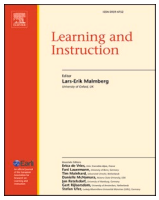
Learning and Instruction 104 (2026) 102391



Contents lists available at ScienceDirect

Learning and Instruction

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



Gender and behavior biases in teachers' perceived academic level of students

Lilas Gurgand [a,*], Hugo Peyre [a,b,c], Xavier Thierry [d], Franck Ramus [a]

[a] Laboratoire De Sciences Cognitives Et Psycholinguistique, Département d'Etudes Cognitives, Ecole Normale Supérieure, EHESS, CNRS, PSL University, Paris, France
[b] Autism Reference Centre of Languedoc-Roussillon CRA-LR, Excellence Centre for Autism and Neurodevelopmental disorders CeAND, Montpellier University Hospital, MUSE University, France
[c] CESP, INSERM U1178, Centre de recherche en Épidémiologie et Santé des Populations, Villejuif, France
[d] French Institute for Demographic Studies (Ined), French Institute for Medical Research and Health (Inserm), French Blood Agency, ELFE Joint Unit, Paris, France

ARTICLE INFO

Keywords:
Gender biases
Teachers' biases
Mathematics
Language
Externalizing behavior

ABSTRACT

Background: Teachers' perception of students' abilities can be influenced by gender and classroom behavior, potentially reinforcing academic disparities.
Aims: This study investigates whether teachers' perception of students' skills in language and mathematics is biased by gender and externalizing behavior. Additionally, it examines how these biases evolve over time from preschool to fourth grade.
Sample: The study uses the French longitudinal birth cohort Elfe, including up to 23,349 observations from 7276 students across three time points (preschool, first grade, and fourth grade) and two domains (mathematics and French language).
Methods: We analyzed teachers' perception of students' abilities in mathematics and language, controlling for students' actual performance and externalizing behaviors.
Results: Teachers overestimated gender differences in both subjects: girls were perceived as overperforming in language and underperforming in mathematics compared to boys, even after accounting for actual performance levels. Biases in mathematics increased from preschool to first grade, aligning with the emergence of actual gender gaps. Additionally, better-behaved students—more often girls—were perceived as more competent overall. This meant that controlling for behavior in the analyses shifted the gender coefficients towards a bias in favor of boys in both language and mathematics.
Conclusions: Teachers' perception of students' abilities are influenced by gender and externalizing behavior, with biases potentially reinforcing gender disparities in mathematics. These findings highlight the need for further research on how such biases influence student outcomes and whether they can be mitigated through targeted interventions.

1. Introduction

It is important for teachers to be able to accurately assess their students' characteristics, notably academic progress and performance. Teachers' ability to assess their students accurately helps them make the learning processes more effective, by adjusting their instructional practice, such as the difficulty of the contents and tasks and some of their interactions with students to match individual needs (Kiuru et al., 2015; Nurmi et al., 2013; Pielmeier et al., 2018). Accurate assessment of students' achievement is also a prerequisite to giving fair grades. These assessments are especially useful as they usually are the primary signal by which parents know about their children's academic skills, thus it is important that they accurately represent children's skills. By accurate assessment, we mean good concordance between a teacher's subjective assessment (e.g., questionnaire-based) of students' level and an objective measure (e.g., test results) of these students' level in the same domain.

Usually, teachers are found to produce reasonably accurate assessments, with a previous meta-analysis suggesting a positive and moderate to high correlation between teachers' judgments of students' academic achievement and students' actual test performance: mean effect size of 0.63 (Südkamp et al., 2012).

Accurate teacher perception of their students' ability is also

* Corresponding author. Laboratoire de Sciences Cognitives et Psycholinguistique, ENS, CNRS, EHESS, 24 rue Lhomond, Paris, 75005, France.
E-mail address: lgurgand@gmail.com (L. Gurgand).

https://doi.org/10.1016/j.learninstruc.2026.102391
Received 4 April 2025; Received in revised form 9 May 2026; Accepted 12 May 2026
Available online 17 May 2026
0959-4752/© 2026 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

particularly important because it can act as a self-fulfilling prophecy for students. For instance, inaccurately high teacher expectations could have a positive impact on students' achievement, and inaccurately low teachers' expectations could have the opposite effect (Gentrup et al., 2020). Thus, underestimation of student's academic achievement could be particularly harmful. However, overestimated assessments might also lead to overconfidence, potentially reducing engagement and study efforts (Dunlosky & Rawson, 2012), which could negatively impact achievement. Yet, the effects of teachers' perceptions may depend not only on whether they are under- or overestimated but also on how teachers provide feedback and frame expectations. For example, overestimation might boost motivation, but it could also foster complacency.

This may be particularly problematic if teachers' accuracy varies depending on specific students' characteristics, i.e., if teachers show a bias for or against specific populations of students. In this context, a teacher's judgment is considered biased if it systematically differs depending on a student's characteristic even after controlling for students' actual ability. Indeed, school grades may be affected by teachers' biases (Zanga & De Gioannis, 2023). For instance, teachers' accuracy of their perception of students' level may depend on student's ability. Studies suggest that teachers are less accurate when judging low-achieving students than high-achieving ones (Paleczek et al., 2017; Urhahne & Wijnia, 2021). More specifically, a study found that low-achieving students tend to be underestimated (Wijnia et al., 2016), although another found opposite results (Bates & Nettelbeck, 2001). Similarly, teachers' judgments have been shown to vary by students' socioeconomic status, with higher-SES children receiving more positive evaluations than their IQ scores would predict, and lower-SES children being judged more negatively (Alvidrez & Weinstein, 1999).

Teachers' perceptions of students may also vary depending on students' behavior in the classroom. Behavior is not intended to be a criterion for evaluating academic achievement; however, some studies indicate that it can influence teachers' assessments. More specifically, teachers tend to confound scholastic and behavioral skills when grading proficiency exams, thus giving lower grades to misbehaved students, above and beyond their actual level (Ferman & Fontes, 2022). Similarly, they tend to underestimate students whom they believe to have lower socio-emotional functioning (Kowalski et al., 2018).

The present study focuses on gender biases in teachers' perceptions of students' skills. A teacher is considered to show gender-biased assessment if their assessment of boys' and girls' performance systematically differs from actual boys' and girls' ability. In other words, regardless of actual gender gaps in academic performance, a teacher showing gender bias shows a perceived gender gap that systematically differs from the actual gender gap. While some studies did not find students' gender to be associated with judgment accuracy (Glascoe, 2001; Ross & Kostuch, 2011; Zhu & Urhahne, 2015), others did find that there was a statistically significant association (Di Liberto et al., 2022; Falch & Naper, 2013; Holder & Kessels, 2017; Meissel et al., 2017; Ready & Wright, 2011), and others suggested that there was an association but only for less experienced teachers (Hofer, 2015, although this was the case for Swiss teachers but not German and Austrian ones). More specifically, when gender differences are found, this can take several forms. Some studies suggest that girls are over-evaluated compared to boys (Di Liberto et al., 2022; Falch & Naper, 2013). Sometimes, it is also suggested that these gender differences vary by domains assessed, such as mathematical or language skills. Indeed, studies focusing on literacy, reading or writing tend to find an overestimation of girls' compared to boys' skills (Meissel et al., 2017; Ready & Wright, 2011), whereas in mathematics, boys are sometimes overestimated compared to girls (Bonefeld et al., 2020; Cimpian et al., 2016; Holder & Kessels, 2017). This might explain why a recent review suggested a null or weak association between student gender and judgment accuracy, but it did not contrast mathematics and literacy results (Urhahne & Wijnia, 2021). In line with this literature, Campbell (2015) studied both math and English skills at the same time and found that boys are more likely to be judged higher than girls in math and vice versa in English. Thus, this literature may suggest that teachers tend to exaggerate stereotypical gender gaps (i.e., boys being better in math and girls in language) (see also Steinmann, 2024). However, Schuster et al. (2021) suggested that pre-service teachers graded girls more generously in mathematics and boys more generously in German (but it was only the case for teachers who believed that gender stereotypes about subject specific abilities are true). Another study in the context of competitive exams in higher education also suggested a bias towards women in male-dominated fields such as math, physics, and philosophy, and a (smaller) bias towards males in literature and foreign languages (Breda & Hillion, 2016; Breda & Ly, 2015). These contrasting findings highlight the importance of distinguishing between different types of bias. Perceptual bias (how teachers judge students' abilities) may tend to reinforce stereotypes and exaggerate gender differences, while biases in grading and selection processes can be shaped by implicit biases or even explicit policies aimed at counteracting gender inequalities.

As previously discussed, students' behavior may influence teachers' perception of students' skills. Since girls tend to display better in-class behaviors and lower externalizing symptoms than boys (Bugler et al., 2015; Fernandez Castelao & Kröner-Herwig, 2014; Kremer et al., 2016), this raises the possibility that girls are consistently overestimated while boys are consistently underestimated. Indeed, several studies suggest that accounting for behavior explains some of the observed gender bias, reducing it when it was in favor of girls, or increasing it towards boys when there was none (Contreras, 2024; Ferman & Fontes, 2022; Robinson-Cimpian et al., 2014). However, a gender effect persists, suggesting that teachers' perception bias is not solely driven by behavior, but may also be influenced by gender-specific factors.

Finally, there are temporal dynamics in the gender differences, especially in mathematics: in France, gender differences to the advantage of boys appear in first grade and tend to increase through elementary school (Eteve et al., 2025). Thus, teacher biases and their perceptions of girls and boys' academic skills may also vary with time.

In this preregistered study (https://osf.io/b9rpt), we studied teachers' perceptions of children's educational achievement in preschool and primary school, in both math and language, using the French birth cohort ELFE. This is especially interesting since, in France, gender differences in math to the advantage of boys have been shown to appear during the first grade of primary school (Fischer & Thierry, 2022; Martinot et al., 2021, 2025). On the contrary, there is a consistent advantage of girls in language and literacy skills (Fischer & Thierry, 2022). Thus, it may be interesting to observe teachers' gender biases across time and domains. Few studies have studied this question at various time points, and fewer across different domains. Furthermore, the present study also investigates to what extent teacher biases result from behavioral differences between boys and girls. Robinson-Cimpian et al. (2014) studied similar questions with children from the United States, and we aim to compare their findings to the French context. More specifically, our hypotheses are the following:

Hypothesis 1. We hypothesize that teachers' perception of pupil ability is influenced by pupil's gender.

a. This predicts that teachers perceive girls as outperforming boys in language skills but as underperforming in mathematics, even when controlling for measured abilities.
b. This effect may become more pronounced over time, as actual gender differences become more noticeable (especially in math).

Hypothesis 2. We hypothesize that teachers incorporate behavior (with girls typically exhibiting better behavior) in addition to abilities when perceiving students' level. This predicts that controlling for children's behavior may modify the effect of gender on teachers' estimations. We do not have strong assumptions as for the direction of the shift, but if the bias becomes more favorable to boys (i.e. increased teachers' perception of boys' skills compared to girls' once controlling for behavior), this would suggest that teachers' estimation of academic ability is biased by pupil behavior, leading to an underestimation of boys' abilities. If on the contrary, the bias becomes more favorable to girls, this would suggest that teachers overcompensate for the more disruptive behavior of boys in their estimation of academic ability, leading to an underestimation of girls' level.

Hypothesis 3. We will further test whether the effect of behavior on teachers' perception differs by gender. Specifically, the relationship between behavior and teacher perceptions may be stronger for one gender than the other. For example, teachers might place more weight on behavior when evaluating boys compared to girls, or vice versa.

2. Material and methods

2.1. Study design

We used data from ELFE (*Étude Longitudinale Française depuis l'Enfance*, French Longitudinal Study since Childhood), a nationwide French longitudinal birth cohort launched in 2011 (Charles et al., 2019). A total of 349 maternity units were randomly selected in metropolitan France out of which 320 agreed to participate. Children and mothers were recruited at birth. Exclusion criteria included multiple births of more than two children, children born before 33 weeks of amenorrhea, mothers aged <18 years, inability to read either French, Arabic, Turkish or English, and parents who were not capable of giving informed consent. The study and each wave of data collection were approved by the national data authority (CNIL) and by either the national advisory committee on information processing in health research (CCTIRS) or the National Statistics Council and the Committee for the protection of persons engaged in research (CPP). Written informed consent was obtained from mothers at the child inclusion. Fathers signed the consent form when present at inclusion or were informed about their right to oppose to their child participation.

2.2. Participants

A total of 18,329 newborns were originally included in the cohort, with a participation rate of 51%. Children were included in the present analysis if they had enough data to create at least one language or mathematics achievement score (as detailed in the next section), if they had at least one teacher's perception measure of their abilities in language or math, and if they were in the grade corresponding to their age. 5071 children had at least one of the two achievement scores in preschool, 4415 in first grade and 2914 in fourth grade. 5017 children had at least one of the two teacher's perception measures in preschool, 4474 in first grade and 2909 in fourth grade. A more complete description of the sample size is presented in Fig. S1.

2.3. French school system

In the French school system, primary school teachers are trained through a single curriculum and can teach both preschool (maternelle) and elementary (élémentaire) levels. In many cases, preschool and elementary classes are located within the same school, so the transition between the two levels does not necessarily involve a major institutional change. Children typically change teachers each school year. At the time of the Elfe cohort, schooling in kindergarten (3 years from age 3) was not compulsory but nearly universal (>97% children schooled from 3). Compulsory education started in grade 1 (age 6). Primary school lasts for 5 years (grades 1 to 5). The grades when school assessments were made in the Elfe cohort were “Moyenne section de maternelle” (2 nd year of kindergarten during the 3rd trimester, age 4.5), “cours préparatoire” (1st grade during the 3rd trimester, age 6.5) and “cours moyen 1ère année’” (4th grade during the 1st trimester, age 9).

Since the Elfe cohort was designed to be representative of the French territory, Elfe children were spread all over continental France, and were therefore usually the only Elfe participant in their school. Their teachers were individually contacted through parents in order to ask them to administer the Elfe school assessments and to answer some questionnaires. Participation rates therefore reflect both families' and teachers' voluntary participation.

2.4. Measures

Fig. 1 presents a graphical overview of the different measurement timepoints.

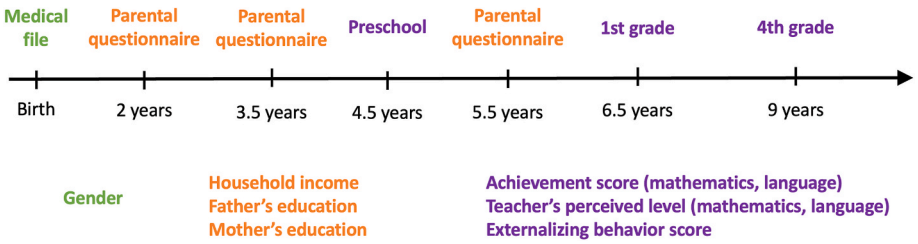


Fig. 1. Graphical overview of the measurement time points.

2.4.1. Teacher's perceived level

In preschool, first grade and fourth grade, the teacher had to answer the following question: "Based on your professional experience, how do you evaluate the child's skills in relation to other children of their age (and not just the children in your class or school)?". This question was asked separately for French language and for mathematics (in preschool, these domains were called "Language" and "Numbers, shapes, sizes"), and each was based on a scale from 1 to 5 (much below average, below average, average, above average, much above average). This score was z-scored, to have a mean of 0 and a standard deviation of 1.

2.4.2. Child's sex registered at birth

This information was obtained from the medical file of the maternity ward. This was coded as −0.5 for boys and 0.5 for girls.

We subsequently use the term Gender and not Sex to align with our focus on teachers' gender biases.

2.4.3. School achievement

Mathematics and French language scores were computed from school data collected in preschool (spring 2016), first grade (spring 2018) and fourth grade (autumn 2020). Questionnaires were sent to the children's schools and data were collected by the teachers during school hours.

In language, there were 35 questions in preschool (among the categories "Letter knowledge", "Phonological odd-one out", "Syllable deletion" and "Picture-word matching"), 45 in first grade (among the categories "Phoneme counting", "Listening comprehension", "Reading comprehension" and "Orthographic choice") and 83 in fourth grade (among "Identifying written words in silent reading of words and non-words: judging homophony", "Sentence comprehension assessment", "Reading comprehension: read a text silently and answer questions in writing", "Complete an incomplete text using a source text to be transformed: shift from present to future tense", "Complete an incomplete text using a source text to be transformed: shift from singular to plural", "Dictation of a sentence").

In mathematics, there were 26 questions in preschool (among the categories "Numerically matching one collection to another without counting", "Developing a mental model to reproduce a collection initially seen and then arithmetically transformed.", "Global comparison of two collections", "Recognizing a quotient from its dictated number", "Recognition of writing with numbers" and "Count a collection and give the result in numerical form"), 32 in first grade (among the categories "Mental calculation of dictated numbers", "Solving a numerical problem and writing down the numerical result", "Calculating sums of money", "Numerical sequence", "Comparison of 2 (or 3) digit numbers with $>$, $<$, $=$ signs") and 36 in fourth grade (among "Calculating dictated numbers: dividing or multiplying two numbers", "Solving a numerical problem and writing down the numerical result", "Number comparison and numerical length estimation", "Numerical sequence: writing the missing number missing in an ordered sequence of numbers").

The Cronbach's alphas for the questionnaires were all greater than or equal to 0.795 (in mathematics, 0.795 in preschool, 0.91 in first grade, 0.85 in fourth grade, and in language, 0.84 in preschool, 0.84 in first grade, 0.94 in fourth grade).

In preschool and first grade, teachers indicated whether each question was correct or incorrect. In fourth grade, the process was similar, except that for some questions, teachers only reported the raw student's answer without evaluating their correctness, and the ELFE cohort team then indicated whether these questions were correct or incorrect. This implies that school achievement scores may be influenced, to some extent, by teachers' gender biases, which is discussed later in the discussion section.

For each age, a language and a mathematics score were computed when at least 10 questions per domain were answered. The score was computed by averaging the responses (0: incorrect response, 1: correct response) within each category, and then averaging them and z-scoring the result. For more information about the tests see supplementary method S1.

2.4.4. Externalizing behavior score

Externalizing behavior refers to outwardly directed behaviors that include impulsivity, hyperactivity, aggression, and difficulty with self-regulation. To evaluate children's behavior in class, we used the externalizing questions of the Strengths and Difficulties Questionnaire (SDQ) filled by the teacher. It was a short version of the full SDQ, with 5 items measuring externalizing symptoms for each grade. In each grade, the externalizing score was the average of these 5 answers.

The 5 items were "Generally obedient, usually does what adults ask", "Can't hold still or squirms constantly", "Easily distracted, has trouble concentrating", "Thinks before acting", "Completes tasks, maintains attention well", and were coded from 1 to 3 (Not true, A little true, Very true). When needed, scales were recoded such that a higher value always indicated higher externalizing symptoms.

The Cronbach's alpha for the externalizing score was 0.83 in reschool, 0.83 in first grade, and 0.82 in fourth grade.

2.4.5. Adjustment variables

In order to control for potentially confounding factors, our models were adjusted on parental education (highest diploma of each parent) and average household income (€/month at 2, 3.5 and 5.5 years, log scale). Ethnicity was not included because it is not collected in French cohorts. We focused on socioeconomic status as a well-established potential confounding variable known to influence both actual and perceived achievement, whereas other variables such as urban versus rural environment have not been shown to have comparable confounding effects in this context.

Education of fathers and mothers: To minimize missing values (and because the correlation between parental education measured at 2, 3.5 and 5.5 years are very high, all $r > 0.99$), education level was the maximum of the available values at 2, 3.5 and 5.5 years.

Log Household income: We removed outliers (income values $>$ mean+4SD) at 2, 3.5 and 5.5 years (because we have previously noticed that outlying values were highly discrepant across time points, suggesting an entry error). To minimize missing values (and because the incomes at the 3 time points were highly correlated, all $r > 0.8$), the income value was the average of the available values at 2, 3.5 and 5.5 years.

2.5. Statistical analyses

All the analyses presented below follow the preregistration plan (https://osf.io/b9rpt).

We ran linear mixed-effect models with a random intercept for each child (because there are repeated measures across the three grades, and children usually have a score in both domains, math and language, thus we wanted to take into account the non-independence of the observations). The specific model used to test each hypothesis is indicated in R syntax below.

2.6. Hypothesis 1

2.6.1. Hypothesis 1A: teachers perceive girls as outperforming boys in language skills but as underperforming in mathematics, even when controlling for measured abilities

The model in R syntax is presented in Table 1.

SES stands for three distinct variables: Household income, Father's education and Mother's education.

In the case where the Child_Gender x Domain coefficient was statistically significant, we planned to run the following analysis separately for Mathematics and French, to better describe the results:

Table 1
Overview of the hypotheses.

Hypothesis	Main Model	Purpose
H1A: Gender differences by subject	Teacher_perceived_level ~ Achievement_Score + Child_Gender + Timepoint + Domain + **Child_Gender x Domain** + SES + (1 | ChildID)	Test if gender effect differs between math and French
H1B: Change over time	Teacher_perceived_level ~ Achievement_Score + Child_Gender + Timepoint + **Child_Gender x Timepoint** + SES + (1 | ChildID)	Test if gender bias changes with time
H2: Role of child behavior	If H1B interaction not significant: Teacher_perceived_level ~ Achievement_Score + **Child_Gender** + Timepoint + Externalizing_Score + SES + (1 | ChildID) If H1B interaction significant: Teacher_perceived_level ~ Achievement_Score + **Child_Gender** + Timepoint + Child_Gender x Timepoint + Externalizing_Score + SES + (1 | ChildID)	Test if gender differences remain after controlling for child behavior
H3: Gender × Behavior interaction	Similar to H2, add: **Child_Gender x Externalizing_Score**	Test if the effect of behavior on perception differs by gender

Teacher_ perceived_level ~ Achievement_ Score + Child_Gender + Timpoint + SES

Hypothesis 1A was tested by the statistical significance of the Child_Gender × Domain interaction, and further predicted opposite signs for the Gender coefficients in the secondary French language and math analyses.

Hypothesis 1B: This effect may become more pronounced over time.

The model in R syntax is presented in Tables 1 and it was run separately for Mathematics and French.

Hypothesis 1 was tested by the statistical significance of the Child_Gender × Time interaction.

2.7. Hypothesis 2: controlling for children's behavior may modify the effect of gender on teachers' estimations

The model in R syntax is presented in Tables 1 and it was run separately for Mathematics and French.

Hypothesis 2 was tested by checking whether the Child_Gender coefficient was statistically different:

- from the Child_Gender coefficient found in 1)A) if the Child_Gender × Timepoint interaction of analysis 1B was not significant.
- from the Child_Gender coefficient found in 1)B) if the Child_Gender × Timepoint interaction of analysis 1B was significant.

2.8. Hypothesis 3: the effect of behavior on teachers' perception differs by gender

The model in R syntax is presented in Tables 1 and it was run separately for Mathematics and French.

Hypothesis 3 predicted that the coefficient of Child_Gender x Externalizing_Score would be significant. In that case, we would then split the analysis between the two genders in order to interpret the interaction.

All analyses were performed using RStudio (Version 2023.06.1) and R (Version 4.2.2). We ran the mixed effect linear models with function *lmer* from lme4 package (version 1.1-35.2).

Because attrition and missing data were associated with socioeconomic status and child achievement, missingness could not reasonably be assumed to be completely at random. To assess the robustness of our findings to missing data handling, we therefore conducted supplementary analyses using multiple imputation. We performed multiple imputation (m = 50 imputed datasets). Imputation was conducted using multiple imputation by chained equations (MICE) in wide format, using a single-level imputation approach. The imputation model included all repeated measures of the variables used in the main analyses in order to maximize the information available to predict missing values: teacher-perceived level in language and mathematics at the three timepoints (6 variables), child achievement scores in language and mathematics at the three timepoints (6 variables), teacher-rated externalizing behavior at the three timepoints (3 variables), household income, mother's education, and father's education. Child sex was not included in the imputation model, as it was fully observed and used only as an explanatory variable in the substantive analyses.

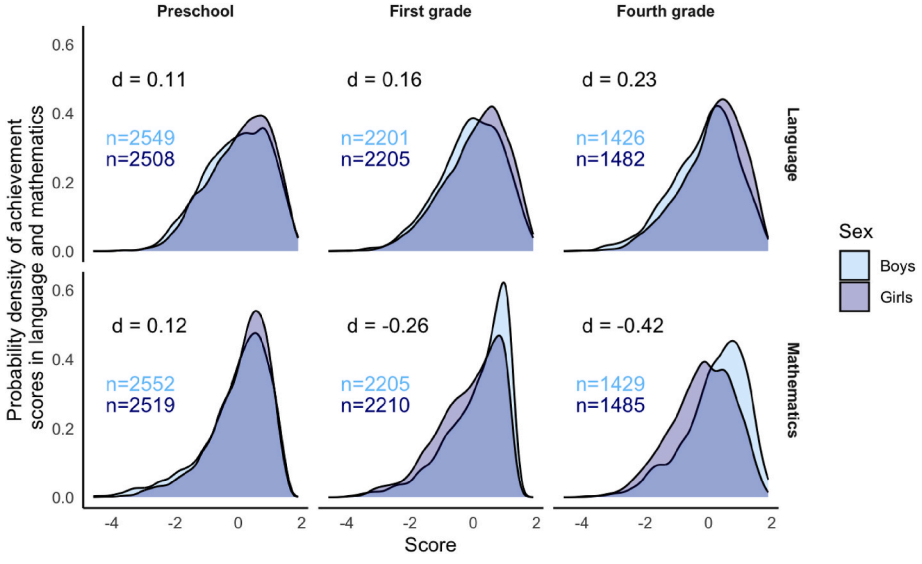


Fig. 2. Distribution of the achievement scores at the three timepoints and in the two domains. The corresponding Cohens' d are written on each plot. A positive d represents higher scores for girls than for boys.

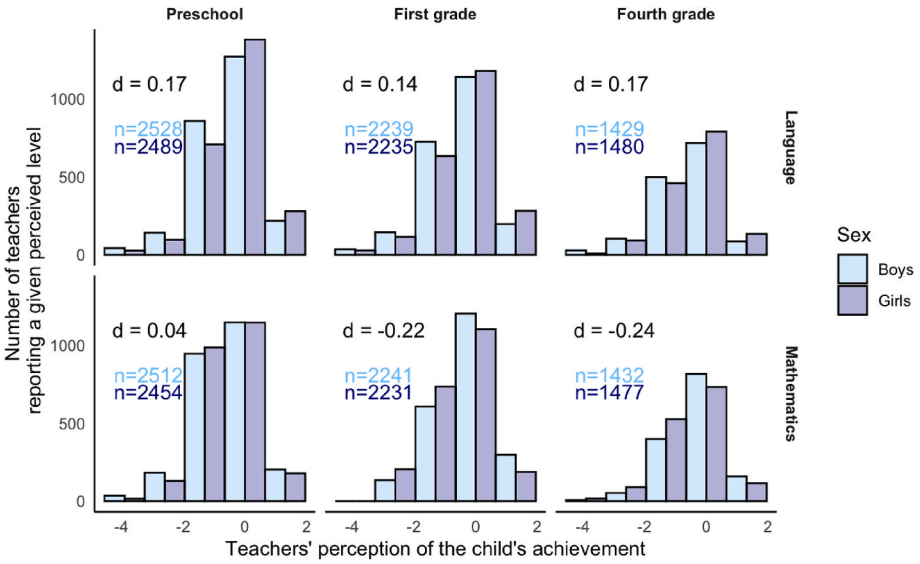


Fig. 3. Distribution of teachers' perceived level of students at the three timepoints and in the two domains. The corresponding Cohens'd are written on each plot. A positive d represents higher scores for girls than for boys.

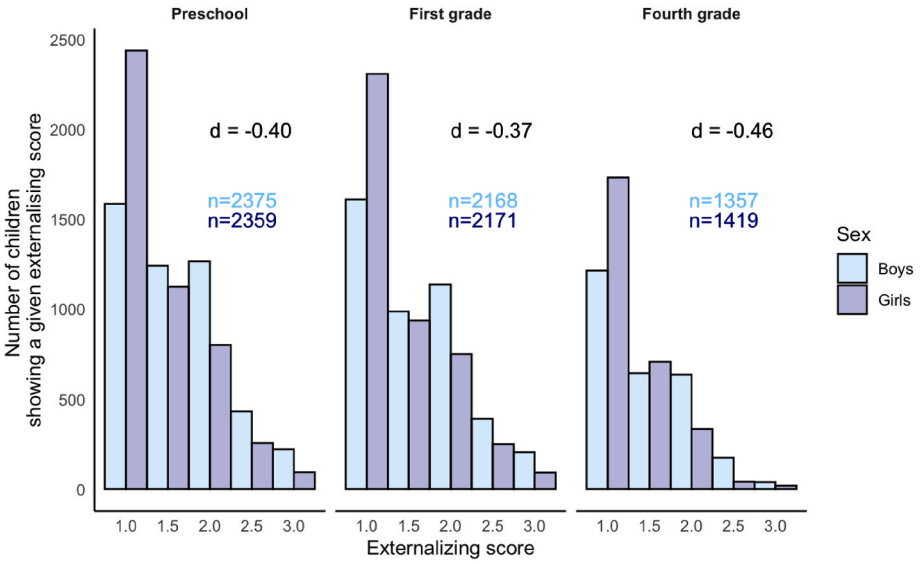


Fig. 4. Distribution of the externalizing scores at the three timepoints. The corresponding Cohens'd are written on each plot. A negative d represents higher scores for boys than for girls.

3. Results

3.1. Descriptive results

The descriptive statistics are available in Table S1. Table S2 presents the same descriptive statistics but for the sample of children with achievement scores values at all timepoints.

Girls tended to outperform boys in language performance in preschool (Cohen's d = 0.11, 95% CI [0.06; 0.17]), first grade (Cohen's d = 0.16, 95% CI [0.1; 0.22]), and fourth grade (Cohen's d = 0.23, 95% CI [0.16; 0.31]). Girls also tended to outperform boys in mathematics in preschool (Cohen's d = 0.12, 95% CI [0.06; 0.17]), whereas boys outperformed girls in mathematics in first grade (Cohen's d = −0.26, 95% CI [−0.32; −0.2]) and fourth grade (Cohen's d = −0.42, 95% CI [−0.49; −0.34]). Fig. 2 displays the distribution of the achievement scores at the three timepoints, in math and in language. Teachers tended to perceive girls as performing better than boys in language in preschool (Cohen's d = 0.17, 95% CI [0.12; 0.23]), first grade (Cohen's d = 0.14, 95% CI [0.09; 0.2]) and fourth grade (Cohen's d = 0.17, 95% CI [0.1; 0.24]). In mathematics, teachers' perception of student skills did not differ between boys and girls in preschool (Cohen's d = 0.04, 95% CI [−0.02; 0.09]), but they tended to perceive boys as performing better in first grade (Cohen's d = −0.22, 95% CI [−0.28; −0.16]) and in fourth grade (Cohen's d = −0.24, 95% CI [−0.31; −0.17]). Fig. 3 displays the distribution of the Teacher's perceived level at the three timepoints, in math and in language. Boys had higher externalizing behavior scores in all grades (in preschool, Cohen's d = −0.4, 95% CI [−0.46; −0.35], in first grade, Cohen's d = −0.37, 95% CI [−0.43; −0.31], in fourth grade, Cohen's d = −0.46, 95% CI [−0.54; −0.38]). Fig. 4 displays the distributions of the externalizing score at the three timepoints. Fig. 5 shows the raw correlation between these variables. As expected, teachers' perception of students' skills and students' achievement are positively correlated, and externalizing scores are negatively correlated to both measures. These correlations are of similar magnitude across the three time points.

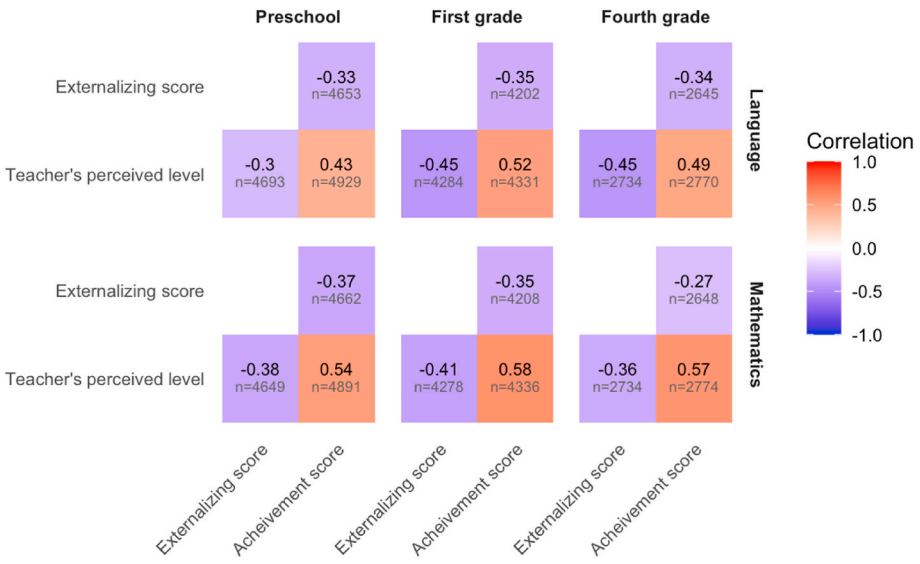


Fig. 5. Raw correlation matrices for teachers' perceived level, externalizing scores and achievement scores, separately by domains and timepoints. All correlations are statistically significant ($p < 0.001$).

3.2. Hypothesis 1

3.2.1. Hypothesis 1A

Consistent with Hypothesis 1A, we found a significant Gender × Domain interaction (beta = −0.17, SE = 0.018, $p < 0.001$), suggesting that gender differences in teachers' perceived level of students' level are not the same across the mathematics and language domains (see Table 2). We thus analyzed separately the mathematics and language domains, and found a significant advantage of boys over girls in teacher's perceived mathematics level (beta = −0.05, SE = 0.016, p = 0.002), and a significant advantage of girls over boys in teachers' perceived language level (beta = 0.10, SE = 0.017, $p < 0.001$), see Table 3.

Table 2
Teacher's perceived level as a function of domain x child gender, controlling for achievement score, income and parental education. N = 23,349. Child's gender is coded −0.5 for boys and 0.5 for girls.

Variables	Beta	SE[1]	95% CI[1]	p-value
Achievement Score	0.37	0.006	0.36, 0.38	<0.001
Child gender	0.11	0.018	0.07, 0.14	<0.001
Timepoint (reference: First grade)				
Fourth grade	−0.03	0.014	−0.05, 0.00	0.040
Preschool	0.02	0.012	−0.01, 0.04	0.14
Domain (reference: Language)				
Math	0.00	0.009	−0.02, 0.02	>0.9
Household income (log)	0.05	0.021	0.01, 0.10	0.014
Mother's education	0.07	0.007	0.05, 0.08	<0.001
Father's education	0.05	0.006	0.03, 0.06	<0.001
Child gender * Domain				
Child gender * Math	−0.17	0.018	−0.21, −0.14	<0.001

3.2.2. Hypothesis 1B

In mathematics, consistent with Hypothesis 1B, the Gender × Time interactions were significant. More specifically, the gender bias, which was globally significant and in favor of boys, increased between preschool and first grade (beta = 0.06, SE = 0.033, p = 0.049), and decreased to a similar extent between first grade and fourth grade, but this latter difference was not statistically significant (beta = 0.06, SE = 0.038, p = 0.13).

In language, the Gender × Time interactions were not significant, suggesting that the gender bias in favor of girls was constant from preschool to 4th grade. More specifically, the effect did not significantly vary between preschool and first grade (beta = 0.05, SE = 0.035, p = 0.2), nor between first grade and fourth grade (beta = 0.00, SE = 0.041, p = 0.98). See Table 4 for the full results, and Fig. 6 for a visualization.

3.3. Hypothesis 2

In mathematics, the gender coefficient initially was −0.09 (SE = 0.026, $p < 0.001$), and further decreased to −0.19 (SE = 0.025, $p < 0.001$) when the externalizing score was added as a covariate. There was a negative association between the externalizing score and teacher's perceived level of mathematics (beta = −0.43, SE = 0.017, $p < 0.001$). See Table 5.

Table 3
Teacher's perceived level in mathematics (left) and language (right) as a function of child gender, controlling for achievement score, timepoint, income and parental education. Child's gender is coded −0.5 for boys and 0.5 for girls.

Variables	H1A, mathematics (N = 11,663)				H1A, language (N = 11,686)			
	Beta	SE[a]	95% CI[a]	p-value	Beta	SE[a]	95% CI[a]	p-value
Achievement Score	0.50	0.008	0.49, 0.52	<0.001	0.41	0.008	0.39, 0.42	<0.001
Child gender	−0.05	0.016	−0.08, −0.02	0.002	0.10	0.017	0.07, 0.14	<0.001
Timepoint (reference: First grade)								
Fourth grade	−0.02	0.019	−0.06, 0.02	0.3	−0.02	0.020	−0.06, 0.02	0.3
Preschool	0.01	0.016	−0.02, 0.04	0.6	0.01	0.017	−0.03, 0.04	0.7
Household income (log)	0.02	0.023	−0.02, 0.07	0.3	0.07	0.024	0.02, 0.12	0.005
Mother's education	0.05	0.008	0.03, 0.06	<0.001	0.07	0.008	0.05, 0.09	<0.001
Father's education	0.04	0.007	0.03, 0.05	<0.001	0.04	0.007	0.03, 0.05	<0.001

[a] SE = Standard Error, CI = Confidence Interval.

Table 4
Teacher's perceived level in mathematics (left) and language (right) as a function of timepoint x child gender, controlling for achievement score, income and parental education. Child's gender is coded −0.5 for boys and 0.5 for girls.

Variables	H1B, mathematics (N = 11,663)				H1B, language (N = 11,686)			
	Beta	SE[a]	95% CI[a]	p-value	Beta	SE[a]	95% CI[a]	p-value
Achievement Score	0.50	0.008	0.49, 0.52	<0.001	0.41	0.008	0.39, 0.42	<0.001
Child gender	−0.09	0.025	−0.14, −0.04	<0.001	0.08	0.027	0.03, 0.13	0.002
Timepoint (reference: First grade)								
Fourth grade	−0.02	0.019	−0.06, 0.02	0.3	−0.02	0.020	−0.06, 0.02	0.3
Preschool	0.01	0.016	−0.02, 0.04	0.6	0.01	0.017	−0.03, 0.04	0.7
Household income (log)	0.02	0.023	−0.02, 0.07	0.3	0.07	0.024	0.02, 0.12	0.005
Mother's education	0.05	0.008	0.03, 0.06	<0.001	0.07	0.008	0.05, 0.09	<0.001
Father's education	0.04	0.007	0.03, 0.05	<0.001	0.04	0.007	0.03, 0.05	<0.001
Child gender * Timepoint								
Child gender * Fourth grade	0.06	0.038	−0.02, 0.13	0.13	0.00	0.041	−0.08, 0.08	0.98
Child gender * Preschool	0.06	0.033	0.00, 0.13	0.049	0.05	0.035	−0.02, 0.12	0.2

[a] SE = Standard Error, CI = Confidence Interval.

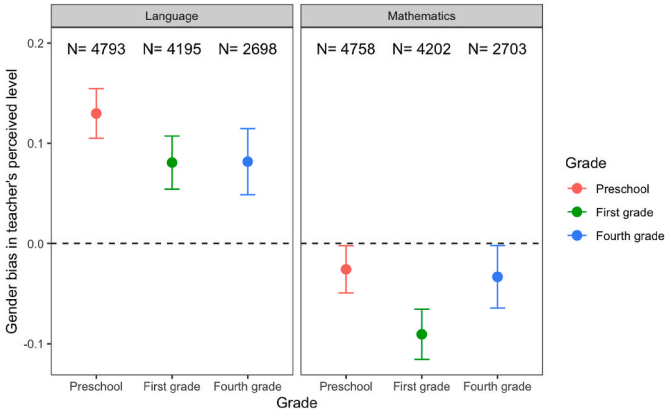


Fig. 6. Gender bias in teachers' perceived level across domains and timepoints. The y axis corresponds to the gender difference in teacher's perception, adjusted for the achievement score and the control variables (income and parental educations), with standard errors. A positive bias is in favor of girls, a negative bias is in favor of boys.

In language, the gender coefficient initially was 0.10 (SE = 0.018, p < 0.001), and further decreased to be statistically non-significant (beta = 0.02, SE = 0.017, p = 0.3) when adding the externalizing score as a covariate. There was a negative association between the externalizing score and teacher's perceived level of language (beta = −0.46, SE = 0.017, p < 0.001). See Table 6. Thus, in both mathematics and language, boys' higher externalizing scores decreases teachers' perception of their level. See Fig. 7 for an illustration.

3.4. Hypothesis 3

In mathematics, there was a significant Gender x Externalizing score interaction (beta = −0.12, SE = 0.031, p < 0.001), revealing that the relation between externalizing score and teachers' perceived level of children was more negative for girls than for boys.

In language, this interaction was not statistically significant (beta = −0.03, SE = 0.033, p = 0.4). See Table 7.

Results from the multiple imputation analyses were broadly consistent with the complete-case analyses (See Table S6–S10 and Fig. S2–S5). In contrast, the evidence was weaker for H1B and H3: in the imputed analyses, the Gender × Time interaction in mathematics was no longer statistically significant (H1B, See Table S7), and the Gender × Externalizing interaction in mathematics was attenuated and became non-significant (p = 0.059) (H3).

4. Discussion

In summary, our findings reveal biases in teachers' perceptions of students' skills, based on both pupils' gender and externalizing behavior. More precisely, as predicted in hypothesis 1, we found that teachers tend to perceive girls as outperforming boys in language skills but as underperforming in mathematics, even when controlling for measured abilities. Regarding hypothesis 2, we found that teachers were influenced by in-class behavior when estimating students' skills: better-behaved students, thus girls more often than boys, were perceived as performing overall better. Finally, while the association between teachers' perceptions of students' skills and behavior was the same for boys and girls when assessing skills in French language, a difference

Table 5
Teacher's perceived level in mathematics as a function of child gender, without (left) and with (right) the externalizing score, controlling for achievement score, income and parental education. The left panel is equivalent to the test of hypothesis 1B reported in Table 3, but the sample is limited to the one with an externalizing score available for better comparison of the coefficients with H2. Child's gender is coded −0.5 for boys and 0.5 for girls.

Variables	H1B (N = 11,016)				H2 (H = 11,016)			
	Beta	SE[a]	95% CI[a]	p-value	Beta	SE[a]	95% CI[a]	p-value
Achievement Score	0.51	0.008	0.49, 0.52	<0.001	0.44	0.008	0.42, 0.45	<0.001
Child gender	−0.09	0.026	−0.14, −0.04	<0.001	−0.19	0.025	−0.24, −0.14	<0.001
Timepoint (reference: First grade)								
Fourth grade	−0.01	0.020	−0.05, 0.02	0.5	−0.06	0.019	−0.10, −0.03	<0.001
Preschool	0.01	0.017	−0.02, 0.04	0.5	0.02	0.016	−0.02, 0.05	0.3
Household income (log)	0.02	0.024	−0.03, 0.06	0.5	0.01	0.023	−0.04, 0.05	0.7
Mother's education	0.05	0.008	0.04, 0.07	<0.001	0.04	0.008	0.03, 0.06	<0.001
Father's education	0.04	0.007	0.02, 0.05	<0.001	0.03	0.007	0.02, 0.04	<0.001
Child gender * Timepoint								
Child gender * Fourth grade	0.06	0.039	−0.02, 0.13	0.2	0.04	0.038	−0.04, 0.11	0.3
Child gender * Preschool	0.06	0.034	−0.01, 0.12	0.084	0.08	0.033	0.01, 0.14	0.020
Externalizing score					−0.43	0.017	−0.47, −0.40	<0.001

[a] SE = Standard Error, CI = Confidence Interval.

Table 6
Teacher' perceived level in language as a function of child gender, without (left) and with (right) the externalizing score, controlling for achievement score, income and parental education. The left panel is equivalent to the test of hypothesis 1A reported in Table 2, but the sample size is limited to the one with an externalizing score available for better comparison of the coefficients with H2. Child's gender is coded −0.5 for boys and 0.5 for girls.

Variables	H1A (N = 11,041)				H2 (N = 11,041)			
	Beta	SE[a]	95% CI[a]	p-value	Beta	SE[1]	95% CI[1]	p-value
Achievement Score	0.41	0.009	0.39, 0.43	<0.001	0.34	0.009	0.33, 0.36	<0.001
Child gender	0.10	0.018	0.07, 0.14	<0.001	0.02	0.017	−0.02, 0.05	0.3
Timepoint (reference: First grade)								
Fourth grade	−0.02	0.021	−0.06, 0.02	0.3	−0.07	0.020	−0.11, −0.03	<0.001
Preschool	0.02	0.018	−0.02, 0.05	0.4	0.02	0.017	−0.01, 0.05	0.3
Household income (log)	0.06	0.025	0.01, 0.11	0.013	0.05	0.024	0.00, 0.09	0.047
Mother's education	0.07	0.009	0.05, 0.09	<0.001	0.06	0.008	0.05, 0.08	<0.001
Father's education	0.04	0.007	0.03, 0.05	<0.001	0.03	0.007	0.02, 0.05	<0.001
Externalizing score					−0.46	0.017	−0.49, −0.43	<0.001

[a] SE = Standard Error, CI = Confidence Interval.

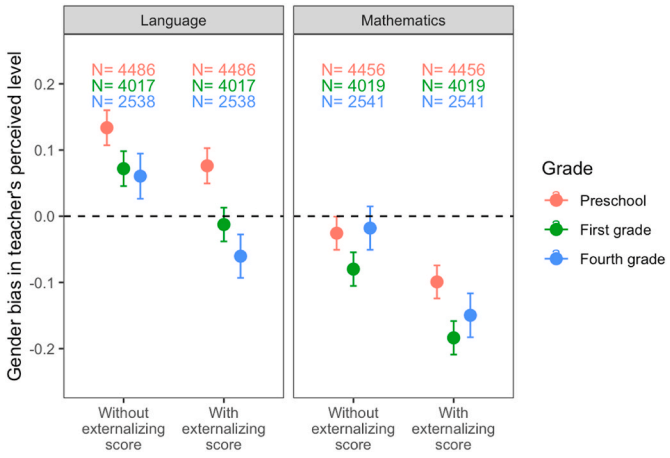


Fig. 7. Gender bias in teachers' perceived level across domains and timepoints, with or without an adjustment on the externalizing score. The y axis corresponds to the gender difference in teacher's perception, adjusted for the achievement score and the control variables (income and parental educations), and with or without adjustment on the externalizing score, with standard errors. A positive bias is in favor of girls, a negative bias is in favor of boys.

emerged in mathematics: teachers' perceptions were more negatively influenced by behavior for girls than for boys.

Regarding the first hypothesis, it seems that teachers tend to overestimate existing gaps. Indeed, girls tended to outperform boys in language, and teachers tended to perceive them as outperforming boys even more than they actually are. We found the opposite pattern in mathematics, with a bias in the perception of mathematics skills to the advantage of boys even when accounting for actual abilities. These results are in line with previous studies suggesting a bias in favor of girls in language and in favor of boys in mathematics (Campbell, 2015; Holder & Kessels, 2017; Ready & Wright, 2011; Robinson-Cimpian et al., 2014). They are also consistent with previous research showing that, while gender stereotypes are often based on real differences, they tend to exaggerate those differences (Eyal & Epley, 2017). However, the same pattern of teacher gender bias amplifying actual gender differences was not observed between grades 1 and 4, so it is not clear at this stage what exactly those fluctuations reflect.

Furthermore, the evolution of these biases over time is noteworthy. In language, the bias in favor of girls remained stable from preschool to fourth grade. However, in mathematics, while the bias was small in preschool, it grew to −0.09 SD in first grade. This is especially interesting as in France, the gender difference to the advantage of boys appears during first grade (Fischer & Thierry, 2022; Martinot et al., 2021, 2025). However, this temporal interaction was no longer statistically significant in the multiple imputation analyses, suggesting that this effect may be modest and sensitive to missing-data assumptions. Accordingly, this pattern should be considered suggestive rather than definitive. Taken together, these findings are consistent with the possibility that teachers' gender biases may track the emergence of actual differences. However, the fact that teachers overestimate these differences could negatively influence girls' future performance, and act as self-reinforcing bias. Indeed, previous studies suggested that inaccurately low teachers' expectations may negatively impact students' achievement (Gentrup et al., 2020; Hollenstein et al., 2024; Lavy & Sand, 2018). Thus, if teachers are biased against girls (or conversely, in favor of boys), this might further reinforce gender differences in

Table 7
Teacher' perceived level in mathematics (left) and language (right) as a function of the interaction between externalizing and child gender, controlling for achievement score, timepoint, income and parental education. Child's gender is coded −0.5 for boys and 0.5 for girls.

Variables	H3, mathematics (N = 11,016)				H3, language (N = 11,041)			
	Beta	SE[a]	95% CI[a]	p-value	Beta	SE[a]	95% CI[a]	p-value
Achievement Score	0.43	0.008	0.42, 0.45	<0.001	0.34	0.009	0.33, 0.36	<0.001
Child gender	0.00	0.054	−0.11, 0.10	>0.9	0.06	0.052	−0.04, 0.16	0.2
Externalizing score	−0.44	0.017	−0.48, −0.41	<0.001	−0.46	0.018	−0.50, −0.43	<0.001
Timepoint (reference: First grade)								
Fourth grade	−0.07	0.019	−0.10, −0.03	<0.001	−0.07	0.020	−0.11, −0.03	<0.001
Preschool	0.01	0.016	−0.02, 0.05	0.4	0.02	0.017	−0.01, 0.05	0.3
Household income (log)	0.01	0.023	−0.04, 0.05	0.8	0.05	0.024	0.00, 0.09	0.049
Mother's education	0.04	0.008	0.03, 0.06	<0.001	0.06	0.008	0.05, 0.08	<0.001
Father's education	0.03	0.007	0.02, 0.04	<0.001	0.03	0.007	0.02, 0.05	<0.001
Child gender * Externalizing score	−0.12	0.031	−0.18, −0.06	<0.001	−0.03	0.033	−0.09, 0.04	0.4
Child gender * Timepoint								
Child gender * Fourth grade	0.02	0.039	−0.05, 0.10	0.5				
Child gender * Preschool	0.08	0.033	0.01, 0.14	0.018				

[a] SE = Standard Error, CI = Confidence Interval.

mathematics, creating a self-perpetuating cycle where lower expectations lead to reduced performance in girls, ultimately reinforcing the very gaps that teachers initially perceived.

Regarding the second hypothesis, we showed that teachers' perceptions were also influenced by pupils' externalizing behavior. Controlling for behavior in the analyses shifts the gender coefficients towards a bias in favor of boys. This means that, when there is a bias in favor of girls, like in language, the bias is reduced and becomes nonsignificant once controlling for behavior. In other words, the teacher gender bias against boys in language is totally accounted for by boys' more disruptive behavior. Conversely, in mathematics, when the preexisting bias was in favor of boys, controlling for behavior further amplified this bias in favor of boys. In this case, the gender bias against girls in mathematics is not accounted for by gender differences in behavior; On the contrary, this gender bias would be even larger if boys and girls behaved the same. As suggested by Robinson-Cimpian et al. (2014), one possible interpretation is that girls must exhibit better behavior than equally achieving boys in order to be perceived as having the same math abilities as boys. However, since girls behave better than boys on average, this suggests that there is a systematic bias against them. Similarly, boys would need to behave as well as girls on average in order to be perceived as equally competent as girls in language.

Finally, behavior, measured by externalizing symptoms, was negatively associated with teachers' perceptions of children's abilities, but this association varied across genders and domains. More specifically, in language, this relation was of similar magnitude for boys and girls. However, in mathematics, the complete-case analyses suggested that this relation was more strongly negative for girls than for boys. This suggests that in mathematics, teachers are not only biased by behavior when assessing students' skills, but they may apply this bias more strongly when assessing girls' than when assessing boys' abilities. However, this interaction was attenuated in the multiple imputation analyses and became borderline non-significant, so this interpretation should be considered tentative. One possible explanation for this pattern, if confirmed, is that, since girls generally display better behavior on average, those who exhibit more disruptive behaviors deviate more from expected norms than similarly behaving boys. As a result, teachers might perceive them as even less capable in mathematics. Interestingly, this pattern does not appear for language. This might be because language is stereotypically viewed as a "feminine" domain, leading teachers to expect both well-behaved and disruptive girls to perform well when considering language outcomes, thereby weakening the association between behavior and perceived skills in language.

This study suggests that teachers are influenced by both gender and externalizing behavior when assessing students' abilities. In additional analyses (Table S3 and S4), we found no evidence that biases varied by teachers' gender. However, this may be due to the small proportion of male teachers in primary school, representing only 6%, 8%, and 17% of our sample in preschool, first grade, and fourth grade, respectively. These biases may extend beyond teachers' perceptions to impact the grades they give students, which in turn may influence not only students' perceptions of their own abilities but also their future academic and career opportunities (Lavy & Sand, 2018), potentially reinforcing gender disparities in certain fields. However, subjective perception of pupils' level and grading may not necessarily follow the same trends. For instance, a study on grades given by French teachers to middle-schoolers found a gender bias in favor of girls in mathematics, and further suggested that this bias helped reduce the gender gap in science track choice in high school (Terrier, 2020). The two results may not be contradictory: on the one hand, teachers' perception of pupils' level is biased by gender stereotypes and behavior; on the other hand, teachers may show a grading bias in favor of girls in mathematics, perhaps in an attempt to deliberately counter gender stereotypes.

It might be tempting to try and draw educational recommendations from our results. For one thing, biased perception is a distortion of reality, so it seems natural to try and correct such distortion and promote a more faithful perception of pupils' level. However, in the present study, we are not talking of grading (which does have obvious consequences) but of teachers' internal perception of pupils' level, interrogated purely for research purposes without any stakes. At this stage, we do not have any evidence that this internal perception has any tangible effect on grading or on pupils' self-perception or learning. We do not even know to what extent this perception might be modifiable, if teachers were made aware of their own biases. Furthermore, given that teachers' perceptions are shaped by broader societal norms, including gender stereotypes, it is possible that these biases could evolve over time, especially as discussions on gender equality become more prominent in education. Whether teacher training and professional development could help correct such biases remains an open question. Therefore, more research may be needed to investigate the causal effects of teachers' biased perception of pupils' level on their future achievement, as well as the extent to which these biases could be reduced.

A major strength of this study is the large sample size, up to 23,349 observations and 7276 individual students in any given analysis. This dataset also enabled us to study teachers' biases over three time points and across two domains (French and mathematics). This allowed us to assess temporal patterns, and study two domains in which the direction and evolution of the gender differences in achievement differ.

One limitation of this study is the generalizability of our findings. The dataset used is a birth cohort, subject to attrition. Thus, wealthier families tend to remain in the cohort longer, and thus children with more timepoints available also tend to have better outcomes in term of achievement and externalizing scores. To address this issue, we conducted supplementary multiple imputation analyses to assess the robustness of the findings to missing data. These analyses yielded broadly similar conclusions for the main effects, but weaker evidence for the more complex interaction effects. This pattern suggests that the central findings are relatively robust, whereas the more fine-grained temporal and moderation effects should be interpreted with greater caution. Future studies should investigate these questions on more demographically representative samples that better reflect the diversity of the broader population.

It is important to acknowledge that school achievement scores may be influenced, to some extent, by teachers' gender biases, as they are responsible for reporting student performance. However, in the worst-case scenario, this would lead to an underestimation of the actual biases in teachers' perceptions, rather than an overestimation. This is because any teacher bias affecting reported school scores would likely align with their perceptions, thereby reducing the observed gap between perceived and actual abilities.

Another limitation is that our measure of externalizing behavior is based on teacher assessments, who may be biased not only in their perception of skills, but also of students' behaviors. If, as we showed for skills' perceptions, teachers overestimate existing gaps, then they might tend to overestimate girls' good behavior and boys' bad behavior. If this is the case, we might be overestimating the gender bias remaining after accounting for behavior in math, and underestimating it in language. However, if teachers are biased by behavior, it should be by the behavior that they perceive rather than the objective behavior of students. Thus, it makes sense to use students' behavior as rated by the teachers in our analyses.

In conclusion, our findings highlight the significant role not only of objective performance but also of gender and behavior in shaping teachers' perception of students' abilities in mathematics and language. These biases, while not necessarily translating directly into grading disparities, may still influence students' academic trajectories, for instance by shaping teacher-student interactions. However, the extent to

which these biases affect actual academic outcomes remains an open question. Future research should further explore the causal links between teachers' perceptions and student achievement, as well as the potential for interventions aimed at mitigating bias. Understanding how teachers form their perception of student skill, and whether these biases can be corrected, remains a crucial step toward fostering a more equitable educational environment.

CRediT authorship contribution statement

Lilas Gurgand: Writing – original draft, Visualization, Software, Formal analysis, Data curation, Conceptualization. **Hugo Peyre:** Writing – review & editing, Validation, Supervision, Formal analysis, Conceptualization. **Xavier Thierry:** Writing – review & editing, Validation, Supervision, Formal analysis, Conceptualization. **Franck Ramus:** Writing – review & editing, Validation, Supervision, Formal analysis, Conceptualization.

Aknowlegment

The Elfe survey is a joint project between the French Institute for Demographic Studies (INED) and the National Institute of Health and Medical Research (INSERM), in partnership with the French blood transfusion service (Etablissement français du sang, EFS), Santé publique France, the National Institute for Statistics and Economic Studies (INSEE), the Direction générale de la santé (DGS, part of the Ministry of Health and Social Affairs), the Direction générale de la prévention des risques (DGPR, Ministry for the Environment), the Direction de la recherche, des études, de l'évaluation et des statistiques (DREES, Ministry of Health and Social Affairs), the Département des études, de la prospective et des statistiques (DEPS, Ministry of Culture), and the Caisse nationale des allocations familiales (CNAF), with the support of the Ministry of Higher Education and Research and the Institut national de la jeunesse et de l'éducation populaire (INJEP). Via the RECONAI platform, it receives a government grant managed by the National Research Agency under the "Investissements d'avenir" programme (ANR-11-EQPX-0038 and ANR-19-COHO-0001). Additional funding came from ANR (ANR-17-EURE-0017 and ANR-10-IDEX-0001-02).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.learninstruc.2026.102391.

Data availability

The authors do not have permission to share data.

References

Alvidrez, J., & Weinstein, R. S. (1999). Early teacher perceptions and later student academic achievement. *Journal of Educational Psychology, 91*(4), 731–746. https://doi.org/10.1037/0022-0663.91.4.731

Bates, C., & Nettelbeck, T. (2001). Primary school teachers' judgements of reading achievement. *Educational Psychology, 21*(2), 177–187. https://doi.org/10.1080/01443410020043878

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

Breda, T., & Hillion, M. (2016). Teaching accreditation exams reveal grading biases favor women in male-dominated disciplines in France. *Science, 353*(6298), 474–478. https://doi.org/10.1126/science.aaf4372

Breda, T., & Ly, S. T. (2015). Professors in core science fields are not always biased against women : Evidence from France. *American Economic Journal: Applied Economics, 7*(4), 53–75. https://doi.org/10.1257/app.20140022

Bugler, M., McGeown, S. P., & St Clair-Thompson, H. (2015). Gender differences in adolescents' academic motivation and classroom behaviour. *Educational Psychology, 35*(5), 541–556. https://doi.org/10.1080/01443410.2013.849325

Campbell, T. (2015). Stereotyped at seven? Biases in teacher judgement of pupils' ability and attainment. *Journal of Social Policy, 44*(3), 517–547. https://doi.org/10.1017/S0047279415000227

Charles, M. A., Thierry, X., Lanoe, J.-L., Bois, C., Dufourg, M.-N., Popa, R., Cheminat, M., Zaros,, C., & Geay, B. (2019). Cohort profile : The French national cohort of children (ELFE) : Birth to 5 years. *International Journal of Epidemiology, 49*(2), 368–369. https://doi.org/10.1093/ije/dyz227

Cimpian, J. R., Lubienski, S. T., Timmer, J. D., Makowski, M. B., & Miller, E. K. (2016). Have gender gaps in math closed? Achievement, teacher perceptions, and learning behaviors across two ECLS-K cohorts. *AERA Open, 2*(4), Article 2332858416673617. https://doi.org/10.1177/2332858416673617

Contreras, D. (2024). Gender differences in grading : Teacher bias or student behaviour? *Education Economics, 32*(6), 762–785. https://doi.org/10.1080/09645292.2023.2252620

Di Liberto, A., Casula, L., & Pau, S. (2022). Grading practices, gender bias and educational outcomes : Evidence from Italy. *Education Economics, 30*(5), 481–508. https://doi.org/10.1080/09645292.2021.2004999. Scopus.

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement : Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction, 22*(4), 271–280. https://doi.org/10.1016/j.learninstruc.2011.08.003

Eteve, Y., Garnero, M., & Paillet, V. (2025). *Évolution des écarts de performances entre filles et garçons en mathématiques, au fil du temps et de la scolarité. DEPP.* Note d'Information. n° 25-04.

Eyal, T., & Epley, N. (2017). Exaggerating accessible differences : When gender stereotypes overestimate actual group differences. *Personality and Social Psychology Bulletin, 43*(9), 1323–1336. https://doi.org/10.1177/0146167217713190

Falch, T., & Naper, L. R. (2013). Educational evaluation schemes and gender gaps in student achievement. *Economics of Education Review, 36*, 12–25. https://doi.org/10.1016/j.econedurev.2013.05.002

Ferman, B., & Fontes, L. F. (2022). Assessing knowledge or classroom behavior? Evidence of teachers' grading bias. *Journal of Public Economics, 216*, Article 104773. https://doi.org/10.1016/j.jpubeco.2022.104773

Fernandez Castelao, C., & Kröner-Herwig, B. (2014). Developmental trajectories and predictors of externalizing behavior : A comparison of girls and boys. *Journal of Youth and Adolescence, 43*(5), 775–789. https://doi.org/10.1007/s10964-013-0011-9

Fischer, J.-P., & Thierry, X. (2022). Boy's math performance, compared to girls', jumps at age 6 (in the ELFE's data at least). *British Journal of Developmental Psychology, 40*(4), 504–519. https://doi.org/10.1111/bjdp.12423

Gentrup, S., Lorenz, G., Kristen, C., & Kogan, I. (2020). Self-fulfilling prophecies in the classroom : Teacher expectations, teacher feedback and student achievement. *Learning and Instruction, 66*, Article 101296. https://doi.org/10.1016/j.learninstruc.2019.101296

Glascoe, F. P. (2001). Can teachers' global ratings identify children with academic problems? *Journal of Developmental and Behavioral Pediatrics, 22*(3), 163.

Hofer, S. I. (2015). Studying gender bias in physics grading : The role of teaching experience and country. *International Journal of Science Education, 37*(17), 2879–2905. https://doi.org/10.1080/09500693.2015.1114190

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

Hollenstein, L., Brühwiler, C., & König, J. (2024). Who benefits the most? Teacher expectations and their relation with students' achievement in mathematics in primary school. *European Journal of Social Psychology, 54*(3), 658–669. https://doi.org/10.1002/ejsp.3035

Kiuru, N., Nurmi, J.-E., Leskinen, E., Torppa, M., Poikkeus, A.-M., Lerkkanen, M.-K., & Niemi, P. (2015). Elementary school teachers adapt their instructional support according to students' academic skills : A variable and person-oriented approach. *International Journal of Behavioral Development, 39*(5), 391–401. https://doi.org/10.1177/0165025415575764

Kowalski, K., Brown, R. D., Pretti-Frontczak, K., Uchida, C., & Sacks, D. F. (2018). The accuracy of teachers' judgments for assessing young children's emerging literacy and math skills. *Psychology in the Schools, 55*(9), 997–1012. https://doi.org/10.1002/pits.22152

Kremer, K. P., Flower, A., Huang, J., & Vaughn, M. G. (2016). Behavior problems and children's academic achievement : A test of growth-curve models with gender and racial differences. *Children and Youth Services Review, 67*, 95–104. https://doi.org/10.1016/j.childyouth.2016.06.003

Lavy, V., & Sand, E. (2018). On the origins of gender gaps in human capital : Short- and long-term consequences of teachers' biases. *Journal of Public Economics, 167*, 263–279. https://doi.org/10.1016/j.jpubeco.2018.09.007

Martinot, P., Colnet, B., Breda, T., Sultan, J., Touitou, L., Huguet, P., Spelke, E., Dehaene-Lambertz, G., Bressoux, P., & Dehaene, S. (2025). Rapid emergence of a maths gender gap in first grade. *Nature*. https://doi.org/10.1038/s41586-025-09126-4

Martinot, P., Dehaene, S., Potier-Watkins, C., Sprenger-Charolles, L., & Ziegler, J. (2021). Qu'apprend-on des évaluations de CP-CE1. *Conseil scientifique de l'éducation nationale*. https://www.reseau-canope.fr/fileadmin/user_upload/Projets/conseil_scientifique_education_nationale/Note_CSEN_2021_03.pdf.

Meissel, K., Meyer, F., Yao, E. S., & Rubie-Davies, C. M. (2017). Subjectivity of teacher judgments : Exploring student characteristics that influence teacher judgments of

student ability. *Teaching and Teacher Education, 65*, 48–60. https://doi.org/10.1016/j.tate.2017.02.021

Nurmi, J.-E., Kiuru, N., Lerkkanen, M.-K., Niemi, P., Poikkeus, A.-M., Ahonen, T., Leskinen, E., & Lyyra, A.-L. (2013). Teachers adapt their instruction in reading according to individual children's literacy skills. *Learning and Individual Differences, 23*, 72–79. https://doi.org/10.1016/j.lindif.2012.07.012

Paleczek, L., Seifert, S., & Gasteiger-Klicpera, B. (2017). Influences on teachers' judgment accuracy of reading abilities on second and third grade students : A multilevel analysis. *Psychology in the Schools, 54*(3), 228–245. https://doi.org/10.1002/pits.21993

Pielmeier, M., Huber, S., & Seidel, T. (2018). Is teacher judgment accuracy of students' characteristics beneficial for verbal teacher-student interactions in classroom? *Teaching and Teacher Education, 76*, 255–266. https://doi.org/10.1016/j.tate.2018.01.002

Ready, D. D., & Wright, D. L. (2011). Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities : The role of child background and classroom context. *American Educational Research Journal, 48*(2), 335–360. https://doi.org/10.3102/0002831210374874

Robinson-Cimpian, J. P., Lubienski, S. T., Ganley, C. M., & Copur-Gencturk, Y. (2014). Teachers' perceptions of students' mathematics proficiency may exacerbate early gender gaps in achievement. *Developmental Psychology, 50*, 1262–1281. https://doi.org/10.1037/a0035073

Ross, J. A., & Kostuch, L. (2011). Consistency of report card grades and external assessments in a Canadian province. *Educational Assessment, Evaluation and Accountability, 23*(2), 159–180. https://doi.org/10.1007/s11092-011-9117-3

Schuster, C., Narciss, S., & Bilz, J. (2021). Well done (for someone of your gender) ! experimental evidence of teachers' stereotype-based shifting standards for test grading and elaborated feedback. *Social Psychology of Education, 24*(3), 809–834. https://doi.org/10.1007/s11218-021-09633-y

Steinmann, I. (2024). Unexpected gender differences in teacher ratings of academic skills and school track recommendations. *Studies In Educational Evaluation, 83*, Article 101384. https://doi.org/10.1016/j.stueduc.2024.101384

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement : A meta-analysis. *Journal of Educational Psychology, 104*(3), 743–762. https://doi.org/10.1037/a0027627

Terrier, C. (2020). Boys lag behind : How teachers' gender biases affect student achievement. *Economics of Education Review, 77*, Article 101981. https://doi.org/10.1016/j.econedurev.2020.101981

Urhahne, D., & Wijnia, L. (2021). A review on the accuracy of teacher judgments. *Educational Research Review, 32*, Article 100374. https://doi.org/10.1016/j.edurev.2020.100374

Wijnia, L., Loyens, S. M. M., Derous, E., & Schmidt, H. G. (2016). University teacher judgments in problem-based learning : Their accuracy and reasoning. *Teaching and Teacher Education, 59*, 203–212. https://doi.org/10.1016/j.tate.2016.06.005

Zanga, G., & De Gioannis, E. (2023). Discrimination in grading : A scoping review of studies on teachers' discrimination in school. *Studies In Educational Evaluation, 78*, Article 101284. https://doi.org/10.1016/j.stueduc.2023.101284

Zhu, M., & Urhahne, D. (2015). Teachers' judgements of students' foreign-language achievement. *European Journal of Psychology of Education, 30*(1), 21–39. https://doi.org/10.1007/s10212-014-0225-6

Lilas Gurgand is a PhD student at Ecole Normale Supérieure, Paris, researching gender differences in educational performance. Hugo Peyre, a researcher and child psychiatrist at Montpellier University Hospital, focuses on cognitive development epidemiology, investigating individual differences in childhood cognitive abilities and their genetic and environmental determinants. Xavier Thierry, a demographer at INED, conducts longitudinal studies on child development, notably through the Elfe study. Franck Ramus, a researcher at ENS, studies cognitive development, academic achievement, and developmental disorders like dyslexia.