Original Article

Shifting Standards When Judging Male and Female Majority and Minority Students' Non-Suitability for the Academic Track?

Ursula Kessels[1] , Ricarda Steinmayr[2], and Anna K. Nishen[2]

[1]Department of Education and Psychology, Freie Universität Berlin, Germany
[2]Institute of Psychology, Technical University of Dortmund, Germany

Abstract: Because tracking recommendations can be impactful for students' educational paths, it is important to understand how they may be influenced by students' social group memberships. The Shifting Standards Model predicts that teachers show both negative and positive biases toward students belonging to stereotyped groups, and that this depends on the framing of the context. In a 2 (student's ethnicity: Turkish vs. German) × 2 (student's gender: male vs. female) × 2 (standard: minimum vs. confirmatory) between-subjects design, $n = 185$ student teachers indicated how much evidence of learning-detrimental behavior they needed to see from a student before suspecting vs. being certain (minimum vs. confirmatory standards) that the student was not suitable for the academic track. The predictions of the Shifting Standards Model were unexpectedly not supported. We discuss these results in relation to the usefulness of parsimonious theoretical arguments as well as an intersectional analysis of tracking recommendations.

Keywords: shifting standards, teacher recommendations, ethnic bias, gender bias, intersectionality

Shifting Standards bei der Beurteilung der Nichteignung für das Gymnasium?

Zusammenfassung: Empfehlungen von Lehrkräften für die weiterführende Schule haben weitreichende Folgen für die betroffenen Schülerinnen und Schüler. Entsprechend relevant ist die Frage, inwiefern die Zugehörigkeit der beurteilten Person zu stereotypisierten Gruppen diese Empfehlungen beeinflusst. Das Shifting Standards Model expliziert, inwiefern Stereotype und Kontexte von Urteilen dazu führen, dass Angehörige stereotypisierter Gruppen entweder einen positiven oder einen negativen Bias erfahren. In einem 2 (Herkunft: türkisch vs. deutsch) × 2 (Geschlecht: männlich vs. weiblich) x 2 (Standard: minimal vs. konfirmatorisch) between-subjects-Design gaben $n = 185$ Lehramtsstudierende an, wie viele Anzeichen von lernhinderlichem Verhalten sie bei einem Schüler/einer Schülerin sehen müssten, bevor sie den Verdacht (Minimalstandard) bzw. die Gewissheit (konfirmatorischer Standard) haben, dass der oder die Schüler/-in nicht für das Gymnasium geeignet sei. Die Vorhersagen des Shifting Standards Model wurden jedoch nicht bestätigt. Die Ergebnisse werden in Bezug auf die Nützlichkeit sparsamer theoretischer Annahmen sowie der intersektionalen Analyse von Übergangsempfehlungen diskutiert.

Schlüsselwörter: Shifting Standards, Übergangsempfehlungen, Migrationshintergrund, Geschlecht, Intersektionalität

Tracking, that is, grouping students by performance, is a controversial educational practice. While it is meant to provide better-tailored lessons (Hattie, 2002), it also reproduces educational inequalities in Germany because of the lower-quality learning environment of the lower tracks (Baumert et al., 2006; Hattie, 2002; Maaz et al., 2008; Retelsdorf et al., 2010; Van Houtte, 2004). For example, the achievement gains are greater in students in the academic track, which has been termed the scissor effect (Becker et al., 2006, 2012; Hattie, 2002; Maaz et al., 2008; Retelsdorf et al., 2012). In Germany, students are tracked between schools after elementary school in the 4th or 6th grade (depending on the federal state), and

the school track they attend impacts their options for post-secondary education (Becker et al., 2016). The highest track is the Gymnasium (academic track), which leads to the university entrance qualification (Abitur). While upward mobility between tracks is possible during mandatory schooling, when students change tracks, it is actually more likely to be downward (Bellenberg, 2012). Students who have attended the Gymnasium are more likely to enroll in higher education than students who have obtained their higher education entrance qualification by a different route (Daniel & Neumann, 2023).

Because the school track has meaningful implications for students' careers, tracking decisions after elementary

school need to be carefully considered. In most federal states, elementary school teachers provide a recommendation for each student, which differs in how binding it is depending on the federal state (Standing Conference of the Ministers of Education and Cultural Affairs [KMK], 2015). To guide teachers in this process, the German government has established guidelines that state that recommendations should take into account "not only performance related to goals established in the curriculum but also the general skills important for success in school" (KMK, 2015, p. 6 [translated by the authors]). Specifically, the recommendation should be based on students' performance, but also their "suitability, affinity, and willingness [...] to work intellectually" (KMK, 2015, p. 5 [translated by the authors]). Indeed, in making their recommendations, teachers do take motivational and social factors, such as demonstrating effort and learning habits (Anders et al., 2010; Kaiser et al., 2013), into account. However, these more subjective indicators for tracking recommendations may be vulnerable to teacher biases (Nishen et al., 2023). This could potentially explain why boys (Federal Statistical Office, 2023) and students with a Turkish immigrant background (Rauch et al., 2016), who are both perceived as lower in such motivational and social factors (Anders et al., 2010; Heyder & Kessels, 2017; Tobisch & Dresel, 2017), are underrepresented in the academic track. In the year 2022, 55% of 20–25-year-old female adults had a school-leaving qualification that entitled them to study at university, compared to only 43% of male adults of the same age group (Federal Statistical Office, 2023). An even larger imbalance is found when comparing students with respect to immigrant background: While 41% of 15-year-old students without an immigrant background attended the Gymnasium in 2012, this was the case for only 26% of students with an immigrant background (Autorengruppe Bildungsberichterstattung, 2016).

The present study seeks to expand our understanding of the ways in which negative performance stereotypes may influence tracking recommendations. We simultaneously consider effects of students' gender and immigrant background, allowing us to map the intersections of these two group memberships. Such an intersectional perspective is helpful in order to uncover how stereotypes affect students who might otherwise be overlooked because they are not prototypical for a given group (Cole, 2009). In the present case, this refers to girls with an immigrant background, about whom we know comparatively little because prior research often, though not exclusively, compares either girls and boys without an immigrant background or boys with and without an immigrant background. In the following, we first describe the existing research on teacher biases by immigrant background, gender, and their interaction. We then introduce the Shifting Standards Model (Biernat, 1995, 2012) and its implications for tracking recommendations. Based on this model, the present experimental study examines whether inducing a minimal standard or a confirmatory standard of judgment will have a differential impact on how students' unsuitability for the academic track is rated, depending on the target student's immigrant background and gender.

Teacher Biases by Immigrant Background, Gender, and Their Intersection

Stereotypes are defined as "beliefs about the characteristics, attributes, and behaviors of members of certain groups" (Hilton & von Hippel, 1996, p. 240). Two of the basic social categories used to categorize people into groups are gender and ethnicity (Macrae & Bodenhausen, 2000). A stereotype is a cognitive schema which guides a person's perception, information processing, and retrieval (Bodenhausen, 1990). Deaux and Lewis (1984) stated that when perceivers know another person's basic category membership, the corresponding stereotype is activated and the perceivers draw conclusions about that person's traits. It follows that when group memberships are known or assumed, stereotypes will likely also influence teachers' expectations for and judgments of students (e.g., Martiny & Froehlich, 2020). Even when controlling performance statistically or experimentally, differential judgments are observed (e.g., Bonefeld et al., 2017; Sprietsma, 2013), which we consider a bias in the present research. Two often-researched dimensions of such biases are ethnic and gender biases, and in the following, we summarize research findings from Germany[1]. Turkish people form a large ethnic minority within Germany (Rauch et al., 2016) and are stereotyped as low in competence (Asbrock, 2010; Bonefeld & Karst, 2020; Froehlich et al.,

---

[1] Note that we consider research on stereotypes of teachers and of student teachers to be of equal importance, as the former group is currently teaching and the latter group is going to teach very soon, so their possibly biased perceptions and reactions will impact future students. From a theoretical viewpoint, we see no grounds for assuming that the relevance and the mechanisms of stereotype-based judgments might differ between these two populations.

2016). Indeed, research has demonstrated that student teachers (Bonefeld & Dickhäuser, 2018) and teachers (Glock, 2016; Lorenz et al., 2016; Sprietsma, 2013; Tobisch & Dresel, 2017) hold lower expectations for and make more negative judgments of students with a Turkish (vs. German) family background. Ethnic biases may vary depending on the students' performance level (e.g., test scores and supposed grades), though no clear pattern has emerged (Bonefeld et al., 2020; Glock, 2016; Glock & Krolak-Schwerdt, 2013; Wenz & Hoenig, 2020). Experimental research on tracking recommendations in particular has also shown that teachers are less likely to recommend students with a Turkish (vs. German) family background to the academic track (Sprietsma, 2013; Tobisch & Dresel, 2017). Teachers' expectations for children's future may also depend on performance, as a negative ethnic bias has been shown to occur only among higher-achieving students (Wenz & Hoenig, 2020). Moreover, student teachers showed a negative ethnic bias regarding higher-achieving students who otherwise confirmed stereotypes (e.g., regarding religion), but student teachers showed no negative ethnic bias regarding lower-achieving students who also confirmed such stereotypes (Klapproth et al., 2018). Moreover, the framing of the tracking decision can also lead to a more lenient evaluation of students with a Turkish background by student teachers (Nishen et al., 2023).

In terms of gender stereotypes, teachers perceive that girls, relative to boys, show more adaptive behaviors, e.g., demonstrating effort and better learning habits (Anders et al., 2010; Heyder & Kessels, 2017; Kuhl & Hannover, 2012), a finding that extends to implicit tests (Glock & Kleen, 2017). An experimental study revealed that teachers ascribed less behavior impeding and more behavior fostering learning to a student labeled female than to a student labeled male when both were described in exactly the same way, thus reflecting the stereotypes of the diligent female student and the lazy and troublesome male student (Heyder & Kessels, 2015). Overall, students' behaviors that foster or impede learning can show their suitability, affinity, and willingness to work intellectually – and should, as per the government guidelines (KMK, 2015), be considered when making tracking recommendations. Indeed, boys receive fewer academic track recommendations (Jürges & Schneider, 2011; Lehmann et al., 1997; Neugebauer, 2011) and are underrepresented in this track (Federal Statistical Office, 2023). In part, this is due to their lower grades in German (Ditton et al., 2005; Lehmann et al., 1997), but also their learning-related behaviors (Neugebauer, 2011). If teachers' judgments of students' positive and negative learning-related behaviors are influenced by stereotypes, this could partially explain the observed underrepresentation. However, a gender bias in tracking recommendations in experimental studies with teachers (Wenz & Hoenig, 2020) and student teachers (Nishen et al., 2023) is often not found. One vignette-based study found that student teachers gave more advantageous recommendations to high-achieving boys as compared to girls when their performance had improved, but this gender bias was inverted among low-achieving students (achievement indicated by supposed grades; Klapproth & Fischer, 2019). Possibly, this bias may not be as pronounced because of positive stereotypes for boys related to inherent brilliance and intelligence (Steinmayr & Spinath, 2009).

Many of the studies described above focus on manipulating either gender or immigrant background (e.g., Glock & Krolak-Schwerdt, 2013; Holder & Kessels, 2017) or controlling for multiple group memberships without testing interactions (e.g., Lorenz et al., 2016; Wenz & Hoenig, 2020). As a result of this one-dimensional approach, knowledge about teacher biases remains incomplete, and more is known about students that are perceived as prototypical for their group (e.g., girls and boys without an immigrant background; Cole, 2009). Intersectionality can be understood as "analytic approaches that consider the meaning and consequences of multiple categories of social group membership" (Cole, 2009, p. 170) and additionally highlights how the consequences of different intersectional group memberships relate to structural inequality and discrimination (Bowleg, 2012; Cole, 2009). Because an intersectional analysis includes explicitly reflecting on inequality and discrimination (Cole, 2009; Else-Quest & Hyde, 2016a), it seems particularly pertinent to apply it to the study of teacher biases, which examines inequality in how students are perceived.

In terms of theory in the context of stereotypes, the shifting standards theory (elaborated below) implies that, if the stereotypes for boys and girls with an immigrant background are sufficiently different, they could be considered two different (sub-)groups that are judged relative to different standards (Biernat, 1995, 2012; Preddie & Biernat, 2021). The lens-based account of intersectional stereotyping (Petsko et al., 2022) also argues that the context can influence whether people attend to stereotypes about a single category (e.g., immigrant background) or multiple categories (e.g., gender and immigrant background). Though we do not examine it in the present study, this means that differential stereotypes about girls and boys with a Turkish immigrant background may be important in some situations, while in others only one of the two "lenses" is used. To date, research on the content of stereotypes about people with a Turkish immigrant background has not specifically examined gender-specific content (Bonefeld & Karst, 2020;

Eckes, 2002). However, one study reports that 10 % of the generated stereotypes about Turkish people were related to male dominance (Kahraman & Knoblich, 2000). This stereotype of patriarchal family structures within Turkish-German communities could lead teachers to expect less from girls than boys with a Turkish immigrant background because they may deduce that (these) girls receive less academic support and are given fewer learning opportunities at home. Alternatively, girls with a Turkish immigrant background may be judged more positively if they are viewed as compliant and more aligned with feminine gender roles. Because femininity is positively associated with perceived academic engagement and achievement among teachers (Heyder & Kessels, 2015; Jones & Myhill, 2004), teachers may assume that girls with a Turkish immigrant background are particularly engaged and may perceive them as being similar to girls without an immigrant background.

Note that these considerations are merely speculations, not hypotheses, given the limited theoretical specifications and empirical intersectional knowledge about the content of stereotypes about specific subgroups. While this limits deducing hypotheses about possible subgroup differences, a number of studies have examined teacher bias as it relates to both male and female students with an immigrant background. Over multiple experimental studies that compared students with a German and a Turkish background, teachers and student teachers rated the German proficiency of students with a Turkish background as compared to those without an immigrant background more negatively for female (Glock & Kleen, 2019) as well as male students (Bonefeld et al., 2021; Glock, 2016; Glock & Krolak-Schwerdt, 2013). It is possible that this only holds for students that are described in vignettes as low-achieving (vs. high-achieving; teachers: Glock, 2016; Kleen & Glock, 2018; student teachers: Glock & Krolak-Schwerdt, 2013; but see also Bonefeld et al., 2020; Wenz & Hoenig, 2020). Another experimental study with student teachers found that whereas the size and direction of teacher bias in judging mathematics performance differed for boys with a Turkish (vs. German) family background depending on their performance in a test, girls with a Turkish (vs. German) background were judged less favorably regardless of their performance (Bonefeld et al., 2020). Interestingly, student engagement was actually judged as higher for girls with a Turkish (vs. German) family background (teachers: Kleen & Glock, 2018; student teachers: Glock & Kleen, 2019), while no difference emerged among teachers' judgments of boys with a Turkish (vs. German) background (Glock, 2016). To date, only one study has examined how tracking recommendations by student teachers may be influenced by gender and immigrant background simultaneously

(Nishen et al., 2023), and this study will be described in detail below. Overall, there is evidence to suggest that ethnic biases may occur toward both boys and girls, but there are some differences in when and how biases occur. This underscores the usefulness of strategically comparing biases directed toward boys and girls with and without an immigration background.

Shifting Standards in Stereotyped Judgments

The present study on teachers' tracking recommendations draws from the shifting standards model (SSM; Biernat, 2012). The basic premise of the SSM is that people shift their standards when evaluating a person according to the social group to which this person belongs and the stereotypes about this group's characteristics. Stereotypes are used as standards against which a given performance is evaluated, since they entail expectations about "the likely mean and range of group members on the attribute" (Biernat, 1995, p. 89). When describing a woman as "very tall," this might imply that she is very tall only compared with an average woman, whereas a man of the same height would be rated as "somewhat tall" (Biernat & Manis, 1994).

The SSM posits that the framing of the specific judgment situation affects whether members of stereotyped groups are judged more leniently or more strictly compared to members from non-stereotyped groups. When making an estimation of a person's performance on an objective scale, which is anchored to an external frame of reference and maintains a constant meaning regardless of who is judging or who is being judged (Biernat & Manis, 1994), the low expectations regarding a member of a negatively stereotyped group will likely translate into the expectation of lower scores (e.g., "he will have 60 % correct answers"). However, the same person could be judged on a subjective scale (e.g., "his performance is good") which allows for "semantic changes of meaning" (Biernat & Manis, 1994, p. 5). In this case, the low expectations regarding a member of a negatively stereotyped group should translate into higher scores because for someone of this particular group, an objectively mediocre performance is perceived as good.

Standard shifts also occur in evaluation situations with specific goals and content. In the SSM, minimum standards and confirmatory standards are considered two ends of a continuum (Biernat et al., 2010; Biernat & Kobrynowicz, 1997). *Minimum standards* (in the following MS) are used when some initial evidence for a required

ability is needed, for example, for a short-list in a selection process, or when giving out a non-exclusive reward (e.g., praise). MS are "expectations for a group and tend to directly reflect stereotypes" (Biernat et al., 2010, p. 855), which implies they are lower for a member of a negatively stereotyped group. *Confirmatory standards* (in the following CS) are used when perceivers seek "a higher level of proof" (Biernat & Kobrynowicz, 1997, p. 546), for instance, when a judgment situation calls for definitive proofs for a required ability, when a zero-sum reward (e.g., an award) is given out, or when a final selection or hiring takes place. CS are "thresholds that reflect certainty that an individual has an attribute" (Biernat et al., 2010, p. 855) and thus are higher for group members stereotyped as deficient. Here, a comparison between low expectations for the negatively stereotyped group and high expectations for the other group takes place (Biernat, 2012).

More lenient MS and stricter CS for members of groups considered to be deficient have been demonstrated in many studies (Biernat & Fuegen, 2001; Biernat & Kobrynowicz, 1997; Biernat & Ma, 2005; Biernat et al., 2008). For example, in one study, women (relative to men) were suspected of being competent for a job position based on fewer behavioral examples of competence (MS) but were confirmed to be competent based on more behavioral examples (CS; Biernat & Kobrynowicz, 1997). Putting the SSM's assumptions to a further test in a series of three studies, Biernat et al. (2010) argued that judgments of *in*competence should follow the inverse pattern of judgments of competence. They found that MS of judging incompetence were lower (suspicion of incompetence is triggered sooner) in one of their studies, and that in all three studies CS were higher for adult groups stereotyped as competent.

Shifting Standards in the Classroom?

Most research on shifting standards has compared the standards applied to adult male and female targets or to adult white and black targets (e.g., Biernat et al., 2010). Only recently, studies applied the shifting standard paradigm to ethnic and gender biases in the school context (Holder & Kessels, 2017; Nishen et al. 2023) and found initial evidence that standard shifts also occur when student teachers judge students. One study (Holder & Kessels, 2017) found girls to be judged as less competent in mathematics on objective scales, but as equally competent on subjective scales compared to boys. In parallel,

boys with a Turkish (vs. German) name were judged as less competent in German on objective scales and as equally able on subjective scales (Holder & Kessels, 2017). Another study (Nishen et al., 2023) tested the variation of the MS versus CS (as in Biernat et al., 2010) when judging a fourth grader's suitability for the academic track. Participants read a short vignette and were asked to indicate how many behaviors of the target student they felt they needed to observe in order to make the recommendation for the academic track, using a behavioral checklist listing ten different behaviors fostering learning (e.g. "prepares himself/herself systematically for tests"). Student teachers were asked to either check the *minimum* (MS) or *total* (CS) number of behaviors fostering learning that are necessary to *suspect* (MS) or *confirm* (CS) that the target student *may be* (MS) or *is* (CS) qualified for the academic track. This study used an intersectional 2 × 2 between-subjects design, varying gender and origin of targets' names (Turkish vs. German). No effects of target gender occurred. However, in line with SSM's predictions, less evidence for positive learning behavior was required for the student with a Turkish name in the MS condition, while in the CS condition, participants tended to require less evidence for the student with a German name (Nishen et al., 2023).

Study Overview and Hypotheses

The SSM is a well-established framework for revealing stereotype-based judgments about adult targets in many fields, mostly focusing on specific competencies or characteristics of different genders or races/ethnicities (Biernat, 2012). Initial studies have supported this framework's predictions regarding gender and ethnic stereotypes about students' performance (Holder & Kessels, 2017) and tracking decisions (Nishen et al, 2023). However, when teachers are reflecting on whether a student might be suitable for the academic track or not, it is very likely they will weigh not only what is in favor but also what speaks against this placement. When judging a student's *un*suitability for the academic track, that is, their incompetence, the SSM predicts the inverse pattern for judgments of competence (Biernat et al., 2010). Therefore, MS of judging unsuitability for the academic track should be lower for students without a Turkish immigrant background (because the suspicion of incompetence is triggered sooner), but CS should be higher (compared to students with a Turkish background). In addition, the higher ratio of girls attending the academic track (Federal Statistical Office, 2023) and the more positive stereotypes about girls regarding learning behavior (e.g. Heyder &

Kessels, 2015) could imply the stereotyping of girls as being especially suitable, that is, being low in incompetence in dealing with the demands of the academic track. However, the ratio between male and female students in the academic track is less imbalanced than that between students with and without a Turkish immigrant background. Further, the ability stereotypes regarding students of Turkish origin are negative (Froehlich et al., 2016), while this is not true for boys (Kessels, 2015). For these reasons and because Nishen and colleagues (2023) did not find any effects involving gender of target in their study on shifting standards in tracking decisions, in the present study gender effects are expected to be smaller than the effects involving ethnicity.

In the present experimental vignette study with German student teachers, MS versus CS were induced when making the tracking judgment. Ethnicity and gender of target students were varied between participants. Following prior research, we applied the behavioral checklist paradigm (Biernat & Kobrynowicz, 1997; Biernat et al., 2008). As we wanted to test the different standards regarding students' *un*suitability for the Gymnasium, the behavioral checklist reflects behaviors that impede students' learning.

If stereotypes are more positive for girls and students without an immigrant background than for boys and students of Turkish origin, an interaction effect of gender and origin of the name with the induced standard should emerge. In the MS condition, student teachers are expected to require less evidence for the unsuitability of girls than for the unsuitability of boys (H1a) and of students with a German-sounding name than for students with a Turkish-sounding name (H2a). In the CS condition, they are expected to require more evidence to confirm the unsuitability of girls compared to boys (H1b) and of students with a German-sounding name compared to students with a Turkish-sounding name (H2b). We will also examine the interplay of gender and ethnicity, though only in an exploratory manner since, to date, intersectional effects have been difficult to deduce from the theoretical basis and have rarely been studied in detail (cf. Nishen et al., 2023).

Methods

Participants

Participants were recruited using the university mailing list for students at a large university in a city of Northwestern Germany. An email invited student teachers to participate in an online study on tracking recommenda-

tions. As an incentive, the voluntary participation in a lottery for vouchers worth 20 € was included. Although both the invitation email and the instruction of the online study clearly stated that only student teachers were the target group for the study, students not in the teacher training program also participated ($n = 125$). These were excluded from the data analyses, as were $n = 7$ persons ticking "other" instead of the different levels of teacher training spelled out in the questionnaire. In addition, we decided to exclude students who were enrolled in the program for teaching children with Special Educational Needs ($n = 82$), as we aimed to test a sample with a comparable conception of regular students' behaviors at grade 4, and the fictitious target student described in the vignette had no special educational needs. The final study sample comprised $n = 185$ student teachers ($n = 26$ elementary school; $n = 30$ secondary school/lower tracks; $n = 129$ secondary school/academic track). The mean age was 24.53 ($SD = 4.24$), the mean study semester was 6.63 ($SD = 4.00$), and 91% spoke German as their native language. Unfortunately, gender of participants was not collected.

Design, Experimental Treatment, and Measures

Participants were randomly assigned to one of eight conditions in a 2 (target student's ethnicity: Turkish vs. German) × 2 (target student's gender: male vs. female) × 2 (standard: minimum vs. confirmatory) between-subjects design. On the first page of the online questionnaire, an "exercise in evaluating elementary school children" was announced. On the second page, participants read that teachers are requested to form an overall picture of a student when making tracking recommendations at the end of elementary school, and that this is not exclusively based on students' grades.

They further read that the guidelines of the KMK ask teachers to consider the child's "suitability, affinity, and willingness [...] to work intellectually" (2015, p. 5 [translated by the authors]) as well as the "general skills important for success in school" (2015, p. 6 [translated by the authors]). As a basis for this, various school-related behaviors that students might exhibit could be considered. Participants were told that they would now engage in a short exercise for rating elementary school students. On the next page, they were asked to imagine one specific example of an elementary student they had to judge. All instructions were presented in German and translated into English for this paper. "[Target's name] is a fourth-grade student in [city]. Considering exclusively his/her GPA, a transition to the Gymnasium would not be

Table 1. Analysis of variance for requested number of behaviors impeding learning

Variable	$F_{(1,177)}$	p-value	Partial η^2
Standard	.70	.405	.004
Target student gender	2.65	.106	.015
Target student ethnicity	.031	.860	.000
Target student gender x standard	.193	.661	.001
Target student ethnicity x standard	5.17*	.024	.028
Target student gender x target student ethnicity	.11	.746	.001
Target student gender x target student ethnicity x standard	.09	.769	<.001

Note. *p < .05.

unambiguously justifiable. How do you proceed when judging [target's name] behavior? In the following, you will see ten behaviors that can be observed in fourth graders."

Following procedures used by Biernat et al. (2010), participants in the MS condition read that we were interested in "the minimum number of behaviors that are necessary to suspect that [the target student] may not be qualified for the Gymnasium." Participants in the CS condition read that we were interested in the "the total number of behaviors that are necessary to confirm that [the target student] is not qualified for the Gymnasium." All participants were asked to review a list of 10 behaviors and to check off "as many or as few behaviors" as the target student would need to engage in to either "give you some inkling or hint that [the target student] may not be qualified for the Gymnasium" (MS condition), or "to confirm that [the target student] is not qualified for the Gymnasium" (CS condition; emphasis in original). In addition to this manipulation, participants were asked to imagine a student who was presented as having a male or female German or Turkish name (names were the same as in Nishen et al., 2023: Tim Menzel, Anna Menzel, Deniz Gül, Selma Gül). Target students' gender was not only indicated by their names, but also by the pronouns used in the description.

The *dependent measure* was the number of behaviors checked (out of ten). These behaviors were taken from a pilot study (Heyder & Kessels, 2015) in which 86 student teachers (77% female; M_{age} = 23.96, SD_{age} = 2.80) rated a pool of 61 behaviors representing possible student behavior as either fostering or impeding learning. Ten behaviors that were unambiguously classified as impeding learning were used for the present study (M > 5.5 and SD < 1.0 on a 7-point Likert scale where 1 = *fosters learning* and 7 = *impedes learning*). The list of behaviors included items such as "copies homework" and "does not file worksheets and loses them" (full list in the Appendix A, Table A1).

Analyses and Results

A target ethnicity × target gender × standards ANOVA on number of behaviors checked as the dependent variable (see Table 1 for all results) yielded no main or interaction effect involving gender of target (all p > .110). Hypotheses 1a and 1b, stating that, in the MS condition, student teachers would require less evidence for girls than for boys, and that in the CS condition, they would require more evidence for confirming the unsuitability of girls than for boys, were not supported by our data. Regarding our exploratory analyses, the non-significant three-way interaction indicates that the effect described below holds for both boys and girls. That is, the pattern of evidence required to decide against an academic track recommendation did not differ for boys and girls with a Turkish name, nor did it differ for boys and girls with a German name.

A significant interaction between target ethnicity and standard emerged, $F_{(1,177)}$ = 5.17, p = .024. Descriptively, MS were lower for a student with a Turkish name (M = 3.12, SD = 1.66) than for a student with a German name (M = 3.62, SD = 1.79), but CS were higher for a student with a Turkish name (M = 3.50, SD = 1.81) than for a student with a German name (M = 2.85, SD = 1.86) (Table 2).

Since the pattern of means directly contradicted the expected direction of effects, four post-hoc Bonferroni-corrected t-tests were conducted (instead of planned contrasts). These t-tests revealed that between the target students with a Turkish or a German name, neither the differences in MS ($t_{(97)}$ = 1.44, p = .608) nor the differences in CS ($t_{(84)}$ = –1.64, p = .420) were significant. Hypotheses H2a and H2b were not supported by the data. Exploratorily, and in order to better understand what might have driven the significant interaction effect reported above, we further tested whether MS made a difference compared to CS within the same ethnic group. Descriptively, for students with a Turkish name, MS was lower (M = 3.12, SD = 1.66) than CS (M = 3.50, SD = 1.81),

Table 2. Means and standard deviations for requested number of behaviors impeding learning by condition

Ethnicity	Gender	n	Standard	M	SD
German name	Boy	28	Minimum	3.39	1.79
		23	Confirmatory	2.61	2.10
	Girl	19	Minimum	3.94	1.77
		23	Confirmatory	3.09	1.59
	total	47	Minimum	3.62	1.79
		46	Confirmatory	2.85	1.86
Turkish name	Boy	26	Minimum	2.84	1.54
		19	Confirmatory	3.42	1.77
	Girl	26	Minimum	3.38	1.77
		21	Confirmatory	3.57	1.89
	total	52	Minimum	3.12	1.66
		40	Confirmatory	3.50	1.81

but this difference was not significant ($t(90) = -1.06$; $p > .999$). Whereas for students with a German name, MS was higher ($M = 3.62$, $SD = 1.79$) than CS ($M = 2.85$, $SD = 1.86$), the Bonferroni-corrected t-test revealed a nonsignificant result ($t(91) = 2.03$, $p = .180$). Thus, the type of standard applied made no difference, indicating that the number of required learning-impeding behaviors did not systematically vary with the gender or the immigrant background of the target student in the different standards conditions.

Discussion

In order to gain a better understanding of the mechanisms of possible ethnic and gender biases in tracking recommendations, the present study tested the specific predictions of the Shifting Standards Model when judging a target person's *in*competence for the highest track. In the study, both gender and immigrant background of students were varied systematically. This permitted us to not only test separate effects of gender stereotypes, but also their interaction, which was another aim of the present study. This intersectional lens allowed us to examine the effects of stereotypes for a subgroup that is not seen as prototypical for the groups to which they belong: girls with a Turkish name.

The present research is the first to test the assumptions from the SSM regarding a possible bias when judging a person's incompetence and apply it to the school context. Following Biernat et al.'s (2010) work, it was expected that MS would be lower, but CS would be higher for students stereotyped as low in *in*competence and therefore, low in *un*suitability to the academic track (i.e., girls, ethnic majority students). However, the results of our experimental study did not confirm the hypotheses derived from the SSM. When student teachers were asked how much evidence of behavior impeding learning they needed in order to only suspect that a girl or a student without an immigrant background may not be qualified for the Gymnasium, they did not require less evidence than when having to be absolutely certain. The number of required learning-impeding behaviors did not systematically vary with the gender or the immigrant background of the target student in the different standards conditions. This also applied to the interaction of standards with gender and immigrant background, which we examined exploratorily. This nonsignificant interaction showed that the interaction of standards and gender did not differ based on the German or Turkish background of the students. Overall, no effects supporting the predictions derived from the SSM were found.

In this respect, our results are in contrast with published studies on the SSM. However, it should be noted that most SSM research has taken the more straightforward way of testing whether members of negatively stereotyped groups are judged more positively on subjective scales or more leniently in an MS condition, and at the same time more negatively on objective scales or more strictly in a CS condition, compared to members of non-stereotyped groups. The evidence from many studies (e.g. Biernat & Kobrynowicz, 1997; Biernat & Manis, 1994) does speak in favour of the theory that perceivers shift their standards when evaluating a person according to the social group to which this person belongs and the associated stereotypes of said group. In the present study, we focused on the assumed incompetence to be incompetent of the student groups that are stereotyped rather positively, or at least not negatively; students who are, as Biernat et al. (2010) put it, "deficient in incompetence"

(p. 859). We considered this way of looking at the standard shift as important, too, because tracking decisions, like any decision, require not only a consideration of what is in favour, but also what speaks against each option. Only few studies so far have taken this less straightforward look at the standard shift when judging target persons. To our knowledge, only the three studies which were jointly presented in Biernat et al. (2010) took this approach, mainly supporting their hypotheses how targets stereotyped as "deficient in incompetence" (p. 859) would be judged relative to those that were stereotyped as incompetent. Even reviewing the above sentences, one might be struck by how overly complex the reasoning about the group, on which the judgment of the group member is supposed to be based, is constructed. Likely, a phrase like "Because of his/her group membership, this student seems deficient in incompetence for meeting the demands of the Gymnasium – so what is confirming this view?" has not crossed anyone's mind in our study. The more straightforward way to think would be "Because of his/her group membership, this student seems competent to meet the demands of the Gymnasium; what could speak against it?", but this much simpler way of reasoning is not what the "reversed" SSM is based on. Taken together, we consider our specific null results as not strong enough to fundamentally question the SSM in general, but to be critical about actually expecting a reverse pattern to occur in a reversely phrased setting. This inversely-constructed setting might reflect complex theoretical deductions from experimental psychologists rather than actually-occuring cognitive processes, thus lacking ecological validity. We want to stress that generally, the SSM has been proven to be useful to explain both the obvious and the masked effects of stereotyped judgments about different negatively stereotyped groups, including in the context of education (Holder & Kessels, 2017; Nishen et al., 2023).

In a similar vein, we want to emphasize that our observation that gender and immigrant background did not interact in our experimental setup does not imply that an intersectional analysis of this topic is not valuable. First, a review of other studies shows that teacher judgments may be impacted by the immigrant background of boys differently than by the immigrant background of girls (Bonefeld et al., 2021; Glock & Kleen, 2019; Kleen & Glock, 2018; but see Nishen et al., 2023), and further complexity has been introduced by interactions with performance level (Bonefeld et al., 2020). Another experimental study conducted with Australian pre-service teachers demonstrates that it is important to examine how teacher bias is influenced by ability level (Matheis et al., 2020). Whereas average-ability boys were seen as more maladjusted than their female peers, gifted children

were judged as more maladjusted irrespective of gender (Matheis et al., 2020).

Additionally, intersectional analyses can help not only predict differences between groups, but can also illuminate how the similarities that may be found between different groups can arise through different underlying processes (Cole, 2009). In this way, intersectionality provides an important impetus to more deeply consider theoretical frameworks. For example, reference groups play a central role in the SSM (Biernat, 1995, 2012). Reworking the theory with an explicitly intersectional approach in mind, researchers could consider under which circumstances new reference groups are used that reflect subgroups of a larger category (e. g., girls and boys with an immigrant background; for an example on race and sexuality, see Preddie & Biernat, 2021). This could then be related to theory and research on the process of subgrouping (Richards & Hewstone, 2001) and intersectional stereotyping more generally (Hall et al., 2019; Petsko et al., 2022). Additionally, the SSM also relates to another important aspect of intersectionality, namely, that social categories are constructed (Bowleg, 2012; Cole, 2009). This is an aspect that the SSM already shares with an intersectional approach, though this could be made more explicit by theorizing and testing how the visibility of a group membership affects standard shifts. Overall, an intersectional analysis of theoretical arguments can provide impetus to be more precise in one's predictions, for example, to understand sources of discrimination, and create new research designs that particularly reflect the experiences of members of multiply oppressed groups (Bowleg, 2012; Cole, 2009; Weber & Parra-Medina, 2003).

When interpreting our results, several limitations should be considered. We do not know whether our results are representative of experienced teachers, as all our participants were still in training. Further, our sample was mostly composed of student teachers preparing to teach at the secondary level and a smaller portion of participants was studying to teach at elementary level, and only the latter are future teachers who would have to give out tracking recommendations. However, for teachers at both school levels, assessments of who is (not) suitable for the Gymnasium are relevant and also fall within their area of competence. While future secondary school teachers do not have to give out tracking recommendations at the end of elementary school, they do know the requirements of the school types (academic or non-academic track) at least as well and should therefore be able to assess a student's (non-) suitability competently. Still, future secondary school teachers have less experience with fourth graders, but are more experienced with children from grade 5 onwards. However, it is unlikely

that this lack of experience with younger children has had an impact regarding the predicted standard shift when making judgments about members of different groups.

Moreover, there is some debate within the literature on intersectionality and its application to research about the type of research that is most suited for this approach (Bowleg, 2008; Cole, 2009; Else-Quest & Hyde, 2016a, 2016b; McCall, 2005). Particularly, quantitative research has been criticized for reflecting a view of social identities as stable (and often binary opposites) rather than shifting over time and between contexts, and for not attending to the heterogeneity within categories (Else-Quest & Hyde, 2016a; McCall, 2005). This includes designs such as that of the present study, which manipulates gender and immigrant background as binary variables. Moreover, the social category of "immigrant background" in Germany has been criticized as imposed by majority-group members to talk about those viewed as "other", not reflecting a self-chosen identity (Will, 2019). In the present study, we examined student teachers' judgments of children rather than the experience of the children themselves. With this goal in mind, we believe it is constructive to use categories that have been demonstrated to be used by teachers in making judgments (e.g., Anders et al., 2010; Bonefeld & Dickhäuser, 2018; Heyder & Kessels, 2017), even though they do not reflect the complexity of the experience within each group. However, when examining the experience of the children themselves, measures encompassing psychological experience within categories (e.g., ethnic identities) could be included to reflect the notion of socially constructed categories in the design. Overall, researchers concur that both quantitative and qualitative approaches can contribute to intersectional research (Bowleg, 2008; Cole, 2009; Else-Quest & Hyde, 2016a). However, an intersectional approach should not only be reflected in the design, but also in theoretical considerations and the interpretation of results, for example, by reflecting on the social constructiveness of the categories used and interpreting results in relation to inequality and discrimination (Cole, 2009; Else-Quest & Hyde, 2016a, 2016b).

Since the differences found in our sample did not reach the level of statistical significance, one might suspect an issue of low power to detect existing effects. A post-hoc power analysis indeed showed that for the significant effect (partial η^2 = .028), a power of β = 0.63 was achieved, which is less than desirable. However, bearing in mind that the differences in our sample were not only not significant, but that the means pattern was in the opposite direction from what the SSM would predict, it is highly unlikely that a larger sample would have led to the confirmation of our hypotheses. Still, future research should include larger and more diverse samples in order to test not only for the effect of target students' charac-

teristics on the outcome variables, but also how these effects might be moderated by teachers' own gender or immigrant background.

Notwithstanding the null results of the present study, given earlier results (Nishen et al., 2023) it is important to consider which standard is used in real-life tracking recommendations. It remains uncertain whether elementary teachers perceive of a recommendation for the academic track as a very limited resource, which would imply a CS, or as something that is plentiful and can be distributed as generously as praise, which would imply an MS. The number of students transferring to the Gymnasium has risen constantly over the last decades (in the federal state where the study took place, from 35 % in 2003 to 42 % in 2021; Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen, 2012, 2022), and the teacher recommendation is binding in only five out of 16 federal states in Germany, and even in these five states, there are ways for parents to opt against the recommendations (KMK, 2015; in the federal state where the study took place, the recommendation is not binding). Even though most teachers will be motivated to give their students accurate and helpful advice, their recommendations do not represent a final decision for or against a school type anymore. This might imply that an MS is increasingly being applied, with the consequence that negatively stereotyped student groups will be judged more leniently than non-stereotyped groups (Nishen et al., 2023). In closing, we hope that the present study can contribute to the discussion around gender and ethnic biases in tracking recommendations by highlighting the practicability of intersectional research design and analyses as well as the need to critically weigh predictions directly drawn from complex theoretical models and the more heuristic cognitive processes that often govern thinking in everyday life.

References

Anders, Y., McElvany, N., & Baumert, J. (2010). Die Einschätzung lernrelevanter Schülermerkmale zum Zeitpunkt des Übergangs von der Grundschule auf die weiterführende Schule: Wie differenziert urteilen Lehrkräfte? [The evaluation of learning-related behaviors at the time of transition to secondary schools: How differentiated are teachers' judgments?]. In K. Maaz, J. Baumert, C. Gresch, & N. McElvany (Eds.), *Der Übergang von der Grundschule in die weiterführende Schule: Leistungsgerechtigkeit und regionale, soziale und ethnisch-kulturelle Disparitäten* (pp. 313–330). Bundesministerium für Bildung und Forschung.

Asbrock, F. (2010). Stereotypes of social groups in Germany in terms of warmth and competence. *Social Psychology, 41*(2), 76–81. https://doi.org/10.1027/1864-9335/a000011

Autorengruppe Bildungsberichterstattung (Ed.). (2016). *Bildung in Deutschland 2016: Ein Indikatoren gestützter Bericht mit einer Analyse zur Bildung und Migration* [Education in Germany 2016: An indicator-based report with an analysis of education and migration]. Bertelsmann.

Baumert, J., Stanat, P., & Watermann, R. (2006). Schulstruktur und die Entstehung differenzieller Lern- und Entwicklungsmilieus [School structure and the development of differential environments for learning and development]. In J. Baumert, P. Stanat, & R. Watermann (Eds.), *Herkunftsbedingte Disparitäten im Bildungswesen: Differenzielle Bildungsprozesse und Probleme der Verteilungsgerechtigkeit* (pp. 95 – 188). VS Verlag für Sozialwissenschaften.

Becker, M., Lüdtke, O., Trautwein, U., & Baumert, J. (2006). Leistungszuwachs in Mathematik: Evidenz für einen Schereneffekt im mehrgliedrigen Schulsystem? [Performance gains in mathematics: Evidence of a scissor effect in the tracked school system?]. *Zeitschrift für Pädagogische Psychologie*, *20*(4), 233 – 242. https://doi.org/10.1024/1010-0652.20.4.233

Becker, M., Lüdtke, O., Trautwein, U., Köller, O., & Baumert, J. (2012). The differential effects of school tracking on psychometric intelligence: Do academic-track schools make students smarter? *Journal of Educational Psychology, 104*(3), 682 – 699. https://doi.org/10.1037/a0027608

Becker, M., Neumann, M., & Dumont, H. (2016). Recent developments in school tracking practices in Germany: An overview and outlook on future trends. *Orbis Scholae*, *10*(3), 9 – 25. https://doi.org/10.14712/23363177.2017.8

Bellenberg, G. (2012). *Schulformwechsel in Deutschland: Durchlässigkeit und Selektion in den 16 Schulsystemen der Bundesländer innerhalb der Sekundarstufe I* [Changing school tracks in Germany: Permeability and selection in the 16 school systems of the federal states regarding middle school]. Bertelsmann Stiftung. Retrieved from http://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/GP_Schulformwechsel_in_Deutschland.pdf

Biernat, M. (1995). The shifting standards model: Implications of stereotype accuracy for social judgment. In Y.-T. Lee, L. J. Jussim, & C. R. McCauley (Eds.), *Stereotype accuracy: Toward appreciating group differences* (pp. 87 – 114). American Psychological Association. https://doi.org/10.1037/10495-004

Biernat, M. (2012). Stereotypes and shifting standards: Forming, communicating, and translating person impressions. In P. Devine & A. Plant (Eds.), *Advances in Experimental Social Psychology* (45th ed., pp. 1 – 60). Elsevier.

Biernat, M., & Fuegen, K. (2001). Shifting standards and the evaluation of competence: Complexity in gender-based judgment and decision making. *Journal of Social Issues*, *57*(4), 707 – 724. https://doi.org/10.1111/0022-4537.00237

Biernat, M., Fuegen, K., & Kobrynowicz, D. (2010). Shifting standards and the inference of incompetence: Effects of formal and informal evaluation tools. *Personality and Social Psychology Bulletin*, *36*(7), 855 – 868. https://doi.org/10.1177/0146167210369483

Biernat, M., & Kobrynowicz, D. (1997). Gender- and race-based standards of competence: Lower minimum standards but higher ability standards for devalued groups. *Journal of Personality and Social Psychology*, *72*(3), 544 – 557. https://doi.org/10.1037//0022-3514.72.3.544

Biernat, M., & Ma, J. E. (2005). Stereotypes and the confirmability of trait concepts. *Personality and Social Psychology Bulletin*, *31*(4), 483 – 495. https://doi.org/10.1177/0146167204271712

Biernat, M., Ma, J. E., & Nario-Redmond, M. R. (2008). Standards to suspect and diagnose stereotypical traits. *Social Cognition*, *26*(3), 288 – 313. https://doi.org/10.1521/soco.2008.26.3.288

Biernat, M., & Manis, M. (1994). Shifting standards and stereotype-based judgments. *Journal of Personality and Social Psychology*, *66*(1), 5 – 20. https://doi.org/10.1037/0022-3514.66.1.5

Bodenhausen, G. V. (1990). Stereotypes as judgmental heuristics: Evidence of circadian variations in discrimination. *Psychological Science, 1*, 319 – 322. https://doi.org/10.1111/j.1467-9280.1990.tb00226.x

Bonefeld, M., & Dickhäuser, O. (2018). (Biased) grading of students' performance: Students' names, performance level, and implicit attitudes. *Frontiers in Psychology, 9*, 481. https://doi.org/10.3389/fpsyg.2018.00481

Bonefeld, M., Dickhäuser, O., Janke, S., Praetorius, A. K., & Dresel, M. (2017). Migrationsbedingte Disparitäten in der Notenvergabe nach dem Übergang auf das Gymnasium [Disparities in grades after transitioning to the academic track based on migration background]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 49*(1), 11 – 23. https://doi.org/10.1026/0049-8637/a000163

Bonefeld, M., Dickhäuser, O., & Karst, K. (2020). Do preservice teachers' judgments and judgment accuracy depend on students' characteristics? The effect of gender and immigration background. *Social Psychology of Education, 23*(1), 189 – 216. https://doi.org/10.1007/s11218-019-09533-2

Bonefeld, M., & Karst, K. (2020). Döner vs. Schweinebraten: Stereotype von (angehenden) Lehrkräften über Personen deutscher und türkischer Herkunft im Vergleich [Döner vs. pig roast: Comparison of stereotypes of (preservice) teachers about people of German and Turkish descent]. In S. Glock, & H. Kleen (Eds.), *Stereotype in der Schule* (pp. 159 – 190). Springer VS.

Bonefeld, M., Kleen, H., & Glock, S. (2021). The effect of the interplay of gender and ethnicity on teachers' judgements: Does the school subject matter? *The Journal of Experimental Education, 90*(4), 818 – 838. https://doi.org/10.1080/00220973.2021.1878991

Bowleg, L. (2008). When black + lesbian + woman ≠ black lesbian woman: The methodological challenges of qualitative and quantitative intersectionality research. *Sex Roles, 59*, 312 – 325. https://doi.org/10.1007/s11199-008-9400-z

Bowleg, L. (2012). The problem with the phrase "women and minorities": Intersectionality, an important theoretical framework for public health. *American Journal of Public Health, 102*(7), 1267 – 1273. https://doi.org/10.2105/AJPH.2012.300750

Cole, E. R. (2009). Intersectionality and research in psychology. *American Psychologist, 64*(3), 170 – 180. https://doi.org/10.1037/a0014564

Daniel, A., & Neumann, M. (2023). Schule und Studium [School and studies]. In T. Hascher, T.S. Idel, & W. Helsper (Eds.), *Handbuch Schulforschung* (pp. 734 – 757). Springer VS. https://doi.org/10.1007/978-3-658-24729-4_73

Deaux, K., & Lewis, L. L. (1984). Structure of gender stereotypes: Interrelationships among components and gender label. *Journal of Personality and Social Psychology, 46*(5), 991 – 1004.

Ditton, H., Krusken, J., & Schauenberg, M. (2005). Bildungsungleichheit: Der Beitrag von Familie und Schule [Education inequality: The contribution of family and school]. *Zeitschrift für Erziehungswissenschaft, 8*(2), 285 – 304. https://doi.org/10.1007/s11618-005-0138-x

Eckes, T. (2002). Paternalistic and envious gender stereotypes: Testing predictions from the stereotype content model. *Sex Roles, 47*(3 – 4), 99 – 114. https://doi.org/10.1023/A:1021020920715

Else-Quest, N. M., & Hyde, J. S. (2016a). Intersectionality in quantitative psychological research: I. Theoretical and epistemological issues. *Psychology of Women Quarterly, 40*(2), 155 – 170. https://doi.org/10.1177/0361684316629797

Else-Quest, N. M., & Hyde, J. S. (2016b). Intersectionality in quantitative psychological research: II. Methods and techniques. *Psychology of Women Quarterly, 40*(3), 319–336. https://doi.org/10.1177/0361684316647953

Federal Statistical Office. (2023). *Gleichstellungsindikatoren* [Gender equality indicators]. Retrieved from https://www.destatis.de/DE/Themen/Querschnitt/Gleichstellungsindikatoren/tab-studienberechtigtenquote-h22.html

Froehlich, L., Martiny, S. E., Deaux, K., & Mok, S. Y. (2016). "It's their responsibility, not ours": Stereotypes about competence and causal attributions for immigrants' academic underperformance. *Social Psychology, 47*(2), 74–86. https://doi.org/10.1027/1864-9335/a000260

Glock, S. (2016). Does ethnicity matter? The impact of stereotypical expectations on in-service teachers' judgments of students. *Social Psychology of Education, 19*(3), 493–509. https://doi.org/10.1007/s11218-016-9349-7

Glock, S., & Kleen, H. (2017). Gender and student misbehavior: Evidence from implicit and explicit measures. *Teaching and Teacher Education, 67*, 93–103. https://doi.org/10.1016/j.tate.2017.05.015

Glock, S., & Kleen, H. (2019). Shifting standards for female ethnic minority students? Evidence from two experimental studies. *Educational Research and Evaluation, 25*(5–6), 229–247. https://doi.org/10.1080/13803611.2019.1701499

Glock, S., & Krolak-Schwerdt, S. (2013). Does nationality matter? The impact of stereotypical expectations on student teachers' judgments. *Social Psychology of Education, 16*(1), 111–127. https://doi.org/10.1007/s11218-012-9197-z

Hall, E. V., Hall, A. V., Galinsky, A. D., & Phillips, K. W. (2019). MOSAIC: A model of stereotyping through associated and intersectional categories. *Academy of Management Review, 44*(3), 643–672. https://doi.org/10.5465/amr.2017.0109

Hattie, J. A. C. (2002). Classroom composition and peer effects. *International Journal of Educational Research, 37*, 449–481. https://doi.org/10.1016/S0883-0355(03)00015-6

Heyder, A., & Kessels, U. (2015). Do teachers equate male and masculine with lower academic engagement? How students' gender enactment triggers gender stereotypes at school. *Social Psychology of Education, 18*, 467–485. https://doi.org/10.1007/s11218-015-9303-0

Heyder, A., & Kessels, U. (2017). Boys don't work? On the psychological benefits of showing low effort in high school. *Sex Roles, 77*(1–2), 72–85. https://doi.org/10.1007/s11199-016-0683-1

Hilton, J. L., & von Hippel, W. (1996). Stereotypes. *Annual Review of Psychology, 47*, 237–271. https://doi.org/10.1146/annurev.psych.47.1.237

Holder, K., & Kessels, U. (2017). Gender and ethnic stereotypes in student teachers' judgments: A new look from a shifting standards perspective. *Social Psychology of Education, 20*(3), 471–490. https://doi.org/10.1007/s11218-017-9384-z

Jones, S., & Myhill, D. (2004). "Troublesome boys" and "compliant girls": Gender identity and perceptions of achievement and underachievement. *British Journal of Sociology of Education, 25*(5), 547–561. https://doi.org/10.1080/0142569042000252044

Jürges, H., & Schneider, K. (2011). Why young boys stumble: Early tracking, age and gender bias in the German school system. *German Economic Review, 12*(4), 371–394. https://doi.org/10.1111/j.1468-0475.2011.00533.x

Kahraman, B., & Knoblich, G. (2000). "Stechen statt Sprechen": Valenz und Aktivierbarkeit von Stereotypen über Türken ["Stabbing, not speaking": Valence and ease of activation of stereotypes about Turks]. *Zeitschrift für Sozialpsychologie, 31*(1), 31–43. https://doi.org/10.1024//0044-3514.31.1.31

Kaiser, J., Retelsdorf, J., Südkamp, A., & Möller, J. (2013). Achievement and engagement: How student characteristics influence teacher judgments. *Learning and Instruction, 28*, 73–84. https://doi.org/10.1016/j.learninstruc.2013.06.001

Kessels, U. (2015). Bridging the gap by enhancing the fit: How stereotypes about STEM clash with stereotypes about girls. *International Journal of Gender, Science and Technology, 7*(2), 280–296. Retrieved from https://genderandset.open.ac.uk/index.php/genderandset/article/view/392

Klapproth, F., & Fischer, B. D. (2019). Preservice teachers' evaluations of students' achievement development in the context of school-track recommendations. *European Journal of Psychology of Education, 34*, 825–846. https://doi.org/10.1007/s10212-018-0405-x

Klapproth, F., Kärchner, H., & Glock, S. (2018). Do students' religion and school absences moderate the effect of ethnic stereotypes on school-placement recommendations? *The Journal of Experimental Education, 86*(2), 173–194. https://doi.org/10.1080/00220973.2017.1293602

Kleen, H., & Glock, S. (2018). A further look into ethnicity: The impact of stereotypical expectations on teachers' judgments of female ethnic minority students. *Social Psychology of Education, 21*, 759–773. https://doi.org/10.1007/s11218-018-9451-0

KMK (Secretary of the Standing Conference of the Ministers of Education and Cultural Affairs). (2015). *Übergang von der Grundschule in Schulen des Sekundarbereichs I und Förderung, Beobachtung und Orientierung in den Jahrgangsstufen 5 und 6 (sog. Orientierungsstufe)* [Transition from elementary school to secondary school and facilitation, observation, and orientation in fifth and sixth grade]. Retrieved from https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2015/2015_02_19-Uebergang_Grundschule-SI-Orientierungsstufe.pdf

Kuhl, P., & Hannover, B. (2012). Differenzielle Benotungen von Mädchen und Jungen: Der Einfluss der von der Lehrkraft eingeschätzten Kompetenz zum selbstgesteuerten Lernen [Differential grading of boys and girls: The influence of teacher-perceived competence for self-regulated learning]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 44*(3), 153–162. https://doi.org/10.1026/0049-8637/a000066

Lehmann, R. H., Peek, R., & Gansfus, R. (1997). *Aspekte der Lernausgangslage und der Lernentwicklung: Bericht über die Erhebung im September 1996 (LAU 5).* [Aspects of the learning baseline and learning development: Report on the data collection in September 1996]. Amt für Schule, Behörde für Schule, Jugend und Berufsbildung.

Lorenz, G., Gentrup, S., Kristen, C., Stanat, P., & Kogan, I. (2016). Stereotype bei Lehrkräften? Eine Untersuchung systematisch verzerrter Lehrererwartungen [Stereotypes in teachers? A study on systematically biased teacher expectations]. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 68*, 89–111. https://doi.org/10.1007/s11577-015-0352-3

Maaz, K., Trautwein, U., Lüdtke, O., & Baumert, J. (2008). Educational transitions and differential learning environments: How explicit between-school tracking contributes to social inequality in educational outcomes. *Child Development Perspectives, 2*(2), 99–106. https://doi.org/10.1111/j.1750-8606.2008.00048.x

Macrae, C. N., & Bodenhausen, G. V. (2000). Social cognition: Thinking categorically about others. *Annual Review of Psychology, 51*(1), 93–120.

Martiny, S. E., & Froehlich, L. (2020). Ein theoretischer und empirischer Überblick über die Entwicklung von Stereotypen und ihre Konsequenzen im Schulkontext [A theoretical and empirical overview of the development of stereotypes and their consequences in the school context]. In S. Glock & H. Kleen (Eds.), *Stereotype in der Schule* (pp. 1–32). Springer.

Matheis, S., Keller, L. K., Kronborg, L., Schmitt, M., & Preckel, F. (2020). Do stereotypes strike twice? Giftedness and gender

stereotypes in preservice teachers' beliefs about student characteristics in Australia. *Asia-Pacific Journal of Teacher Education, 48*(2), 213–232. https://doi.org/10.1080/1359866X.2019.1576029

McCall, L. (2005). The complexity of intersectionality. *Signs, 30*(3), 1771–1800. https://doi.org/10.1086/426800

Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen. (2012). *Statistik-TELEGRAMM 2011/12* [Statistics Telegram 2011/12]. Retrieved from https://www.schulministerium.nrw/sites/default/files/documents/StatTelegramm2011.pdf

Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen. (2022). *Statistik-TELEGRAMM 2021/22* [Statistics Telegram 2021/22]. Retrieved from https://www.schulministerium.nrw/system/files/media/document/file/statistiktelegramm-2021.pdf

Neugebauer, M. (2011). Werden Jungen von Lehrerinnen bei den Übergangsempfehlungen für das Gymnasium benachteiligt? Eine Analyse auf Basis der IGLU-Daten [Are boys disadvantaged by female teachers' tracking recommendations for the academic track? An ananlysis based on PIRLS data]. In A. Hadjar (Ed.), *Geschlechtsspezifische Bildungsungleichheiten* (pp. 235–260). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-92779-4_10

Nishen, A.K., Corcoran, K., Holder, K., & Kessels, U. (2023). When ethnic minority students are judged as more suitable for the highest school track: A shifting standards experiment. *European Journal of Psychology of Education, 38*, 367–387. http://doi.org/10.1007/s10212-021-00595-5

Petsko, C. D., Rosette, A. S., & Bodenhausen, G. V. (2022). Through the looking glass: A lens-based account of intersectional stereotyping. *Journal of Personality and Social Psychology, 123*(4), 763–787. https://doi.org/10.1037/pspi0000382

Preddie, J.P., & Biernat, M. (2021). More than the sum of its parts: Intersections of sexual orientation and race as they influence perceptions of group similarity and stereotype content. *Sex Roles, 84*, 554–573. https://doi.org/10.1007/s11199-020-01185-3

Rauch, D., Mang, J., Hartig, H., & Haag, N. (2016). Naturwissenschaftliche Kompetenz von Schülerinnen und Schülern mit Zuwanderungshintergrund [Science competency in students with a migration background]. In K. Reiss, C. Salzer, A. Schiepe-Tiska, E. Klieme, & O. Köller (Eds.), *PISA 2015: Eine Studie zwischen Kontinuität und Innovation* (pp. 317–347). Waxmann.

Retelsdorf, J., Becker, M., Köller, O., & Möller, J. (2012). Reading development in a tracked school system: A longitudinal study over 3 years using propensity score matching. *British Journal of Educational Psychology, 82*(4), 647–671. https://doi.org/10.1111/j.2044-8279.2011.02051.x

Retelsdorf, J., Butler, R., Streblow, L., & Schiefele, U. (2010). Teachers' goal orientations for teaching: Associations with instructional practices, interest in teaching, and burnout. *Learning and Instruction, 20*(1), 30–46. https://doi.org/10.1016/j.learn instruc.2009.01.001

Richards, Z., & Hewstone, M. (2001). Subtyping and subgrouping: Processes for the prevention and promotion of stereotype change. *Personality and Social Psychology Review, 5*(1), 52–73. https://doi.org/10.1207/S15327957PSPR0501_4

Sprietsma, M. (2013). Discrimination in grading: Experimental evidence from primary school teachers. *Empirical Economics, 45*(1), 523–538. https://doi.org/10.1007/s00181-012-0609-x

Steinmayr, R., & Spinath, B. (2009). What explains boys' stronger confidence in their intelligence? *Sex Roles, 61*(9–10), 736–749. https://doi.org/10.1007/s11199-009-9675-8

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*(4), 731–752. https://doi.org/10.1007/s11218-017-9392-z

Van Houtte, M. (2004). Tracking effects on school achievement: A quantitative explanation in terms of the academic culture of school staff. *American Journal of Education, 110*(4), 354–388. https://doi.org/10.1086/422790

Weber, L., & Parra-Medina, D. (2003). Intersectionality and women's health: Charting a path to eliminating health disparities. In M. Texler Segal, V. Demos, & J. J. Kronenfeld (Eds.), *Gender perspectives on health and medicine* (Advances in Gender Research, Volume 7) (pp. 181–230). Emerald Group Publishing.

Wenz, S. E., & Hoenig, K. (2020). Ethnic and social class discrimination in education: Experimental evidence from Germany. *Research in Social Stratification and Mobility, 65*, 100461. https://doi.org/10.1016/j.rssm.2019.100461

Will, A. K. (2019). The German statistical category "migration background": Historical roots, revisions and shortcomings. *Ethnicities, 19*(3), 535–557. https://doi.org/10.1177/1468796819833437

ORCID
Ursula Kessels
https://orcid.org/0000-0002-1764-9442

Prof. Dr. Ursula Kessels
Fachbereich Erziehungswissenschaft und Psychologie
Freie Universität Berlin
Habelschwerdter Allee 45
14195 Berlin
Germany
ursula.kessels@fu-berlin.de

Appendix A

Table A1. Complete list of example behaviors in the behavioral checklist (from Heyder and Kessels (2015): Behavior impeding learning

- spends lessons counting the minutes left till the end of class using a tally sheet
- looks up the solutions at the end of the book before completing the tasks by themselves
- does not file worksheets and loses them
- never reads books
- secretly reads comics under their school desk during lessons
- does not pay attention in class
- copies homework
- sends text messages during lessons
- spends about 6 hours per day watching TV
- forgets their school supplies at home