Teaching and Teacher Education 123 (2023) 103985

Contents lists available at ScienceDirect

Teaching and Teacher Education

journal homepage: www.elsevier.com/locate/tate

Review article

Are tracking recommendations biased? A review of teachers' role in the creation of inequalities in tracking decisions☆

Anatolia Batruch [a], Sara Geven [b, *], Emma Kessenich [c], Herman G. van de Werfhorst [c]

[a] Laboratoire de Psychologie Sociale de l'Université de Lausanne (UnilaPS), Université de Lausanne, Batiment Géopolis, Office: 5136 UNIL-Mouline, Chavannes-près-Renens, 1015, Lausanne, Switzerland
[b] Department of Sociology, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV, Amsterdam, Netherlands
[c] Department of Political and Social Sciences, European University Institute, Florence, Villa Sanfelice, Via Dei Roccettini, 3, 50014, San Domenico di Fiesole, Florence, Italy

HIGHLIGHTS

- Accounting for performance, track recommendations are biased against students from lower SES backgrounds.
- Evidence for ethnic biases in teacher tracking recommendation is more mixed.
- Student, teacher and parental characteristics affect tracking recommendations but cannot explain the biases.
- To combat the biases, research should focus on institutional and situational moderators of the biases.

ARTICLE INFO

Article history:
Received 17 June 2022
Received in revised form
13 October 2022
Accepted 8 December 2022
Available online 31 December 2022

Keywords:
Teachers
Educational inequality
Ability tracking
Teacher expectations
Track recommendations

ABSTRACT

Sorting students on the basis of their academic performance into hierarchically ordered curriculums (i.e., between-school tracking) is common practice in various educational systems. International studies show that this form of tracking is associated with increased educational inequalities. As track placement is often based on teacher recommendations, biased track recommendations may contribute to this inequality. To shed light on the role that teachers play in the reproduction of inequalities in school, we conducted a systematic review of 27 recent articles on teachers' between-school tracking recommendations and students' socio-economic or ethnic background. We find that teacher recommendations are biased against students from disadvantaged socio-economic backgrounds, yet evidence with respect to ethnic biases is more mixed. While student, parent, teacher, and contextual factors seem to play a role in tracking recommendations, they cannot account for the biases in tracking recommendations. We discuss promising areas for future studies and argue that research on institutional moderators may have more potential than research on psychological mediators to effectively reduce bias in educational institutions.

☆ Notes: Anatolia Batruch and Sara Geven contributed equally to the manuscript and share first authorship. This study was supported by the Dutch Research Council (NWO) through a Veni grant awarded to Sara Geven (#016.Veni.195.125) and a Vici grant awarded to Herman van de Werfhorst (#453-14-017). We would like to thank the reviewers for their helpful suggestions. Some sections in this review study are based on an extended narrative review study commissioned by the Dutch Ministry of Education, Culture and Sciences (see https://pure.uva.nl/ws/files/33184714/blg_864911.pdf). However, these sections were updated, rewritten, and edited prior to submission.

* Corresponding author.
E-mail addresses: anatolia.batruch@unil.ch (A. Batruch), s.a.j.geven@uva.nl (S. Geven), kes.emma@googlemail.com (E. Kessenich), herman.vandewerfhorst@eui.eu (H.G. van de Werfhorst).

https://doi.org/10.1016/j.tate.2022.103985

Contents

1. Introduction

In most educational systems, students are sorted into different educational programs on the basis of their academic performance. The objective of this so-called tracking is to enhance the efficiency of education, as it enables schools and teachers to tailor the pace and content of the educational material to fit students' academic needs (Van de Werfhorst & Mijs, 2010). Whereas in the United States students tend to be tracked for specific courses (i.e., within-school tracking), in multiple continental European countries, students are tracked into entirely different schools or classrooms for their full curriculum (i.e., between-school tracking) (Chmielewski, 2014). Because the consequences of tracking policies can be extensive for both students and societies (e.g., labor market structure), the effects of tracking have garnered the attention of many researchers in the social and educational sciences (Gamoran, 2009).

1.1. Inequality in tracking

One of the most prolific research areas in tracking has focused on its impact on the reproduction of social inequalities. Various studies suggest that educational systems characterized by higher levels of between-school tracking are marked by stronger socio-economic and ethnic inequalities in achievement and attainment (Brunello & Checchi, 2007; Terrin & Triventi, 2022; Van de Werfhorst & Mijs, 2010). Specifically, students from disadvantaged backgrounds consistently attend lower ability tracks (Organization for Economic Co-operation and Development [OECD], 2011).

Traditionally, educational inequalities in tracking have often been explained by the lower ability levels of disadvantaged groups, as well as their choice for less ambitious educational pathways (Kloosterman, Ruiter, De Graaf, & Kraaykamp, 2009; Van de Werfhorst & Hofstede, 2007). However, educational inequalities in tracking could potentially have more systemic causes (Dumont, Klinge, & Maaz, 2019; Esser, 2016). In many between-school tracking countries, teachers' track recommendations play an important role in students' track placements (Boone & Van Houtte, 2013; Dumont et al., 2019; Pietsch & Stubbe, 2007), and recently, researchers have started to consider biases in these recommendations as a third potential mechanism underlying inequalities in educational attainment and achievement in educational systems characterized by between-school tracking (Dumont et al., 2019; Esser, 2016).

1.2. Inaccuracy of teachers' judgements

Although there are no systematic reviews that directly examine biases in (between-school) tracking recommendations (Wenz & Hoenig, 2020), there are systematic review and meta-analytic studies on teacher judgements and expectations, which are arguably related to tracking recommendations. These studies demonstrate that teachers' judgements and expectations are correlated with students' academic abilities, yet a significant part of the variance remains unexplained. For example, a meta-analysis on 73 studies shows an overall mean effect size of 0.63 between teacher judgements and student achievement at standardized test scores (Südkamp, Kaiser, & Möller, 2012). This implies that teacher judgements do not perfectly resemble student performance, and more than 40% of the variance may be attributed to teachers'

reliance on non-academic information. A more recent review study concludes that the accuracy of teacher judgements tends to be quite high with respect to students' academic achievement and intelligence, yet is substantially lower for outcomes that can indirectly contribute to students' educational success, such as judgements about student creativity, memory, meta-cognition, social skills and learning motivation (Urhahne & Wijnia, 2020). This can have important implications for biases in track recommendations, as teachers often base their recommendations on more than academic achievement and intelligence alone (Boone & Van Houtte, 2013; Vanlommel & Schildkamp, 2019). Moreover, another recent review study finds that teacher expectations are related to students' socio-economic status (Wang, Rubie-Davies, & Meissel, 2018). This finding is corroborated by a narrative review that also specifically considers teacher track recommendations (Geven, Batruch, & van de Werfhorst, 2018). Together these findings suggest that in countries in which teachers' judgements and expectations are used to make tracking decisions, teachers may contribute to the creation of educational inequalities by making systematic errors for or against students' belonging to specific socio-economic and/or ethnic/migrant groups.

1.3. Present study

Despite its interest to social and educational scientists, and a vast number of individual studies on the topic, no study has yet systematically consolidated the findings on socio-economic and ethnic biases in teacher tracking recommendations. To gain an overview of the existence and sources of socio-economic and ethnic biases in teacher track recommendations, we conduct a systematic review in which we aim to answer the following two research questions: 1) Do students' social and/or ethnic background impact teachers' between-school track recommendations, and 2) Which factors can potentially account for social and/or ethnic biases in these tracking recommendations? Track recommendations refer to a teacher's placement of students into ability groups or programs. We define biases as systematic discrepancies in teachers' recommendations for equally performing students from different socio-economic and/or ethnic backgrounds (Axt & Lai, 2019). In other words, bias occurs when teachers do not formulate similar tracking recommendation for same-ability (i.e., comparable grades/standardized test scores) students from different social groups.

To examine how a student's social and/or ethnic or migration background impact teachers' between-school tracking recommendations, we conduct a systematic review of 27 recent (from 2000 onward) observational and experimental quantitative studies. To shed light on the sources of the potential social and/or ethnic biases in track recommendations, we review factors that are included in the reviewed studies that could potentially explain the biases. We distinguish three types of factors: individual student and parent-related characteristics, teacher-related characteristics, and contextual factors.

We opt for a systematic review rather than a meta-analysis, because of the high heterogeneity of the literature in terms of: design, data, context, methods, and the type and number of ability tracks as well as SES and ethnic/migrant groups that are being distinguished and studied. Given the high numbers of moderators and the rather low number of studies meeting the inclusion criteria (27), we decided that a meta-analysis would not be the most appropriate choice of analysis. Moreover, various studies did not provide (sufficient) details to discern comparable (standardized) effect sizes (e.g., standard deviations of the crucial dependent and independent variables).

2. Methods

2.1. Search and selection procedure

We conducted a systematic literature search using the Web of Science database as well as ERIC and APA PsycInfo in ORCID. These databases were chosen for their advanced search options and their coverage of studies from a wide range of disciplines including psychology, educational sciences, sociology, and economics, as our topic of interest combines insights from different fields. We used a concatenated search string to trawl through the abstracts of all available documents that were published since January 1, 2000 and that were written in English: AB=((“teacher* recommend*") OR (“school*placement recommend*") OR (track* AND teacher*) OR (allocat* AND teacher*) OR (“school*placement” AND teacher*) OR (“teacher* expect*") OR (“teacher* bias*") OR (“teacher* judg*")). The concatenation was necessary as the terms used by scholars to refer to track recommendations vary widely.

We restrict ourselves to studies published from 2000 onwards, as there have been considerable educational reforms in tracking institutions in various European countries between the 1950s and the 1990s. For example, multiple countries (e.g., England, Finland, France and Sweden) moved from an early-tracking system to a comprehensive system (Van de Werfhorst, 2019). Conversely, some Eastern and Central European countries changed to a more between-school tracking system after the fall of communism. By restricting our review to papers published from 2000 onwards, we focus on the literature that studies track recommendation in clearly solidified systems of education.

In addition to selecting studies written in English since 2000, we used the following in- and exclusion criteria:

1) The use of empirical data (i.e., no reviews or meta-analyses);
2) The analysis of between-school track recommendations (i.e., no within-school track recommendations for, for example, advanced mathematics or gifted programmes; no analysis of student's or parent's track choice rather than teachers' recommendation);
3) The inclusion of a measure of a student's socio-economic status (SES) and/or ethnic/migration background, measured at the student level[1];
4) The inclusion of academic performance in terms of standardized test scores or grades, either as an experimentally controlled condition or an observational measure;
5) The complete reporting of results.[2]

The Web of Science search was conducted in June 2020 and yielded a total of 2771 results. All results were saved and screened in a two-step process, using the Rayyan systematic review software. In the first screening phase, all studies were categorized as either ‘irrelevant’ or ‘potentially relevant’, based on whether the titles and, where necessary, abstracts fitted our inclusion criteria. In this stage, studies were only excluded if they were clearly irrelevant content-wise (e.g., studies that matched the search criteria because they employed an eye-tracking methodology instead of being on school tracking). This was followed by a second screening phase, in which those 422 studies identified as being ‘potentially relevant’ were coded inductively, based on the abstract and, if necessary, the methods section. In this second phase, qualitative studies were

[1] One study was excluded as student SES was measured at the neighborhood level.

[2] When this was not the case, the authors of the study were contacted. Studies were only excluded if, even upon request, the reporting of results was insufficient.

excluded (N = 71) as well as quantitative studies that did not consider track recommendations as one of their dependent variables (N = 309). The remaining 42 articles' full texts were evaluated by three of the four authors. Twenty of these studies were considered to meet the selection criteria.

To get a more comprehensive list of studies on teacher track recommendations, we conducted a second search in ERIC and APA PsychInfo in January 2022, using the same concatenated search string and inclusion criteria. After removing the articles that had already been included in our first search, we ended up with 364 additional studies. 319 of these studies were excluded on the basis of the titles and abstracts by two of the authors. The remaining 45 full texts were evaluated and one of them was found to meet the inclusion criteria.

Finally, we expanded our selection of studies by inspecting the full reference lists of the 21 studies included. After eliminating duplicates, we followed the same procedure to identify studies that used track recommendations as a dependent variable, were written in English, and published in 2000 or later. This led to an additional 37 studies that were assessed against all the inclusion criteria on the basis of their full texts. Six of these studies met the criteria and were included. In total, the review involves 27 studies analysing between-school track recommendation, published between 2007 and 2020.

In Table 1 we present an overview of the included studies. Out of the 27 studies, 17 present findings based on observational data, nine use experimental data, and one study relies on both. Eight studies are set in Germany, seven in the Netherlands, four in Luxembourg, two in France, two in Belgium, one in Hungary, one in Switzerland, and two in both Germany and Luxembourg. Thus, all included studies are from European countries, which is a likely consequence of our focus on between-school ability tracking. Studies also differ in how they account for student performance. Of the nine experimental studies, two manipulate information on student grades, two on test scores, four on test scores and grades, and one manipulates student performance by using the same essay for different student profiles. Of the 17 observational studies, twelve account for test scores, three account for student grades, and two for student grades and test scores. It is important to consider the measurement of student performance, especially for observational studies, because teacher-assigned grades may already subsume teacher biases. That is, teachers could assign different grades to similarly performing students from different backgrounds (Autin, Batruch, & Butera, 2019).

3. Empirical findings on biases in teacher track recommendation

3.1. Socio-economic status of students

Overall, findings on tracking recommendations are consistent with respect to student SES. Of the 19 studies that report findings on SES biases, 13 show that teachers provide higher tracking recommendation for students from high-SES backgrounds than for equally performing students from low-SES backgrounds.[3] For example, an observational study on the track recommendations for more than 11,000 French students shows that SES disparities remain after accounting for students' school marks and repeated school years (Barg, 2013). An observational study on a sample of 500 Dutch primary school teachers also finds little teacher heterogeneity in the SES bias in teacher track recommendations: accounting for student performance on standardized academic tests, *all* teachers give lower track recommendations to students from lower socioeconomic backgrounds (Timmermans, Kuyper, & Werf, 2015). Using information on nine Dutch cohorts between 1995 and 2014, these researchers also show that SES biases in teacher track recommendation have remained stable over time (Timmermans, de Boer, Amsing, & van der Werf, 2018). Finally, two experiments in Switzerland reveal that teachers (and students playing the role of teachers) find the academic track more suitable and the vocational track less suitable for high-SES students whose school performance is slightly below official standards than for low-SES students with the same performance levels (Batruch, Autin, Bataillard, & Butera, 2019).

Three of the 19 studies that include SES as a predictor of teacher track recommendations find SES biases for some, but not all indicators of SES. More specifically, the observational studies by Barg (2015), Boone and Van Houtte (2013), and Feron, Schils, and Ter Weel (2016) find (some) support for SES biases when using occupational-related measures of SES, but not when using educational measures. Of these studies, Barg (2015) and Boone and Van Houtte (2013) measure student performance by teacher-assigned grades.

Another observational study using parental occupational status as a measure of SES also finds no significant SES bias in teacher track recommendation when analyzing a sample of 374 grade-4 students in Southern Germany (Niklas & Schneider, 2017). However, aside from students' test performance, the authors also account for parental reports on a student's home learning environment. When the authors do not account for this, students from higher SES backgrounds are more likely to be recommended to the highest track.

Only two of the 19 studies that report on SES biases in teacher track recommendation find no support for such biases. The first study involves two within-participant experiments with respectively 54 and 60 primary school teachers in Luxembourg (Glock, Krolak-Schwerdt, Klapproth, & Böhmer, 2012); and the second study relies on observational data of 2731 6th graders in Luxembourg (Klapproth, Glock, Böhmer, Krolak-Schwerdt, & Martin, 2012). In both these studies SES is measured by parental occupational status. The observational study accounts for student performance by including both grades and test scores.

A vast majority of the studies thus find support for SES biases in teacher track recommendations. Exceptions are potentially related to the measurement of SES (i.e., using parental education versus parental occupation) or the educational context in which the study is set. Interestingly, the two studies that did not find SES biases were both conducted in Luxembourg. Like most other contexts included in the review (i.e., Belgium, Germany, the Netherlands and Switzerland), Luxembourgish students are tracked at a relatively young age (age 13). However, in Luxembourg, it is a council that decides about a student's track recommendation (Glock et al., 2012). This council includes the primary school teacher(s), but also the school inspector and secondary school teachers who may not know the students. This could cause the home and/or socioeconomic situation of the student to play a less central role in tracking decisions, and may even lead primary school teachers to focus less on these factors when shaping their own recommendation for students.

While most studies find SES biases in teacher track recommendations, it is important to note that some of the effects are small in magnitude (e.g., Driessen, Sleegers, & Smit, 2008;

[3] This includes the study by Dumont et al. (2019). They find SES disparities in teacher track recommendations in Germany after accounting for student test scores, yet these disparities disappear after also accounting for teacher-assigned grades. However, teacher biases may also affect teacher-assigned grades, such that teacher biases in tracking recommendations (partly) reflect teacher biases in grading.

Table 1
Overview of included studies.

	Meta data				Variables					Effect, accounting for student performance[a]	
#	Author(s), year	Study design[b]	Country	Sample	Relevant dependent variable(s)	SES	Ethnic/Migration background	Student performance	Explanatory factors included[c]	SES	Ethnic/Migration background
1	Barg, 2015	Obs.: panel; multi-nomial logistic regression, SE adjusted for school-level clustering	FR	11,623 grade-9 students	Retention, general or vocational track recommendation	Parental occupation, 6 categories; parental educational attainment, 3 categories	Parents' countries of birth (with/without migration background)	GPA of school grades from grades 8 & 9	Stud. n.i. Fam. + TS n.i. Prej. n.i. %Ab n.i. %Ses n.i. %Eth n.i. Inst n.i.	+ parental occupation; n.s. parental education 16 perc. points dif btw EGP I and working class for general vs vocational track	NA
2	Barg, 2013	Obs.: panel; logistic regression, SE adjusted for school-level clustering	FR	11,667 students in lower secondary school	General or vocational track recommendation	Parental occupation, 5 categories	Parents' nationality and country of birth (French, European migrant, Non-European migrant, Mixed)	Average of school grades from grades 8 & 9	Stud. n.i. Fam. + TS n.i. Prej. n.i. %Ab n.i. %Ses n.i %Eth n.i. Inst n.i.	+ y-std dif btw EGP I and EGP V-VII = 0.58; y-std dif btw EGP IV and EGP V-VII = 0.17	+ y-std dif btw students without and with migration bg = 0.10 −0.17
3	Batruch et al., 2019	Exp.: case vignettes	CH	Study 1 : 99 university students, Study 2 : 70 pre- and in-service teachers; Study 3 : 160 university students	Teachers' rating on a 7-point scale which school track (lower, higher) is most suitable for a student	Stereotypical high- vs low-class name; parental occupation; extra-curricular activities	NA	School grades	Stud. i.s. Fam. n.i. TS n.i. Prej. n.i. %Ab n.i. %Ses n.i. %Eth n.i. Inst. +[b]	+ Cohen's d of 0.6 for low track suitability; 0.5 for high track suitabiltiy	NA
4	Boone & Van Houtte, 2013	Obs.: cross-sectional, logistic regression	Flanders, BE	1339 grade-6 students, in 53 primary schools; due to missing data analyses on 544 cases	Practically or theoretically oriented track recommendation	Parental occupation, 8-point scale recoded into 4 categories; maternal educational attainment, 3 categories	Two groups based on country of birth of maternal grandmother (Belgian or West European origin/Non-West European origin)	GPA of school grades from grade 5	Stud. n.i. Fam. n.i. TS n.i. Prej. n.i. %Ab n.i %Ses n.i. %Eth n.i. Inst n.i.	+ parental occupation; n.s. education mother Odds working class 83%, odds middle class 64% lower than upper middle class	n.s.
5	Boone et al., 2018	Obs.: cross-sectional, logistic multi-level (students nested in classes)	Antwerp & Ghent, BE	1049 grade-6 students, in 36 primary schools	Academic or practical track recommendation; Latin or science track recommendation	Parental occupation, ISEI	Maternal grandmother's place of birth (with/without migration background)	Standardized Raven test score	Stud. n.i. Fam. n.i. TS n.i. Prej. n.i. %Ab - %Ses n.s. %Eth n.s. Inst n.i.	+ 28 percentage points dif in probability for academic track btw 25th and 75th SES percentile	− 14 percentage points difference in probability for academic track
6	Bruneau et al., 2020	Exp.: case vignette	HU	Final sample 29 (study 1) and 161 (study 2) pre-service teachers	Teachers' rating (0 −100) of appropriateness of three different tracks per student (high, middle, low)	NA	Stereotypical Roma vs non-Roma name	Student scores on two competence tests and five subject tests. In study 1a Roma and non-Roma profiles had similar means, in study 1 b Roma and non-Roma profiles had identical means across subject tests, and also for tests	Stud. n.i. Fam. n.i. TS n.i. Prej. +[b], n.s. %Ab n.i. %Ses n.i. %Eth n.i. Inst n.i.	NA	n.s./- Cohen's d for low track suitability 0.26, and for high track suitability 0.37

(*continued on next page*)

Table 1 (*continued*)

	Meta data				Variables					Effect, accounting for student performance[a]	
#	Author(s), year	Study design[b]	Country	Sample	Relevant dependent variable(s)	SES	Ethnic/Migration background	Student performance	Explanatory factors included[c]	SES	Ethnic/Migration background
								in math, physics history, Hungarian and English			
7	Caro et al., 2009	Obs.: panel, logistic multi-level (students nested in classes)	Berlin, DE	2242 students (grade-4 onwards), in classes	Academic or non-academic track recommendation	Parental educational level; parental vocational training; parental occupation, ISEI; composite SES measure	Citizenship, mother tongue, language spoken at home, countries of birth of students and parents (native German, German with migration background, foreign)	Standardized test scores (basic cognitive skills tests); math test achievement level; growth in math test achievement	Stud. n.i. Fam. n.i. TS n.i. Prej. n.i. %Ab - %Ses n.i. %Eth + Inst n.i.	+ 1-SD increase in SES relates to 1.8 increase in probability for academic track	+ Probability for academic track 1.65 (with migration bg) and 1.24 (foreign students) higher
8	De Boer et al., 2010	Obs.: panel, multi-level regression (students nested in schools)	NL	11,040 students (grade-7 onwards), in 112 secondary schools;	Teacher bias in track recommendation, i.e. degree to which track recommendation reflects prior test scores	Parental education, 7-point scale	Parents' country of birth (with/without migration background)	School-leaving test score; IQ score	Stud. + Fam. + TS n.i. Prej. n.i. %Ab n.i. %Ses n.i. %Eth n.i. Inst n.i.	+ β of 0.054	n.s.
9	Driessen et al., 2008	Obs.: cross-sectional, multi-level regression (students nested in classes)	NL	7883 grade-6 students, in 519 classes	Teachers' track recommendation, 5 categories	Parental educational level, 4-point scale	Six groups, unclear on which information this is based (Native Dutch, mixed, Surinamese, Turkish, Moroccan, other)	Standardized test scores	Stud. +, - Fam. + TS n.i. Prej. n.i. %Ab - %Ses n.s. %Eth n.s. Inst n.i.	+ One unit increase in parental education relates to 0.1 increase in recommendation (SD recommendation = 1.2, SD parental education not reported)	+ for Turkish, Moroccan, Other; n.s. for Mixed, Surinamese/ Antillean Dif native Dutch and Turkish (0.1), Moroccan (0.12), other (0.21) (SD recommendation = 1.2)
10	Dumont et al., 2019	Obs.: cross-sectional, logistic regression (KHB), clustered SEs	Berlin, DE	3935 grade-6 students, in 87 primary schools;	Recommendation to highest track or not	Parental education, binary; parental occupation, ISEI	Parents' country of birth (with/without migration background)	Standardized test scores in reading and math; weighted average of students' grades	Stud. n.i. Fam. + TS n.i. Prej. n.i. %Ab n.i. %Ses n.i. %Eth n.i. Inst n.i.	+/n.s. after also accounting for grades APE parental occu 2 perc. points, APE parental edu 7 perc. points (z-standardized predictors)	+ APE migrant bg 3 perc. points
11	Feron et al., 2016	Obs.: panel, ordered probit regression	NL	4500 children (grade-6 onwards) in one province (Limburg)	Teacher bias in track recommendation, i.e. the degree to which teachers' track recommendations reflect prior test score	Parental education, 4 categories; employment status (employed, unemployed, sick, other);	Region of birth for child and parents (Limburg, NL, abroad)	School-leaving test score	Stud. n.i. Fam. n.i. TS n.i. Prej. n.i. %Ab n.i. %Ses n.i. %Eth n.i. Inst n.i.	n.s. parental education; +/n.s. employment status 0.09 lower probability for a recommendation > test score for students with unemployed or sick (v.s. employed) mother	n.s.
12	Glock et al., 2015	Exp.: case vignettes	DE, LU	48 pre-service teachers, 16 primary school teachers	Discrepancy between track recommendation and expected recommendation based on student profile, ranging from 0 (no discrepancy) to 5	Vignette contained information about parents' occupations	Nationality parents (majority, Luxembourgish/ minority, Portugese)	School grades; standardized test scores	Stud. i.s. Fam. i.s. TS n.i. Prej. n.i. %Ab n.i. %Ses n.i. %Eth n.i. Inst n.i.	NA	?

13	Glock et al., 2013	Exp.: case vignettes	LU	Study 1 : 54 primary school teachers; Study 2 : 60 primary school teachers	Recommendation to highest track or not; ratings of the probability of successful attendance in the highest track	Vignette contained information about parents' SES	Name signaling ethnic background and language spoken at home (majority, Luxembourgish/ minority, Portuguese)	School grades; standardized test scores	Stud. Fam. TS Prej. %Ab %Ses %Eth Inst	i.s. n.i. n.i. n.i. n.i. n.i. n.i. n.i.	NA	– Partial eta squared Recommendation: 0.23; partial eta squared success prob.: 0.55 (study 1) - 0.34 (study 2)
14	Glock et al., 2012	Exp.: case vignettes	LU	Study 1 : 54 primary school teachers; Study 2 : 60 primary school teachers	Recommendation to highest track or not	Father's occupation, binary	Language spoken at home (with/without migration background)	School grades; standardized test scores	Stud. Fam. TS Prej. %Ab %Ses %Eth Inst	i.s. n.i. n.i. n.i. n.i. n.i. n.i. +[b], n.s.	n.s.	-/n.s. Probability highest track 3 times higher for student without migration bg (study 1, low accountability condition)
15	Klapproth et al., 2012	Obs.: cross-sectional, multi-level regression (students nested in classes)	LU	2731 grade-6 students, in 211 classes in 104 primary schools	Academic or vocational track recommendation	Parental occupation, ISEI	Students' nationality (Luxembourgish/ Portuguese/Other foreign)	School grades in French, German Mathematics; standardized test score in French German, Mathematics in grade 6	Stud. Fam. TS Prej. %Ab %Ses %Eth Inst	+, n.s. + n.i. n.i. n.i. n.i. n.i. n.i.	n.s.	- for Portugese; n.s. for other foreign odds for academic track 27% lower for Portuguese
16	Klapproth et al., 2018	Exp.: case vignettes	DE	72 pre-service teachers	Recommendation to highest track or not	NA	Stereotypical German or Turkish names; religion (Muslim versus Christian) (Study 1)	School grades (per-subject and GPA)	Stud. Fam. TS Prej. %Ab %Ses %Eth Inst	+, i.s. n.i. n.i. n.i. n.i. n.i. n.i. n.i	NA	– Odds for high track 1.86 (exp 1) and 1.79 (exp 2) higher for German student
17	Klapproth & Fischer, 2020	Exp.: case vignettes	DE	102 primary school teachers	Recommendation to highest track or not	NA	Stereotypical male German or Turkish names	School grades (for unknown subjects)	Stud. Fam. TS Prej. %Ab %Ses %Eth Inst	i.s. n.i. n.i. n.i. n.i. n.i. n.i. n.i.	NA	Main effect: n.s. Interaction effect: Ethnicity X grades X development
18	Krolak-Schwerdt et al., 2018	Mixed: Exp. and Obs., ordinal regression (with adjusted and un-adjusted SE for class-level clustering)	DE (Study 1), LU (Study 2)	Study 1: 56 teachers and their students; Study 2: 54 teachers and grade-6 students from 199 classes	Teachers' track recommendation, 3 and 2 categories	NA	Parental and student country of birth (Obs, study 1) and info on nationality (exp, study 1); student nationality (exp and obs, study 2) (with/without migration background)	Grades of main subjects (Study 1 & 2); Standardized achievement test scores (Study 2 only);	Stud. Fam. TS Prej. %Ab %Ses %Eth Inst	+ + n.i n.i. n.i. n.i. n.i. n.i.	NA	-, n.s. Study 1: Odds for higher track 1.5 times higher for majority student; Study 2: Odds for highest track 2.3–2.4 higher for majority student
19	Lüdemann & Schwerdt, 2013	Obs.: cross-sectional, multi-nomial regression adjusted SE for class-level clustering	DE	3436 grade-4 students (West-Germany; states that track at age 10)	Track recommendation, 3 categories	Number of books at home, 5 categories; household income, 6 categories; highest parental educational	Parents' countries of birth (with/without migration background)	Reading and mathematics test performances; standardized scores on subscales of cognitive ability test (IQ)	Stud. Fam. TS Prej. %Ab %Ses %Eth Inst	n.i. n.i. n.i. n.i. n.i. n.i. n.i. n.i.	+ Effect size not reported	- (males before accounting for SES)/n.s. Immigrant males 6.3 perc. points more likely to receive general track and 5.8 perc. points less likely to receive high track

(*continued on next page*)

Table 1 (*continued*)

	Meta data				Variables					Effect, accounting for student performance[a]	
#	Author(s), year	Study design[b]	Country	Sample	Relevant dependent variable(s)	SES	Ethnic/Migration background	Student performance	Explanatory factors included[c]	SES	Ethnic/Migration background
						degree, 3 categories					
20	Niklas & Schneider, 2017	Obs.: panel, multi-nomial regression	DE (South)	374 children	Track recommendation, 3 categories	Parental occupation, occupational prestige	Child's and parents' countries of birth (with/without migration background)	Standardized test scores in reading, spelling and math; intelligence score	Stud. n.i. Fam. + TS n.i. Prej. n.i. %Ab n.i. %Ses n.i. %Eth n.i. Inst n.i.	n.s./+ (when home environment is not accounted for) Effect size when home environment is not accounted for is not reported	n.s.
21	Pietsch & Stubbe, 2007	Obs.: cross-sectional, logistic regression	DE	6763 grade-4 students	Recommendation to highest track or not	Parental occupation, 6 categories (EGP); family gross income, 6 categories	NA	Standardized test score	Stud. n.i. Fam. n.i. TS n.i. Prej. n.i. %Ab n.i. %Ses n.i. %Eth n.i. Inst n.i.	+ Likelihood highest track > likelihood not highest track when reading score EGP 1 student ≥551 and reading score EGP VII student ≥601 (score mean 500, SD; 100)	NA
22	Pit-ten Cate et al., 2016	Exp.: case vignettes, longitudinal	LU	38 primary school teachers	Accuracy (correct, incorrect) of teachers' track recommendation	Vignette contained information about parents' SES	Nationality (Luxembourgish vs minority)	School grades; standardized test scores	Stud. i.s. Fam. i.s. TS n.i. Prej. n.i. %Ab n.i. %Ses n.i. %Eth n.i. Inst +[b], n.s.	NA	?
23	Sprietsma, 2013	Exp.: essays with manipulated names	DE	88 primary school teachers (4th grade) from 58 schools	Track recommendation, 3 categories	NA	Student name signaling ethnic background (German/Turkish)	Performance kept stable, because essay was the same for different students	Stud. n.i. Fam. n.i. TS n.i. Prej. n.s. %Ab n.i. %Ses n.i. %Eth n.i. Inst n.i.	NA	-/n.s. 11% lower probability to receive highest track for Turkish students
24	Timmermans et al., 2016	Obs.: cross-sectional, multi-level regression (students nested in classes)	NL	5316 grade-6 students, in 469 classes	Teachers' track recommendation on interval scale, ranging from 0.5 to 5	Combination parental education (3 categories) and ethnicity (Dutch, Turkish,Moroccan, other)	See SES	School-leaving test score; additional standardized test scores	Stud. +, -, n.s. Fam. n.i TS n.s. Prej. n.i. %Ab n.i. %Ses n.i. Eth n.i. Inst n.i.	+ β low SES is btw −0.06 (Turkish/Moroccan) and −0.11 (Dutch)	n.s.
25	Timmermans et al., 2018	Obs.: repeated cross-sections, multi-level regression (students nested in cohorts nested in schools)	NL	Grade-6 students in 9 cohorts in primary schools (>250 schools and >5000 students per cohort)	Teachers' track recommendation on interval scale, ranging from 1 to 11	Parents' highest level of education, 4-point scale	Parents' countries of birth (native Dutch, mixed, migrant)	School-leaving test score	Stud. n.i. Fam. n.i. TS n.i. Prej. n.i. %Ab n.i. %Ses n.i. %Eth n.i. Inst n.i.	+ Unit increase in parental edu relates to 0.19 increase in recommendation (SD btw. 2.5 and 2.9)	+/n.s.

26	Timmermans et al., 2015	Obs.: cross-sectional, multi-level regression (students nested in teachers)	NL	7550 grade-6 students in 500 classes	Teachers' track recommendation on interval scale, ranging from 1 to 5	Combination parental education (3 categories) and ethnicity (Dutch, Turkish, Moroccan, other)	See SES	School-leaving test score; additional standardized test scores	Stud. Fam. TS Prej. %Ab %Ses %Eth Inst	n.i. n.i. n.i. n.i. + + n.i n.i	+ High (v.s. middle): 0.08 low (v.s middle): btw −0.06 and −0.11 (SD recommendation = 1.3)	n.s.
27	Timmermans et al., 2019	Obs.: cross-sectional, multi-level regression (students nested in classes)	NL	9881 grade-6 students, in 485 primary schools	Teachers' track recommendation on interval scale, ranging from 1.5 to 6	Parental education, 3 categories	Parents' country of birth (with/without migration background)	School-leaving test score; additional standardized test scores	Stud. Fam. TS Prej. %Ab %Ses %Eth Inst	n.i. n.i. +, n.s. n.i n.i. n.i. n.i. n.i.	+ β low (vs. middle): −0.104 β high (vs. middle): 0.071	+

a + a positive effect ($p < 0.05$) on receiving a high track recommendation, and/or a negative effect on receiving a low track recommendations for students from more advantaged SES backgrounds (SES) or students with a migration/ethnic minority background; - a negative effect ($p < 0.05$) on receiving a high track recommendation, and/or positive effect on receiving a low track recommendations for students from more advantaged SES backgrounds (SES) or students with a migration/ethnic minority background; n.s. effect is not statistically significant; ? unclear whether there is a negative or a positive bias towards students with a migration background (study use level of accuracy as a dependent variable). For statistically significant effects, we report on the reported (standardized) effect sizes, if available.
b Obs, refers to observational data, Exp. Refers to experimental study;
c Factors included in the quantitative analyses of teacher track recommendations:
Stud : student pro-school behaviors and/or attitudes,
Fam: family-related variables (e.g., supportiveness of the home environment, parental school involvement, educational preferences),
TS: teacher-student relationship quality
Prej. : teacher stereotypes and prejudices
%Ab: average ability of students in the class/school,
%Ses: average SES or share of high-SES students in the class/school
%Eth: share of ethnic minority students in the class/school,
Inst: tracking Institutions
+ a positive effect ($p<0.05$) on receiving a high track recommendation, and/or negative effect on receiving a low track recommendation,
$+^{b}$ a statistically significant effect in the hypothesized direction on biases/accuracy in track recommendations (e.g., interaction effect between the variable with student SES and/or ethnicity),
- a negative effect ($p<0.05$) on receiving a high track recommendation, and/or positive effect on receiving a low track recommendation
$-^{b}$ a statistically significant effect in the opposite direction as hypothesized on biases/accuracy in track recommendations (e.g., interaction effect between the variable with student SES and/or ethnicity);
i.s.: included as a stable factor in the experimental set-up;
n.i.: not included as a variable in the study;
n.s.: not statistically significant

Timmermans, Boer, & Werf, 2016). Moreover, a recent working paper warned for measurement error in standardized test scores, leading to a potential overestimation of the SES bias in teacher track recommendations in observational studies in the Netherlands (Van Huizen, 2021). More specifically, student scores in the final (end-of-school) standardized test may not accurately reflect students' learning skills. The SES bias in track recommendation would be overestimated if biases are examined by comparing teacher recommendations to student scores on this final test. It is important to mention that most of the included Dutch studies do not only account for the score on this final test, but also account for additional standardized achievement or intelligence test scores (i.e., De Boer, Bosker, & van der Werf, 2010; Driessen et al., 2008; Timmermans et al., 2015, 2016, 2019). This will likely attenuate the problem.

3.2. Ethnicity of students

Findings are less consistent when considering teacher biases with respect to a student's ethnic or migration background. Of the 24 studies that report results on teacher biases by ethnic or migration background, 9 find (some) evidence suggesting that track recommendations are biased *against* students with a migration or ethnic minority background (Boone, Thys, Van Avermaet, & Van Houtte, 2018; Bruneau, Szekeres, Kteily, Tropp, & Kende, 2020; Glock et al., 2012, 2013; Klapproth et al., 2012, 2018; Krolak-Schwerdt, Hörstermann, Glock, & Böhmer, 2018; Lüdemann & Schwerdt, 2013; Sprietsma, 2013), 6 find no support for biases (Boone & Van Houtte, 2013; De Boer et al., 2010; Feron et al., 2016; Niklas & Schneider, 2017; Timmermans et al., 2015, 2016), and another 6 find (some) support that teacher track recommendations are biased in *favor* of students with a migration or ethnic minority background (Barg, 2013; Caro, Lenkeit, Lehmann, & Schwippert, 2009; Driessen et al., 2008; Dumont et al., 2019; Timmermans et al., 2018, 2019). There are also two studies that show that teacher track recommendations are less accurate for students from migration backgrounds (Glock, Krolak-Schwerdt, & Pit-ten Cate, 2015; Pit-ten Cate, Krolak-Schwerdt, & Glock, 2016). This inaccuracy may both reflect biases in favor of, or against students with a migration or ethnic minority background (Glock et al., 2015; Pit-ten Cate et al., 2016). Finally, one study finds no main effect of student ethnic background, yet reports that teachers' reliance on student performance and performance development vary by student ethnic background (Klapproth & Fischer, 2020).

The inconsistent findings with respect to ethnic or migration background could partly be due to differences in research design. Given that most of the included studies in our review rely on observational data, there seems to be an overrepresentation of experimental research designs among the studies finding biases against students from minority groups. More specifically, five of the nine studies use an experimental design (Bruneau et al., 2020; Glock et al., 2012, 2013; Klapproth, Kärchner, & Glock, 2018; Sprietsma, 2013), one study uses both experimental and observational data (Krolak-Schwerdt et al., 2018), and three studies solely rely on observational data (Boone et al., 2018; Klapproth et al., 2012; Lüdemann & Schwerdt, 2013). It is important to note that student SES is not accounted for in three of these experiments (Bruneau et al., 2020; Klapproth et al., 2018; Krolak-Schwerdt et al., 2018; Sprietsma, 2013), one of these observational studies (Boone et al., 2018) as well as the study relying on both observational and experimental data (Krolak-Schwerdt et al., 2018). Moreover, another of the observational studies only finds support for ethnic biases against male students before, but not after, accounting for SES, the language spoken at home, and whether a student attended pre-primary education (Lüdemann & Schwerdt, 2013). Interestingly, several observational studies that control for both student performance and SES find that teachers evaluate students from disadvantaged minority groups more *positively* than students from the native majority (Barg, 2013; Caro et al., 2009; Driessen et al., 2008; Dumont et al., 2019; Timmermans et al., 2018, 2019).

Aside from design and the inclusion of student SES, experimental and observational studies also tend to vary in their measurement of migration background/ethnicity. While observational studies mostly rely on the country of birth of the student and/or (grand)parents (Barg, 2013, p. 201; Boone et al., 2018; Boone & Van Houtte, 2013; Caro et al., 2009; De Boer et al., 2010; Dumont et al., 2019; Feron et al., 2016; Krolak-Schwerdt et al., 2018; Lüdemann & Schwerdt, 2013; Niklas & Schneider, 2017; Timmermans et al., 2018), experimental studies often use indirect measures and try to signal a student's ethnicity or migration background by a student's name (Bruneau et al., 2020; Klapproth et al., 2018; Klapproth & Fischer, 2020; Sprietsma, 2013), language spoken at home (Glock et al., 2012) or both (Glock, Krolak-Schwerdt, Klapproth, & Böhmer, 2013). Only three experimental studies seem to directly display a student's nationality (Glock et al., 2015; Krolak-Schwerdt et al., 2018; Pit-ten Cate et al., 2016). While the explicit mentioning of a student's nationality could foster social desirability in experimental studies, research shows that ethnicity effects may be conflated with SES effects in studies that solely use names to signal a student's ethnic background (Wenz & Hoenig, 2020). This is because the names that are intended to signal an ethnic minority background may (also) signal a disadvantaged SES background.

Findings may also be inconsistent because teachers tend to sometimes over- and other times underestimate students from ethnic minority groups, leading to positive, negative, as well as null findings. Experimental research in Germany and Luxembourg indicates that track recommendations for students from ethnic minority groups are more likely to be 'inaccurate' than those for students from the ethnic majority (Glock et al., 2015; Pit-ten Cate et al., 2016)). Inaccuracies include track recommendations that are either higher or lower than expected on the basis of a student's academic profile. Possibly, teachers overestimate some, while underestimating other ethnic minority students. This may also be because ethnic minority students are a highly diverse group, with different ethnic origins (attached to different stereotypes) and social status.

Related to this, inconsistent findings may stem from large variations *across* teachers in how they evaluate students from ethnic minority groups as compared to students from the native majority. A Dutch study shows that some teachers tend to give *higher* track recommendations to Turkish, Moroccan, and other foreign students from low socioeconomic backgrounds, while others give *lower* track recommendations to these minority groups (Timmermans et al., 2015). These effects cancel each other out, leading to a non-significant overall effect of student migration background on teacher track recommendation.

Finally, differences in findings may be due to differences in study context. Included studies that report results on biases in track recommendations by ethnic or migration background are conducted in the Flanders part of Belgium (2), France (1), Germany (9), Hungary (1), Luxembourg (5), and the Netherlands (7); with studies finding support for biases against students from minority backgrounds being set in Flanders (1), Germany (4), Hungary (1), and Luxembourg (3). These countries vary in their ethnic composition and migration histories, political and societal climate with respect to minorities, as well as educational institutions. For example, the study on ethnic biases in teacher recommendations in Hungary involves biases against students with a Roma background, a group that is subject to blatant stereotypes (Bruneau et al., 2020).

Aside from the country in which the study is set, the year in which the data is collected may play a role. For example, one study

shows that in the Netherlands, ethnic biases in track recommendations have changed between 1995 and 2014 (Timmermans et al., 2018). In 1995, track recommendations were on average *higher* for students from ethnic minority groups than for equally performing ethnic majority students, however this difference reduced over time, and eventually disappeared. The authors note that this might be due to (1) equity-related policies gradually starting to focus less on ethnic inequalities in the Netherlands, and/or (2) Dutch society becoming less tolerant towards members of minority groups over this period.

The measurement of student performance does not seem to explain differences in study findings with respect to ethnic biases in tracking recommendations. For example, of the observational studies that find no support for ethnic biases, some measure student performance by grades (Boone & Van Houtte, 2013), while others only account for test scores (Timmermans et al., 2015)

4. Explaining inequality in track recommendations

So far our review shows that a majority of studies find biases in track recommendations against students from disadvantaged SES backgrounds, while suggesting that evidence is mixed with respect to biases by students' ethnic background. In this section, we review the papers included in the systematic review to shed light on the reasons *why* teacher tracking recommendations are biased. We only review articles included in the systematic review because our selection criteria (e.g., objective SES or ethnicity, control for grades) are also relevant for examining sources of bias. Based on the studies included in the review, we distinguish between three types of factors that may contribute to biases: student- and parent-related factors (e.g., students' school behavior and attitudes, and parental involvement or teachers' perception thereof), teacher-related factors (e.g., teacher prejudices and stereotypes) and contextual factors (e.g., school- or class-level composition or institutional characteristics).

4.1. Student- and parent-related factors (- or teachers' perception thereof)

4.1.1. Student school behavior and attitudes

Whether intentionally or not, teachers base their track recommendations on student behaviour and attitudes in class. Some scholars even argue that teacher track recommendations are accurate to the extent they are explained by either a student's performance, ability, or *motivation* or *effort* in school (e.g., De Boer et al., 2010). Of the 27 studies included in our systematic review, six studies report findings on the relation between students' school behaviour and attitudes and track recommendations (De Boer et al., 2010; Driessen et al., 2008; Glock et al., 2012; Klapproth et al., 2012, 2018; Krolak-Schwerdt et al., 2018; Timmermans et al., 2016). Moreover, some experimental studies account for student school behaviour and/or attitudes by including it as a stable factor in the experimental set-up (Batruch et al., 2019; Glock et al., 2013, 2015; Klapproth & Fischer, 2020; Pit-ten Cate et al., 2016). Finally, one study sheds light on the potential impact of student behavior and attitudes by using (1) an additional questionnaire asking teachers directly about the student characteristics they consider when forming track recommendations, and (2) supplementary qualitative focus groups with teachers on this topic (Boone & Van Houtte, 2013).

Most studies accounting for student behavior or attitudes include a measure for working habits, effort, or motivation, and find that this is positively related to teacher track recommendations. For example, research in Luxembourg and Germany indicates that teachers provide higher track recommendations when they have a more positive perception of a student's working behaviour (Krolak-Schwerdt et al., 2018). This relationship exists while controlling for a student's school performance (i.e., grades, and in Luxembourg also test scores) and nationality or migration background. Similarly, De Boer et al. (2010) find that when Dutch teachers evaluate students' achievement motivation more positively, they are more likely to recommend a track that exceeds the student's shown performance and achievement motivation (i.e., overestimate a student).

Two studies in the Netherlands (Driessen et al., 2008; Timmermans et al., 2016), and one study in Luxembourg (Klapproth et al., 2012), move beyond a single indicator for school behaviour and attitudes. The study in Luxembourg shows that students' reliability and accuracy are positively related to teacher track recommendations, yet find no statistically significant relationship between students' achievement/learning motivation and teacher track recommendations (Klapproth et al., 2012). The two Dutch studies reveal that teacher track recommendations are slightly higher for students who are perceived to be more self-confident and have better study attitudes or work habits (e.g., work hard, plan better), but surprisingly *lower* for students who are perceived to exhibit more socially accepted behaviors (e.g., stick to the class rules) (Driessen et al., 2008; Timmermans et al., 2016). Timmermans et al. (2016) also find that a teacher's perceptions of a student's popularity is not predictive of teacher track recommendations. Moreover, they show that teachers vary in the extent to which they take into account their perceptions of a student's self-confidence, work habits and social behavior. For example, some teachers weigh self-confidence and social behaviour positively, while others weigh these factors negatively. Finally, teachers seem to evaluate student performance by a student's class behaviour and attitudes. More specifically, performance has a stronger impact on a teacher's recommendation when a student is perceived to have more positive work habits.

Research in Flanders suggests that some teachers consciously rely on student behaviour and attitudes when forming track recommendations: in a questionnaire 69 percent of the 390 teachers explicitly reported to base track recommendations on student attitudes and behaviours (Boone & Van Houtte, 2013). Additional focus groups with a subset of seven teachers reveal that teachers for example take into account independence, planning capacity, responsibility, and punctuality.

Although teachers may (explicitly) rely on school behaviours and attitudes when formulating track recommendations, research also suggests that these factors play a minor role, and are far less important than cognitive competencies (Driessen et al., 2008; Timmermans et al., 2016). Nevertheless, teachers' reliance on students' school behaviour and attitudes could still, *in theory*, contribute to socio-economic and/or ethnic biases in tracking recommendations.

Timmermans et al. (2016) explicitly examine whether teachers' reliance on student behaviour and attitudes explain the higher track recommendations for students from advantaged SES backgrounds and find no support for this. Similarly, the study by Driessen et al. (2008) shows that the effect of parental education on track recommendation hardly changes after accounting for teachers' perceptions of student behaviour and attitudes. Relatedly, other studies that account for teachers' perceptions of student working behaviour or motivation, still find SES or ethnic gaps in track recommendations after accounting for these perceptions (De Boer et al., 2010; Krolak-Schwerdt et al., 2018).

SES or ethnic biases in track recommendations may also vary by a student's school attitudes or behaviour. Klapproth et al. (2018) expect that the ethnic gap in a teacher's track recommendations is contingent on student absenteeism. Ethnic minority students

who show high absence rates may confirm the stereotype of ethnic minority students as poor academic performers, leading to an activation of the ethnic stereotype, and higher ethnic discrepancies in track recommendations. To test this idea, the authors conduct a vignette experiment among 95 preservice teachers in Germany. In the vignette experiment, respondents are asked whether they are in favor of placing a hypothetical male student in the highest secondary school track. The GPA, ethnicity (i.e., Turkish or German), and absence rates of the hypothetical student are experimentally manipulated. The authors do not find clear support for their hypothesis. Overall, students with a higher GPA, a German background, and low absence rates are more likely to be assigned to the highest track. In line with the hypothesis, higher absence rates are related to a lower likelihood for a high track-recommendation for Turkish students with a high GPA than for German students with a high GPA. However, among students with a low or medium GPA, high absence rates only decrease the probability to be recommended to a high track for German, but not for Turkish students.

4.1.2. Parent-related factors

Nine of the 27 studies that are included in the systematic review report findings on how characteristics of the home environment, such as parental school support and aspirations (or teachers' perception thereof), are related to teacher track recommendations (Barg, 2013, 2015; Boone & Van Houtte, 2013; De Boer et al., 2010; Driessen et al., 2008; Dumont et al., 2019; Klapproth et al., 2012, 2018; Krolak-Schwerdt et al., 2018; Niklas & Schneider, 2017). Some experimental studies also account for characteristics of the home environment by including it as a stable factor in the experimental set-up (Glock et al., 2015; Pit-ten Cate et al., 2016).

Studies suggest that parental support and school involvement - or a teacher's perception thereof - relate to higher track recommendations. For example, a study in France shows a positive association between parents' involvement in parent associations and teacher track recommendations (Barg, 2013). Research among German primary school teachers indicates that teacher track recommendations are higher when teachers perceive parents to provide more support in problems that occur in school (Krolak-Schwerdt et al., 2018). Similarly, a Dutch study indicates that teachers provide higher track recommendations to students who live in homes that teachers perceive to be supportive (e.g., homes in which learning and curiosity are stimulated) (Driessen et al., 2008).

The fact that teachers seem to provide higher track recommendations to students when they perceive their parents to be more involved in school could partly explain SES and ethnic biases in tracking recommendations, as research suggests that teachers tend to hold more positive perceptions with respect to the school involvement of parents from advantaged backgrounds (Bakker, Denessen, & Brus-Laeven, 2007). However, Driessen et al. (2008) find that the positive relationship between parental SES and teacher tracking recommendation in the Netherlands hardly changes after accounting for teachers' perception of the home environment.

Parents from advantaged socioeconomic backgrounds may also be more likely to question the tracking decisions of the school or to successfully exert (implicit) 'pressure' on track recommendations. Studies find that parental aspirations and preferences for their child's educational attainment positively relate to track recommendations (De Boer et al., 2010; Dumont et al., 2019; Klapproth et al., 2012), yet De Boer et al. (2010) still find SES biases in track recommendations after accounting for parental aspirations.[4] The findings by Dumont et al. (2019) do suggest that the relatively high aspirations among parents with a migration background partly explain the higher track recommendations for these students. That is, after accounting for parental aspirations, Dumont et al. (2019) do not find a positive effect anymore of being a student with a migration background on track recommendations.

Research among around 11,000 students in France shows that SES differences in track recommendations are heavily reduced (or disappear) when accounting for families' school track requests (Barg, 2013, 2015). Barg (2013) suggests that schools may expect that (upper) middle class parents will object to relatively 'low' recommendations, and try to avoid such objections by giving higher track recommendations to students from (upper) middle class backgrounds (Barg, 2013, 2015). It should be noted that, in France, schools explicitly take into account parental wishes in their track recommendations (Barg, 2013). In a first stage, parents are asked to request a track for their child; subsequently, the school staff recommends a track; and, finally, the family can reject the recommendation.

In sum, findings are mixed with respect to the role of (teachers' perception of) family-related factors in biases in tracking recommendations. In analyses that account for teacher perceptions of parental school support or involvement, SES disparities in teacher track recommendations remain, indicating that this cannot explain SES biases in track recommendations. Some studies do indicate that parents from advantaged socioeconomic backgrounds exert more implicit pressure on teachers, and/or are more likely to requests for higher tracks, leading to biases in track recommendations. However, current findings are still inconclusive.

4.2. Teacher-related factors

4.2.1. Student-teacher relationship

Two studies with large samples ($N_{Study\ 1} = 5316$; $N_{Study\ 2} = 9881$) investigate whether the student-teacher relationship as perceived by the teacher could be a factor that explains biased tracking recommendations in the Netherlands (Timmermans et al., 2016, 2019). The first study finds no effect of teacher-student relationship quality on tracking recommendations this, and the effects of SES and ethnicity remain the same after accounting for relationship quality (Timmermans et al., 2016). The second study examines three separate dimensions of student-teacher relationship: closeness, conflict, and dependency (Timmermans et al., 2019). Only one dimension (i.e., dependency) correlates with tracking recommendations and none interacts with student ethnicity or SES. These results suggest that teacher perceptions of their relationship with students is not a prime candidate to explain biased recommendations.

4.2.2. Teacher stereotypes and prejudices

In the psychological literature, prejudices, and stereotypes are different theoretical constructs. Prejudices are the general negative evaluation of a social group or individual based on their group membership (Crandall & Schaller, 2005), whereas stereotypes are thought to be cognitive schemas used by perceivers to process information about others, which manifest as positive or negative beliefs about traits or behaviors associated with certain social groups (Al Ramiah, Hewstone, Dovidio, & Penner, 2010). Several studies find relations between teachers' stereotypes and prejudicial attitudes and student outcomes (Denessen, Hornstra, van den Bergh, & Bijlstra, 2022; Pit-ten Cate & Glock, 2019). However, few

[4] The other two studies do not reveal whether SES differences in parental aspirations can account for SES biases in teacher track recommendations. In the study by Dumont et al. (2019) SES disparities in teacher track recommendations already disappear after accounting for teacher-assigned grades, and before parental aspirations are added to the model. In the study by Klapproth et al. (2012) student SES and migration background are not included in a model that includes parental aspirations.

studies directly test their effects on tracking recommendations. Notably, five experimental articles in our review shed light on the extent to which biases in tracking recommendation depend on teachers' stereotypes and prejudices. Two of those are focused on prejudicial attitudes (Bruneau et al., 2020; Sprietsma, 2013), while the other three studies claim to provide some indirect evidence for the existence of non-descript stereotypes (i.e., the authors attribute their results to teachers' stereotypes but do not specify which specific stereotype —the trait or behavior— that is associated with the target group; Glock et al., 2013; Glock et al., 2015; Klapproth and Fischer, 2020). It should be noted that in the latter category, experimental studies suffer from low power. We would therefore advise caution when interpreting these results.

Among the two that directly focus on prejudice, one is a study on 88 German teachers who were asked to grade 10 essays supposedly produced by German or Turkish students, and to formulate track recommendations for these students. Teachers also had to fill in feeling thermometers about 12 social groups, including Germans and Turks. By including essay and teacher fixed effects in the analyses, essay quality and teacher severity is accounted for (Sprietsma, 2013). Whereas, the author finds support for inequality in grading (10% of a standard deviation in test scores worse for Turkish students around the passing grade) and in track recommendation (11% lower probability for Turkish students to receive a recommendation for the higher track with the same essay) and more positive attitudes towards Germans than Turks (8.5 attitude gap), neither teachers' personal characteristics (age, gender, etc.), nor their prejudicial attitudes explain the ethnic bias in grades or recommendations.

The second study involves a study on 161 Hungarian pre-service teachers who responded to questions about blatant and subtle dehumanization and who filled out feeling thermometers about Roma Hungarians (Bruneau et al., 2020). Six weeks later, the pre-service teachers were contacted again supposedly for another study about track recommendations and were asked to assess how suitable 22 different students were for the low, middle and high school track on a scale ranging from 0 to 100. Student profiles included student names (either typically Roma or non-Roma) and competence and subject test scores. The study shows that Roma students are perceived as marginally better suited for the low track and marginally less well suited for the high track. The authors then created a discrimination score by averaging the tendency to favor placing Roma over non-Roma in the low track with the tendency to favor placing non-Roma over Roma in the high track. They regressed this discrimination score on teachers' blatant dehumanization, subtle dehumanization, and prejudice. The study only finds blatant dehumanization to be significantly related to discrimination. However, the authors note that these results may not necessarily generalize to other minority groups and/or countries because Roma people may be an extreme example as they are the target group of blatant prejudice in Hungary.

Another research line, developed by researchers from Luxembourg and Germany, focuses on experiments testing whether non-descript stereotypes act as a mechanism underlying biased tracking decisions. In one study, the authors created nine student vignettes with information taken from a real database containing the profiles of 2696 Luxembourgish students that had been tracked and followed in subsequent years (Glock et al., 2015). Based on this database, the authors assessed the "accuracy" of tracking recommendations for each student profile, given how well the real students fared after being tracked. The nine profiles were then shown to 48 pre-service teachers who decided which school track was most suitable for each student. They also provided their confidence in the decision. Participants then engaged in a recognition task: they were presented with either correct or incorrect information about the student profiles and asked to flag false information. Results show that tracking decisions are less accurate for ethnic minority students than for ethnic majority students, and that teachers feel less confident about the recommendation for ethnic minority students. The lower accuracy in the tracking decisions for ethnic minority students are due to participants being less able to recognize false information about minority students' grades. The authors interpret the result as evidence for stereotypes by arguing that if teachers hold stereotypes, they focus less on ethnic minority students' grades and therefore are less likely to recognize false information.

In two experimental studies, Glock et al. (2013) examine whether the consistency of a student's grade interacts with student ethnicity to produce more biased tracking recommendations. The authors' hypothesized that inconsistent student profiles require deeper information processing which should lead to less stereotyping. To test this, Luxembourgian primary-school teachers ($N_{Study1} = 54$; $N_{Study2} = 60$) were shown 16 fictitious students' files from native or immigrant (i.e., Portuguese name) students with either consistent or inconsistent information (i.e., standardized test scores did or did not match teacher grades). While native students receive higher track recommendations than immigrant students, the authors find no or marginal ($p = 0.09$) interaction effects with case consistency.

Using a similar experimental design, Klapproth and Fischer (2020) study the effects of achievement level and development (improving vs. declining) and students' ethnicity (German vs. Turkish). 16 fictitious male student files with varying information were presented to 102 German primary-school teachers. The authors find that higher achievement and grade improvement are related to higher tracking recommendation. They find no main effect of ethnicity, but a significant 3-way interaction, such that teachers are affected by German students' achievement development to a larger degree when students have lower grades. This effect is reversed for Turkish students: teachers are less affected by Turkish students' achievement development when their grades are lower. Moreover, when German students improve their grades, their actual level of achievement was less important than when their grades decline. In contrast, for Turkish students, teachers rely less on actual achievement to determine tracking recommendations when students are declining rather than improving. The authors suggest that this result may indicate the presence of stereotypes, which lead teachers to rely more on potential (of improvement) rather than on actual abilities for German students, specifically when descriptions fitted an ethnic stereotype.

Overall, these studies indicate that prejudices or stereotypes could partly explain biases in tracking recommendations. However, as mentioned before, we would recommend caution when interpreting these studies, as they are often underpowered, and the last three do not measure non-descript stereotypes directly.

We do want to note a recent study — that appeared after our literature search was conducted — that does provide more convincing support for the role of implicit stereotypes in biases against students with a migration background in teacher track recommendations (Carlana, La Ferrara, & Pinotti, 2022). In this study, the authors rely on data collected in Northern Italy in which the implicit stereotypes of 1384 math and literature teachers was measured using an implicit association test (IAT). Math and literature teachers usually have an influential say in the track recommendation for students in Italy. The IAT scores were linked to administrative data from more than 23,000 students. Accounting for student test scores and SES, they find that when a teacher scores one standard deviation higher on the IAT, the probability of an immigrant student to be recommended to the vocational track is 2.5 percentage points higher, while the probability to be

recommended to the top track is 1.2 percentage points lower.

4.3. Contextual factors

There are relatively few studies that examine how elements of the larger school- or classroom context relates to track recommendations and/or biases herein.

4.3.1. Composition of the student body

One aspect of the class context that has been considered, is the composition of the student body. Four of the 27 studies include a measure of the ability composition of the class (i.e., average score on standardized test(s)), as well as (an) indicator(s) capturing the share of students from disadvantaged backgrounds (Boone et al., 2018; Caro et al., 2009; Driessen et al., 2008; Timmermans et al., 2015). With respect to the latter, specific measures differ. One study includes separate measures for the SES and ethnic composition of the student body (Boone et al., 2018), a second study uses one measure that combines both aspects (i.e., percentage of native disadvantaged pupils and percentage of minority disadvantaged pupils; Driessen et al., 2008), and the two other studies only focus on one of these aspects (SES: Timmermans et al., 2015; share of students with a migration background: Caro et al., 2009).

Findings on the relationship between the composition of the student body and teacher track recommendation are mixed. With respect to the academic ability composition of the class, three studies find a negative association with track recommendations after accounting for a student's individual academic ability level, whereas one study finds a positive relation. A negative association implies that students attending a class characterized by high academic ability level receive *lower* track recommendations than their equally performing peers in a class characterized by a low academic ability level (Boone et al., 2018; Caro et al., 2009; Driessen et al., 2008). A positive association implies that this group of students receive *higher* track recommendations (Timmermans et al., 2015).

With respect to the share of disadvantaged students in the student population, two studies find little support that this relates to track recommendations. More specifically, a study in Flanders finds no relationship for the average parental occupational status or the share of ethnic minority students in class (Boone et al., 2018). Similarly, Driessen finds that neither the percentage of ethnic minority students from disadvantaged SES backgrounds, nor the percentage of native Dutch students from disadvantaged SES backgrounds contribute to explaining variance in teacher track recommendations in the Netherlands. However, Timmermans et al. (2015) find that students who attend a class with fewer children from low-SES backgrounds receive higher track recommendations in the Netherlands. Caro et al. (2009) report that students are more likely to be recommended to the academic track (Gymnasium) in German classes with a higher share of students with a migration background.

Some of these differences may be explained by the country in which the studies are set. More specifically, Boone et al. (2018) suggest that the SES composition of the student body may play a larger role in the Dutch context than in the Flemish context. In the Netherlands, but not in Flanders, track recommendations are binding. Hence, parents pressuring for high track recommendations may be more pronounced in the Netherlands, especially in high SES schools. Another aspect that could contribute to the different findings are the variables included in the models. All three studies account for individual student performance and SES, but Boone et al. (2018), Caro et al. (2009), and Timmermans et al. (2015) additionally include indicators of the average cognitive ability level of the classroom when examining the socio and/or ethnic composition of the class. In the study by Driessen et al. (2008), class-level ability is negatively associated with track recommendations, yet positively associated with class-level SES. Hence, not accounting for class-level ability may potentially suppress a positive effect of class-level SES effect.

Both Boone et al. (2018) and Timmermans et al. (2015) examine whether the student composition of a class is related to biases in teacher track recommendations, and find no support for this. More specifically, Boone et al. (2018) show that neither the ethnic bias, nor the SES bias depend on the ethnic composition of the class in Flanders. Timmermans et al. (2015) include interactions between average student ability and the share of students with low educated parents in class on the one hand and student gender and socio-ethnic background on the other hand. None of these cross-level interactions is statistically significant.

4.3.2. Institutional features

Three studies in our systematic review pay attention to the potential role of institutional features in biases in tracking recommendations. Two studies examine the role of accountability induced by the institutional context (Glock et al., 2012; Pit-ten Cate et al., 2016). In the first study, teachers were randomly assigned to three different experimental conditions (Glock et al., 2012). In the first condition, teachers had sole responsibility for their tracking decisions and decisions had a large impact on the future educational and occupational careers of students. In the second condition, teachers had to advice a colleague on his/her track recommendations, without further commitment. In the third condition, teachers had to prepare tracking decisions for a council, and were informed that the final tracking decision would be made by the council. This last condition corresponds with the actual tracking procedure in Luxembourg. The first experiment shows that student nationality only impacts teacher track recommendations in the second condition, and that teachers also feel less accountable for their decision in this condition. However, these findings are not replicated in the second experiment in which teachers were asked to think out loud during the experiment.

In the second study, the authors hypothesize accountability to play a role in increasing individuals' motivation to invest effort in the tracking decision. They use an experimental longitudinal design to test this hypothesis (Pit-ten Cate et al., 2016). 38 school teachers from Luxembourg were asked to make tracking decisions at three separate points in time for hypothetical students of ethnic minority vs. majority background. After making tracking decisions for the first set of case vignettes, participants were asked to answer how accountable they felt for their decision on a 7-point scale. This scale was used to render accountability salient in the mind of the participants. Participants then reviewed another set of vignettes and were asked to come back six months later. The researchers assessed the accuracy of the track recommendations by using the same index as the one used in Glock et al., 2015 (see p.19) that was based on information taken from a real database of 2696 Luxembourgish students that had been tracked and followed in subsequent years. Additionally, the researchers assessed the extent to which participants were overconfident when making a wrong decision or underconfident when making a right decision. The results of this experiment reveal that the average level of accuracy for the tracking decisions is high. Nevertheless, tracking decisions for ethnic majority students have a higher accuracy than those for ethnic minority students. Accountability seems to reduce these ethnic differences: after respondents completed questions about accountability, the accuracy of the tracking decisions for ethnic minority students increases. The results also indicate that participants are more likely to be overconfident in their decision for

ethnic minority students before answering the accountability questions than after. Just after responding to the accountability questions (i.e., at time 2), participants' level of confidence is more in line with the actual accuracy of the decision. There are no significant differences in accuracy for ethnic minority students between Time 2 and Time 3 (i.e., six months later); however, accuracy for ethnic majority students improved at Time 3, creating again a discrepancy in accuracy between both groups.

Finally, Batruch et al. (2019) examine whether biases in tracking decisions are impacted by directing evaluators either towards the 'selective function' (of classifying students) or the 'educational function' (of helping all students) of schooling. One of their experiments tests whether manipulating a target student's socioeconomic status as well as the school's function (selection vs. educational) results in differences in the tracking decisions of student participants playing the role of teachers ($N = 160$). The results indicate that for the higher track, the high-SES pupil in the selection condition is considered the most suitable, followed by the high-SES pupil in the educational condition, next the low-SES pupil in the educational condition, and finally the low-SES pupil in the selection condition. The order is reversed with respect to suitability for the lower track. These findings imply that SES biases in track recommendations are larger when the selection function is salient than when the educational function is salient.

The findings of Batruch et al. (2019) are in line with findings from an earlier study on teachers' grading of students. In this study, 455 students playing the role of teachers were asked to assess a dictation test which was supposedly produced by a low or a high-SES student (Autin et al., 2019; experiment 3). To assess the test, participants had to use either a selective assessment method (i.e., grading) or an educational assessment method (i.e., providing comments). When participants used the selective assessment method, they found more mistakes in the low-SES than in the high-SES condition for the same test. This difference was not found when participants used an educational assessment method.

Both the studies by Batruch et al. (2019) and Autin et al. (2019; experiment 3) suggest that institutional selection tools such as tracking may induce SES biases in teacher evaluations. These findings highlight that contexts in which evaluators are encouraged to focus on selecting students may enhance biases.

4.4. Summary and discussion of the sources of inequality in track recommendations

Studies on potential explanations for inequality in teacher track recommendations tend to be linked to: (1) student- and family-related factors, (2) teacher-related factors, (3) contextual factors. With respect to the first explanation, teachers' perception of students' school behavior and attitudes are related to track recommendations, yet do not seem to explain socioeconomic biases herein. As for evidence concerning parental attributes, findings are mixed. Research shows that teachers *perceive* higher SES parents to be more involved in school, and other studies find that positive perceptions of parental involvement are related to higher track recommendations. However, the few studies that *explicitly* examine whether teacher perceptions of parental support or involvement can explain socioeconomic inequality in teacher track recommendations, find little support for this potential pathway. Some studies do indicate that parents from advantaged socioeconomic backgrounds exert more implicit pressure on teachers, and/or are more likely to requests for higher tracks, leading to biases in track recommendations. However, current findings are still inconclusive.

The few studies on teacher-related factors and biases in tracking recommendation have been confined to teacher-student relationship quality and teacher stereotypes and prejudicial attitudes. However, these studies are small in number, and the ones examining teacher stereotypes and prejudice often suffer from small samples and/or only provide indirect evidence for the existence of stereotypes. An exception is the study conducted on Roma students in Hungary, but the target group suffers from such negative stereotypes, it remains unknown whether these effects can be generalizable to other groups (Bruneau et al., 2020). Another exception is the study by Carlana et al. (2022) showing the role of implicit stereotypes in biases in track recommendations by student migration background in Northern Italy. Moreover, research finds support for the effect of stereotypes and prejudice on teacher grades (Alesina, Carlana, Ferrara, & Pinotti, 2018; Van den Bergh, Denessen, Hornstra, Voeten, & Holland, 2010). These findings together suggest that they may play a role in tracking recommendations.

There are very few studies examining the role of contextual factors such as classroom compositions or institutional policies on bias in teacher tracking recommendation. At first glance, results appear promising when it comes to institutional features. However, it should be noted that samples sizes are small.

5. Future directions

While various studies examine ethnic and SES biases in teacher track recommendations, only few of them consider the possible mechanisms underlying these biases. Most existing studies that do shed light on this, include potential mechanisms at the level of the individual student, including student attitudes, behavior, and the home situation or parental involvement. Our review suggests that these factors seem unable to mediate (or explain) biases in tracking recommendations. Moreover, research finds considerable variation in tracking recommendations, as well as biases herein, across teachers and schools (e.g., Timmermans et al., 2015; Timmermans & Rubie-Davies, 2018). Nevertheless, specific characteristics of the teacher (e.g., stereotypes) or the larger educational context are hardly considered in research on biases in track recommendations.

We argue that research into contextual characteristics could be an important avenue for future research. Various theoretical accounts have highlighted the importance of the larger educational context (e.g., schools and educational systems) in the reproduction of social inequalities. These theories propose that educational institutions are cultural contexts that shape the way in which teachers behave, and that contribute to the (re)production of social inequalities (Adams, Biernat, Branscombe, Crandall, & Wrightsman, 2008; Geven, Wiborg, Fish, & van de Werfhorst, 2021; Stephens, Fryberg, Markus, Johnson, & Covarrubias, 2012).

Research on the role of the educational context would be especially important, as it may provide the best point of departure for policies to reduce bias in tracking recommendations. This is important as there are no known methods that can consistently decrease *individual* (implicit) prejudices or biases in the long-term in real world settings (FitzGerald, Martin, Berner, & Hurst, 2019; Forscher et al., 2019), and most studies only show modest short-term effects in laboratory settings (Paluck & Green, 2009; Paluck, Porat, Clark, & Green, 2021). Therefore, investigating how the context induces or mitigates the expression of prejudice could be more useful. Moreover, as presented in the section on contextual characteristics, initial findings suggest that institutions can play a role in the expression of prejudices (Adams et al., 2008; Batruch et al., 2019; Glock et al., 2015).

5.1. Macro-context: research on institutional policies

One possible avenue is studying the impact of institutional policies. In this review, we included studies from different

European countries that vary in their institutional settings (i.e., Germany (10), Netherlands (7), Luxembourg (5), France and Belgium (2), and Switzerland and Hungary (1)). As noted earlier, differences in the national institutional context may impact (biases in) track recommendation decisions. This especially pertains to differences with respect to ability tracking institutions (e.g., age of tracking, binding tracking) (c.f. Geven, et al,. 2018). For example, in several German states and the Flanders region of Belgium, recommendations are not binding, and secondary schools cannot reject students on the basis of their recommendation. On the one hand, this could reduce teachers' sense of accountability and enhance biases in tracking recommendations (Glock et al., 2012). On the other hand, parents may put less pressure on tracking decisions in contexts in which recommendations are not binding, which can reduce bias (Boone et al., 2018; Geven, et al., 2018). Aside from country differences in tracking institutions, other national institutional features, such as policies aiming to support socially disadvantaged groups or to prevent discrimination, could influence teachers' attitudes and behavior towards students from disadvantaged groups.

A recent study indeed suggests the potential relevance of (nation-wide) institutions for teacher expectations. Geven et al. (2021) use a vignette experiment to compare teacher expectations in New York City, Oslo, and Amsterdam. While teachers in different contexts rely on the same student traits to form expectations, they weigh these traits differently. In Amsterdam - a context characterized by early between-school tracking and intense standardized testing to assess student ability - teachers rely more on academic performance. In Oslo, teachers base their expectations more on student SES, seemingly because they make more inferences about student performance on the basis of student SES, especially compared to NYC teachers. The authors note that this latter finding may also reflect teachers' relative blindness to actual educational inequalities in the United States.

Biases in track recommendations may also depend on specific *school-level* institutions that guide teachers' selection decisions (Batruch et al., 2019). Especially in countries in which schools receive a lot of autonomy to organize their work, (in)formal track recommendation procedures may vary substantially across school. In her dissertation, Thys (2018) examines how a school's explicit attention for track allocation is related to biases in track recommendations in 32 schools in two cities in Flanders (Antwerp and Ghent). Explicit attention for track allocation in schools was measured by asking teachers about the opportunities their school provided to professionalize and cooperate with respect to track allocations, and the clarity of the school's vision with respect to ability tracking. Results indicated that students from advantaged socio-economic backgrounds were more likely to receive a recommendation for the academic track, especially in schools in which the explicit attention for track allocation was higher. Results suggest that in these schools, teachers are more likely to take into account student characteristics other than their cognitive performance. This may explain the higher socio-economic inequality in these schools.

Given that tracking recommendations seem to differ across institutional contexts, we argue that it is important for future research to pay attention to institutional features at both the national- and the school-level. In this light, we also want to note that our review is limited to studies set in Europe, and that this may hinder its transferability to non-European countries.

What are potentially low-cost interventions at the institutional level to reduce biases in schools, and track recommendations in particular? In the next section we present a couple of examples of studies that provide cues as how to intervene on the context to diminish bias. Even though this work was not conducted in educational settings, it could be promising for future interventional research in educational settings.

5.2. Research providing avenues for contextual-level interventions to reduce the expression of bias in schools

A few recent experiments conducted outside of the educational setting highlight that decision-making contexts can impact the extent to which participants discriminate in a social judgement task (Adams, Biernat, Branscombe, Crandall, & Wrightsman, 2008; Axt, Casola & Nosek, 2019; Lai & Banaji, 2020). Seven pre-registered studies ($N > 7000$) reveal that asking participants to avoid a potential bias for one social category reduces the bias for that category in an academic judgement task (i.e., selecting honor society applicants based on academic credentials) (Axt, et al., 2019). However, it does not reduce biases in other social categories, and the strategy proves ineffective when statements included more than one social category. In eight other studies ($N > 7000$), the authors replicate the finding that warning participants for biases reduces those biases. Moreover, they find that not revealing an applicant's social category (i.e., blinding application) is an effective method to reduce bias. Other situational techniques (e.g., manipulating time constraints or motivation) do reduce the number of judgement errors, and therefore partly the number of biases, yet do not reduce the *share* of biases in judgement errors (Axt & Lai, 2019). A future study could investigate whether biases in track recommendations can be reduced by for example involving third-party teachers who formulate tracking recommendations on the basis of anonymized students' records.

Uhlmann and Cohen (2005) also provide an interesting demonstration of how bias can manifest in specific contexts. They asked participants to select one of two candidates, Michael or Michelle, for promotion to the position of police chief. Michael and Michelle had identical dossiers, except that one candidate was known for his/her practical knowledge (street-smart), while the other was known for being formally educated (book-smart). The gender of the candidate possessing the competence was reversed in the other condition. After reviewing the candidates, participants were more likely to select Michael rather than Michelle. When asked why, participants mentioned either the importance of formal training or being streetsmart for the job depending in which condition Michael was. This shows that biases can encourage individuals to change the importance of certain criteria after the fact. In a second study, participants were asked to report on what qualities were important to being a police chief before selecting a candidate. When the criteria were selected before, the bias disappeared. This finding suggests that constraining decision-makers to clear and predefined (objective) criteria can powerfully reduce bias (Lai & Banaji, 2020). In the context of educational institutions, this could mean for instance testing whether basing track recommendations on standardized (competency) indicators that have to be interpreted on the basis of clearly *predefined* criteria is effective in reducing bias (c.f., Vanlommel and Schildkamp, 2019).

6. Conclusion

In this paper, we conducted a systematic review to answer two main questions: 1) Are track recommendations biased against students from disadvantaged socio-economic and/or ethnic backgrounds? 2) Which factors account for social and ethnic biases in teacher tracking recommendations? To answer the first, we

conducted a review of 27 recent articles (from 2000 onward) and concluded that recommendations are biased against students from disadvantaged socio-economic backgrounds, even when performance is controlled for. Evidence with respect to ethnic biases was more mixed.

To answer the second question, we reviewed the articles in our systematic review with respect to factors that could explain biased teacher recommendations. Student and family characteristics seem to affect track recommendations, yet do not seem to explain biases in teacher track recommendation. In other words, they do not mediate the relationship between student SES or ethnic background and track recommendations. Although there is some preliminary evidence suggesting that teachers' stereotypes and prejudices can affect biases in tracking recommendation, we believe more robust research is needed before concluding on this matter. Irrespective of this, research (outside of the educational setting) has not yet found methods to reduce stereotypes and prejudices in the long-term.

Given these findings, we believe that work on the impact of institutions is most promising, not only for making innovative scientific contributions (i.e., how institutions can shape individual behavior), but also for finding concrete pathways to reduce bias in tracking. Indeed, it is likely easier to design policies that change the institutional conditions under which teachers take decisions than to change students' personal characteristics, home environments or teachers' long-held beliefs. It is with that in mind that we proposed a few ideas for future research that go in this direction.

Social psychological research has too often constructed its theoretical models in a societal and institutional vacuum and may therefore have been irrelevant to policy-makers (Pettigrew, 2001). If educational psychological research hopes to substantially contribute to policies, then the field should consider broadening their understanding of inequality in education by incorporating the situational conditions that shape how actors think and behave.

Credit author statement

Anatolia Batruch: Conceptualization, Methodology, Data curation and Investigation, Writing (Writing – original draft preparation and review/editing). Sara Geven: Conceptualization, Methodology, Data curation and Investigation, Writing (Writing – original draft preparation and review/editing), Funding acquisition. Emma Kessenich: Data curation and Investigation, Writing (review/editing). Herman van de Werfhorst: Writing (review/editing), Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The current study involves a review study. More detailed notes on study selection/exclusion are available upon request.

References

Adams, G., Biernat, M., Branscombe, N. R., Crandall, C. S., & Wrightsman, L. S. (2008). Beyond prejudice: Toward a sociocultural psychology of racism and oppression. In G. Adams, M. Biernat, N. R. Branscombe, C. S. Crandall, & L. S. Wrightsman (Eds.), *Commemorating Brown: The social psychology of racism and discrimination* (pp. 215–246). Washington, DC: American Psychological Association. https://doi.org/10.1037/11681-012.

Al Ramiah, A., Hewstone, M., Dovidio, J. F., & Penner, L. A. (2010). The social psychology of discrimination: Theory, measurement and consequences. In L. Bond, F. McGinnity, & H. Russell (Eds.), *Making equality count: Irish and international research measuring equality and discrimination* (pp. 84–112). Dublin: The Liffey Press.

Alesina, A., Carlana, M., Ferrara, E. L., & Pinotti, P. (2018). *Revealing stereotypes: Evidence from immigrants in schools*. https://doi.org/10.3386/w25333. NBER Working Paper 25333.

Autin, F., Batruch, A., & Butera, F. (2019). The function of selection of assessment leads evaluators to artificially create the social class achievement gap. *Journal of Educational Psychology, 111*(4), 717.

Axt, J. R., Casola, G., & Nosek, B. A. (2019). *Reducing social judgment biases may require identifying the potential source of bias. Personality and Social Psychology Bulletin*. Advance online publication. https://doi.org/10.1177/0146167218814003

Axt, J. R., & Lai, C. K. (2019). Reducing discrimination: A bias versus noise perspective. *Journal of Personality and Social Psychology, 117*(1), 26. https://doi.org/10.1037/pspa0000153

Bakker, J., Denessen, E., & Brus-Laeven, M. (2007). Socio-economic background, parental involvement and teacher perceptions of these in relation to pupil achievement. *Educational Studies, 33*(2), 177–192. https://doi.org/10.1080/03055690601068345

Barg, K. (2013). The influence of students' social background and parental involvement on teachers' school track choices: Reasons and consequences. *European Sociological Review, 29*(3), 565–579. https://doi.org/10.1093/esr/jcr104

Barg, K. (2015). Educational choice and cultural capital: Examining social stratification within an institutionalized dialogue between family and school. *Sociology, 49*(6), 1113–1132.

Batruch, A., Autin, F., Bataillard, F., & Butera, F. (2019). School selection and the social class divide: How tracking contributes to the reproduction of inequalities. *Personality and Social Psychology Bulletin, 45*(3), 477–490. https://doi.org/10.1177/0146167218791804

Boone, S., Thys, S., Van Avermaet, P., & Van Houtte, M. (2018). Class composition as a frame of reference for teachers? The influence of class context on teacher recommendations. *British Educational Research Journal, 44*(2), 274–293. https://doi.org/10.1002/berj.3328

Boone, S., & Van Houtte, M. (2013). Why are teacher recommendations at the transition from primary to secondary education socially biased? A mixed-methods research. *British Journal of Sociology of Education, 34*(1), 20–38. https://doi.org/10.1080/01425692.2012.704720

Bruneau, E., Szekeres, H., Kteily, N., Tropp, L. R., & Kende, A. (2020). Beyond dislike: Blatant dehumanization predicts teacher discrimination. *Group Processes & Intergroup Relations, 23*(4), 560–577. https://doi.org/10.1177/1368430219845462

Brunello, G., & Checchi, D. (2007). Does school tracking affect equality of opportunity? New international evidence. *Economic Policy, 22*(52), 782–861. https://doi.org/10.1111/j.1468-0327.2007.00189.x

Carlana, M., La Ferrara, E., & Pinotti, P. (2022). Implicit stereotypes in teachers' track recommendations. *AEA Papers and Proceedings, 112*, 409–414.

Caro, D. H., Lenkeit, J., Lehmann, R., & Schwippert, K. (2009). The role of academic achievement growth in school track recommendations. *Studies In Educational Evaluation, 35*(4), 183–192. https://doi.org/10.1016/j.stueduc.2009.12.002

Chmielewski, A. K. (2014). An international comparison of achievement inequality in within-and between-school tracking systems. *American Journal of Education, 120*(3), 293–324. https://doi.org/10.1086/675529

De Boer, H., Bosker, R. J., & van der Werf, M. P. (2010). Sustainability of teacher expectation bias effects on long-term student performance. *Journal of Educational Psychology, 102*(1), 168. https://doi.org/10.1037/a0017289

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

Driessen, G., Sleegers, P., & Smit, F. (2008). The transition from primary to secondary education: Meritocracy and ethnicity. *European Sociological Review, 24*(4), 527–542. https://doi.org/10.1093/esr/jcn018

Dumont, H., Klinge, D., & Maaz, K. (2019). The many (subtle) ways parents game the system: Mixed-method evidence on the transition into secondary-school tracks in Germany. *Sociology of Education, 92*(2), 199–228. https://doi.org/10.1177/0038040719838223

Esser, H. (2016). The model of ability tracking-theoretical expectations and empirical findings on how educational systems impact on educational success and inequality. In H.-P. Blossfeld, S. Buchholz, J. Skopek, & M. Triventi (Eds.), *Models of secondary education and social inequality: An international comparison* (pp. 25–44). Edward Elgar Publishing Limited. https://doi.org/10.4337/9781785367267.00009.

Feron, E., Schils, T., & Ter Weel, B. (2016). Does the teacher beat the test? The value of the teacher's assessment in predicting student ability. *Economist, 164*(4), 391–418. https://doi.org/10.1007/s10645-016-9278-z

FitzGerald, C., Martin, A., Berner, D., & Hurst, S. (2019). Interventions designed to reduce implicit prejudices and implicit stereotypes in real world contexts: a systematic review. *BMC Psychology, 7*(1), 1–12.

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

Gamoran, A. (2009). *Tracking and inequality: New directions for research and practice* (pp. 231–246). Routledge.

Geven, S., Batruch, A., & van de Werfhorst, H. G. (2018). *Inequality in teacher judgements, expectations and track recommendations: A review study.*

Amsterdam: Universiteit van Amsterdam.
Geven, S., Wiborg, Ø. N., Fish, R. E., & van de Werfhorst, H. G. (2021). How teachers form educational expectations for students: a comparative factorial survey experiment in three institutional contexts. *Social Science Research, 100*, 102599.
Glock, S., Krolak-Schwerdt, S., Klapproth, F., & Böhmer, M. (2012). Improving teachers' judgments: Accountability affects teachers' tracking decisions. *International Journal of Technology and Inclusive Education, 1*, 89–98.
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*(4), 555–573. https://doi.org/10.1007/s11218-013-9227-5
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*(2), 169–188. https://doi.org/10.1007/s10212-014-0237-2
Klapproth, F., & Fischer, B. D. (2020). Achievement development is less important for school-placement recommendations when students are stereotyped. *Social Psychology of Education: International Journal*. https://doi.org/10.1007/s11218-020-09593-9
Klapproth, F., Glock, S., Böhmer, M., Krolak-Schwerdt, S., & Martin, R. (2012). School placement decisions in Luxembourg: Do teachers meet the Education Ministry's standards? *Literacy Information and Computer Education Journal, 1*, 765–771.
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
Kloosterman, R., Ruiter, S., De Graaf, P. M., & Kraaykamp, G. (2009). Parental education, children's performance and the transition to higher secondary education: Trends in primary and secondary effects over five Dutch school cohorts (1965–99). *British Journal of Sociology, 60*(2), 377–398. https://doi.org/10.1111/j.1468-4446.2009.01235.x
Krolak-Schwerdt, S., Hörstermann, T., Glock, S., & Böhmer, I. (2018). Teachers' assessments of students' achievements: The ecological validity of studies using case vignettes. *The Journal of Experimental Education, 86*(4), 515–529. https://doi.org/10.1080/00220973.2017.1370686
Lüdemann, E., & Schwerdt, G. (2013). Migration background and educational tracking. *Journal of Population Economics, 26*(2), 455–481. https://doi.org/10.1007/s00148-012-0414-z
Niklas, F., & Schneider, W. (2017). Home learning environment and development of child competencies from kindergarten until the end of elementary school. *Contemporary Educational Psychology, 49*, 263–274. https://doi.org/10.1016/j.cedpsych.2017.03.006
Paluck, E. L., & Green, D. P. (2009). Prejudice reduction: What works? A review and assessment of research and practice. *Annual review of psychology, 60*, 339–367.
Paluck, E. L., Porat, R., Clark, C. S., & Green, D. P. (2021). Prejudice reduction: Progress and challenges. *Annual review of psychology, 72*, 533–560.
Pietsch, M., & Stubbe, T. C. (2007). Inequality in the transition from primary to secondary school: School choices and educational disparities in Germany. *European Educational Research Journal, 6*(4), 424–445. https://doi.org/10.1146/annurev.psych.60.110707.163607
Pit-ten Cate, & Glock, S. (2019). Teachers' implicit attitudes toward students from different social groups: A meta-analysis. *Frontiers in psychology, 10*, 2832.
Pit-ten Cate, I. M., Krolak-Schwerdt, S., & Glock, S. (2016). Accuracy of teachers' tracking decisions: Short-and long-term effects of accountability. *European Journal of Psychology of Education, 31*(2), 225–243. https://doi.org/10.1007/s10212-015-0259-4
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
Stephens, N. M., Fryberg, S. A., Markus, H. R., Johnson, C. S., & Covarrubias, R. (2012). Unseen disadvantage: How American universities' focus on independence undermines the academic performance of first-generation college students. *Journal of Personality and Social Psychology, 102*(6), 1178–1197.
Südkamp, A., Kaiser, J., & Möller, J. (2012). *Accuracy of teachers' judgments of students' academic achievement: A meta-analysis*. American Psychological Association. https://doi.org/10.1037/a0027627
Terrin, E., & Triventi, M. (2022). The Effect of School Tracking on Student Achievement and Inequality: A Meta-Analysis. *Review of Educational Research, 00346543221100850*.
Timmermans, A. C., Boer, H., & Werf, M. P. (2016). An investigation of the relationship between teachers' expectations and teachers' perceptions of student attributes. *Social Psychology of Education, 19*(2), 217–240. https://doi.org/10.1007/s11218-015-9326-6
Timmermans, A. C., de Boer, H., Amsing, H. T. A., & van der Werf, M. P. C. (2018). Track recommendation bias: Gender, migration background and SES bias over a 20-year period in the Dutch context. *British Educational Research Journal, 44*(5), 847–874. https://doi.org/10.1002/berj.3470
Timmermans, A. C., Kuyper, H., & Werf, G. (2015). Accurate, inaccurate, or biased teacher expectations: Do Dutch teachers differ in their expectations at the end of primary education? *British Journal of Educational Psychology, 85*(4), 459–478. https://doi.org/10.1111/bjep.12087
Timmermans, A. C., & Rubie-Davies, C. M. (2018). Do teachers differ in the level of expectations or in the extent to which they differentiate in expectations? Relations between teacher-level expectations, teacher background and beliefs, and subsequent student performance. *Educational Research and Evaluation, 24*(3–5), 241–263.
Timmermans, A. C., van der Werf, M. P. C. G., & Rubie-Davies, C. M. (2019). The interpersonal character of teacher expectations: The perceived teacher-student relationship as an antecedent of teachers' track recommendations. *Journal of School Psychology, 73*, 114–130. https://doi.org/10.1016/j.jsp.2019.02.004
Urhahne, D., & Wijnia, L. (2020). A review on the accuracy of teacher judgments. *Educational Research Review*. , Article 100374. https://doi.org/10.1016/j.edurev.2020.100374
Van den Bergh, L., Denessen, E., Hornstra, L., Voeten, M., & Holland, R. W. (2010). The implicit prejudiced attitudes of teachers: Relations to teacher expectations and the ethnic achievement gap. *American Educational Research Journal, 47*(2), 497–527.
Van Huizen, T. (2021). Teacher bias or measurement error bias? Evidence from track recommendations. *USE Working Paper Series, 21*(13).
Vanlommel, K., & Schildkamp, K. (2019). How do teachers make sense of data in the context of high-stakes decision making? *American Educational Research Journal, 56*(3), 792–821. https://doi.org/10.3102/0002831218803891
Van de Werfhorst, H. G. (2019). Early tracking and social inequality in educational attainment: Educational reforms in 21 European countries. *American Journal of Education, 126*(1), 65–99. https://doi.org/10.1111/j.1468-4446.2007.00157.x
Van de Werfhorst, H. G., & Hofstede, S. (2007). Cultural capital or relative risk aversion? Two mechanisms for educational inequality compared 1. *British Journal of Sociology, 58*(3), 391–415.
Van de Werfhorst, H. G., & Mijs, J. J. (2010). Achievement inequality and the institutional structure of educational systems: A comparative perspective. *Annual Review of Sociology, 36*, 407–428. https://doi.org/10.1146/annurev.soc.012809.102538
Wang, S., Rubie-Davies, C. M., & Meissel, K. (2018). A systematic review of the teacher expectation literature over the past 30 years. *Educational Research and Evaluation, 24*(3–5), 124–179. https://doi.org/10.1080/13803611.2018.1548798
Wenz, S. E., & Hoenig, K. (2020). Ethnic and social class discrimination in education: Experimental evidence from Germany. *Research in Social Stratification and Mobility, 65*, Article 100461. https://doi.org/10.1016/j.rssm.2019.100461