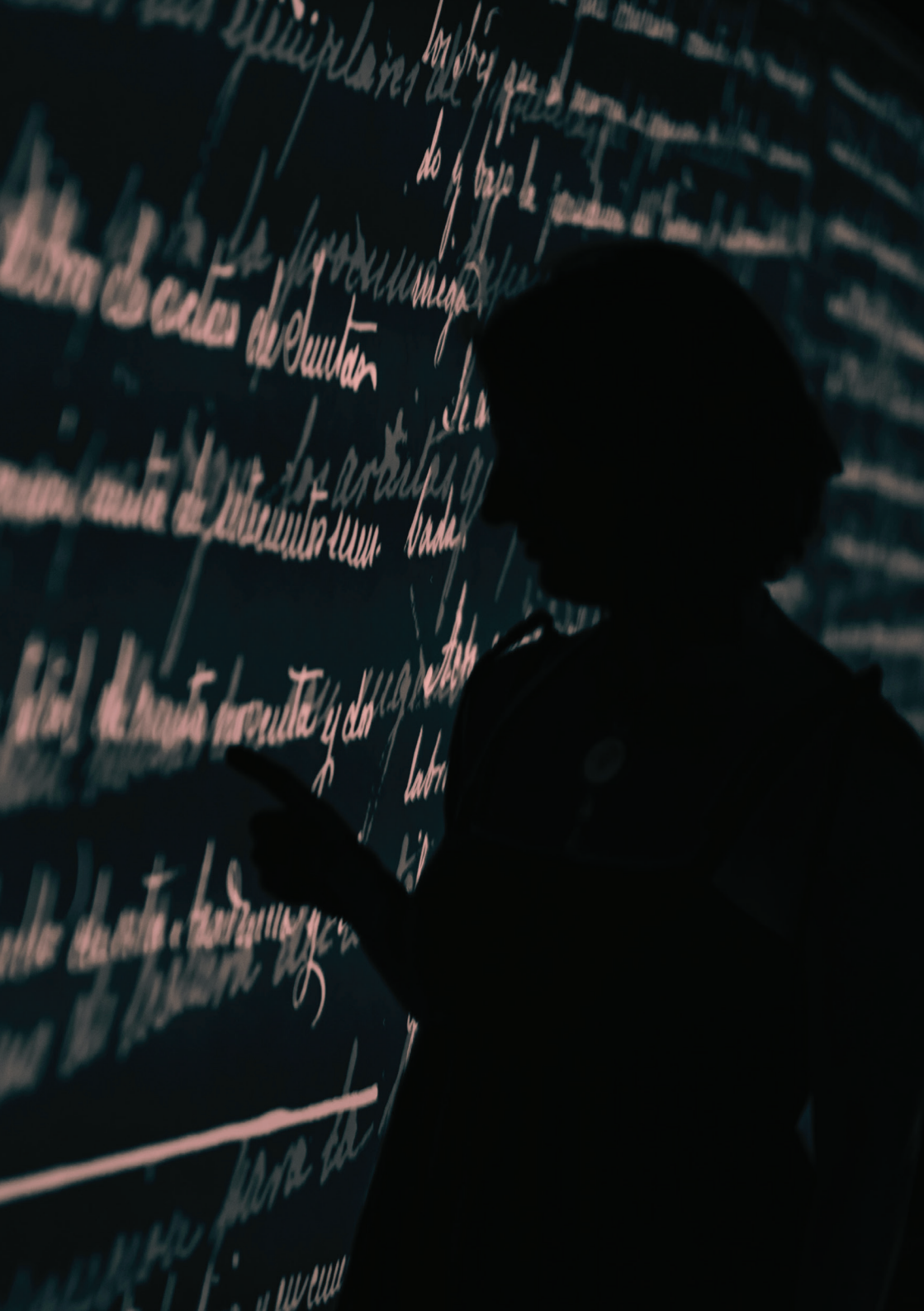
Parents, teachers, and media: agents of biased socialization

Kroes, A.D.A.



Chapter 4

Gender and Ethnicity: Bias in Written Evaluations of Primary School Pupils

Antoinette D. A. Kroes, Marleen G. Groeneveld, Meike M. de Boer, Lotte D. Van der Pol, Judi Mesman

In preparation for publication.

Abstract

Teacher bias can affect pupils' (future) wellbeing and achievements, but research findings on the existence of bias based on pupil characteristics such as gender and ethnicity have been inconsistent. The aim of this study was to investigate whether gender and ethnicity are associated with differences between written evaluations by primary school teachers on report cards. Written evaluations were coded for valence, performance-relatedness, and language abstraction. In Study 1 (N = 203) on ethnic majority pupils, results indicated that girls were evaluated more positively than boys in relation to performance, while boys were evaluated more negatively than girls in relation to other attributes, like character traits and behavior. In Study 2 (N = 136), results indicated that ethnic majority girls were evaluated more positively than all other groups. Ethnic minority girls received more negative performance-related evaluations than other pupils did. Results are discussed in terms of their implications for the school context.

Keywords: linguistic intergroup bias, gender, ethnicity, primary school, teacher bias

Primary school pupils regularly receive evaluations from their teachers through common practices like grading, verbal feedback, and written comments. These evaluations are an expression of teachers' expectations which affect pupils' (future) achievement, socio-psychological, and behavioral outcomes, making them an important research area (Wang et al., 2018). Besides past achievement, pupil evaluations can be informed by various pupil factors, like (perceived) effort, and pupil characteristics (Rubie-Davies, 2015). When teacher expectations are based on pupil characteristics independent of their actual achievement, this indicates bias. This has for instance been found consistently for socioeconomic status (SES), with teachers having lower expectations of pupils with low SES than of those with high SES (Timmermans et al., 2018). However, findings on the relation between expectations and pupil gender and ethnicity have been inconsistent (Geven et al., 2018; Wang et al., 2018), indicating that more research is necessary. In this paper, we will investigate Dutch teachers' written pupil evaluations for possible gender and ethnicity-based bias. We will examine differences between boys and girls, and between ethnic minority and majority pupils along three components of written teacher comments. Firstly, the proportions of positive and negative remarks. Some studies show that the proportions of positive and negative evaluations are affected by pupil gender and ethnicity (Ni & Li, 2013; Rojek et al., 2019). Secondly, the proportion of remarks related and unrelated to performance. Previous research has shown that teachers use more performance-related feedback when they have high expectations of a pupil (Gentrup et al., 2020). Lastly, we investigate language abstraction, which has been found to be a good measure of implicit bias (Beukeboom, 2014; Menegatti et al., 2017; Semin & Fiedler, 1988).

Bias in teacher expectations

There is a long history of research into teacher expectations and their effect on pupil achievement and wellbeing (Wang et al., 2018). Teacher

expectations can be defined as "inferences that teachers make about the present and future academic achievement and general classroom behavior of students" (Good & Brophy, 1997, p. 79). According to this definition, teacher expectations are expressed in many different forms, and teacher bias has been found in many of these expressions, like in implicit and explicit attitudes (Glock & Klapproth, 2017; Glock & Kleen, 2017), and in teacher behaviors and school processes like verbal feedback and turn giving (Bašaragin & Savic, 2019; Denessen et al., 2020; Gentrup et al., 2020), track recommendations (Geven et al., 2018; Timmermans et al., 2018), and evaluations of pupil achievement and behavior (Shepherd, 2011; Van Ewijk, 2011). Teacher expectations are positively related to pupil outcomes like achievement, motivation, and perceived self-efficacy, and this relation is especially strong for high expectations (Gentrup et al., 2020; Rubie-Davies, 2015; Wang et al., 2018). For instance, a pupil that receives more demanding feedback gets more opportunities for optimal development (Denessen et al., 2020), this can simultaneously increase their motivation, which in turn indirectly improves performance (Jussim, 2009). There is evidence for long-term effects of teacher expectations on pupil achievement, indicating a self-fulfilling prophecy effect (De Boer et al., 2010; Gentrup et al., 2020). This is especially true for pupils from marginalized groups, as they are impacted more by the effects of teacher expectations (Wang et al., 2018). It is therefore important to stimulate high expectations in teachers (Rubie-Davies, 2015), as Weinstein, Gregory, and Strambler (2004) have stated: "all but high expectations are, by definition, inaccurate, given that achievement scores reflect what has been taught, potential is not measurable, and the goal of education is to foster academic growth" (p. 513).

Differences in teacher expectations can stem from different sources, like institutional settings, classroom and school composition, and teacher traits (Geven et al., 2018). Biases that exist in a broader societal context, like gender and ethnicity based bias, can be (unconsciously) reproduced by teachers (Myhill & Jones, 2006). Pupil demographics have therefore

often been studied in relation to teacher bias; with gender, ethnicity, and socioeconomic status (SES) being studied the most (Wang et al., 2018). While positive effects for pupil SES on teacher bias are found quite consistently, the relations between teacher bias and, respectively, pupil gender and ethnicity are less clear (Geven et al., 2018; Wang et al., 2018).

Gender bias

Findings for gender bias in teacher expectations have been inconsistent internationally (Wang et al., 2018), and in the Netherlands specifically (Geven et al., 2018). Gender stereotypes about behaviors and characteristics may constitute a basis for biased teacher expectations. On the one hand, girls are perceived to have better work habits and to be more motivated than boys (Glock & Klapproth, 2017; Myhill & Jones, 2006; Timmermans et al., 2016, 2018). On the other hand boys are seen as more disruptive, competitive, dominant, and outspoken (Frawley, 2005). The majority of teachers takes behavioral pupil characteristics into account in their evaluations (Inspectorate of Education, 2014; McMillan, 2019). This may be why girls generally receive higher grades than boys on non-standardized tests, independent from actual achievement (Voyer & Voyer, 2014). Conversely, classroom observations have shown that while boys are disciplined more, they have a higher status than girls, which is reflected in being allowed more speaking time and receiving more helpful feedback (Bašaragin & Savic, 2019; Bassi et al., 2016; Frawley, 2005).

Gender bias is likely affected by cultural differences. The Netherlands score quite high on various international indices of gender equality (Dutch Central Bureau of Statistics, 2022a), possibly causing a decrease in gender bias. For example, a longitudinal Dutch study on track recommendations between 1995 and 2014 shows a decrease in positive bias towards girls (Timmermans et al., 2018). In the final year of primary school, Dutch pupils receive a track recommendation by their teacher, which assigns pupils to one of six secondary school tracks. These tracks are based on cognitive skills, running from low level vocational education

to pre-academic education (for a more extensive explanation of the Dutch educational system, see Stevens et al., 2019). More recent data shows that gender bias in track recommendations appeared to be absent after 2014, and in the last few years (2019-2021) has actually reversed, with boys receiving higher recommendations than girls, independent from achievement (Dutch Central Bureau of Statistics, 2022a). The inconsistencies and changes over time in findings of gender bias in teacher expectations indicate the necessity of further investigation.

Ethnic bias

The ethnic/racial background of pupils may be another characteristic that leads to bias in teacher expectations. Bias could be fueled by stereotypical beliefs and attitudes about ethnic outgroups. To what extent teachers show ethnic bias has been the subject of ongoing research debates (Geven et al., 2018; Mason et al., 2014; Stevens et al., 2019). Systematic reviews indicate that most studies find negative evaluation bias against ethnic minority groups in North America and Europe when compared to the ethnic majority (Childs & Wooten, 2023; Tenenbaum & Ruck, 2007; Wang et al., 2018). In a smaller number of studies from both regions, no negative ethnic bias is reported. Several studies indicate that pupils from ethnic minority groups are not evaluated more negatively or positively on average, but that teachers are found to be less accurate in their evaluations of ethnic minority pupils than in those of ethnic majority pupils (Geven et al., 2018; Mason et al., 2014; Tenenbaum & Ruck, 2007; Wang et al., 2018). In the Netherlands, various large scale quantitative studies have shown no ethnicity-based bias against ethnic minority pupils in teachers' expectations (Stevens et al., 2011, 2019). However, smaller scale studies that used implicit measures, did find ethnicity-based bias (Van den Bergh et al., 2010; Van Ewijk, 2011), as did ethnographic research in Dutch classrooms (Weiner, 2015, 2016). This indicates that research may benefit from the investigation of a larger variety of practices, using qualitative and/or implicit measures.

In various studies that found no ethnicity based bias in teacher expectations, negative bias against ethnic minority pupils was explained by socio-economic factors (De Boer et al., 2010; Rubie-Davies & Peterson, 2016; Stevens et al., 2019). However, SES and ethnicity show high levels of overlap (Stevens et al., 2019). Ethnicity and SES indicators intersect and interact, and treating these categories as distinct and separate does not offer a true representation of social processes (Gillborn et al., 2018; Stevens et al., 2019). The overlap of SES and ethnicity is illustrated by differential effects of SES on the attainment of academic proficiency for ethnic minority pupils compared to ethnic majority pupils, which has been recorded in many countries (OECD, 2018). Attempting to disentangle indicators of SES from ethnicity and immigration status in statistical analyses, while these concepts are intertwined, overlooks socially constituted injustice (Gillborn et al., 2018; Stevens et al., 2019). In the current study, we will therefore match pupils on parents' educational level.

Research outcomes for both gender and ethnicity bias in teacher expectations are inconclusive; information on the intersection of ethnicity and gender is even less clear. Some studies have found a similar gender gap for ethnic minority pupils, with girls being evaluated more positively than boys (Farris & de Jong, 2014), while others have found the opposite, with negative bias against ethnic minority girls (Kleen & Glock, 2018). Several studies indicate that pupils from ethnic minority groups are evaluated more negatively than the ethnic majority group, regardless of gender (Glock & Klapproth, 2017; Menegatti et al., 2017). More research may provide insights into how pupil characteristics relate to teacher bias.

Written evaluations

In this paper, we will focus on written evaluations of pupils, which are an expression of teacher expectations (Wang et al., 2018). Including written evaluations (also called teacher comments, written feedback, or narrative evaluations) on report cards is common practice, especially for primary

school pupils (Hollingsworth et al., 2019). These comments can convey information that goes beyond grades, clarifying what pupils learned, the progress made, and what is left to learn (Hattie & Timperley, 2007). According to previous studies, there are several important characteristics that written evaluations should meet to stimulate development for all pupils, independent from past performance. Firstly, evaluations should mainly be aimed at performance, including specific tasks and task processes (Guskey, 2019; Hattie & Timperley, 2007). Evaluations concerning personal attributes, like "good girl" and "great effort", are unlikely to be effective, although they are more commonly given than performance-related evaluations (Hattie & Timperley, 2007). Secondly, pupils benefit from evaluations that include statements with positive and negative valence (Guskey, 2019; Hattie & Timperley, 2007; Hyland & Hyland, 2006). Evaluations with positive valence (e.g., "She is diligent"), can contribute to motivation, interest, self-efficacy, and teacher-pupil relationships. However, positive evaluations can also have negative effects. For instance, pupils can interpret praise as an expression of low teacher expectations (Hattie & Timperley, 2007). Evaluations with negative valence (e.g., "She is not able to multiply numbers over 12") are beneficial when they are constructive, related to specific tasks and task processes, indicate what needs to be improved, and providing guidance on how to improve (Guskey, 2019; Hattie & Timperley, 2007; Hyland & Hyland, 2006). However, in a qualitative analysis of written evaluations, it was found that there are often very few negative comments on report cards, the few negative comments relate mainly to behavior, and they are often formulated in very general terms, like "try harder" or "a consistent effort is needed" (Hattie & Peddie, 2003). These types of evaluations offer little guidance for pupils and their parents on how to improve on specific tasks. Moreover, pupils interpret these evaluations as impersonal, and can find them unhelpful, confusing, and discouraging (Hyland, 2013). When evaluations are not made in relation to a specific task, both positive and negative evaluations may imply that performance is not the result of the

pupils' efforts, but of a stable ability or trait (Cimpian, 2010). This can lower motivation and effort, thus affecting (future) performance (Hattie & Timperley, 2007). In summary, written evaluations should primarily be performance-related, consist of guidance for improvement, and should include positive and negative remarks for all pupils, independent from the actual performance and expectations of future achievement (Gentrup et al., 2020; Guskey, 2019; Hattie & Timperley, 2007).

Subtle forms of bias in written evaluations appear to be added by evaluators outside of awareness (Beukeboom, 2014; Rojek et al., 2019). This makes written evaluations interesting for the investigation of implicit teacher bias. Written evaluations are forms of archival and authentic data (Ni & Li, 2013), which gives them a higher ecological validity than measures that include fictional students, and (digital) implicit bias tests. However, not many studies into written evaluations on report cards exist (Hollingsworth et al., 2019; Ni & Li, 2013). In a study focusing on written evaluations of non-academic areas, teachers showed ethnic bias (Ni & Li, 2013). Compared to White pupils, Black pupils received significantly more negative comments, while Asian pupils received significantly more positive comments on behavior. Gender was not considered. To the best of our knowledge, similar quantitative research on performance-related written comments does not exist. In a qualitative study on teacher comments, it was found that lower performing pupils received less comments related to performance overall, and more related to effort, resulting in a similar overall number of positive comments (Hattie & Peddie, 2003). Effects of gender and ethnicity were not investigated. In studies involving classroom observations, verbal teacher feedback has been investigated. In one study, boys received more negative comments than girls (Bassi et al., 2016), in another study teachers gave more performance-related and more positive feedback to boys (Bašaragin & Savic, 2019). In a third study, teachers were more likely to give performance-related verbal feedback to pupils of whom they had inaccurate (i.e., not based on actual performance) high

expectations in general, and more positive performance-related feedback specifically, compared to pupils of whom they had inaccurate low expectations (Gentrup et al., 2020). Effects of pupil gender and ethnicity were not investigated. Therefore, investigating whether the proportion of performance-related teacher comments differ based on gender and ethnicity can help in determining whether bias is present.

Linguistic bias

Language, for instance in written evaluations, is a tool that regulates cognitive and motivational processes between a sender and a receiver (Semin, 2000). Cognitive processes entail the transfer of meaning, for instance, when an event is described, or instructions are given. Motivational processes entail the way in which messages are affected by wishes and desires, like the desire to portray someone as positive or negative (Beukeboom, 2014). These processes affect the words people choose to use, and take place largely outside of awareness. Hence, people's choice of words shows their implicit bias, which makes written evaluations an interesting source for bias research.

Bias in word choice on written evaluations has been found for both gender and ethnicity (Biernat et al., 2012; Rojek et al., 2019). For instance, in a study comparing written evaluations of medical students, it was found that ethnic majority students were more often than ethnic minority students described with positive competency-related words like "outstanding" and "impressive", and with positive words related to personal attributes like "mature" and "sophisticated" (Rojek et al., 2019).

Besides focusing on the meaning of words, linguistic bias can be investigated by the degree of linguistic abstraction that is used. Imagine a boy scribbling on a worksheet, and compare the following two sentences: "he scribbled on his paper" and "he is messy". The first sentence is more concrete, signaling an incident, or temporary event, while the latter is more abstract, signaling a stable, typical characteristic of the boy (Beukeboom, 2014; Maass et al., 1989; Menegatti et al., 2017; Wigboldus et

al., 2000, 2005). Language abstraction, like communication, is thought to be affected by both cognitive and motivational mechanisms (Beukeboom, 2014; Semin, 2000). The cognitive mechanism is driven by expectations; people tend to use language that is more abstract when behavior is expectancy-consistent, while expectancy-inconsistent behaviors are described at a lower level of abstraction. The motivational mechanism is formed by the desire to protect social identities, for instance by portraying a member of a marginalized group as negative, even when this is in contrast with cognition. Linguistic abstraction can be investigated with the Linguistic Category Model (LCM), which sorts verbs and adjectives in four categories with different levels of abstraction (see the method section of Study 1). Language abstraction is considered a sound implicit measure of bias, and has been used to that end in various contexts (e.g., Menegatti et al., 2017; Menegatti & Rubini, 2017; Prati et al., 2015; Schoel et al., 2014). According to cognitive mechanisms underlying the LCM, when teachers have biased expectations about girls (e.g., "girls are better behaved than boys"), they will use more abstract words when evaluating a girl who is well behaved. According to the motivational mechanism, teachers will be more likely to use abstract language for positive evaluations of the dominant social group, while using more concrete language for positive evaluations of marginalized and underrepresented groups (Beukeboom, 2014; Semin, 2000).

By investigating language abstraction, possible implicit bias in teacher evaluations can be detected, and previous inconsistencies in research findings can be informed using this method. In an Italian context, Menegatti et al. (2017) showed that girls were described with more abstract positive terms and more concrete negative terms than boys, revealing that positive attributes of girls were expected whereas negative attributes were seen as exceptional. The same was observed for students with non-migrant origins over students with migrant origins. Reversely, boys and students from migrant origins were more often described with abstract negative language and concrete positive language, implying

stable negative and situational positive characteristics. By investigating similar associations in a Dutch sample, we can both add to the existing body of knowledge on linguistic bias and teacher bias, and we can see whether cultural differences between Italy and the Netherlands arise.

Study 1

The aim of Study 1 was to investigate the relation between teacher evaluations and pupil gender. Because positive bias towards girls has been reported regularly (Wang et al., 2018), especially when concerning attributes like behavior and effort (Glock & Klapproth, 2017), and because report cards have been found to contain mainly evaluations in relation to behavior and effort, we expected that: 1.1A) girls receive more positive evaluations than boys, both related and unrelated to performance; 1.1B) girls receive more performance-related evaluations than boys, both with positive and negative valence; 1.2) girls are evaluated with more abstract positive terms and more concrete negative terms than boys.

Method

Sample

This study is part of the longitudinal project "Girls in Science", which examines adolescents' gender socialization in the family and school context. This study reports on data from the first wave of the project. Opposite-sex couples with at least two children from the Western part of the Netherlands were eligible for participation. Two groups of families were recruited, with children in two different age groups. Exclusion criteria were severe physical or mental disabilities of a family member, divorced/separated families, single-parent families, families with a non-biological parent, and parents raised outside the Netherlands. The first group had participated in a previous longitudinal project (Boys will be Boys?; see Hallers-Haalboom et al., 2017) and were selected from

municipality records when their second-born child was approximately 1 year old. At the moment of first participation, there were 1,249 eligible families, out of which 390 participated (31%). For the current study, families were invited to participate again when their second-born was approximately 10 years old. Due to dropout during the previous project, 345 families were invited. In total, 233 families were willing to participate, six of which were excluded due to divorce, emigration, or decease of a family member. Ultimately, 144 families participated (66%). Data collection for this group took place in 2020 and 2021, and had to be paused for several months due to COVID-19, likely affecting the rate of participation. The second group was recruited for the current study and were eligible when their second-born was approximately 12 years old. The maximum age difference with the first-born was 36 months. Municipality records were used to select 2,988 families who received a written invitation by mail. In total, 164 families (5%) in this group fully participated. In total across both groups, 308 families participated.

Children were eligible to participate in the current study if they were in primary school and the family could provide a report card which included written evaluations by the teacher (N = 213). Some children (n = 8) were excluded because no written remarks were codable, 1 child was excluded because the report card was from a lower grade, and 1 child was excluded because the remarks on the report card were unreadable, resulting in a final sample of 203 (104 boys). The sample contained 33 sibling-pairs. The pupils attended 155 different schools. For pupils who attended the same school, we checked whether the report cards could have been written by the same teacher. There were 5 duos of pupils whose report cards were written by the same teacher, one of these duos was a sibling pair. Removing these duos from the sample did not make any significant changes to the results, so the duos were included in the data analysis. All pupils were in sixth, seventh, or eight grade of primary school. These are the final grades of Dutch primary school; pupils are

typically aged 9-13 in these grades. Most parents received higher education (mothers: 88%, fathers: 78%).

Procedure

Each family was visited once by one or two trained graduate or undergraduate researchers. Report cards were first collected during home visits, where they were photographed by the visiting researcher. Due to COVID-19, the home visits were continued online in 2020 and subsequently report cards were collected through digital uploading by the parents of the children. Families completed several questionnaires and observational tasks before, during and after these visits. Families received gift certificates after completing the tasks. Informed consent was obtained from all participating children and their parents. Ethical approval was provided by the research ethics committee of the researchers' host institute.

Instruments

Written teacher evaluations were transcribed from pseudonymized photographs into Microsoft Excel files. The evaluations were then coded. In the coding process, separate units were identified for each written evaluation. Units were defined as (parts of) sentences containing a (implied) subject and a predicate. For instance, sentences like: "You are sporty and know a lot about exercise" were separated into two units: 1) "You are sporty", 2) "(and) know a lot about exercise", even when the subject was not repeated. Each unit was coded in three aspects: (1) performance-relatedness, (2) valence, and (3) language abstraction. Units were excluded from coding when they: (A) described a neutral situation ("You had music lessons on Tuesday"); (B) related to someone else or to multiple pupils ("The class made a nice play"); (C) characterized the teacher ("I look forward to next year"). In total, 24% of units were excluded from coding on language abstraction.

Valence was coded for each unit. Positive ("You are very good at collaborating") and negative ("She often doubts herself") units were further analyzed. Neutral units were excluded. A small number of units was excluded because they were ambivalent (< 1%; "You do not enjoy being in the spotlight very much"). The relative number of positive units was calculated as the percentage of positive units out of the total number of included units.

Performance-relatedness was a binary category. A unit received a score of 1 if the unit directly related to scholastic skills ("You are good at spelling"). If a unit related to other categories, like behavior ("You do your best") or perceived characteristics ("You are sweet"), a score of 0 was given. The relative number of performance-related units was calculated as the percentage of performance-related units out of the total number of included units.

Language abstraction was coded using the LCM (Semin & Fiedler, 1988), adapted for the current purpose in accordance with Menegatti et al. (2017) and Watson and Gallois (2002). This model consists of four categories, three of which are verb categories, and the last and most abstract category consists of adjectives. All units in which the (inferred) subject was or concerned the pupil were coded as one of the four categories. If a unit contained an *adjective* (ADJ), the unit was coded as the most abstract category ("You are an incredible boy"; "You are a creative thinker"). If a unit did not contain an adjective, the main verb was coded. The most concrete and objective verb category in the LCM is formed by *descriptive action verbs* (DAVs), which refer to observable actions with a beginning and end ("You wrote one book report"; "She giggles in class"). The next, slightly less concrete verb category are the *interpretative action verbs* (IAVs). IAVs are again actions with a beginning and end, but are slightly less objective than DAVs in the sense that they are not unambiguously observable and have a positive or negative connotation, i.e., require interpretation ("You help the teacher"; "You ridicule others"). The next and least concrete verb category, *state verbs* (SVs), do not refer to an action

but rather express an enduring emotional or cognitive state ("She loves her friends"; "He understands it"). Each unit received a score for language abstraction accordingly (DAV = 1, IAV = 2, SV = 3, ADJ = 4). An average score of language abstraction was calculated for positive and negative units, that either were related or unrelated to the student's performance, resulting in four mean scores. The possible range was 1 to 4, with a higher score representing a higher level of abstraction. Additional examples of the four linguistic categories, combined with valence and performance-relatedness, can be found in Table 4.1.

Table 4.1

Examples of Positive and Negative, Performance-Related and Performance-Unrelated Units at each Abstraction Level.

	Abstr. level	Positive	Negative
Performance-related	1 - low	You watched the videos.	You did not finish your bookmark.
	2	You anticipate your opponent.	You forget the letter 'n'.
	3	You know the times tables.	You do not understand it.
	4 - high	Your essay was very strong.	Your spelling is sloppy.
Performance-unrelated	1 - low	You raise your hand in class.	You did not bring your presentation.
	2	You show perseverance.	You can't concentrate.
	3	You enjoy school.	You dislike working together.
	4 - high	You are a funny boy.	You should have a more confident attitude.

Note. The abstraction levels relate to 4 linguistic categories: 1) Descriptive Action Verbs (DAV), 2) Interpretative Action Verbs (IAV), 3) State Verbs (SV), 4) Adjectives (ADJ). The relevant verbs and adjectives have been underlined.

The first and third author developed a codebook (in Dutch, available through the supplemental materials, see Appendix D). Subsequently four graduate students were trained and independently coded a subset of data

(15 report cards, consisting of 284 units, 5% of the total number of units). The coders first selected the units and coded the performance-relatedness (0/1) and valence (+/-) for all report cards. Agreement with the head coders was >95%. Coders then coded abstraction (1-4), for which intercoder reliability was satisfactory (Cohen's κ = .77 - .88).

Data analysis

A priori power analysis with G*power 3.1 (Faul et al., 2007) showed that with power of .80 and α set at 0.05 a total sample of at least N = 82 was required to detect the smallest effects of interest found in previous studies, i.e., an effect size of η_p^2 = .07 (Gentrup et al., 2020; Menegatti et al., 2017). Further analyses were carried out using SPSS 27.

Welch's t-tests and a split-plot ANOVA were used to investigate the effect of gender on the relative number of positive units and performance-related units. Part of the sample received solely positive units (n = 71 pupils, 40 girls), resulting in missing values for the mean abstraction score of negative units. Excluding these pupils not only decreases the power of the study, but it also excludes precisely those pupils that were evaluated most positively. Following previous studies, missing values for negative abstraction have therefore been replaced with a score of zero, which signifies a very concrete level of negative abstraction (Menegatti et al., 2017; Rubini, Moscatelli, Albarello, et al., 2007; Rubini, Moscatelli, & Palmonari, 2007). A split-plot ANOVA was used to investigate the effects of gender (boy, girl), and valence (positive, negative), on language abstraction in written evaluations. Part of the sample (n = 60, 37 boys) received only units unrelated to performance, 4 girls received only performance-related units. These scores were not replaced. As effect size we report $\widehat{\omega}_p^2$, formulas and full ANOVA tables can be found in the supplemental materials (see Appendix B and D). Marginal effects were investigated through pairwise comparisons of the estimated marginal means. SPSS 27 calculates an adjusted p-level for multiple comparisons, so that significance can be accepted when $p < .05$.

All variables were inspected for outliers, for groups $n \geq 100$ defined as values more than 3.3 *SD* above or below the mean, and for groups $n < 100$ defined as values 2.58 *SD* above or below the mean (Tabachnick & Fidell, 2007a). Outliers on main variables were winsorized, meaning they were brought closer to the distribution while maintaining the same rank. Results after winsorizing are reported.

Results

The mean proportions of positive, negative, performance-related, and performance-unrelated units can be found in Table 4.2. The effect of gender on the relative number of positive units and performance-related units was investigated. The assumption of homogeneity of variances was violated for both positive units and performance-related units, as assessed by Levene's test for equality of variances ($p < .001$, and $p = .023$ respectively), so Welch's *t*-tests were used. Girls' report cards contained a significantly higher percentage of positive units than boys report cards, mean difference = 7.81, 95% CI [3.06, 12.55], $t(180.79) = 3.24$, $p = .001$, Cohen's $d = .45$, 95% CI [0.17, 0.73]. Girls' report cards also contained a significantly higher percentage of performance-related units than boys' report cards, mean difference = 6.90, 95% CI [1.02, 12.78], $t(184.18) = 2.32$, $p = .022$, Cohen's $d = .33$, 95% CI [0.05, 0.60]. Because each unit was either positive or negative, the effect of gender for negative units was practically the same as the effect for positive units. Similarly, the effect of gender was practically the same for performance-related and performance-unrelated units.

To investigate possible interactions of gender, valence, and performance-relatedness, a split-plot ANOVA was used. No three-way interaction was found. Direct effects of valence ($F(1, 201) = 799.98$, $p < .001$, $\widehat{\omega}_p^2 = .797$), and performance-relatedness, ($F(1, 201) = 383.32$, $p < .001$, $\widehat{\omega}_p^2 = .653$) indicated that report cards contained more positive than negative units, and more performance-unrelated than performance-related units.

There was a significant two-way interaction between performance relatedness and valence, $F(1, 201) = 143.55$, $p < .001$, $\widehat{\omega}_p^2 = .413$, which is not further discussed because it is not relevant for the hypotheses of this paper. Significant two-way interactions between gender and valence ($F(1, 201) = 10.22$, $p = .002$, $\widehat{\omega}_p^2 = .043$), and between gender and performance ($F(1, 201) = 5.81$, $p = .017$, $\widehat{\omega}_p^2 = .023$), confirmed the findings from the t-tests. Pairwise comparisons of marginal effects revealed that girls received significantly more positive performance-related units than boys, $F(1, 201) = 8.36$, $p = .004$, $\widehat{\omega}_p^2 = .035$, mean difference = 7.66, 95% CI [2.44, 12.88]. Additionally, girls received significantly less performance-unrelated negative units than boys, $F(1, 201) = 10.14$, $p = .002$, $\widehat{\omega}_p^2 = .043$, mean difference = -7.09, 95% CI [-11.47, -2.70]. There was no significant difference between boys and girls for positive performance-unrelated units, nor for negative performance-related units.

Means and standard deviations of positive and negative abstraction scores are reported in Table 4.3. A split-plot ANOVA was carried out to investigate the effects of gender (boy, girl), and valence (positive, negative), on language abstraction in written evaluations. In line with previous research (Menegatti et al., 2017), a main effect for valence was found, indicating that positive units had a higher mean abstraction score than negative units ($F(1, 201) = 233.16$, $p < .001$, $\widehat{\omega}_p^2 = .534$). For our hypothesis, an interaction effect of gender and valence was relevant, but no interaction was present. There was a main effect of gender, $F(1, 201) = 3.91$, $p = .048$, $\widehat{\omega}_p^2 = .014$; indicating that units on boys' report cards were slightly more abstract on average, regardless of valence. Pairwise comparisons indicated that for negative units, the effect of gender was in the expected direction, with boys receiving more abstract negative units than girls. However, this effect was not significant ($F(1, 201) = 3.06$, $p = .082$, $\widehat{\omega}_p^2 = .010$), and the 95% CI for the mean difference of 0.37 included zero, [-.05, 0.80].

Table 4.2

Mean (Standard Deviation) Proportions for Positive and Negative Units Related and Unrelated to Performance (Study 1)

		Boys		Girls			Total	
	n	104		99			203	
Total	positive	79.85	(20.23)	87.66	(13.55)	**	83.66	(17.69)
	negative	20.12	(20.20)	12.20	(13.42)	**	16.27	(17.64)
	perf.-related	17.67	(18.28)	24.57	(23.70)	*	21.03	(21.33)
	perf.-unrelated	82.30	(18.28)	75.32	(23.63)	*	78.90	(21.30)
Positive	perf.-related	13.73	(14.65)	21.39	(22.46)	**	17.46	(19.21)
	perf.-unrelated	66.01	(25.07)	65.74	(24.75)		65.88	(24.86)
Negative	perf.-related	3.63	(6.67)	2.86	(5.42)		3.35	(6.09)
	perf.-unrelated	16.29	(18.85)	9.20	(11.90)	**	12.83	(16.20)

Note. The differences between the top four rows have been tested with Welch's *t*-tests. The differences between the bottom four rows have been tested through the estimated marginal means. Significant differences between boys and girls have been denoted with asterisks.
* $p < .050$, ** $p < .010$

Table 4.3

Mean (Standard Deviation) Abstraction Scores for Positive and Negative Units for Male and Female Pupils (Study 1)

Variable	Boys	Girls
n	104	99
Positive units	3.64 (*0.29*)	3.60 (*0.31*)
Negative units	2.13 (*1.49*)	1.76 (*1.55*)

Note. A higher score signifies a higher level of abstraction. Scores for negative units could range between 0 and 4, scores for positive units could range between 1 and 4.

Conclusion Study 1

Girls received more positive units and more performance-related units than boys did. Girls especially received more positive performance-related evaluations than boys, partially confirming Hypotheses 1.1A and 1.1B. We expected girls to also receive more performance-related negative evaluations than boys, but there was no significant difference. Additionally, we expected girls to also receive more positive performance-unrelated evaluations than boys, but this was not the case. However, boys did receive more negative performance-unrelated evaluations, which could be seen as other side of the same coin. These results indicate positive bias towards girls. We expected girls to be evaluated with more abstract positive terms and more concrete negative terms than boys, but we did not find this in the data. Instead, we found that boys were evaluated slightly more abstract in general. Hypothesis 1. 2 was therefore rejected.

Study 2

In this second study, we investigated the relation of pupil ethnicity with teacher evaluations, and explored the intersection of ethnicity and gender. We will use the term ‘ethnic majority’ to indicate White people who were born in the Netherlands, and whose parents and grandparents were born in the Netherlands. Majority refers to group size, as this is the largest ethnic group in the Netherlands. We will use the term ‘ethnic minority’ to indicate people who were not born in the Netherlands, and/or whose (grand)parents were not born in the Netherlands; following the definition used by the Dutch Central Bureau of Statistics (2016). Ethnic minority pupils do not form one homogenous ethnic group, but rather belong to many different smaller ethnic groups. However, ethnic minority pupils share the status of a (often marginalized) outgroup in relation to the dominant majority. Any person from an ethnic minority background can be seen as 'the Other' by members of the dominant ethnic group,

without distinctions regarding specific backgrounds (Jensen, 2011). This generalized otherness of ethnic minority individuals is also reflected in governmental categorization of citizens, when people are seen as either 'native' or 'non-native' (Van Schie, 2018). The simple fact of being an ethnic minority, regardless of which specific one, has been shown to be relevant for opportunities, performance, and wellbeing across different domains of education (e.g., Kleen & Glock, 2018; Menegatti et al., 2017; OECD, 2018).

While findings on the effect of pupil ethnicity on teacher bias have been inconsistent, negative bias towards ethnic minority pupils has been most common (Wang et al., 2018). Therefore, we hypothesized that: 2.1A) ethnic majority pupils would receive more positive evaluations than ethnic minority pupils, both related and unrelated to performance; 2.1B) ethnic majority pupils would receive more performance-related evaluations than ethnic minority pupils, both positive and negative; 2.2) ethnic majority pupils are evaluated with more abstract positive terms and more concrete negative terms than ethnic minority pupils. Additionally, because findings on the relation between pupil gender and ethnicity have been inconsistent (compare for instance, Farris & de Jong, 2014; Glock & Klapproth, 2017; Kleen & Glock, 2018; Menegatti et al., 2017), we explored the possible interactions between ethnicity and gender, and the main variables.

Method

Sample

For the ethnic majority group, a subsample of the Study 1 participants was used (n = 92, 47 boys), including 15 sibling-pairs. Parents of children from ethnic minority groups were recruited for participation through social workers, (weekend) schools, social media, and the personal network of the researchers. To participate, parents were asked to confirm that their child belonged to an ethnic minority group (the exact questions and definitions used can be found in the questionnaire in the supplemental materials).

Parents were eligible to participate if they had a child of whom they could share a report card that was written during sixth, seventh, or eight grade of primary school. The report card had to be written during the same period as the ethnic majority group (2016-2022). Out of 81 parents who indicated to be willing to participate, 51 were able to send a report card and fill in the questionnaire. Some participants were excluded because the report card was from a lower grade (n = 4), and some because none of the written remarks were codable for language abstraction (n = 3). This ultimately resulted in a sample of 44 (24 boys). The total sample size of Study 2 was 136. In the ethnic minority sample, adoption was not an exclusion criterion; 2 of the pupils were adopted. This sample contained 3 sibling-pairs. The education level of the parents in the ethnic minority sample was slightly higher than the distribution in the Dutch population, with 12% of parents having received low, 58% medium, and 30% higher-level education. The ethnic majority sample was selected based on the educational level of mothers and fathers, matched as closely to the ethnic minority sample as possible. Differences were assessed with t-tests. There were no significant differences between the two samples for parental education level ($t(58.20) = 1.63$, $p = .109$). The entire sample of pupils attended 111 different schools. For pupils who attended the same school, we checked whether the report cards could have been written by the same teacher. This was the case for one pair of (nonrelated) ethnic majority pupils. Excluding one of these pupils from the analyses did not result in any significant differences, so results are reported including both pupils.

Procedure

The data of ethnic minority pupils was collected through a short online questionnaire in Dutch. Some questionnaires were administered on paper or by telephone, for parents who were unable to complete the questionnaire online due to language or digital barriers (n = 5). Report cards were photographed and sent digitally. Participants received a small gift after participation was completed. Ethical approval for this research

was provided by the research ethics committee of the researchers' host institute. As the ethnic majority pupils were a subsample of Study 1, the procedure for them was identical as the one described there.

Instruments

Written teacher evaluations were coded for valence, performance-relatedness, and language abstraction as described in Study 1. As names were deleted from the text, coders were unaware of ethnicity while coding.

One of the report cards included a small portion written in Frisian, which was translated by a native speaker. Part of the sample received solely positive units (n = 40), resulting in missing values for the mean abstraction score of negative units. In the ethnic minority group 8 pupils did not receive negative remarks (4 girls), and in the ethnic majority group 32 pupils (19 girls). More pupils in the ethnic majority group received only positive units than pupils in the ethnic minority group (assessed with Welch's t-test; $t(102.22) = 2.15$, mean difference 17%, CI 95% [0.1, 31.9], $p = .034$, Cohen's $d = .37$, 95% CI [0.01, 0.73]).

Ethnicity was a binary variable (i.e., belonging to the ethnic majority or not). Ethnicity was assigned by self-determination, as well as by the birth country of the pupil, their parents, and grandparents. Following the governmental definition of having a migration background at the time of the study, pupils were included if they or at least one of their (grand)parents was born outside of the Netherlands (Dutch Central Bureau of Statistics, 2016; Van Schie, 2018). There were 6 pupils with a first-generation migration background (i.e., born outside the Netherlands), 31 with a second, and 5 with a third-generation migration background; 2 pupils were adopted. Countries and regions of origin apart from the Netherlands were Suriname (n=13), Morocco (n=7), Turkey (n=6), South Asia (n=6), Middle East (n=5), East Asia (n=5), Europe (n=4), Southeast Asia (n=3), North America (n=1), Lesser Antilles (n=1), and West Africa (n=1).

Data Analysis

A priori power analysis with G*power 3.1 (Faul et al., 2007) showed that with power of .80 and α set at 0.05 a total sample of at least $N = 72$ was required to detect the smallest effects of interest found in previous studies, i.e., $\eta_p^2 = .04$ (Gentrup et al., 2020; Menegatti et al., 2017). Further analyses were carried out using SPSS 27. The effects of ethnicity and gender on the main variables were investigated with between-group and split-plot ANOVAs. Full ANOVA tables can be found in the supplementary materials. Marginal effects were investigated through pairwise comparisons of the estimated marginal means. SPSS 27 calculates an adjusted *p*-level for multiple comparisons, so that significance can be accepted when $p < .05$ and no further adjustments have to be made to the *p*-value.

All variables were inspected for outliers, for groups $n < 100$ defined as values more than 2.58 *SD* above or below the mean (Tabachnick & Fidell, 2007a). Outliers on main variables were winsorized, meaning they were brought closer to the distribution while maintaining the same rank. Results after winsorizing are reported. According to the distinction between "western" and "non-western" migrant origins (Roeleveld et al., 2011; Van Schie, 2018), 5 ethnic minority pupils (3 girls) had a 'western' background. Out of these pupils, 4 did not receive any negative remarks on their report cards (2 girls). This group is too small to be included as a separate level for ethnicity in the ANOVAs. However, we did run all the analyses with these 5 pupils excluded, yielding similar results. Full results excluding "western" ethnic minority pupils can be found in the supplemental materials.

Results

The effects of ethnicity and gender on the proportion of positive units were investigated with a between-group ANOVA. Means and standard deviations are reported in Table 4.4. There was a main effect of ethnicity

Table 4.4

Mean (SD) Proportions of Positive and Performance-Related Units on Report Cards of Male and Female Pupils from Ethnic Minority and Ethnic Majority Backgrounds (Study 2)

		Ethnic minority						Ethnic majority					
		Boys		Girls		Total		Boys		Girls		Total	
	n	24		20		44		47		45		92	
Total	positive	73.39	(20.35)	73.04	(21.56)	73.23	(20.66)	78.04	(20.92)	88.79	(13.02)*	83.30	(18.23)
	negative	26.61	(20.35)	26.75	(21.55)	26.68	(20.66)	21.95	(20.92)	10.96	(12.91)	16.58	(18.23)
	performance-related	19.61	(16.97)	25.22	(19.91)	22.16	(18.36)	17.61	(16.58)	23.19	(23.39)	20.40	(20.39)
	performance-unrelated	80.39	(16.97)	74.57	(19.78)	77.74	(18.32)	82.28	(16.87)	76.03	(24.83)	79.22	(21.26)
Positive	performance-related	14.95	(12.90)	16.17	(15.37)	15.50	(13.92)	14.55	(14.83)	20.86	(22.82)	17.64	(19.31)
	performance-unrelated	58.42	(25.45)	56.87	(21.26)	57.71	(23.39)	63.50	(25.76)	67.35	(25.31)	65.38	(25.48)
Negative	performance-related	4.61	(7.56)	9.05	(11.24)*	6.63	(9.56)	2.98	(5.64)	2.02	(3.81)	2.51	(4.83)
	performance-unrelated	21.90	(18.29)	17.70	(18.02)	19.99	(18.08)	18.75	(19.88)	8.62	(10.77)*	13.80	(16.78)

Note. Values marked with * differ significantly from the other values in the same row.

on the total number of positive units, $F(1, 132) = 8.86$, $p = .003$, $\widehat{\omega}_p^2 = .055$, with ethnic majority pupils receiving more positive units than ethnic minority pupils, mean difference = 10.21, 95% CI = [3.42, 16.99]. Pairwise comparisons revealed that ethnic majority girls received more positive units than all other groups ($p < .010$, $\widehat{\omega}_p^2$ ranging between .046 and .061, and mean differences ranging between 10.74 and 15.75). There were no differences between the other groups. This indicates an interaction between gender and ethnicity, but this interaction was not significant ($F(1, 132) = 2.62$, $p = .108$, $\widehat{\omega}_p^2 = .012$). As the effect size could be interpreted as a small effect (Kirk, 1996), non-significance could be due to lack of power. The effects of ethnicity and gender on the proportion of performance-related units was investigated with a between-groups ANOVA as well, there were no significant effects.

To further investigate possible interactions between ethnicity, gender, valence, and performance-relatedness of the written evaluations, a four-way split-plot ANOVA was used. There were no significant four-way or three-way interactions. Like in Study 1, there were significant main effects of valence ($F(1, 132) = 279.55$, $p < .001$, $\widehat{\omega}_p^2 = .676$) and performance ($F(1, 132) = 244.61$, $p < .001$, $\widehat{\omega}_p^2 = .646$), indicating that report cards contained more positive than negative units, and more performance-unrelated than performance-related units. There was a significant two-way interaction between performance relatedness and valence, $F(1, 132) = 143.55$, $p < .001$, $\widehat{\omega}_p^2 = .413$, which is not further discussed because it is not relevant for the hypotheses of this paper.

There was a significant two-way interaction between valence and ethnicity, $F(1, 132) = 9.01$, $p = .003$, $\widehat{\omega}_p^2 = .057$, confirming the results from the between-group ANOVA. Pairwise comparisons revealed that ethnic minority girls received significantly more negative performance-related units than all other groups ($p < .05$, mean differences ranging between 4. 44 and 7.03). These are negative remarks that concern performance (e.g., "Your fine motor skills are weak"). Additionally, ethnic majority girls received less negative performance-unrelated units than all other groups

(p < .05, mean differences ranging between 9.08 and 13.28). These are negative remarks that concern mainly personal attributes (e.g., "You are a messy kid"). No other marginal effects were found.

A split-plot ANOVA was carried out for each model to investigate the effects of ethnicity (minority, majority), gender (boy, girl), and valence (positive, negative), on language abstraction in written evaluations. Means and standard deviations of positive and negative abstraction scores are reported in Table 4.5.

Table 4.5

Mean (SD) Abstraction Scores for Positive and Negative Units for Male and Female Pupils from Ethnic Minority and Ethnic Majority Backgrounds (Study 2)

	Ethnic minority		Ethnic majority	
Variable	Boys	Girls	Boys	Girls
N	24	20	47	45
Positive units	3.62 (0.38)	3.59 (0.24)	3.58 (0.33)	3.56 (0.33)
Negative units	2.74 (1.31)	2.35 (1.23)	2.15 (1.45)	1.60 (1.48)

Note. A higher score signifies a higher level of abstraction. Scores for negative units could range between 0 and 4, scores for positive units could range between 1 and 4.

There was a main effect of valence, indicating that positive units had a higher mean abstraction score than negative units in general, $F(1, 132) = 101.55$, $p < .001$, $\widehat{\omega}_p^2 = .287$. There was a significant main effect for ethnicity, $F(1, 132) = 7.54$, $p = .007$, $\widehat{\omega}_p^2 = .047$, which was qualified by the significant two-way interaction between ethnicity and valence, $F(1, 132) = 5.37$, $p = .022$, $\widehat{\omega}_p^2 = .017$. There was a significant simple main effect of ethnicity for negative units, but not for positive units. Negative units on report cards of ethnic minority pupils were more abstract than negative units on report cards of ethnic majority pupils ($F(1, 132) = 6.71$, $p = .011$, $\widehat{\omega}_p^2 = .040$, mean difference = 0.67, 95% CI [0.16, 1.18]). Pairwise comparisons revealed that the negative units received by ethnic majority girls were significantly

more concrete than those received by ethnic minority boys, $p = .002$, and ethnic minority girls, $p = .048$. Additionally, they appeared to be more concrete than those received by ethnic majority boys, but this difference was not significant, $p = .059$. There was no difference between ethnic minority girls and ethnic majority boys, nor between ethnic minority boys and ethnic majority boys. There were no significant effects for positive units.

Conclusion Study 2

The report cards of ethnic majority girls contained more positive units than those of all other groups, which indicates a positive bias towards them. This effect was present overall, but not at the marginal level. This indicates that the difference in positive units is not specifically driven by performance-related or -unrelated evaluations. Thus, Hypothesis 2.1A is partially accepted. Report cards of ethnic majority girls contained less performance-unrelated negative units than all other groups. Additionally, report cards of ethnic majority pupils more often contained solely positive units than report cards of ethnic minority pupils.

There was no difference in the number of performance-related units in general. Report cards of ethnic minority girls contained more performance-related negative units than all other groups. This result was unexpected, Hypothesis 2.1B is rejected.

We expected that ethnic minority pupils would be evaluated with more concrete positive terms, and more abstract negative terms than ethnic majority pupils. We found that negative units on ethnic minority pupils' report cards are indeed more abstract than those on ethnic majority pupils' report cards. Pairwise comparisons revealed that this difference was only significant when ethnic minority pupils were compared to ethnic majority girls. Hypothesis 2.2 is partially accepted.

General discussion

In the present studies we investigated teacher bias through the valence, performance-relatedness, and language abstraction of written pupil evaluations. Generally, patterns of teacher bias driven by pupil gender and ethnicity have indicated that girls and ethnic majority pupils are advantaged by positive bias, but these findings have not been consistent across different studies (Geven et al., 2018; Wang et al., 2018). Written evaluations have been found to be a useful but underused source to investigate intergroup bias (Beukeboom, 2014; Menegatti et al., 2017; Ni & Li, 2013). Our findings indicate positive bias towards ethnic majority girls.

Valence and performance-relatedness

According to previous studies, holding high expectations is beneficial for all pupils, but especially those from marginalized groups (Rubie-Davies, 2015; Wang et al., 2018; Weinstein et al., 2004). Effective evaluations should include both positive and negative comments, focusing on performance and offering clear guidance on how pupils can improve their achievements in the future (Guskey, 2019; Hattie & Timperley, 2007; Hyland & Hyland, 2006). If all report cards would satisfy these ideas, it could be argued that we should not find differences between pupils in the proportions of positive/negative and performance-related/unrelated remarks. However, this was not what we found.

In the current studies, we found that ethnic majority girls received more positive evaluations than all other groups. In Study 1, the difference between (ethnic majority) boys and girls for the proportion of positive statements was explained by the performance-related statements. In Study 2, ethnic majority girls received more positive evaluations overall, but not at the performance-related (e.g., "Your project had a logical structure") and -unrelated (e.g., "You are a clever gal") level specifically. While ethnic majority girls received both more positive performance-related and performance-unrelated evaluations than all other groups, these differences were not significant. This is likely due to lack of power.

Additionally, we found that ethnic majority girls received the least negative performance-unrelated evaluations (e.g., "You should wait your turn"). This fits the "good girls, bad boys" stereotype, which encompasses the idea that girls are better behaved, more motivated, and have better work habits than boys (Glock & Klapproth, 2017; Glock & Kleen, 2017; Myhill & Jones, 2006; Timmermans et al., 2016, 2018; Voyer & Voyer, 2014). Overall, positive bias towards girls in teacher expectations has been found regularly in the past (Myhill & Jones, 2006; Voyer & Voyer, 2014). Our findings suggest that positive gender-based bias only works for ethnic majority girls, which is consistent with some other studies (Glock & Klapproth, 2017; Menegatti et al., 2017), but not with studies in which ethnic minority girls were evaluated more positively than ethnic minority boys (Farris & de Jong, 2014), or vice versa (Kleen & Glock, 2018).

We found that report cards of ethnic minority girls contained more negative performance-related statements than the report cards of other pupils. This could be interpreted in two ways. Firstly, we could interpret this as an indicator of positive bias towards ethnic minority girls compared to other groups, as teachers tend to give more performance-related feedback to pupils of whom they have higher expectations (Gentrup et al., 2020), and negative feedback can be very beneficial for improving performance (Hattie & Timperley, 2007). However, as performance-related feedback is not given often on the report cards in our studies, and negative performance-related feedback is especially rare, it appears that the way Dutch teachers write evaluations does not concur with the guidelines that stem from research (Hattie & Peddie, 2003; Hattie & Timperley, 2007; Hyland, 2013; Hyland & Hyland, 2006). In that sense, the higher number of negative performance-related comments on the report cards of ethnic minority girls can be an indicator of bias against them.

Language abstraction

According to the Linguistic Category Model (LCM), a higher level of abstraction implies stable and (stereo)typical characteristics, suiting expectations held by the person who wrote the message, while a lower level of abstraction implies unexpected, situational, and temporary events (Beukeboom, 2014; Maass et al., 1989; Wigboldus et al., 2000). Based on the LCM, we expected that girls would be evaluated with more abstract positive terms and more concrete negative terms than boys; and that ethnic minority pupils would be evaluated with more concrete positive terms, and more abstract negative terms than ethnic majority pupils (cf. Menegatti et al., 2017). In Study 2, ethnic minority pupils were evaluated with more abstract negative terms than ethnic majority girls were. This linguistic difference can be driven by ethnicity-based biased expectations through cognitive linguistic processes that take place outside of teachers' awareness, and by motivational linguistic processes, because teachers may be implicitly motivated to retain ethnicity-based social structures (Beukeboom, 2014; Semin, 2000). We did not find any effect on statements with positive valence, nor did we find any gender effects. While the mean abstraction scores did indicate that ethnic majority girls were evaluated with more concrete negative terms than ethnic majority boys were, this difference was not significant. There are several possible explanations for our unexpected results.

Firstly, the differences could be explained by culturally specific gender bias and gender stereotypes. For example, Italians have been found to show a somewhat higher endorsement of some gender stereotypes, and a somewhat lower level of agreement with gender equality norms than the Dutch (Halman et al., 2022).

Secondly, it appears that Dutch teachers used more adjectives than Italian teachers, especially for positive evaluations. This is reflected in the very high mean abstraction score in our studies, which is around 3.60. In the studies by Menegatti et al. (2017), the mean abstraction score for positive evaluations is considerably lower, around 2.90. While it is known

that there are cultural differences in the use of adjectives and verbs, these differences are generally small between speakers/writers from Western countries like Italy and the Netherlands (Maass et al., 2006). It would be interesting to examine why Dutch teachers are more inclined to use adjectives when writing evaluations.

Lastly, the mean abstraction scores are strongly affected by the number of report cards that did not contain any statements of positive or negative valence. Following previous studies of language abstraction, when a pupil did not receive any statements in either category, they received a score of zero, indicating a very concrete score (Menegatti et al., 2017; Rubini, Moscatelli, Albarello, et al., 2007; Rubini, Moscatelli, & Palmonari, 2007). In line with the LCM, this means that a pupil who received no negative statements receives a negative abstraction score of zero, signaling that negative evaluations are unexpected. In our study, a considerable number of pupils received no negative statements, and this was true for significantly more ethnic majority pupils than ethnic minority pupils. In the Italian study, this pattern was present too, but additionally there were pupils who received no positive statements, and these were almost exclusively ethnic minority pupils, and ethnic minority girls more so than boys were. Furthermore, the differences in missing values between groups were far larger in the study by Menegatti et al. (2017), artificially increasing the effect of gender and ethnicity on language abstraction.

Implications for the school context

Effective evaluations focus on performance and include both positive and negative comments (Guskey, 2019; Hattie & Timperley, 2007; Hyland & Hyland, 2006). However, we found that in general, over 80% of the remarks on the report cards were positive, and around 80% of the remarks were unrelated to performance. The large number of positive statements, especially unrelated to performance, is not in the best interest of pupils. Many of these statements concerned personal attributes, a type of praise

that is not beneficial for performance (Hattie & Timperley, 2007). Both positive and negative statements on personal attributes, especially when they are made up of generic language, may foster the idea that performance and behavior are mainly related to stable traits and characteristics, and not to the effort made by pupils (Cimpian, 2010). This is especially true when these statements are made at the highest level of abstraction (Menegatti et al., 2017), which was generally the case in our studies. The notion that performance and behavior are mainly related to stable attributes can have adverse effects on the self-esteem and motivation of pupils. It is therefore important that teachers learn about these effects of language. While it is hard to inhibit linguistic bias in direct verbal interaction (Beukeboom, 2014), teachers can be trained to use more concrete language when writing evaluations (Menegatti et al., 2017). Moreover, the most effective evaluations contain both positive and especially negative evaluations, providing pupils with specific information on what they should improve and how they can improve it (Guskey, 2019; Hattie & Timperley, 2007).

The absence of performance-related statements and negative statements on many report cards indicates that perhaps teachers are not effectively prepared during teacher training to write effective evaluations. Additionally, the proportions of valence and performance-relatedness pose normative questions about what goals teachers have with their evaluations. Research shows that there are gaps in our knowledge about how teachers form their evaluations, and about teachers' beliefs and perceptions of what should be included in effective evaluations (McMillan, 2019). Research into teachers' evaluative processes could clarify how they employ written evaluations and what types of messages they aim to convey. Several studies indicate that pupils and parents regularly find that the standards to which the child is measured are unclear, and that teacher comments are not understood (Hattie & Peddie, 2003; Hyland, 2013; Tuten, 2007). When evaluative standards are vague, this can increase the effect of teacher bias (Quinn,

2020). Educational research, schools, teachers, pupils, and parents could benefit from discussions about how evaluations should be formed and what they mean (Hattie & Peddie, 2003; McMillan, 2019).

Besides improvements upon evaluative processes, gender- and ethnicity-based bias can be combatted through various types of interventions. For instance, bias has been shown to decrease through intergroup contact, as intergroup contact increases knowledge and empathy towards members of other groups (Martin et al., 2017; Pettigrew et al., 2011). Additionally, intergroup contact leads to seeing others more as individuals than as a homogenous group. Some research indicates that this also decreases linguistic bias (Prati et al., 2015).

Limitations and future directions

Our studies and analyses had some limitations. It was impossible to investigate the relation of valence, performance-relatedness, and language abstraction with actual performance. This is in part due to the Dutch system, in which schools are allowed to choose between a variety of standardized tests from multiple commercial providers, schools are free to decide how many standardized tests they want to conduct (Ministerie van Onderwijs, Cultuur en Wetenschap, 2014). The only mandatory test is a school leavers' test, which is made in the final year of primary school, and results in a track recommendation for secondary school. Research has shown that the test scores across test providers are incomparable (Van Baars, 2022). Additionally, not all schools share the standardized test results on pupils' report cards. While we did have data of the school leavers' test for some pupils in Study 1 (n = 68), these results came from five different test providers. For the most popular standardized test (n = 40), we found no correlations between test score and the main variables (valence, performance-relatedness, and language abstraction, p > .450), indicating that differences between written evaluations were not the result of differences in performance. In future studies, it would be beneficial either to conduct a standardized cognitive

test among participating pupils, or to recruit report cards that include results on standardized tests by the same provider.

The sample was too small to find differences between different ethnic minority groups, or differences between pupils with a first-, second-, or third generation immigration background. Although ethnic minority groups are connected through their marginalized status, this may mask bias that is specific to certain ethnic minority groups (Gillborn et al., 2018). We did run the analyses excluding "western" ethnic minority pupils (e.g., European, North American), yielding similar results, but with larger effect sizes for the proportions of positive and performance-related evaluations (which can be found in the supplemental materials). These results indicate that teachers may not (only) differentiate between pupils based on majority-minority status, but also based on specific ethnic minority background. In the future, it may be worthwhile to employ a different recruiting strategy, for instance by recruiting through schools instead of through families. This could result in a larger and more representative sample.

We did not investigate the content of evaluations beyond their valence and performance-relatedness. This means we can only draw limited conclusions about what our results mean. In Study 2, ethnic minority girls received the highest number of negative performance-related remarks. This type of evaluation can be very useful for future achievement, but this is only true when the evaluations are clear and specific. However, studies have shown that teacher comments are usually formulated in vague and general terms, thus not benefiting the pupil (Hattie & Peddie, 2003; Hyland, 2013). Content analysis would be necessary to evaluate whether the negative performance-related statements written by the teachers were useful. Additionally, previous studies into written evaluations have shown that bias is present at the content level, through differential word use based on gender and ethnicity (Biernat et al., 2012; Rojek et al., 2019).

In future research, content analysis could be employed to investigate what types of performance-related and performance-unrelated aspects

are discussed exactly, and how these aspects are associated with gender and ethnicity. Teachers' perceptions of pupil behavior (Mason et al., 2014; Myhill & Jones, 2006) and reactions to pupil behavior (Bašaragin & Savic, 2019; Frawley, 2005) are known to show bias. Combining research on written evaluations with observational studies can help disentangle what parts of evaluations are based on bias and what parts are based on actual pupil behavior.

Conclusion

By using a combination of methods to study bias through linguistic habits of teachers, we have gained important new insights into how gender- and ethnicity-based bias is present in teacher evaluations. As ethnic majority girls were evaluated more positively in written teacher evaluations, our study indicates that boys and ethnic minority pupils are disadvantaged. Teacher bias can have detrimental as well as uplifting effects on many aspects of the (future) functioning of pupils. Our study emphasizes the need for continuous attention to the fair treatment of pupils.