain geographical region in Germany.[5] Although there are apparently typical upper and lower class Turkish names, we have reason to believe that German teachers are simply not familiar enough with Turkish culture to recognize the difference. Therefore, we cannot vary class background independently of migration background by name manipulation. Instead, we assume that a Turkish name indicates a class background that is comparable to that of German lower class names. Accordingly, differences between German lower class and Turkish names are interpreted as evidence of ethnic discrimination in the following discussion, whereas differences between the two German name groups are interpreted as class discrimination. The alternative would be a design based on comprehensive vignettes that explicitly include child background characteristics. We decided against the use of vignettes because these create a more artificial setting in which teachers are more likely to ask themselves why the researcher provided them with this information. In contrast, names make for a much more subliminal stimulus.

The names underwent a manipulation check using a separate sample of elementary school teachers ($N = 75$), who were asked to guess the migration and class background (upper, middle, or lower class) of each name. As intended, the vast majority of teachers indicates that Murat (69%) and Ayse (70%) have a Turkish background, whereas Jakob (94%), Sophie (97%), Justin (99%), and Jacqueline (92%) are virtually unanimously identified as German.[6] Remarkably, a sizable minority identifies Murat (23%) and Ayse (24%) as German.

Regarding social class, Sophie and Jakob are believed to come from families with an upper class (Sophie: 53%, Jakob: 59%) or middle class (Sophie: 40%, Jakob: 36%) background. As expected, for Jacqueline and Justin the pattern is reversed: these names are perceived predominantly as names held by children with a lower social class background (Jacqueline: 71%, Justin: 67%). However, a sizable minority of teachers categorizes them as middle class (Jacqueline: 21%, Justin: 25%) or even upper class (both 8%). The Turkish names are also perceived as being most likely to be names from students with a lower social class background (Ayse: 52%, Murat: 53%), followed by middle class (Ayse: 39%, Murat: 36%) and upper class (Ayse: 9%, Murat: 11%). Yet, fewer teachers report to perceive Ayse and Murat as lower class than Jacqueline and Justin.

Overall, teachers tend to perceive the selected names as we had intended. However, we identified potential problems[7] : if a sizable minority of participating teachers really perceive students with Turkish names as ethnically German, and students with lower class names as having a middle class background, our estimates of both ethnic discrimination and social class discrimination would be downwardly biased. Also, the estimate of ethnic discrimination would be downwardly biased if teachers perceived students with Turkish names as having a higher social class background than students with ethnic majority lower class names.

To better understand how severe these potential problems really are for our estimates and how they compare to other studies, consider this: first, our estimates are biased if and only if the numbers above deviate from the perception teachers in the population have. So, if in the population, for instance, the same sizable minority of teachers perceives students with a Turkish immigrant background not as having a Turkish

[3] Since we do not know how many teachers eventually received our invitation, we cannot calculate a precise response rate.
[4] Due to restrictions by the state's Ministry of Education, which had to approve our study, we were not allowed to ask teachers the name of their school in the questionnaire. This unfortunately means that we are unable to account for clustering of teachers by schools in our analysis.
[5] The German names were selected from a list compiled by Kube (2009).
[6] Although Justin and Jacqueline are not traditional German names, foreign names are popular among German families with a lower socioeconomic background. Evidently, the teachers in our sample recognize this fact.
[7] Thanks also to one of the reviewers, who suggested a closer look at these numbers than we took in the initial submission.

background but as ethnic majority German, then the numbers above do not indicate bias in our estimates. However, we do not know this number. Secondly, our numbers are similar to those reported for the perception of names of Blacks and Whites in the US: Gaddis (2017a) finds congruent perceptions of 87.3% for first names held by Whites and 75.0% for first names held by Blacks. That doesn't mean that our numbers are fine, but that in other countries and cultures, similar rates and differences are found.

Thirdly, and maybe most importantly, a closer look at the data reveals that the deviations on all three dimensions—ethnicity signal of Turkish names, class signal of lower class names, and class signal of Turkish names—are highly correlated. It is virtually the same group of teachers who deviates on all three items, except for the additional some 20% that declare to perceive the Turkish names as middle or upper class. The behavior of this group of teachers could well be a manifestation of social desirability bias instead of a real difference in perception. Admittedly, we do not know whether the deviations are due to socially desirable behavior by some teachers. We also don't know whether this social desirability bias, should it indeed be the explanation of the pattern we find, may also influence teachers' behavior in the field and, thus, *not* bias our estimates of discrimination.

5.4. Questionnaire

Teachers had to answer each item of the online questionnaire and could not go back once they had left a page. This was done to prevent teachers from skipping back and changing their evaluation of the child's performance as a reaction to items in the following questionnaire. At the beginning of the survey, teachers were presented with the essay, information about the specifics of the assignment, and the information that it was written by a ten year old fourth grader with a particular name. They were then asked to evaluate the child's performance with the following items:

1. "Which grade would you give [name of child] for this essay?" Teachers could assign German grades from 1 (best) to 6 (worst), including plus and minus signs to differentiate further between full grades, so that the full scale had 15 categories.
2. "How likely is it that [name of child] can keep up in German lessons at the *Gymnasium* with this performance?", rated on a scale of 1 to 5.

The essay was visible for each of these items. Afterwards, teachers answered a few questions about their teaching experience, the ethnic and social composition of their class, and their own social and ethnic background. On average, teachers took 12 min to complete the survey.

6. Analytic strategy

6.1. Essay grading

In order to assess discrimination in grading we model the grade as given by the teacher with two different models of which the following is the simpler one:

$$Y_i = \alpha + Q_i\beta_1 + F_i\beta_2 + T_i\beta_3 + U_i\beta_4 + \boldsymbol{C}_i\boldsymbol{\gamma} + \varepsilon_i \quad (1)$$

where Y is the grade assigned to essay i, Q captures the quality of the essay (good = 1), F identifies the gender of the name attached to the essay (female = 1), T distinguishes between Turkish and German names (Turkish = 1), and U stands for the social class associated with a German name (upper middle class = 1). Q, F, T, and U are dummy variables, $\boldsymbol{C}$ is a vector of controls. ε is an error term with the usual properties in an OLS scenario. To assess the sensitivity of the standard errors, we also estimated models featuring heteroskedasticity-robust standard errors. This did not change the significance levels of any of the parameters.

The second model features interactions testing whether name effects depend on the quality of the essay and looks as follows

$$Y_i = \alpha + Q_i\beta_1 + F_i\beta_2 + T_i\beta_3 + U_i\beta_4 + (Q_iT_i)\beta_5 + (Q_iU_i)\beta_6 + \boldsymbol{C}_i\boldsymbol{\gamma} + \varepsilon_i \quad (2)$$

Now, β_1 captures the difference between bad and good essay among German lower class names, β_2 still captures overall gender differences, β_3 estimates the difference between Turkish and German lower class names when the essay is bad, and β_4 the one between German lower class and upper middle class names for the bad essay. Finally, β_5 and β_6 estimate whether the effects estimated by β_3 and β_4 are any different when the essay quality is good instead.

The participating teachers graded the essays according to a usual 15 point German grading scale, $Y = \{1, 1-, \ldots, 5-, 6\}$, turned into a scale ranging from 0 (worst grade, German 6) to 14 (best grade, German 1), $Y = \{0, 1, \ldots, 13, 14\}$. Empirically, teachers assigned grades from 2 (German 5) to 12 (German 2+). For an easier and substantively more meaningful interpretation of our regression estimates, we standardized the grades and, thus, Y has a mean of 0 and a standard deviation (SD) of 1 in both the restricted and the full sample.

6.2. Teachers' expectations

German elementary school teachers' expectations about the future development of students' abilities and skills are crucial for students' success in the education system. Since the *Gymnasium* is the highest track available in all federal states, we focus on teachers' estimation of the likelihood that the child can keep up in German lessons at the *Gymnasium*.

Discrimination in expectations is assessed by modeling the probability of having a teacher assigning a particular likelihood of success. This ordinal variable originally has five categories with endpoints labeled "very unlikely" (1) and "very likely" (5), respectively. We model this ordinal dependent variable using an *ordinal logit model* (OLM) (Long, 1997).[8]

For both essays, teachers are hesitant to assign high likelihoods of success (see Table 1). In part, this is probably due to the average to low grades teachers gave the essays—as expected, grades and expectations are correlated ($r = 0.51$, $p < 0.001$). On the other hand, teachers have admittedly little information about the child's true ability after only one short essay, and German elementary school teachers tend to be risk averse when making track recommendations (Maaz et al., 2008). Thus, we should expect them to be cautious in their estimation of children's potential. To address its skewed distribution, we recoded Y. It now has three categories ($J = 3$), and is linked to the measurement model of the OLM as follows:

$$Y_i = \begin{cases} 1 \Rightarrow 1(\text{"very unlikely"}) & \text{if} \tau_0 = -\infty \leq Y_i^* < \tau_1 \\ 2 \Rightarrow 2 & \text{if} \tau_1 \leq Y_i^* < \tau_2 \\ 3 \Rightarrow 3, 4, 5(\text{"very likely"}) & \text{if} \tau_2 \leq Y_i^* < \tau_3 = \infty \end{cases} \quad (3)$$

where τ_1 through τ_{J-1} are cutpoints estimated in the OLM (Long, 1997).

Model specifications for teachers' expectations and the underlying latent variable in the OLM, Y_i^*, are the same as those for grades (see Eq. (1) and Eq. (2)).

For both model specifications, we test the *parallel regression assumption*, using the Wald test suggested by Brant (1990). Results suggest that the assumption holds for both models.

To foster interpretation of the results from this non-linear model and to address problems of group comparisons (Allison, 1999; Karlson, Holm, & Breen, 2012; Mood, 2010), we calculate and plot the probabilities of falling into the different categories of the dependent

[8] We also ran all models as linear regressions using the original Likert scale, as well as logistic regressions using a dichotomized variable combining categories 1 and 2 versus 3 to 5. Substantively, this does not alter our conclusions as discussed below.

Table 1
Teachers' expectations, dependent on essay quality.

Likelihood of keeping up at the *Gymnasium*	Essay quality				Total		
	Good		Bad		No.	Col %	Cum %
	No.	Col %	No.	Col %			
1 (very unlikely)	22	19.5	58	46.8	80	33.8	33.8
2	32	28.3	40	32.3	72	30.4	64.1
3	42	37.2	21	16.9	63	26.6	90.7
4	17	15.0	3	2.4	20	8.4	99.2
5 (very likely)	0	0.0	2	1.6	2	0.8	100.0
Total	113	100.0	124	100.0	237	100.0	

variable. Group differences are calculated as average marginal effects (AMEs) on the probability for a specific change in one of the independent variables averaging over all observations.

6.3. Sample of analysis

To increase external validity and arrive at a more accurate estimation of discrimination in the actual school context, we follow the convincing arguments in Heckman (1998, p. 102) and estimate all models using both the full and a restricted sample. We focus on the restricted sample that only includes teachers who report to have students with an immigrant background in their classes and, thus, have the opportunity to discriminate against them. Results for the full sample can be found in the appendix.

29 teachers did not report having students with an immigrant background in their class. We lose nine further cases because we control for background variables, of which some have missing data. The variables we control are the teacher's gender, immigrant background, and teaching experience, as well as the education of the teacher's parents. Thus, the restricted sample shrinks to $N = 199$, while the full sample has $N = 223$ observations.[9]

7. Results

7.1. Grading

The essay that was pretested as "good" received rather average grades, with a mean of 7.22 (SD = 1.97) on the original 14-point scale – corresponding to a "satisfactory" grade according to German grading conventions. However, it is significantly better than the bad essay that has a mean of 5.55, just above a passing grade (SD = 1.88; $t = 6.68$, $p < 0.001$). Table 2 shows basic summary statistics for each of the six names by essay quality on the original scale. Although there is some variation, no clear patterns are visible.

This impression is supported by the regressions we ran using standardized grades, as can be seen in Table 3. Excepting the coefficient for essay quality, β_1, no coefficient turns out significant on conventional levels. The same holds for the linear combinations that allow to test whether grades differ between groups for the good essay. The results for the interaction effects with essay quality also show that there are no group differences in returns to the good essay compared with the bad essay. Thus, we find no evidence of ethnic or social class discrimination in essay grading and reject hypotheses 1_a and 1_b.

Table 2
Summary statistics of grades by child's name and essay quality.

	N	Mean	SD	Median	Min	Max
Good essay						
Jakob	18	6.72	1.87	7	3	10
Sophie	16	7.56	2.16	8	4	10
Justin	21	7.43	1.94	7	4	11
Jaqueline	21	6.90	1.61	6	5	12
Murat	19	6.95	2.46	8	2	11
Ayse	18	7.83	1.76	8	5	12
Total	113	7.22	1.97	7	2	12
Bad essay						
Jakob	24	5.71	2.14	6	2	11
Sophie	22	5.45	2.06	6	2	10
Justin	12	4.83	1.40	5	2	7
Jaqueline	18	6.11	1.68	6	3	9
Murat	19	5.16	1.89	5	2	9
Ayse	29	5.69	1.81	5	3	9
Total	124	5.55	1.88	6	2	11

Table 3
Regression of essay grades on randomized variables (shown) and controls (not shown).

		Model 1.1		Model 1.2	
Essay quality: good	β_1	0.80**	(0.13)	0.78**	(0.23)
Gender: female	β_2	0.18	(0.13)	0.19	(0.13)
Name: Turkish	β_3	−0.12	(0.16)	−0.17	(0.23)
Name: German upper class	β_4	−0.07	(0.16)	−0.04	(0.24)
Turkish × good quality	β_5			0.12	(0.32)
Upper × good quality	β_6			−0.07	(0.33)
Observations		199		199	
R^2		0.211		0.213	

Standard errors in parentheses.
$^{\dagger}p < 0.10$.
$^{*}p < 0.05$.
** $p < 0.01$.
Model contains controls for teacher characteristics.

7.2. Expectations

In contrast to the results for grades, estimates of ethnic discrimination and social class discrimination in teachers' expectations show effects in the hypothesized directions and lower *p*-values (see Table 4). However, in the simpler model 2.1, two of the three contrasts between group estimates fail to reach conventional levels of statistical significance. Averaging over essay quality, children whose name indicates a Turkish background are perceived to be less likely to succeed at the *Gymnasium* than children with a German lower class background ($\beta_3 = -0.42$, $p = 0.205$; AME: -0.08, $p = 0.205$).[10] The difference between children whose names indicate German upper middle class vs. German lower class turns out to be of similar size ($\beta_4 = 0.52$, $p = 0.139$; AME: 0.11, $p = 0.134$). Only the contrast between Turkish names and German upper middle class names reaches conventional levels of statistical significance and is larger in size than the two preceding contrasts ($\beta_3 - \beta_4 = -0.94$, $p = 0.006$; AME: -0.20, $p = 0.004$). In sum, this simple model only shows significant group differences when ethnic and social class signals are combined, that is, when ethnic discrimination and social class discrimination add up.

The interaction effects between essay quality and child's background (Table 4, model 2.2) show that group differences depend on essay quality. If the essay is bad, there is no significant difference in

[9] We also ran all models with the full sample and without controls for teacher characteristics, with similar results and substantively unaltered conclusions. See the supplementary material at https://osf.io/43exq/ for the results.
[10] Average Marginal Effects (AMEs) for the probability of being judged to be successful at the highest track ($Y_i = 3$). See the supplementary files at https://osf.io/43exq/ for all AMEs of all outcomes.

Table 4
Ordinal logistic regression of teachers' expectations on randomized variables (shown) and controls (not shown).

		Model 2.1		Model 2.2	
Essay quality: good	β_1	1.17**	(0.28)	1.27**	(0.49)
Gender: female	β_2	0.03	(0.28)	−0.02	(0.28)
Name: Turkish	β_3	−0.42	(0.33)	−0.04	(0.48)
Name: German upper class	β_4	0.52	(0.35)	0.31	(0.48)
Turkish × good quality	β_5			−0.79	(0.66)
Upper × good quality	β_6			0.76	(0.74)
τ_1		−0.35	(0.48)	−0.33	(0.53)
τ_2		1.05*	(0.48)	1.10*	(0.54)
Observations		199		199	
Log Likelihood		−203.46		−201.02	

Standard errors in parentheses.
$^{\dagger}p < 0.10$.
* $p < 0.05$.
** $p < 0.01$.
Model contains controls for teacher characteristics.

estimated expectations for either contrast between the three groups ($\beta_3 = -0.04,\ p = 0.933$; AME: $-0.007,\ p = 0.933$; $\beta_4 = 0.31,\ p = 0.514$; AME: 0.06, $p = 0.508$; $\beta_3 - \beta_4 = -0.35,\ p = 0.412$; $AME: -0.06,\ p = 0.412$).

However, when the essay is better, group differences in teachers' expectations are larger and the corresponding *p*-values are smaller: Turkish children have lower odds than German children of comparable social class background ($\beta_3 + \beta_5 = -0.83, p = 0.067$; AME: $-0.19, p = 0.059$) to be trusted to succeed at the *Gymnasium*, and German lower class children in turn have lower odds than German upper middle class children ($\beta_4 + \beta_6 = 1.07, p = 0.059$; AME: $0.24, p = 0.042$). Consequently, the difference between Turkish children and German upper middle class children is highest ($(\beta_3 + \beta_5) - (\beta_4 + \beta_6) = -1.90, p = 0.001$; AME: $-0.43, p < 0.001$).

Thus, we find no evidence for discrimination for the bad essay, but we do find evidence for discrimination on the basis of social class, as evidenced by the contrast between German upper and lower class names, and ethnicity, captured by the contrast between Turkish and German lower class names for the better essay. These results suggest that hypotheses 2_a and 2_b hold only conditional on essay quality.

Another interesting result is that in this model, essay quality is a significant predictor of teachers' expectations towards both groups of German children (lower class: $\beta_1 = 1.27, p = 0.009$; AME: $0.28, p = 0.005$; upper middle class: $\beta_1 + \beta_6 = 2.03, p < 0.000$; AME: $.46, p < 0.000$), but *not* Turkish children ($\beta_1 + \beta_5 = 0.48, p = 0.286$; AME: $0.09, p < 0.291$), for whom returns are significantly lower than for upper middle class children ($\beta_5 - \beta_6 = -1.55, p = 0.032$). The difference in returns for Turkish children and lower class children, captured by β_5, is not significant on conventional levels.

To get a more vivid impression of the effect sizes, we visualize the key results of model 2.2 in Fig. 3. The left panel shows predicted probabilities for falling into the category with the highest likelihood of success at the *Gymnasium* as assigned by the teachers for the three groups of students whose contrasts define social class discrimination and ethnic discrimination. The right panel shows differences in the corresponding probabilities for the three contrasts.

8. Discussion and conclusion

Even though in most countries social class is a better and more robust predictor of educational achievement than ethnicity or race, research on discrimination in education has mainly focused on the latter. We explicitly assess discrimination through teachers by virtue of both a student's ethnic and social class background.

Our major contribution is that, in the present study, we discuss and address the confounding of ethnic discrimination and social class discrimination in experimental designs using names as treatments. In observational studies on ethnic discrimination it is good practice to include some measure of social class. However, experimental research on ethnic discrimination often disregards social class as key confounder and simply compares outcomes of stimuli signaling ethnic minority and ethnic majority. We formalize and address this problem: using DAGs, we show that the common practice of using names to send varying ethnic signals will, under certain conditions, lead to biased estimates of ethnic discrimination in the sense of a causal effect of an ethnic signal. To disentangle ethnic discrimination from social class discrimination, we compare outcomes for names that signal different ethnicity but not different social class background.

We further contribute to the literature by discussing and addressing other common shortcomings of experiments on discrimination in education: we rely on a state-wide random sample of schools, focus on teachers who actually teach immigrants, and take into account potential effect heterogeneity across the distribution of observed performance. This is achieved by an experiment featuring a 2 × 2 × 3 factorial design, varying essay quality, gender, and social and ethnic class background. Examining discrimination in both grades and track recommendations also allows to indirectly test different theories of discrimination.

In contrast to Sprietsma (2013), whose design is similar to that of the present study, we do not find evidence for ethnic or social class discrimination in *grading*. However, the results for teachers' expectations are more complex: a simple model that averages over essays of different quality does not provide direct evidence for either ethnic or social class discrimination. When we contrast names that vary on both ethnic and social class signals, differences are larger and statistically significant on conventional levels. This suggests that social class plays a role in discrimination against Turkish students; ignoring that names may confound ethnic and social class signals would have led to an upwardly biased estimate of ethnic discrimination.

A model featuring interaction effects of essay quality and student background reveals that there is indeed discrimination on the basis of both social class and ethnicity but only for the better of two essays (cf. Bonefeld & Dickhäuser, 2018). This can be interpreted as evidence for differential returns to essay quality: returns are lowest and not statistically significant for Turkish students, higher for German lower class students, and highest for German upper middle class students. This finding also shows that it is critical at which point in the ability distribution discrimination is assessed.

Taken together, the results provide evidence against simpler models of ingroup-favoritism or outgroup derogation, such as social identity theory (Tajfel, 1982; Tajfel & Turner, 1986) or Becker (1971)'s model of taste discrimination.

Other theories receive more support: a statistical discrimination model with group-specific reliabilities (Aigner & Cain, 1977)—lowest reliability for Turkish students, followed by German lower middle class, and upper middle class students—is in line with the observed interaction with essay quality. Such a model may also feature teachers that are risk-averse (Maaz et al., 2008) and hold negatively biased stereotypes about Turkish students and those from lower social class families (Wenz et al., 2016).

The findings also appear to be in line with the mechanism proposed by the continuum model (Fiske et al., 1999; Fiske & Neuberg, 1990). The better essay that turned out to be rather average might have been not bad enough to move teachers from a category-based response based on stereotypes to a piecemeal-integration of individuating information in case of upper middle class students. Conversely, it might have not

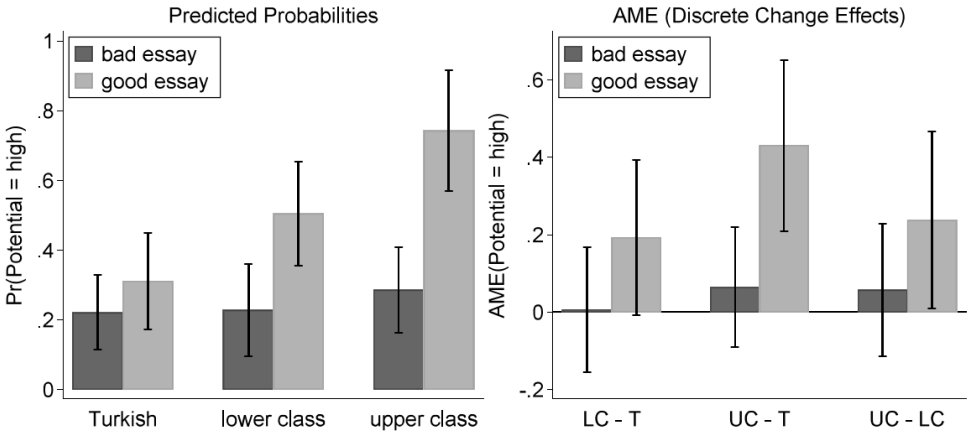

Fig. 3. Left panel: Predicted probabilities for a high likelihood of success at the *Gymnasium*, dependent on name and essay quality. Right panel: Average marginal effects in probabilities for each of the three contrasts, dependent on essay quality—lower class vs. Turkish (LC – T; + 0.01 and + 0.19), upper class vs. Turkish (UC – T; + 0.06 and + 0.43), upper class vs. lower class (UC – LC; + 0.06 and + 0.24), with 95% confidence bars. Calculations are based on model 2.2 in Table 4.

been good enough to foster the same process for the Turkish students and they, too, were treated according to the teachers' stereotypes. However, since the bad essay turned out to be really bad, teachers might have turned from a category-based judgment to a more individuating judgment in case of upper middle class students and, thus, have graded them as bad as Turkish and lower class students.

Similarly, based on aversive racism theory we could also explain the observed pattern: in case of the supposedly good but really rather average essay, teachers might have taken advantage of the ambiguity of the situation and treated students according to their stereotypes and prejudices.

One possibility to distinguish between different theoretical mechanisms might be to assess discrimination additionally for an excellent essay. While models of statistical discrimination that rely on group-specific reliabilities would predict an even larger gap between students from different social and ethnic groups, applying the mechanisms from the continuum model or aversive racism theory probably leads to the opposite prediction of less discrimination on the basis of these factors.

Sociologists study ethnic and social class discrimination not only because it is considered unfair or unjust by many, but also as a mechanism of ethnic and social stratification. Such a perspective requires an understanding of the conditions under which discrimination as an individual-level causal effect aggregates to group discrimination and, thus, inequality. While simple models of statistical discrimination cannot explain inequality between groups, models that receive the most support by our findings—models that feature group-specific reliabilities, risk-averse teachers, and biased stereotypes (Aigner & Cain, 1977; England & Lewin, 1989)—can. From the less formalized continuum model and aversive racism theory it is less clear to make such a prediction, but should teachers follow the mechanisms suggested by these theories it would be difficult to explain how the individual level causal effects should disappear on the group level.

The experiment presented in this chapter was designed to address various shortcomings and desiderata of previous experimental research in education. But, of course, it has itself several limitations that should be addressed in future research. We discuss them in the remainder.

While we did undertake several steps to increase the external validity of our findings, it might be improved further by making the experimental situation more realistic. The essay grading task should resemble quite closely real-life situations of grading a manifest performance in German elementary schools. However, estimating the likelihood of success at different secondary school tracks and making corresponding recommendations requires more individuating and diagnostic information that teachers typically have in real life—even though this information is still incomplete. Thus, providing more diagnostic information about the fictitious student under evaluation seems necessary to come closer to an unbiased estimate of discrimination in track recommendations. Our experiment was certainly closer to a lab experiment than to a field experiment in this regard.

$N = 237$ teachers in the whole sample and $N = 199$ in the analysis sample did not provide enough statistical power for investigating fully interacted models for the $2 \times 2 \times 3$ factorial design. On the backdrop of gender inequalities in education among immigrants (Fleischmann et al., 2014) and teachers' differential attitudes towards boys and girls with immigrant background (Glock & Klapproth, 2017), interaction effects of ethnic and social class background with gender should be examined.

Furthermore, fully disentangling ethnic from social class discrimination would require including a signal for Turkish upper class background. Only if both ethnicity and social class vary independently of one another can we estimate all main effects and interaction effects. The experimental design in the present paper allows identifying ethnic discrimination only among students from lower class families, while social class discrimination is identified for German students without immigrant background only. Using names—as a rather subliminal stimulus—to signal an upper class background of Turkish immigrants in Germany might prove very difficult, though, because teachers—and other relevant decision makers and gate keepers—cannot distinguish between lower class and upper class names of Turkish immigrants.

We have theorized about discrimination in track recommendations more generally and argued that expectations determine not only track recommendations but also many other important decisions in education such as ability grouping within tracks and may turn into self-fulfilling prophecies. However, since we have not explicitly asked teachers which track they would recommend based on the observed performance, we can only hypothesize about potential differences to our findings for expectations of future performance.

Discrimination in an outcome explicitly asking for track recommendations would have probably been higher, since variables such as parental support and involvement should play an even more important role than for the more narrow question on future performance in one subject. If anything, the more narrow question should reduce the effect of students' social and ethnic background and render our estimate of discrimination conservative. However, this remains speculative unless empirically addressed.

Also, the question we asked and examined does not ask about actual behavior or behavioral intention. It could be the result of either stereotyping or, in case of aversive racism, maybe applied prejudice. Here, too, the question which track the teacher would recommend, would have been a very interesting outcome to look at.

Most importantly, even if we had asked the participating teachers to recommend a secondary school track, our study's design dictated that, in their decision, teachers had to rely on one essay only. However, in the real world, teachers know much more about a single student and, thus, should have to rely less on their stereotypes in their estimation of the probability of success at the different tracks.

Looking at two different outcomes in education, we were able to provide indirect evidence about theoretical mechanisms of discrimination in education. For a more direct test of different theories of discrimination and their proposed mechanisms, direct measures of, for example, stereotypes and prejudices would be needed.

Despite these limitations, in conclusion, we think that our study provides evidence for ethnic and social discrimination by elementary school teachers when it comes to estimating students' future potential. It shows that it is possible to build on experimental research on discrimination in education to provide methodologically sounder and theoretically more enlightening empirical insights into an important mechanism of ethnic and social stratification.

Conflict of interest

None declared.

Acknowledgements

The authors would like to thank Anne Landhäußer, Thorsten Schneider, Johannes Keller, Daniel Klein, Clemens Kroneberg, and many others for valuable support and feedback. For valuable feedback we would also like to thank participants at the 2010 RC28 session 10 at the XVII ISA World Congress in Gothenburg, the 2010 BiKS Summer School, the 2010 ECSR conference in Bamberg, the 2011 NORFACE conference in London, the WZB CO:STA meeting on 8 October 2012, the 2013 Spring Metting of the DGS section "Modellbildung und Simulation" in Konstanz, and the ASA 2013 Annual Meeting in New York City. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. For helpful comments we would also like to thank the editors of the RSSM special issue and two anonymous reviewers.

Appendix A. Tables for the full and restricted sample

Table A.1
Regression of essay grades on essay quality, child's gender and child's background: comparison of full and restricted sample.

	Restricted sample				Full sample			
	Model 1.1		Model 1.2		Model 1.1		Model 1.2	
Essay quality: good	0.80**	(0.13)	0.78**	(0.23)	0.85**	(0.12)	0.85**	(0.22)
Gender: female	0.18	(0.13)	0.19	(0.13)	0.20	(0.12)	0.20	(0.12)
Name: German upper middle class	− 0.07	(0.16)	− 0.04	(0.24)	0.01	(0.15)	0.05	(0.22)
Name: Turkish	− 0.12	(0.16)	− 0.17	(0.23)	− 0.07	(0.15)	− 0.11	(0.22)
Teaching experience in years	− 0.02**	(0.01)	− 0.02**	(0.01)	− 0.02**	(0.01)	− 0.02**	(0.01)
Gender teacher: male	0.39*	(0.20)	0.38†	(0.20)	0.51**	(0.18)	0.51**	(0.18)
Parental education: CASMIN 2	− 0.10	(0.16)	− 0.10	(0.16)	− 0.19	(0.15)	− 0.18	(0.15)
Parental education: CASMIN 3	0.03	(0.17)	0.04	(0.17)	0.01	(0.16)	0.00	(0.16)
Teacher has migration background	− 0.11	(0.25)	− 0.12	(0.25)	− 0.21	(0.23)	− 0.21	(0.23)
Upper × good quality			− 0.07	(0.33)			− 0.09	(0.30)
Turkish × good quality			0.12	(0.32)			0.11	(0.30)
Constant	− 0.16	(0.23)	− 0.15	(0.25)	− 0.20	(0.21)	− 0.19	(0.23)
Observations	199		199		223		223	
R^2	0.211		0.213		0.238		0.240	

Standard errors in parentheses.
† $p < 0.10$.
* $p < 0.05$.
** $p < 0.01$.
Restricted sample includes only teachers who teach migrants.

Table A.2
Ordinal logistic regression of expectations on essay quality, child's gender and child's background: comparison of full and restricted sample.

	Restricted sample				Full sample			
	Model 2.1		Model 2.2		Model 2.1		Model 2.2	
Essay quality: good	1.17**	(0.28)	1.27**	(0.49)	1.37**	(0.27)	1.37**	(0.47)
Gender: female	0.03	(0.28)	− 0.02	(0.28)	0.02	(0.26)	− 0.01	(0.27)
Name: German upper middle class	0.52	(0.35)	0.31	(0.48)	0.64†	(0.33)	0.36	(0.46)
Name: Turkish	− 0.42	(0.33)	− 0.04	(0.48)	− 0.33	(0.32)	− 0.06	(0.45)
Teaching experience in years	− 0.02	(0.01)	− 0.02	(0.01)	− 0.01	(0.01)	− 0.01	(0.01)
Gender teacher: male	− 0.06	(0.40)	− 0.01	(0.40)	0.09	(0.38)	0.10	(0.38)
Parental education: CASMIN 2	− 0.10	(0.33)	− 0.16	(0.34)	0.02	(0.31)	− 0.01	(0.31)
Parental education: CASMIN 3	0.34	(0.35)	0.34	(0.35)	0.49	(0.34)	0.52	(0.34)
Teacher has migration background	0.72	(0.55)	0.80	(0.55)	0.62	(0.52)	0.65	(0.52)
Upper × good quality			0.76	(0.74)			0.86	(0.69)

(continued on next page)

Table A.2 *(continued)*

	Restricted sample				Full sample			
	Model 2.1		Model 2.2		Model 2.1		Model 2.2	
Turkish × good quality			− 0.79	(0.66)			− 0.60	(0.63)
τ_1	− 0.35	(0.48)	− 0.33	(0.53)	− 0.07	(0.44)	− 0.06	(0.50)
τ_2	1.05*	(0.48)	1.10*	(0.54)	1.40**	(0.46)	1.41**	(0.51)
Observations	199		199		223		223	
Log likelihood	− 203.46		− 201.02		− 224.43		− 221.97	

Standard errors in parentheses.
$^{\dagger}p < 0.10$.
* $p < 0.05$.
** $p < 0.01$.
Restricted sample includes only teachers who teach migrants.

References

Aigner, D., & Cain, G. (1977). Statistical theories of discrimination in labor markets. *Industrial and Labor Relations Review, 30*, 175–187.

Allison, P. D. (1999). Comparing logit and probit coefficients across groups. *Sociological Methods & Research, 28*, 186–208.

Allport, G. W. (1954). *The nature of prejudice*. Addison-Wesley.

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

Becker, G. (1971). *The economics of discrimination* (2nd ed.). University of Chicago Press.

Becker, R., & Beck, M. (2012). Herkunftseffekte oder statistische Diskriminierung von Migrantenkindern in der Primarstufe? In R. Becker, & H. Solga (Eds.). *Soziologische Bildungsforschung number 52 in Kölner Zeitschrift für Soziologie und Sozialpsychologie Sonderhefte* (pp. 137–163). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-00120-9_6.

Berk, R. A. (1983). An introduction to sample selection bias in sociological data. *American Sociological Review, 48*, 386–398. https://doi.org/10.2307/2095230.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal?. A field experiment on labor market discrimination. *The American Economic Review, 94*, 991–1013. https://doi.org/10.1257/0002828042002561.

Blalock, H. M. J. (1967). *Toward a theory of minority-group relations*. John Wiley & Sons.

Blank, R. M., Dabady, M., & Citro, C. F. (Eds.). (2004). *Measuring racial discrimination*. The National Academies Press.

Blumer, H. (1958). Race prejudice as a sense of group position. *The Pacific Sociological Review, 1*, 3–7. https://doi.org/10.2307/1388607.

Bonefeld, M., & Dickhäuser, O. (2018). (Biased) Grading of Students' Performance: Students' Names, Performance Level, and Implicit Attitudes. *Frontiers in Psychology, 9*. https://doi.org/10.3389/fpsyg.2018.00481.

Bos, W., Tarelli, I., Bremerich-Vos, A., & Schwippert, K. (Eds.). (2012). *IGLU 2011: Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich*. Waxmann.

Bos, W., Wendt, H., Köller, O., & Selter, C. (Eds.). (2012). *TIMSS 2011: Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich*. Waxmann.

Brant, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics, 46*, 1171–1178.

Breen, R., & Jonsson, J. O. (2005). Inequality of opportunity in comparative perspective: Recent research on educational attainment and social mobility. *Annual Review of Sociology*, 223–243. https://doi.org/10.2307/29737718.

Campbell, D. T. (1965). Ethnocentric and other altruistic motives. In D. Levine (Ed.). *Nebraska symposium* (pp. 283–311). Lincoln, NE, USA: University of Nebraska Press.

Croizet, J.-C. (2008). The pernicious relationship between merit assessment and discrimination in education. In G. Adams, M. Biernat, & N. R. Branscombe (Eds.). *Commemorating brown: The social psychology of racism and discrimination* (pp. 153–172). American Psychological Association.

Deming, D. J., Yuchtman, N., Abulafi, A., Goldin, C., & Katz, L. F. (2016). The value of postsecondary credentials in the labor market: An experimental study. *American Economic Review, 106*, 778–806. https://doi.org/10.1257/aer.20141757.

Diehl, C., & Fick, P. (2016). Ethnische Diskriminierung im deutschen Bildungssystem. In C. Diehl, C. Hunkler, & C. Kristen (Eds.). *Ethnische Ungleichheiten im Bildungsverlauf* (pp. 243–286). Springer. https://doi.org/10.1007/978-3-658-04322-3_6.

Ditton, H. (2013). Wer geht auf die Hauptschule? Primäre und sekundäre Effekte der sozialen Herkunft beim Übergang nach der Grundschule. *Zeitschrift für Erziehungswissenschaft, 16*, 731–749. https://doi.org/10.1007/s11618-013-0440-y.

Ditton, H., Krüsken, J., & Schauenberg, M. (2005). Bildungsungleichheit - der Beitrag von Familie und Schule. *Zeitschrift für Erziehungswissenschaft, 8*, 285–304. https://doi.org/10.1007/s11618-005-0138-x.

Dovidio, J. F., & Gaertner, S. L. (2008). New directions in aversive racism research: Persistence and pervasiveness. In C. Willis-Esqueda (Ed.). *Motivational aspects of prejudice and racism the Nebraska symposium on motivation* (pp. 43–67). Springer New York. https://doi.org/10.1007/978-0-387-73233-6_3.

Dustmann, C., Frattini, T., & Lanzara, G. (2012). Educational achievement of second-generation immigrants: An international comparison. *Economic Policy, 27*, 143–185. https://doi.org/10.1111/j.1468-0327.2011.00275.x.

Elwert, F., & Winship, C. (2014). Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable. *Annual Review of Sociology, 40*, 31–53. https://doi.org/10.1146/annurev-soc-071913-043455.

England, P., & Lewin, P. (1989). Economic and sociological views of discrimination in labor markets: Persistence or demise? *Sociological Spectrum, 9*, 239–257.

Tobisch, A., & Dresel, M. (2017). Negatively or positively biased? Dependencies of teachers' judgments and expectations based on students' ethnic and social backgrounds. *Social Psychology of Education, 20*, 731–752. https://doi.org/10.1007/s11218-017-9392-z.

van Ewijk, R. (2011). Same work, lower grade? Student ethnicity and teachers' subjective assessments. *Economics of Education Review, 30*, 1045–1058. https://doi.org/10.1016/j.econedurev.2011.05.008.

Farkas, G. (2003). Racial disparities and discrimination in education: What do we know, how do we know it, and what do we need to know? *The Teachers College Record, 105*, 1119–1146.

Figlio, D. N. (2005). *Names, expectations and the Black-White test score gap*. National Bureau of Economic Research Working Paper 11195.

Fiske, S. T. (1998). Stereotyping, prejudice, and discrimination. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.). *The handbook of social psychology*(4th ed.). McGraw-Hill.

Fiske, S. T., Lin, M., & Neuberg, S. L. (1999). The continuum model. Ten years later. In S. Chaiken, & Y. Trope (Eds.). *Dual process theories in social psychology* (pp. 231–254). Guilford.

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. *Advances in Experimental Social Psychology, 23*, 1–74. https://doi.org/10.1016/S0065-2601(08)60317-2.

Fleischmann, F., Kristen, C., Heath, A. F., Brinbaum, Y., Deboosere, P., Granato, N., Jonsson, J. O., Kilpi-Jakonen, E., Lorenz, G., Lutz, A. C., Mos, D., Mutarrak, R., Phalet, K., Rothon, C., Rudolphi, F., & Werfhorst, H. G.v.d. (2014). Gender inequalities in the education of the second generation in Western countries. *Sociology of Education, 87*, 143–170. https://doi.org/10.1177/0038040714537836.

Fryer, R. G., Jr., & Levitt, S. D. (2004). The causes and consequences of distinctively black names. *The Quarterly Journal of Economics, 119*, 767–805.

Gaddis, S. M. (2015). Discrimination in the credential society: An audit study of race and college selectivity in the labor market. *Social Forces, 93*, 1451–1479. https://doi.org/10.1093/sf/sou111.

Gaddis, S. M. (2017a). How black are Lakisha and Jamal? Racial perceptions from names used in correspondence audit studies. *Sociological Science, 4*, 469–489. https://doi.org/10.15195/v4.a19.

Gaddis, S. M. (2017b). Racial/ethnic perceptions from Hispanic names: Selecting names to test for discrimination. *Socius, 3*. https://doi.org/10.1177/2378023117737193237802311773719323780231177371932.

Gaertner, S. L., & Dovidio, J. F. (1986). The aversive form of racism. In J. F. Dovidio, & S. L. Gaertner (Eds.). *Prejudice, discrimination, and racism* (pp. 61–89). Academic Press.

Gangl, M. (2010). Causal inference in sociological research. *Annual Review of Sociology, 36*, 21–47. https://doi.org/10.1146/annurev.soc.012809.102702.

Glock, S., & Klapproth, F. (2017). Bad boys, good girls? Implicit and explicit attitudes toward ethnic minority students among elementary and secondary school teachers. *Studies in Educational Evaluation, 53*, 77–86. https://doi.org/10.1016/j.stueduc.2017.04.002.

Glock, S., Krolak-Schwerdt, S., & Pit-ten Cate, I. M. (2015). Are school placement recommendations accurate? The effect of students' ethnicity on teachers' judgments and recognition memory. *European Journal of Psychology of Education, 30*, 169–188. https://doi.org/10.1007/s10212-014-0237-2.

Glock, S., Krolak-Schwerdt, S., Klapproth, F., & Böhmer, M. (2013). Beyond judgment bias: How students' ethnicity and academic profile consistency influence teachers' tracking judgments. *Social Psychology of Education, 16*, 555–573. https://doi.org/10.1007/s11218-013-9227-5.

Greiner, D. J., & Rubin, D. B. (2010). Causal effects of perceived immutable characteristics. *Review of Economics and Statistics, 93*, 775–785. https://doi.org/10.1162/REST_a_00110.
Gresch, C. (2012). *Der Übergang in die Sekundarstufe I. Leistungsbeurteilung, Bildungsaspiration und rechtlicher Kontext bei Kindern mit Migrationshintergrund*. Springer VS.
Hanna, R. N., & Linden, L. L. (2012). Discrimination in grading. *American Economic Journal: Economic Policy, 4*, 146–168. https://doi.org/10.1257/pol.4.4.146.
Heath, A. F., Rothon, C., & Kilpi, E. (2008). The second generation in Western Europe: Education, unemployment, and occupational attainment. *Annual Review of Sociology, 34*, 211–235. https://doi.org/10.1146/annurev.soc.34.040507.134728.
Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica, 47*, 153–161. https://doi.org/10.2307/1912352.
Heckman, J. J. (1998). Detecting discrimination. *The Journal of Economic Perspectives, 12*, 101–116. https://doi.org/10.2307/2646964.
Heckman, J. J., & Siegelman, P. (1993). The urban institute audit studies: Their methods and findings. In M. Fix, & R. J. Struyk (Eds.). *Clear and convincing evidence: Measurement of discrimination in America* (pp. 187–258). The Urban Institute Press.
Hernán, M. A., Hernández-Díaz, S., & Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology, 15*, 615–625. https://doi.org/10.1097/01.ede.0000135174.63482.43.
Hernán, M. A., & Robins, J. M. (2020). *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC. In press https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/.
Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*, 945–960. https://doi.org/10.2307/2289064.
Holzer, H., & Ludwig, J. (2003). Measuring discrimination in education: Are methodologies from labor and markets useful? *The Teachers College Record, 105*, 1147–1178.
Imai, K., Tingley, D., & Yamamoto, T. (2013). Experimental designs for identifying causal mechanisms. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 176*, 5–51. https://doi.org/10.1111/j.1467-985X.2012.01032.x/full.
Imbens, G. W. (2004). Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review. *Review of Economics and Statistics, 86*(1), 4–29. https://doi.org/10.1162/003465304323023651.
Jacquemet, N., & Yannelis, C. (2012). Indiscriminate discrimination: A correspondence test for ethnic homophily in the Chicago labor market. *Labour Economics, 19*, 824–832. https://doi.org/10.1016/j.labeco.2012.08.004.
Jussim, L., Eccles, J., & Madon, S. (1996). Social perception, social stereotypes, and teacher expectations: Accuracy and the quest for the powerful self-fulfilling prophecy. *Advances in Experimental Social Psychology, 28*, 281–388.
Jussim, L., & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review, 9*, 131–155. https://doi.org/10.1207/s15327957pspr0902_3.
Kaiser, J., Südkamp, A., & Möller, J. (2016). The effects of student characteristics on teachers' judgment accuracy: Disentangling ethnicity, minority status, and achievement. *Journal of Educational Psychology*. https://doi.org/10.1037/edu0000156.
Kalter, F. (2008). Ethnische Ungleichheit auf dem Arbeitsmarkt. In M. Abraham, & T. Hinz (Eds.). *Arbeitsmarktsoziologie* (pp. 303–332). (2nd ed.). VS Verlag für Sozialwissenschaften.
Karlson, K. B., Holm, A., & Breen, R. (2012). Comparing regression coefficients between same-sample nested models using logit and probit a new method. *Sociological Methodology, 42*, 286–313. https://doi.org/10.1177/0081175012444861.
Kiss, D. (2013). Are immigrants and girls graded worse? Results of a matching approach. *Education Economics, 21*, 447–463. https://doi.org/10.1080/09645292.2011.585019.
Kogan, I. (2007). A study of immigrants' employment careers in West Germany using the sequence analysis technique. *Social Science Research, 36*, 491–511. https://doi.org/10.1016/j.ssresearch.2006.03.004.
Kristen, C. (2006a). *Ethnische Diskriminierung im deutschen Schulsystem? Theoretische Überlegungen und empirische Ergebnisse. Discussion Paper SP IV 2006-601 Wissenschaftszentrum Berlin für Sozialforschung*. Arbeitsstelle Interkulturelle Konflikte und Gesellschaftliche Integration.
Kristen, C. (2006b). Ethnische Diskriminierung in der Grundschule? Die Vergabe von Noten und Bildungsempfehlungen. *Kölner Zeitschrift Für Soziologie Und Sozialpsychologie, 58*, 79–97. https://doi.org/10.1007/s11575-006-0004-y.
Kube, J. (2009). *Vornamensforschung: Fragebogenuntersuchung bei Lehrerinnen und Lehrern, ob Vorurteile bezüglich spezifischer Vornamen von Grundschülern und davon abgeleitete erwartete spezifische Persönlichkeitsmerkmale vorliegen*. Oldenburg, Germany: Carl von Ossietzky-Universität Oldenburg. Unpublished Master's thesis.
Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. SAGE.
Lorenz, G., Gentrup, S., Kristen, C., Stanat, P., & Kogan, I. (2016). Stereotype bei Lehrkräften? Eine Untersuchung systematisch verzerrter Lehrererwartungen. *Kölner Zeitschrift Für Soziologie Und Sozialpsychologie, 68*, 89–111. https://doi.org/10.1007/s11577-015-0352-3.
Lüdemann, E., & Schwerdt, G. (2013). Migration background and educational tracking. *Journal of Population Economics, 26*, 455–481. https://doi.org/10.1007/s00148-012-0414-z.
Maaz, K., Neumann, M., Trautwein, U., Wendt, W., Lehmann, R., & Baumert, J. (2008). Der Übergang von der Grundschule in die weiterführende Schule: Die Rolle von Schüler- und Klassenmerkmalen beim Einschätzen der individuellen Lernkompetenz durch die Lehrkräfte. *Schweizerische Zeitschrift Für Bildungswissenschaften, 30*, 519–548.
Marks, G. N. (2005a). Accounting for immigrant non-immigrant differences in reading and mathematics in twenty countries. *Ethnic and Racial Studies, 28*, 925–946. https://doi.org/10.1080/01419870500158943.
Marks, G. N. (2005b). Cross-national differences and accounting for social class inequalities in education. *International Sociology, 20*, 483–505. https://doi.org/10.1177/0268580905058328.
Mickelson, R. (2003). When are racial disparities in education the result of racial discrimination? A social science perspective. *The Teachers College Record, 105*, 1052–1086.
Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European Sociological Review, 26*, 67–82. https://doi.org/10.1093/esr/jcp006.
Myrdal, G. (1944). *An American dilemma: The Negro problem and modern democracy, volume I*. Transaction Publishers.
Neumark, D. (2012). Detecting discrimination in audit and correspondence studies. *Journal of Human Resources, 47*, 1128–1157. https://doi.org/10.3368/jhr.47.4.1128.
Oaxaca, R. (1973). Male–female wage differentials in urban labor markets. *International Economic Review, 14*, 693–709. https://doi.org/10.2307/2525981.
OECD (2016). *PISA 2015 results (volume I): Excellence and equity in education*. OECD Publishing https://doi.org/10.1787/9789264266490-en.
Pager, D., & Shepherd, H. (2008). The sociology of discrimination: Racial discrimination in employment, housing, credit, and consumer markets. *Annual Review of Sociology, 34*, 181–209. https://doi.org/10.1146/annurev.soc.33.040406.131740.
Pearl, J. (2001). Direct and indirect effects. *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*, 411–420.
Pearl, J. (2009). *Causality* (2nd ed.). Cambridge University Press.
Pearl, J. (2014). Interpretation and identification of causal mediation. *Psychological Methods, 19*, 459–481. https://doi.org/10.1037/a0036434.
Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer (Pap/psc ed.)*. John Wiley and Sons Ltd.
Phelps, E. S. (1972). The statistical theory of racism and sexism. *American Economic Review, 62*, 659–661.
Quillian, L. (2006). New approaches to understanding racial prejudice and discrimination. *Annual Review of Sociology, 32*, 299–328. https://doi.org/10.1146/annurev.soc.32.061604.123132.
Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology, 66*, 688–701. https://doi.org/10.1037/h0037350.
Rubin, D. B. (1978). Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics, 6*(1), 34–58. http://www.jstor.org/stable/2958688.
Rubin, D. B. (1986). Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association, 81*, 961–962. https://doi.org/10.2307/2289065.
Schneider, T. (2011). Die Bedeutung der sozialen Herkunft und des Migrationshintergrundes für Lehrerurteile am Beispiel der Grundschulempfehlung. *Zeitschrift Für Erziehungswissenschaft, 14*, 371–396. https://doi.org/10.1007/s11618-011-0221-4.
Schulze, A., & Schiener, J. (2011). Lehrerurteile und Bildungsgerechtigkeit. *Zeitschrift Für Soziologie Der Erziehung Und Sozialisation* 31.
Sen, M., & Wasow, O. (2016). Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. *Annual Review of Political Science, 19*, 499–522. https://doi.org/10.1146/annurev-polisci-032015-010015.
Sherif, M. (1966). *Group conflict and co-operation: Their social psychology*. London: Routledge & Kegan Paul.
Sherif, M., Harvey, O. J., White, B. J., Hood, W. R., & Sherif, C. W. (1961). *Intergroup conflict and cooperation: The Robbers' Cave experiment, volume 10*. Norman, OK: University of Oklahoma Book Exchange.
Sprietsma, M. (2013). Discrimination in grading: Experimental evidence from primary school teachers. *Empirical Economics, 45*, 523–538. https://doi.org/10.1007/s00181-012-0609-x.
Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*, 797–811. https://doi.org/10.1037/0022-3514.69.5.797.
Sumner, W. G. (1906). *Folkways: A study of mores, manners, customs, and morals*. Ginn.
Tajfel, H. (1982). *Social identity and intergroup relations*. Cambridge University Press.
Tajfel, H., & Turner, J. C. (1986). The social identity theory of intergroup behavior. In S. Worchel, & W. G. Austin (Eds.). *Psychology of intergroup relations*. Nelson-Hall Publishers.

VanderWeele, T. J., & Hernán, M. A. (2012). Causal effects and natural laws: Towards a conceptualization of causal counterfactuals for nonmanipulable exposures, with application to the effects of race and sex. In C. Berzuini, P. Dawid, & L. Bernardinelli (Eds.). *Causality* (pp. 101–113). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119945710.ch9/summary.

Wang, X., & Sobel, M. E. (2013). New perspectives on causal mediation analysis. In S. L. Morgan (Ed.). *Handbook of causal analysis for social research handbooks of sociology and social research* (pp. 215–242). Springer Netherlands. https://doi.org/10.1007/978-94-007-6094-3_12.

Weichselbaumer, D. (2016). *Discrimination against female migrants wearing headscarves*. IZA IZA Discussion Paper 10217.

Wendt, H., Bos, W., Köller, O., Schwippert, K., & Kasper, D. (Eds.). (2016). *TIMSS 2015: Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich.*. Waxmann.

Wenz, S. E., Olczyk, M., & Lorenz, G. (2016). *Measuring teachers' stereotypes in the NEPS*. Leibniz Institute for Educational Trajectories NEPS Survey Paper 3.