A teacher like me: the role of teacher gender representation and gender stereotypes in education

Doornkamp, L.

CHAPTER **FIVE**

Understanding gender bias in teachers' grading: The role of gender stereotypical beliefs

Previously published as: Doornkamp, L., Van der Pol, L.D., Groeneveld, S., Mesman, J., Endendijk, J.J, & Groeneveld, M.G. (2022). Understanding Gender Bias in Teachers' Grading: The Role of Gender Stereotypical Beliefs. *Teaching and Teacher Education, 118*, 103826.

ABSTRACT

This study experimentally tested the influence of secondary school students' gender on Dutch language and math teachers' grading (N = 358) and examined the role of teachers' gender and gender stereotypes in gender grading bias. Results showed that grading, on average, was not gender biased. However, differences between teachers' gender grading bias were related to their gender stereotypes. Although we found no direct effect of teachers' gender on grading bias, for math we found an indirect effect through their gender stereotypes. This study provides evidence for the role of teachers' gender stereotypes in gender grading bias that thus far had only been assumed.

5.1 INTRODUCTION

Gender bias in teachers' grading refers to the differences in grades between male and female students with the same level of academic skills (Protivínský & Münich, 2018). As grades are interpreted as objective measures of students' level of academic skills, gender biases in teachers' grading can have far-reaching consequences. For instance, gender bias in teachers' grading was found to be associated with students' enrollment in advanced level math courses which has long-term implications for students' future careers (Lavy & Sand, 2018).

Despite the multitude of studies on the topic of gender grading bias, still little is known about the causal effect of students' gender on teachers' grading as experimental designs to study the effect are rare (Hanna & Linden, 2012; Hinnerich et al., 2011). Further, although scholars urged that the role of teachers' characteristics in gender grading bias should be explored (Falch & Naper, 2013; Lavy, 2008; Matějů & Smith, 2015; Protivínský & Münich, 2018), thus far, only a few studies empirically tested the association between teachers' gender and gender grading bias (Lavy, 2008; Lindahl, 2016). Moreover, the role of teachers' implicit gender stereotypes has often been assumed, but not empirically tested, to underly gender grading bias (Breda & Ly, 2015; Burgess & Greaves, 2013; Gibbons & Chevalier, 2008; Lavy, 2008; Lavy & Sand, 2018; Matějů & Smith, 2015).

In the current study we use a field experiment among Dutch teachers in training in the subjects Dutch language and math to contribute to the understanding of gender grading bias. We aim to answer the research question: *to what extent is teachers' grading gender biased and what is the role of teachers' gender and teachers' implicit gender stereotypes in gender grading bias?* By manipulating gendered names on tests, we examine the extent to which teachers' grading is influenced by students' perceived gender. We further investigate whether gender grading bias is associated with teachers' gender and implicit gender stereotypes and we propose and empirically test a model that explains the potential influence of the gender of the teacher on gender grading bias through teachers' implicit gender-typed associations and gender-typed expectations. In doing so, we are able to make inferences about causality, about the role of teacher characteristics and about a mechanism underlying gender grading bias.

5.2 GENDER GRADING BIAS

The literature on gender grading bias consists of several studies without a clear-cut theory that binds them together. Studies on the phenomenon of gender grading bias date back to at least 1967 (Caldwell & Hartnett, 1967). Since then, the topic has been studied in the fields of economics, education, and sociology, each using different terminologies to refer to the same phenomenon. Gender grading bias (Berg et al., 2020), gender grading gap (Falch & Naper, 2013), gender disparities in grading (Cornwell et al., 2013), gender bias

in examination (Stobart et al., 1992), evaluation bias (Breda & Ly, 2015), assessment bias (Lindahl, 2016), and teacher discrimination (Lavy, 2008) all refer to teachers' systematic over- or underassessment of a group of students based on students' gender. The literature on gender grading bias treats gender as a binary concept as it describes teachers' differential grading of male versus female students (e.g., Protivínský & Münich, 2018). We are aware that gender can entail more than binary categories male and female. However, societies generally are still organized in this binary way, categorizing humans as either male or female. This categorization lies at the basis of gender stereotypes and the accompanying gender discriminative behaviors in everyday life (Hyde et al., 2019). As we study the role of gender stereotypes in teachers' actual grading behavior, we follow the gender binary approach.

The vast majority of studies on the topic used secondary data and relied on the comparison of students' exam scores obtained from non-blind grading (generally the teacher is the grader and is aware of student gender) to students' exam scores obtained from blind grading (the assessor is not aware of student gender, these scores are interpreted as objective measures of students' academic level of skills). If the gender differences in test scores awarded by teachers (non-blind grading) were larger than the gender differences in blindly graded standardized entrance exams (Protivínský & Münich, 2018), central exit exams (Falch & Naper, 2013), externally graded state exams (Lavy, 2008), or national tests (Lindahl, 2016), these studies concluded that there is gender grading bias. Generally, blind grading produces lower test scores for all students (male and female) than non-blind (teacher) grading (Hinnerich et al., 2011). More importantly, whereas blind grading shows no or little differences in average male and female test scores, teachers do seem to differentiate between male and female students in their grading, i.e., gender grading bias (e.g., Protivínský & Münich, 2018).

The differences in male and female scores as a result of teachers' grading is generally to the disadvantage of male students, meaning that male students systematically receive lower scores than female students in teacher grading (Protivínský & Münich, 2018). Tens of previous studies found evidence for this grading bias against male students across several subjects, in different countries, and for different educational levels (for an overview see Protivínský & Münich, 2018). Whereas only a handful of studies found biased grading against female students in different subjects like math, but also in humanities and English (Breda & Ly, 2015; Gibbons & Chevalier, 2008; Lavy & Sand, 2018), or did not find gender grading bias at all (Hanna & Linden, 2012; Hinnerich et al., 2011). Notably, studies that did not find gender bias in grading were the only few studies that have experimentally tested the influence of students' gender on teachers' grading (Hanna & Linden, 2012; Hinnerich et al., 2011). Hence, little is known about whether there is a causal relation between students' gender and teachers' grading. Insight in cause-effect provides the basis for methods to reduce grading bias. In all, most of the evidence on the topic of gender grading bias points in the direction of a grading bias in teachers' grading against male students, we therefore hypothesize:

H1: Perceived male gender of students is negatively associated with teachers' grading.

5.2.1 THE ROLE OF TEACHERS' GENDER AND GENDER STEREOTYPES IN GENDER GRADING BIAS

The role of teacher characteristics in gender grading bias has often been suggested in the literature (e.g., Matějů & Smith, 2015), but only a handful of studies have empirically tested the associations between teacher characteristics and gender grading bias (Lavy, 2008; Lavy & Sand, 2018; Lindahl, 2016). Teachers' gender, age, experience, number of sons and daughters, and marital status were found to be related to gender grading bias, indicating that gender grading bias is indeed sensitive to teacher characteristics (Lavy, 2008; Lavy & Sand, 2018).

Teachers' own gender (whether they classify themselves as male or female) can play an important role in gender grading bias. In schools, some subjects are stereotyped as typically masculine (e.g., math) or typically feminine (e.g., languages). Although the magnitude of the 'math = masculine' gender stereotype can differ across cultures, the content of the stereotype generally is the same (Breda et al., 2020; Nosek et al., 2009). A female teacher in a stereotypically masculine subject (e.g., a female math teacher) may have accumulated counter-gender stereotypic experiences that affect her gender bias in grading (Lavy & Sand, 2018). Because of her personal experience that a woman can perform well in a subject that is stereotyped as masculine, a female math teacher can have weaker stereotypic ideas and beliefs regarding male and female students in math (Bussey & Bandura, 1999). These experiences might even result in counter-stereotypic beliefs that might turn into counter-stereotypic behaviors (Crisp et al., 2009; Endendijk et al., 2013). In general, though, experiences affect the strength of gender stereotypes but not the direction of the stereotype (i.e., gender stereotypes are flexible but remain gender stereotypic) (Charlesworth & Banaji, 2021; Xu & Meier, 2021). Therefore, counter-stereotypic experiences are more likely to result in weaker gender stereotypical beliefs and behaviors (for a meta-analysis see Lenton et al., 2009). Following this logic, female teachers would have weaker gender grading bias in math than male teachers, and male teachers would have weaker gender grading bias in languages than female teachers (Lavy, 2008).

H2a: Female math teachers display less gender grading bias than male math teachers.

H2b: Male Dutch language teachers display less gender grading bias than female Dutch language teachers.

Indeed, teachers' gender stereotypes are presumed to underly gender grading bias (Breda & Ly, 2015; Hanna & Linden, 2012; Hinnerich et al., 2011; Lavy, 2008; Lavy & Sand, 2018;

Lindahl, 2016; Matějů & Smith, 2015). Gender stereotypes are socially constructed ideas that provide descriptions of what men and women are like, and prescriptions about what men and women should be like (Ellemers, 2018). The content of gender stereotypes is assumed to be similar across different cultures, however it should be noted that recent research demonstrated that culture can shape how men and women are perceived based on gender stereotypes (Obioma et al., 2021). Obioma et al. (2021) showed that the content of the gender stereotype 'men are agentic, women are communal' is similar in Germany and Nigeria. However, in their study, in Germany men were perceived as more agentic than women while men and women were perceived as equally communal. In Nigeria, men and women were perceived as equally agentic while women were perceived as more communal than men (Obioma et al., 2021). Notwithstanding differences in perception based on gender stereotypes, individuals who have strong gender stereotypes make a strong distinction between men and women and between typically masculine and typically feminine attributes and behaviors (i.e., gender-typed individuals/gender-schematic individuals) (Bem, 1981). When confronted with someone else's behavior, these individuals are likely to perceive the behavior in terms of congruence with their ideas about and expectations of males and females (Bem, 1981). Behavior in line with these ideas and expectations are rewarded, whereas incongruent behavior is penalized. Consequently, the distinction between males and females based on gender stereotypes can result in different evaluations of the same behavior of males and females (e.g., gender grading bias) (e.g., Burgess & Borgida, 1999). Following this logic, teachers who have strong gender stereotypes are likely to make a strong distinction about what is typical and appropriate for male and female students. As a result, the same behavior in males and females is evaluated differently which makes gender bias in grading more likely.

An individual can have implicit and explicit gender stereotypes. Implicit gender stereotypes refer to the *automatic and less controllable* distinctions individuals make, whereas explicit gender stereotypes refer to individuals *purposeful and controllable* distinguishment about what is typical of and appropriate for males and females (Fazio & Olson, 2003). Implicit and explicit gender stereotypes are only weakly correlated, in part because of explicit measures' sensitivity to social desirability or lack of insight in one's own stereotypes (Fazio & Olson, 2003). Teachers can report that they do not distinguish between male and female students, whereas observations or implicit measures can prove otherwise (Jones & Myhill, 2004; Nürnberger et al., 2016). In a country such as the Netherlands, in which gender equality is considered important, implicit measures of gender stereotypes could provide a better understanding of the (implicit) role of gender stereotypes in teachers' grading behaviors.

In this study, we distinguish between implicit gender-typed associations and implicit gender-typed expectations. Both are unconscious cognitive processes through which one differentiated between male and female students. Implicit gender-typed associations refer to teachers' automatic and unconscious associations between math and masculinity, and between languages and femininity. Generally, people tend to associate math with

masculinity and languages with femininity based on the traditional belief that math is the domain in which men can excel and that languages is the domain in which women can excel (Nosek et al., 2002b). As a result, teachers can have stereotypical perceptions of male and female students' capabilities in math which can turn in to different assessments of male and female students (Riegle-Crumb & Humphries, 2012). Math teachers who have strong associations between math and masculinity and between language and femininity make a strong distinction between males' and females' capabilities in the subjects, which might result in a stronger gender grading bias. On the other hand, teachers who have weak associations between gender and subjects might be more equal in their grading of male and female students.

In a related vein, implicit gender-typed expectations refer to teachers' automatic and unconscious expectations regarding male and female students' talent and effort based on gender stereotypes. Teachers can expect female students to have less talent in math than male students (e.g., Espinoza et al., 2014; Heller et al., 2001; Riegle-Crumb & Humphries, 2012), and male students to have less talent in languages than female students (e.g., Schmenk, 2004; Siegle & Reis, 1998). On the other hand, based on negative stereotypes regarding male students' interest, motivation, and achievements in schools (Brown & Stone, 2016), teachers can expect less effort from male students than from female students (Heyder & Kessels, 2017; Jones & Myhill, 2004; Siegle & Reis, 1998). Again, teachers who differ in their expectations of male and female students might display gender bias in their grading.

Indeed, literature on gender bias in job applications shows that negative expectations can lead to stricter assessments and positive expectations can lead to more lenient assessments (Powell, 1986). These positive or negative expectations can be informed by gender stereotypes. As a result, female applicants have less chance of being assessed positively by recruiters for jobs that are perceived as typically masculine (e.g., management positions), and male applicants have less chance of being assessed positively by recruiters for jobs that are perceived as typically feminine (e.g., communication positions) (Cole et al., 2004). Following this logic, teachers who have negative expectations regarding male or female students' talent or effort in a subject, might assess these students more strictly, and vice versa. At the same time, teachers with more equal expectations regarding male and female students' talent and effort, might assess male and female students more equally.

In sum, we expect that the stronger a teacher perceives male and female students as different in terms of their suitability, talent and effort in math or languages, the stronger the teacher shows differential grading of male and female students (Dasgupta, 2004; Gansen, 2019). Vice versa, we expect teachers with weaker gender-typed associations and gender-typed expectations regarding students' talent and effort to have weaker gender grading bias.

> **H3:** Weaker implicit gender-typed associations and weaker gender-typed expectations regarding students' talent and effort are related to weaker gender grading bias.

5.2.2 TEACHERS' GENDER, GENDER STEREOTYPES, AND GENDER GRADING BIAS

Based on the literature discussed thus far it is likely that implicit gender-typed associations and gender-typed expectations at least partly mediate the relation between gender of the teacher and gender grading bias. That is, the implicit associations between gender and math and languages may be weaker among female math teachers and male language teachers compared to male math teachers and female language teachers (Martin & Dinella, 2012; Smeding, 2012). Consequently, these weaker implicit gender-typed associations help shape biased expectations regarding male and female students' talent and effort in the subjects (Denessen et al., 2022; Muntoni & Retelsdorf, 2018; Van den Bergh et al., 2010). Teachers' expectations are shaped by a number of co-occurring factors including teachers' implicit associations (Denessen et al., 2022). Teachers' expectations may in turn be related to gender grading bias. The proposed mediation model is depicted in Figure 5.1.

> **H4:** The association between teachers' gender and gender grading bias is mediated by teachers' implicit gender-typed associations and implicit gender-typed expectations.

Female math teachers may have weaker gender-typed associations, that are in turn related to weaker gender-typed expectations regarding students' talent and effort, and weaker gender grading bias than male math teachers. Male Dutch language teachers may have weaker gender-typed associations, that are in turn related to weaker gender-typed expectations regarding students' talent and effort, and weaker gender grading bias than female Dutch language teachers.

Our hypotheses concentrated on teachers' gender and implicit gender stereotypes in relation to gender grading bias without hypothesizing the direction of effects (i.e., effects on grading bias against male or female students). Because the examination of psychological mechanisms that could underly teachers' gender bias in grading is still largely uncharted territory, we decided to explore the full spectrum of teachers' gender stereotypes and gender grading bias, without excluding participants with contra-stereotypical associations, expectations, or grading biases a priori. We thus focused on the extent to which the distinction between male and female students based on gender stereotypes can result in differential grading of male and female students. However, our data allows us to explore the direction of the (mediating) effects of implicit gender stereotypes both on grading biases against male and female students in Dutch language and math. These additional explorations will provide a complete picture of the role of implicit gender stereotypes in gender grading bias.

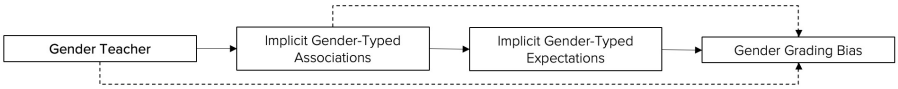

Figure 5.1. The Indirect Effect of Teachers' Gender on Gender Grading Bias

5.3 METHOD

This study's field experiment is part of the longitudinal research project 'Girls in Science' that examines gender socialization in the family and school context in the Netherlands from 2017 until 2022. The experimental data for this study were collected between February 2020 and August 2021 among teachers in training in the subjects Dutch language and math in secondary education. For the sake of legibility, in the following sections we use the term 'teacher/teachers' to refer to the participating teachers in training in our study.

5.3.1 SAMPLE

The teachers who participated in this study were recruited through their trainers. Trainers from all the 18 institutions that provide regular teacher training programs for secondary education in Dutch language and math in the Netherlands were invited to cooperate by e-mail (trainers from online institutions were not invited). Thirty-two trainers from 15 institutions agreed to cooperate, the others did not participate because of time issues or they did not respond to the invitation. The trainers either asked their students (the teachers) to participate in the research or facilitated the researcher to ask the students (teachers) to participate.

Based on the calls, 159 Dutch language teachers and 268 math teachers signed up to participate in the research. In all, 28 cases for Dutch language and 41 cases for math were regarded as missing data because their evaluations got lost in the Dutch mail system (2 Dutch language, 5 math), they did not send back their evaluations (24 Dutch language teachers, 26 math), they did not complete the evaluation (1 Dutch language, 5 math), or they did not complete the questionnaire (1 Dutch language, 5 math). Only the participants with scores on all relevant variables were included in the study, resulting in N = 131 for Dutch language (mean age = 33.73, *SD* = 12.098, 76% female), and N = 227 for math (mean age = 29.44, *SD* = 12.027, 56% female). Participants that were excluded from the analyses did not deviate on background characteristics from the participants that we included.

The distribution of male and female teachers across math and Dutch language is skewed, but similar to the gender distribution in Dutch secondary schools. Dutch language teachers in the Netherlands are disproportionately female (72%). For math, the distribution is slightly more balanced with 58% male and 42% female teachers (numbers based on open educational data 2018: DUO 2020).

Both the Dutch language and math teachers performed internships at secondary schools for 10 hours (Dutch language *SD* = 8.53, math *SD* = 7.49) a week. For Dutch language, 27% of

the participants received their training on a university (first degree) whereas 73% received their training on a university for applied sciences (second degree). Furthermore, 41% of the Dutch language teachers had experience with teaching. For math, 15% received their training at university, whereas the other 85% received it at a university for applied sciences. In math, 37% of the teachers in training had experience with teaching.

5.3.2 STUDY DESIGN

The participating teachers each (1) graded a set of 24 tests made by last-year students in higher secondary education, (2) assessed each test on invested talent and effort, and (3) filled in an online questionnaire. Data collection took place at home and participation required approximately 2.5 hours. In return for their participation, teachers received individual feedback on their grading practices, the results of the study, and a small gift.

The student names above the 24 tests were manipulated. Originally, 10 female students and 14 male students completed the Dutch language tests and 14 female students and 10 male students completed the math tests. The 24 tests were divided into two groups with the same average scores; one group was given names that are generally considered as male, whereas the other group was given names that are generally considered as female. We tried to maximize the ecological validity of our experiment within the limitations of its practical feasibility. To create conditions that are transferable to the real-life exam evaluation situation, we designed a grading experiment using real school exams with real students' answers. The study design is depicted in Figure 5.2.

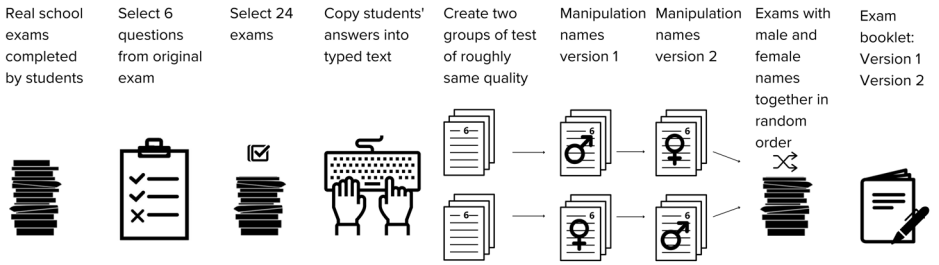

Figure 5.2. Study Design

The Dutch language and math school exams and students' answers to the exams that were used in the experiment were retrieved from a Dutch language and a math teacher from different Dutch secondary schools in the western region of the Netherlands. Both exams were conducted in 2019 by students in the final year of their secondary education.

We selected six open-answer questions from both school exams (Dutch language and math) to avoid asking too much time from the participants. We used the following two

criteria in the selection of these questions: (1) there had to be ample room for participants' own interpretation in their assessment of the given answers to the question, and (2) there was sufficient between-student range in the points given for answers to a question by the original assessor to ensure that the quality of the answers to the question differed.

From the groups of students who originally took the Dutch language exam (N = 39) and the math exam (N = 45), 24 students for each subject were selected to mimic real-life situations in which teachers must evaluate the tests of one class with approximately 25 students. The selected students' answers (including mistakes, misspellings, and notes) were copied into typed text and printed to eliminate any reference to the gender of the original student through for example handwriting.

To ensure that we could form two groups of tests (one group with names that are generally considered as male and one group with names that are generally considered as female) of equal quality in terms of answers given to the selected questions, we recruited 10 experienced teachers for each subject to blindly grade the 24 tests (i.e., without student names). Based on these blind evaluations, two groups were created with 11 tests each.

We used common Dutch female names (8), Dutch male names (8), Turkish female names (4), and Turkish male names (4) that were randomly distributed over the groups of tests (for an overview of the names see Appendix A). To reflect the Dutch multicultural society, we used Turkish names in addition to Dutch names as people with a Turkish migration background are, after the native Dutch, the largest ethnic-cultural group in Dutch society (CBS, 2020). To control for an unforeseen effect of differences in students' answers we created two versions. The tests with male names in the first version had female names in the second version, and the tests with female names in the first version had male names in the second version. The versions were randomly distributed across the participants. Furthermore, to control for grading order and grading fatigue, the order of tests was randomly distributed and different for the two versions. Additionally, the first test in the set alternated within each version.

In addition to the evaluation of the tests, participants were asked to estimate students' talent and effort based on each test. The 24 tests, including these questions, were presented in a booklet. A short introduction with general information about the experiment was added, but nothing was said about the gender grading bias component. To prevent participants from being primed, participants were told that the study was on the topic of grading accuracy in general. After completing their participation in the study, teachers in training were fully informed about the goal of the study.

The booklet also included an answer model provided by the original assessors, with a brief description of the main elements of a correct answer for each question. The maximum number of points per question for a good answer (which was similar to the original assessment) was given to the participants, but not further specified. The maximum total number of points for the Dutch language exam was 10 whereas for math it was 26.

During the evaluations of the tests, participants were asked to write the name of the

student on a provided form. This was done to ensure that participants read the name of the student and thus became aware of the student's gender (the manipulation). Ethical approval for this study was provided by the Ethics Committee of the Faculty of Governance and Global Affairs (University Leiden).

5.3.3 MEASURES

Gender grading bias

Gender grading bias refers to bias in evaluations against either male or female students. The mean scores for the tests with a male name (11) and the tests with a female name (11) were calculated based on the total number of points rewarded for the answers to the six questions. Then, gender grading bias was calculated by subtracting the mean male test score from the mean female test score. As a result, gender grading bias is a continuous variable including negative and positive scores. The negative scores indicate a gender grading bias against females, whereas positive scores indicate a gender grading bias against males.

Implicit gender-typed expectations

The gender stereotypical expectations of the teachers were measured for two aspects: talent for the subject and invested effort. After the teacher evaluated a single test, the teacher was asked to estimate the (fictious) student's talent and effort on a scale from 1 to 10 (i.e., on a scale from 1 to 10 to what extent do you think this student has talent for the subject/invested time in preparing the test). We did not put any explicit emphasis on gender of the student.

Similar to the gender grading bias measure, mean scores were calculated for the tests with male names (11) and the tests with female names (11). Then, gender bias in expectations was calculated by subtracting the mean score for the male tests from the mean score on the female tests. Again, negative scores indicate a bias against females (males have more talent/put more effort than females), whereas positive scores indicate a bias against males (females have more talent/put more effort than males).

Implicit gender-typed associations

The implicit gender-typed associations of the teachers were assessed with the arts-science Implicit Association Test (IAT). The IAT was built in LimeSurvey, an online questionnaire tool, and was based on the task on the Harvard project Implicit demonstration website and the Nosek et al. (2002a) paper. The IAT measures participants' implicit association between female and male attributes and school related concepts 'languages' and 'science'. During the IAT, participants were requested to sort words into groups by pressing keys. In congruent blocks, female names (e.g., 'Julia') and stereotypically feminine words (e.g., 'Dutch language') needed to be sorted under the 'Female & Language' category and male names (e.g., 'Daan') and stereotypically masculine words (e.g., 'math') needed to be sorted under

the ‘Male & Science’ category. In incongruent blocks female names and stereotypically masculine words needed to be sorted under the ‘Female & Science’ category and male names and stereotypically feminine words needed to be sorted under the ‘Male & Languages’ category. The reaction time and accuracy scores were recorded by LimeSurvey. Each participant’s level of implicit gender-typed associations was determined by calculating differences in reaction time and accuracy in congruent and incongruent blocks resulting in a d-score (using scoring algorithm by Greenwald et al., 2003). A positive d-score represents stronger implicit gender-typed associations (e.g., math is for males, Dutch language for females). Negative d-scores represent counter-gender-typed associations (e.g., math is for females, Dutch language for males). To reduce possible order effects of the presentation of congruent and incongruent blocks, two versions of the IAT were used, one in which the congruent block was administrated first and one in which the incongruent block was administrated first (Nosek et al., 2005).

5.3.4 DATA ANALYSES

We used SPSS Statistics version 27 for data inspection and further analyses. All variables were inspected for possible outliers that were defined as values more than 3.29 SD under or above the mean (Tabachnick & Fidell, 2012). Two outliers were identified and winsorized by giving them the most extreme not outlying value (Tabachnick & Fidell, 2012). All continuous variables were normally distributed.

To examine whether teachers’ grading bias is negatively associated with male students we plotted the distribution of the scores of the gender grading bias variable in histograms for Dutch language and math. Additional one sample t-tests were performed to test whether the mean scores were significantly different from zero (hypothesis 1). Differences between male and female teachers’ gender grading bias in Dutch language and math were examined using independent t-tests (hypothesis 2).

To test hypotheses 3 and 4, the serial multiple mediation model (Model 6) in PROCESS was run to incorporate bootstrapping techniques for estimating indirect effects (Hayes, 2018). Using this model we estimated the effects of implicit gender-typed associations and expectations on gender grading bias controlled for teachers’ gender (hypothesis 3), as well as the indirect effect of teachers’ gender on gender grading bias through implicit gender-typed associations and expectations (hypothesis 4). We ran the model for expectations regarding students’ talent and effort separately. Indirect effects are significant when the confidence intervals do not include zero.

The main analyses provided insight in the strength of the associations between the main variables, but lacked information about the direction of the effects of implicit gender stereotypes on grading biases against either male or female students. To explore these directions, scatterplots with regression lines were produced and additional post hoc analyses were performed with separate outcome variables for grading bias against male students and grading bias against female students (results are presented in appendix B).

The gender grading bias variable in the main analyses included scores from -3 to +3 with negative scores representing grading bias against females and positive scores representing grading bias against male students. In the post hoc analyses, the gender grading bias variable was separated for a grading bias against male students and a grading bias against female students. In the variable for grading bias against male students, we gave teachers with biases against female students score 0. In the variable for grading bias against female students, we gave teachers with biases against male students score 0. This corresponds with the method of Liben and Bigler (2002) to infer stereotypical attitudes towards others, rather than contra-stereotypical attitudes, using questionnaire data. A higher score on the variables in the post hoc analyses indicated a stronger grading bias against either male or female students.

To test the robustness of our results we conducted the analyses with the control variables age, experience, and test version.

Table 5.1. Mean Scores and Pearson Correlations on Central Variables of the Study

			Mean (SD)	Min, Max	Pearson Correlations			
					1	2	3	4
Dutch	1	Gender Grading Bias	0.023 (0.533)	-1.14, 1.64				
Language	2	Gender Teacher	0.760 (0.427)	0, 1	-0.025			
N = 131	3	Gender-typed Associations	0.488 (0.424)	-0.51, 1.35	0.069	0.512**		
	4	Expectations Talent	0.038 (0.490)	-1.39, 1.55	0.727**	-0.050	0.008	
	5	Expectations Effort	0.114 (0.499)	-1.36, 1.25	0.410**	-0.119	0.004	0.404**
Math	1	Gender Grading Bias	0.019 (0.828)	-2.36, 2.64				
N = 227	2	Gender Teacher	0.560 (0.497)	0, 1	0.012			
	3	Gender-typed Associations	0.386 (0.431)	-0.74, 1.33	-0.031	-0.311**		
	4	Expectations Talent	-0.129 (0.480)	-1.64, 1.09	0.442**	-0.024	-0.119	
	5	Expectations Effort	0.284 (0.517)	-1.27, 1.82	0.116	0.004	0.122	-0.042

Note: Gender, 0 = male, 1 = female, Expectations Talent = Implicit gender-typed expectations regarding students' talent, Expectations Effort = Implicit gender-typed expectations regarding students' effort. ** p<0.01, * p<0.05

5.4 RESULTS

5.4.1 GENDER GRADING BIAS ON THE AGGREGATE LEVEL

Table 5.1 presents the means, standard deviations and range of gender grading bias in Dutch language and math. The distributions of the scores on gender grading bias are depicted in Figure 5.3. Negative scores indicate a grading bias against female students, positive scores indicate a grading bias against male students. In Dutch language, the mean gender bias in grading is 0.023 (*SD* = 0.533), indicating that Dutch language teachers on average did not differentiate in their grading based on students' gender ($t(130) = 0.497$, $p = 0.620$, $d = 0.043$). In math, the average score on the gender grading bias variable is 0.019 (*SD* = 0.828), indicating that math teachers did not differentiate in their grading of male and female students either ($t(226) = 0.338$, $p = 0.736$, $d = 0.022$. Thus, hypothesis 1 should be rejected.

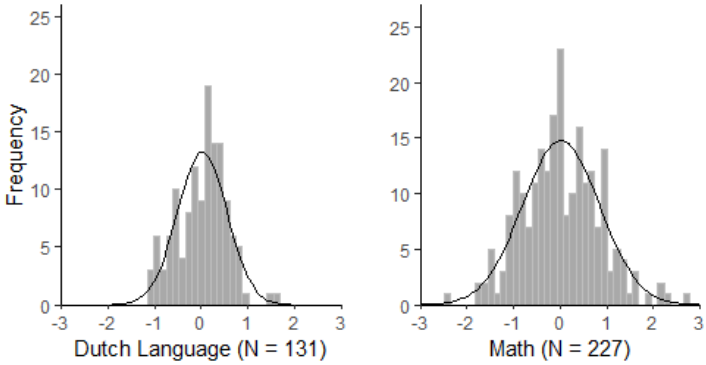

Figure 5.3. Histograms Gender Grading Bias

Nevertheless, Table 5.1 and Figure 5.3 show that there is ample variation in gender grading biases on the individual level. In Dutch language, teachers' grading varies from a bias against female students, rewarding male students up to 1.14 point higher than female students, to a bias against male students, rewarding female students up to 1.64 points higher than female students (on a scale from 1 to 10). In math, in the most extreme cases math teachers graded male students up to 2.36 points higher than female students, or graded female students up to 2.64 points higher than male students (on a scale from 1 to 26).

5.4.2 TEACHERS' GENDER AND GENDER GRADING BIAS

Next, we tested whether the variation in teachers' gender grading bias is associated with teachers' gender. Table 5.1 reveals that teachers' gender grading bias is not correlated with teachers' gender. Independent sample t-tests confirm that gender grading bias among male Dutch language teachers (M = 0.047, SD = 0.519) is not significantly different from the gender grading bias among female Dutch language teachers (M = 0.016, SD = 0.540), $t(129)$ = 0.283, p = 0.616, d = 0.058. Also for math, gender grading bias among male teachers (M = 0.007, SD = 0.838) is not significantly different from gender grading bias among female teachers (M = 0.027, SD = 0.824), $t(225)$ = -0.180, p = 0.858, d = -0.024. Therefore, hypothesis 2 should be rejected.

5.4.3 TEACHERS' IMPLICIT GENDER STEREOTYPES AND GENDER GRADING BIAS

Table 5.1 presents the mean scores and correlations of teachers' implicit gender-typed associations and implicit gender-typed expectations. On average, teachers in both subjects tend to associate math with masculinity and languages with femininity (Dutch language $t(130)$ = 13.158, p = 0.000, d = 1.150, math $t(226)$ = 13.487, p = 0.000, d = 0.895. Further, both Dutch language and math teachers expect female students to invest more effort in preparing an exam than male students (Dutch language $t(130)$ = 2.622, p = 0.010, d = 0.229, math $t(226)$ = 8.279, p = 0.000, d = 0.549). Additionally, math teachers expect male students to have

more talent in math than female students ($t(226) = -4.043$, $p = 0.000$, $d = -0.268$). Table 5.1 shows that teachers' gender-typed associations are not correlated with gender grading bias in either Dutch language or math. However, teachers' gender-typed expectations were positively correlated with gender grading bias. In Dutch language, both gender-typed expectations regarding students' talent and effort were positively associated with teachers' gender bias in grading. In math, only teachers' expectations regarding students' talent were positively associated with gender grading bias.

Models 3 and 6 in Table 5.2 present the results of regressing teachers' implicit gender-typed associations and implicit gender-typed expectations regarding students' talent (top half of Table 5.2) and effort (bottom half of Table 5.2) on gender grading bias controlled for teachers' gender (hypothesis 3). For Dutch language, Model 3 confirms that teachers' implicit gender-typed associations are not related to gender grading bias. The extent to which teachers associate math with masculinity and languages with femininity is not related to gender bias in teachers' grading. This Model also confirms that teachers' gender-typed expectations regarding students' talent and effort are positively associated with gender grading bias. Figure 5.4 indicates that teachers who expect male students to have more talent in Dutch language than female students, give male students more points for their answers than female students, despite the equal quality of the tests. Teachers who expect female students to have more talent in Dutch language than male students, appear to give female students more points for their answers than male students. The same is true for the relation between teachers' gender-typed expectations regarding students' effort and gender grading bias. Post-hoc analyses confirm this image (see appendix B).

For math, Model 6 in Table 5.2 shows a similar pattern. Teachers' implicit gender-typed associations are not related to gender grading bias, neither are teachers' gender-typed expectations regarding students' effort. However, teachers' gender-typed expectations regarding students' talent is positively associated with gender grading bias. Also for math, Figure 5.4 indicates that the positive relation between teachers' expectations and gender grading bias works both ways: teachers who expect male students to have more talent in math compared to female students, give male students more points for their answers compared to female students. Teachers who expect female students to have more talent in math than male students, give female students more point for their answers compared to male students. Post-hoc analyses confirm this pattern regarding the association between teachers' expectations of students' talent and gender grading bias. Furthermore, the post-hoc analyses show that teachers' gender-typed expectations regarding students' effort are significantly positively associated with a grading bias against female students, but not with a grading bias against male students. Teachers who expect male students to put more effort in preparing for an exam than female students give male students more points for their answers than female students, whereas teachers who expect female students to put more effort in preparing for an exam than male students do not give female students more points for their answers than male students (see appendix B).

Table 5.2. Double Mediation of Implicit Gender-typed Associations and Expectations regarding Talent (top-half) and Effort (bottom-half) in the Relation Between Teachers' Gender and Gender Grading Bias

	Dutch Language (N = 131)			Math (N = 227)		
	(1) IA	(2) IE	(3) GGB	(4) IA	(5) IE	(6) GGB
Constant	0.099 (0.066)	0.076 (0.089)	-0.027 (0.067)	0.539 (0.041)	-0.032 (0.064)	0.063 (0.100)
Gender Teacher	0.509** (0.075)	-0.084 (0.118)	-0.036 (0.088)	-0.270** (0.055)	-0.065 (0.067)	0.055 (0.106)
Implicit Associations		0.053 (0.119)	0.098 (0.089)		-0.155* (0.077)	0.063 (0.122)
Expectations Talent			0.789** (0.066)			0.772** (0.105)
R^2	0.262**	0.004	0.534**	0.097**	0.018	0.197**
Constant	0.099 (0.066)	0.210 (0.090)	-0.059 (0.091)	0.539 (0.041)	0.194 (0.069)	0.001 (0.112)
Gender Teacher	0.509** (0.075)	-0.192 (0.119)	-0.018 (0.119)	-0.270** (0.055)	0.048 (0.073)	-0.005 (0.117)
Implicit Associations		0.104 (0.120)	0.094 (0.118)		0.163 (0.084)	-0.089 (0.135)
Expectations Effort			0.436** (0.087)			0.194 (0.108)
R^2	0.262**	0.020	0.173**	0.097**	0.017	0.015

Note: Bootstrap estimates derived from 10 000 bootstrap samples along with 95% CIs are reported. Standard Deviations in parentheses. IA = Implicit Gender-typed Associations, IE = Implicit Gender-typed Expectations, GGB = Gender Grading Bias, ** p<0.01, * p<0.05

In all, hypothesis 3 can only partly be accepted. Implicit gender-typed associations are not directly associated with gender grading bias. However, teachers' gender-typed expectations regarding students' talent is related to gender grading bias. Evidence for a role of gender-typed expectations regarding students' invested effort is mixed.

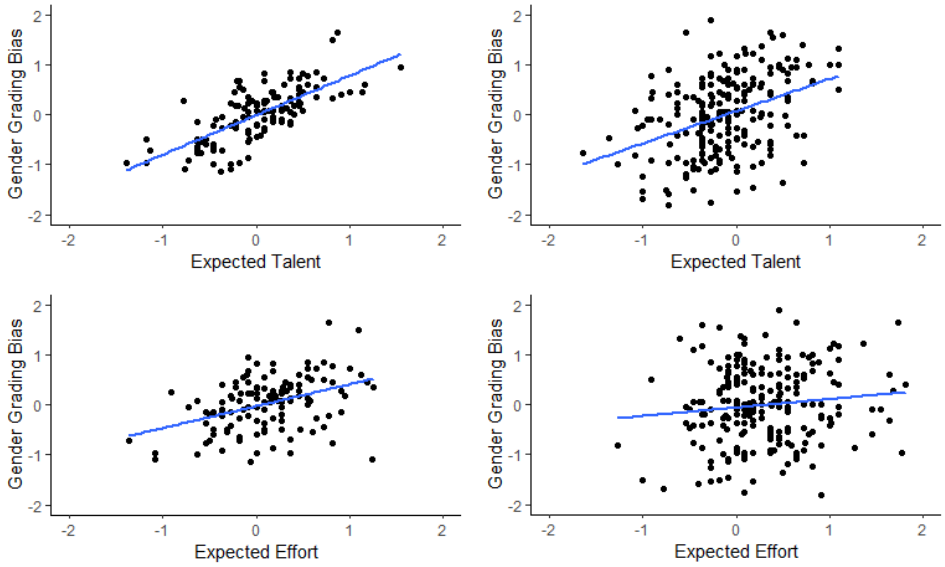

Figure 5.4. Plots of the Relation between Gender-Typed Expectations and Gender Grading Bias
Note: negative scores indicate biases against female students, positive scores indicate biases against male students.

5.4.4 THE INDIRECT RELATION BETWEEN GENDER AND GENDER GRADING BIAS

Table 5.1 reveals that teachers' gender is correlated with teachers' gender-typed associations in Dutch language and math. In Dutch language, gender and gender-typed associations are positively correlated, indicating that male Dutch language teachers have weaker gender-typed associations than female Dutch language teachers. In math, gender and gender-typed associations are negatively correlated, indicating that female math teachers have weaker gender-typed associations than male math teachers.

Despite the absence of a direct effect of teachers' gender on their gender grading bias, we estimated the indirect effect of teachers' gender on gender grading bias because a correlation between X and Y is not necessarily needed as a precondition for a mediation effect (Hayes, 2018). Table 5.2 presents the findings. Models 1 and 4 include the findings of regressing teachers' gender on their implicit gender-typed associations. Models 2 and 5 present the findings of regressing both teachers' gender and implicit gender-typed associations on their gender-typed expectations regarding students' talent (top half

of Table 5.2) and effort (bottom half of Table 5.2). As mentioned above, Models 3 and 6 show the results of regressing gender, implicit gender-typed associations and gender-typed expectations on gender grading bias. The mediation effects of implicit gender-typed associations and gender-typed expectations, as well as the double mediation effect and the total effect of the model are presented in Table 5.3.

For Dutch language, the findings are presented in the first three Models. Model 1 confirms that implicit gender-typed associations are stronger among female Dutch language teachers compared to male Dutch language teachers. As expected, male Dutch language teachers have less stereotypic associations between gender and math and languages. Model 2 shows that both gender and implicit gender-typed associations are not related to gender-typed expectations regarding students' talent and effort. Contrary to what was expected, strong implicit gender-typed associations are not related to making a stronger distinction between male and female students regarding their talent and effort in Dutch language. As described above, Model 3 shows that teachers' gender-typed expectations are related to gender grading bias.

For math, the findings are presented in the last three Models. Model 4 confirms that implicit gender-typed associations are stronger among male math teachers compared to their female counterparts. As expected, female math teachers have milder gender-typed associations than male math teachers. Model 5 reveals that teachers' gender is not associated with gender-typed associations regarding students' talent and effort in math. Teachers' implicit gender-typed associations are negatively related to their gender-typed expectations regarding students' talent, but not related to teachers' expectations regarding students' effort. As described above, Model 3 shows that only teachers' gender-typed expectations regarding students' talent are positively related to gender grading bias.

Table 5.3 summarizes the effects of teachers' gender on gender grading bias, through implicit gender-typed associations and expectations. For Dutch language, no indirect effects were found for implicit gender-typed associations and gender-typed expectations (or both). For math, implicit gender-typed associations and expectations could not explain differences in male and female teachers' gender grading bias on their own. However, a positive double mediation effect was found: differences in male and female teachers' gender grading bias can be explained through differences in male and female teachers' implicit gender-typed associations and expectations regarding students' talent. Post-hoc analyses reveal that teachers' implicit gender-typed associations and expectations regarding students' talent can explain differences in male and female math teachers' grading biases against male students and female students. Both for the grading bias against male students and the grading bias against female students applies that female math teachers have weaker associations between math and masculinity, that is related to expecting approximately the same of male and female students in terms of talent, which is related to having weaker (closer to zero) gender grading bias. Furthermore, the post-hoc analyses show that implicit gender-typed associations and expectations regarding students' effort can explain differences in male

and female teachers grading biases against female students. Female math teachers have weaker gender-typed associations that lead to expecting male students to put less effort in math which is related to having a stronger grading bias against female students (see appendix B). In all, hypothesis 4 should be accepted in the subject math, but rejected in the subject Dutch language.

Table 5.3. Indirect Effects of Gender on Gender Grading Bias

	Dutch Language			Math		
	Effect Size (SE)	LLCI	ULCI	Effect Size (SE)	LLCI	ULCI
Total Mediated Effect	0.004 (0.098)	-0.192	0.192	-0.035 (0.055)	-0.145	0.073
Mediation through IA	0.050 (0.050)	-0.051	0.146	-0.017 (0.032)	-0.084	0.046
Mediation through IE	-0.066 (0.086)	-0.239	0.098	-0.050 (0.053)	-0.155	0.053
Double mediation	0.021 (0.042)	-0.059	0.110	0.032 (0.019)	0.000	0.073
Total Mediated Effect	-0.013 (0.087)	-0.187	0.155	0.025 (0.042)	-0.055	0.111
Mediation through IA	0.048 (0.068)	-0.090	0.182	0.024 (0.039)	-0.052	0.107
Mediation through IE	-0.084 (0.050)	-0.186	0.011	0.009 (0.017)	-0.022	0.051
Double mediation	0.023 (0.032)	-0.032	0.095	-0.009 (0.007)	-0.026	0.002

Note: Total Mediated Effect is the sum of the indirect effects of X>Y. Size of bootstrap sample for calculations of the indirect effect = 10 000. LLCI = bootstrapped lower level confidence interval; ULCI = bootstrapped upper level confidence interval, 95% confidence interval, Effect size is significant when LLCI and ULCI do not include 0.

Finally, the pattern of results for the model was similar when controlling for teachers' age, experience (Lavy, 2008), and the version of the tests that were graded by the teachers, though it should be noted that for math the significance level of the effect of implicit gender-typed associations on gender-typed expectations regarding students' talent and the double mediation effect were no longer significant on the 0.05 level, but on the 0.1 level.

5.5 CONCLUSION AND DISCUSSION

By means of a field experiment among teachers in training in the subjects Dutch language and math in secondary education in the Netherlands, this study tested the influence of students' gender on teachers' grading and examined the role of teachers' gender and implicit gender stereotypes in gender grading bias. We found that teachers' grading, on average, was not gender biased. However, we did find that individual variation in gender grading bias is associated with teachers' implicit expectations of male and female students' talent and effort. Further, we found an indirect effect of teachers' gender on gender grading bias in math via teachers' implicit gender-typed associations and expectations.

Contrary to what we expected (e.g., Protivínský & Münich, 2018), in our field experiment we did not find an association between perceived male gender of students and teachers' grading. Instead, we found that teachers' grading, on average, did not differ between male

and female students. This finding is similar to the few studies that also experimentally tested the influence of students' gender on teachers' grading (Hanna & Linden, 2012; Hinnerich et al., 2011), suggesting that students' gender and teachers' grading might not be *causally* related. Previous studies that found grading bias against male students, all used comparative designs to examine gender grading bias, i.e., comparing scores on tests that were blindly graded to scores on tests that were graded by a teacher (e.g., Protivínský & Münich, 2018). The observed bias against male students in teachers' grading in these studies might not be a direct result of students' gender but rather a reflection of teachers' beliefs about male and female students. Indeed, one of these previous studies found that an effect of grading bias against male students disappeared when controlling for teachers' reports on students' attitudes towards learning (Cornwell et al., 2013). Although Cornwell et al. (2013) concluded that differences in male and female students' behavior could fully explain gender bias in teachers' grading, based on the results in our study we argue that it could have been a gender bias in teachers' perception of male and female students' behavior that explained gender grading bias.

Although we did not find a gender bias in teachers' grading on average, we did find ample variation in gender grading bias between teachers. Our study showed that the variation in gender grading bias was associated with variation in teachers' expectations of male and female students' talent and effort despite the equal quality of the tests of male and female students. Teachers with more equal expectations of male and female students displayed less gender bias in grading. Teachers who differed in their expectations of male and female students displayed more grading bias against both male and female students, despite the equal quality of their tests. Our finding confirmed previous research that demonstrated that the extent to which a person distinguishes between males and females based on generalized preconceptions is related to their discriminative behaviors against males and females (e.g., Burgess & Borgida, 1999; Dasgupta, 2004), and that different expectations can lead to different assessments (Muntoni & Retelsdorf, 2018; Powell, 1986).

Additionally, to get a complete picture of the role of implicit gender-typed expectations in gender grading bias our study explored the full spectrum of gender-typed expectations including stereotypical and contra-stereotypical expectations. Based on previous research we expected gender stereotypes to differ mainly in strength, and to lesser extent in direction (e.g., Lenton et al., 2009). However, the post-hoc analyses showed that teachers' implicit gender-typed expectations differed in both strength and direction, and that both stereotypical and contra-stereotypical expectations were related to discriminative behaviors. Dutch language and math teachers' grading biases against male or female students were in line with the content of their biases in expectations. Despite the equal quality of the tests, in both subjects, teachers who expected male students to have more talent, gave male students more points for their answers. Teachers who expected female students to have more talent, gave female students more points for their answers. The same was true for Dutch language teachers' expectations regarding students' effort: expecting more effort

from male students was related to rewarding male students more points for their answers, and expecting more effort from female students was related to rewarding female students more points for their answers. The explorations support previous research in the family context that found that fathers' stereotypical and contra-stereotypical expectations of males and females are related to opposite differential treatments of males and females (Endendijk et al., 2013). It should be noted that, in math, only teachers' contra-stereotypical expectations of students' effort were associated with gender biases against female students. Math teachers who had the contra-stereotypical expectation that male students put more effort in the preparation of a test, gave male students more points for their answers, whereas math teachers who expected female students to put more effort, did not gave female students more points for their answers. Putting effort is gender stereotyped as something female students generally do, and male students generally do not do (e.g., Heyder and Kessels, 2007). Therefore, expecting students to have put effort could be praised and result in positive effects in teachers' grading for male students but not for female students. This finding supports literature on double standards for men and women in, among others, parenting roles (Deutsch & Saxon, 1998), on the work floor (Lyness & Heilman, 2006), and in engaging in sexual behaviors (Endendijk et al., 2020; Sagebin Bordini & Sperb, 2013).

Furthermore, similar to previous research (Martin & Dinella, 2012; Smeding, 2012), our results showed that female teachers in a stereotypically masculine subject (math) and male teachers in a stereotypically feminine subject (languages) had weaker associations between math and masculinity and between language and femininity than male math teachers and female Dutch language teachers. Our finding supports the assumption that an individual's gender stereotypes are flexible and that these gender stereotypes can become less strict by experiencing or observing rejection of the gender stereotype (Bussey & Bandura, 1999; Charlesworth & Banaji, 2021; Groeneveld et al., 2021). Though it should be noted that some scholars have argued that gender stereotypes are resistant to change (Fiske, 2017; Haines et al., 2016). Additionally, previous studies showed that implicit gender stereotypes were similar among males and females (Banaji & Greenwald, 1995; Rudman & Kilianski, 2000), or somewhat stronger among females than males (Endendijk et al., 2013; Nosek et al., 2002a). Our study showed that both males and females can have weaker implicit gender stereotypes and that these seem to be dependent on someone's own experiences rather than someone's gender perse.

We found that these weaker implicit gender-typed associations among female math teachers were related to weaker gender-typed expectations of students and that this mechanism could explain weaker gender grading biases among female math teachers than male math teachers. Explorations of the direction of the mediating effect showed that, in math, female teachers' weaker gender stereotypical associations (math is for men and women, languages is for men and women) were related to weaker gender-stereotypical expectations of students' talent (males and females are talented in math). This finding supports previous findings that teachers' stereotypes and expectations are positively

related (Muntoni & Retelsdorf, 2018). Consequently, weaker (more equal) gender-typed expectations regarding students' talent resulted in a weaker grading bias against both males and females. Weaker gender-typed expectations regarding students' effort could only explain differences between male and female math teachers' grading bias against female students, but it should be noted that this effect size was fairly small.

To our knowledge, our study is the first to find evidence for a psychological mechanism underlying gender grading bias. However, we did not find evidence for the mechanism in Dutch language teachers. Male Dutch language teachers did have weaker gender-typed associations than female Dutch language teachers, but these were not related to weaker gender-typed expectations. This may be because gender stereotypes regarding students' talent in language are not as persuasive as in math. Large government campaigns were directed at enhancing female participation in math, but not at enhancing male participation in languages. Furthermore, many famous Dutch writers are male, providing convincing contra-gender stereotypical examples in Dutch language. It thus seems that gender salience is important for effects of implicit gender stereotypes in gender grading bias to occur (Keiser et al., 2002).

The findings of this study should be interpreted in light of its limitations. The balancing between the maximalization of this experiment's ecological validity and practical feasibility resulted in the exclusion of real-life student-teacher interactions that could play an important role in gender grading bias (Falch & Naper, 2013). Moreover, the study design resulted in the selection of two experimental groups of tests that were not of the exact same quality. The group of tests with male names were of approximately the same quality as the group of tests with female names based on the experts' blind grading. However, as the tests in the two groups were not identical in terms of students' answers, differences in the students' answers might have resulted into differences between the two groups that affect gender grading bias. Further, although implicit measures are encouraged to study differential treatment of male and female students (Denessen et al., 2022), in particular the IAT is criticized for its validity and reliability in predicting discriminative behaviors (Blanton et al., 2009). However, for the current study the Arts-Science version of the IAT was highly relevant because of the theoretical link between the content of this test and our outcome measure, i.e., gender grading bias in the subjects math and languages. Despite the limitations of the use of different and real-student answers and the implicit measures of gender stereotypes, these components of this study's design contributed in particular to the participants not being aware of what the research was about. Therefore, social desirability and conscious behaviors, which participants often have when they participate in social research (Wulff & Villadsen, 2019), were probably prevented.

In all, the results in our study question previous research on the role of students' behaviors in gender grading bias that relied on teacher reports (Cornwell et al., 2013). Our study encourages future research to be conscious of gender bias in teacher reports when relying on teacher reports to explain differences between male and female students. Furthermore,

we recommend further exploration of the psychological mechanisms underlying gender grading bias and empirically test the role of gender stereotypes and teachers' gendered beliefs. Future research could tap into the role of teachers' explicit gender stereotypes, or the extent to which teachers' gendered beliefs and gender grading bias are causally related to unravel the processes that can result in gender biases in teachers' grading.

Further, our results have implications for teacher training programs. Teacher training programs might want to contribute to increasing awareness of teachers (in training) about the role of their own gender-typed expectations of male and female students in grading practices. Improving teachers' self-awareness can stimulate their quality of teaching (Schussler et al., 2010). We are working on translating the setup of this study's experiment into an educational tool directed at increasing teachers' awareness of their own gender (grading) bias and improving teachers agency in equal treatment of male and female students. Teachers achieve agency (i.e., actions that are assumed to contribute to good and meaningful education (Biesta et al., 2015) when they are able to intentionally choose behaviors instead of just behave in a routinized matter without considering alternatives (Priestley et al., 2015). Awareness of one's own gender-typed expectations of male and female students and gender grading bias, might help teachers to tackle automatic thoughts and behaviors and to treat males and females more equally, both in interaction with and evaluation of students. Indeed, previous research demonstrated that awareness of one's own gender bias can contribute to reducing gender bias in an education setting (Carnes et al., 2015; Devine et al., 2017; Forscher et al., 2017; Girod et al., 2016). It should be noted that it is crucial that gender equality is valued by the individual teacher as well as by the (school) collective (Biesta et al., 2015). This can be a complicating factor as the extent to which gender inequality is perceived (see for instance Kinias & Kim, 2011), gender equality in education is valued, and ideas about the routes towards equal treatment can differ between individual teachers and across cultures (see Cardona López et al., 2018). Besides the importance of discussions between teachers, teacher trainers, and policy makers 'to get everyone on the same page' (Cardona López et al., 2018), we believe that courses that increase awareness of teachers' own gender biases are critical as gender bias often happens unconsciously and its consequences are therefore often overlooked or underestimated. Teacher training programs could be a suitable place to increase teachers' (in training) self-awareness of gender stereotypes and agency in equal treatment of boys and girls (De Boer et al., 2019).

To conclude, our study pointed out that rather than a causal relation between students' gender and teachers grading, the role of teachers' implicit gendered beliefs about male and female students should be considered. Not only were teachers' implicit gender-typed expectations related to gender grading bias, implicit gender-typed associations and expectations together could explain differences in male and female math teachers' grading biases. Students' gender and teachers' gender do not directly translate in differential grading practices, instead, teachers' beliefs about males and females can affect their perception and assessment of male and female students.

5.6 ACKNOWLEDGEMENTS

The authors would like to thank Arthur Pormes, Anneke Wurth and Peter Kop for their assistance in the design of the field experiment. Further, the authors would like to thank Anne Floor Lubbers for her effort during the experiment's data collection.

5.7 APPENDIX

5.6.1 APPENDIX A – MANIPULATION

	Dutch male names	Dutch female names	Turkish male names	Turkish female names
1	Bastiaan Driessen	Sara Wolters	Yusuf Koc	Leyla Bulut
2	Thomas van der Horst	Meike Groen	Murat Aslan	Sevda Avci
3	Daniël van der Pol	Juliette Kramer	Hakan Yilmaz	Kubra Demir
4	Benjamin de Lange	Madelief Scholten	Ahmet Cicek	Merve Günes
5	Pieter Smeets	Fien de Ruiter		
6	Teun Bakker	Emma Hoekstra		
7	Lars Willemsen	Tess van Veen		
8	Florian van Dijk	Olivia Maas		

5.6.2 APPENDIX B – BIAS AGAINST MALE OR FEMALE STUDENTS

Table: Gender Grading Bias Against Male students.

Double Mediation of Implicit Gender-typed Associations and Gender-typed Expectations in the Relation Between Teachers' Gender and Gender Grading Bias against Male Students

	Dutch Language (N = 131)			Math (N = 227)		
	(1) IA	(2) IE	(3) GGB	(4) IA	(5) IE	(6) GGB
Constant	0.099 (0.066)	0.076 (0.089)	0.191 (0.045)	0.539 (0.041)	-0.032 (0.064)	0.398 (0.062)
Gender Teacher	0.509** (0.075)	-0.084 (0.118)	-0.028 (0.059)	-0.270** (0.055)	-0.065 (0.067)	0.011 (0.066)
Implicit Associations		0.053 (0.119)	0.086 (0.060)		-0.155* (0.077)	-0.036 (0.076)
Expectations Talent			0.363** (0.044)			0.430** (0.065)
R^2	0.262**	0.004	0.356**	0.097**	0.018	0.170**
Constant	0.099 (0.066)	0.210 (0.090)	0.167 (0.052)	0.539 (0.041)	0.194 (0.069)	0.370 (0.069)
Gender Teacher	0.509** (0.075)	-0.192 (0.119)	-0.012 (0.068)	-0.270** (0.055)	0.048 (0.073)	-0.020 (0.071)
Implicit Associations		0.104 (0.120)	0.080 (0.068)		0.163 (0.084)	-0.114 (0.083)
Expectations Effort			0.243** (0.050)			0.072 (0.066)
R^2	0.262**	0.020	0.171**	0.097**	0.017	0.012

Note: Standard Deviations in parentheses. IA = Implicit Gender-typed Associations, IE = Implicit Gender-typed Expectations, GGB = Gender Grading Bias, ** p<0.01, * p<0.05

Table: Gender Grading Bias against Males

Indirect Effects of Gender on Gender Grading Bias against Male Students

	Dutch Language			Math		
	Effect Size (SE)	LLCI	ULCI	Effect Size (SE)	LLCI	ULCI
Total Mediated Effect	0.023 (0.050)	-0.079	0.117	0.000 (0.033)	-0.065	0.065
Mediation through IA	0.044 (0.031)	-0.018	0.103	0.010 (0.019)	-0.028	0.048
Mediation through IT	-0.031 (0.041)	-0.116	0.044	-0.028 (0.031)	-0.093	0.030
Double mediation	0.010 (0.020)	-0.027	0.052	0.018 (0.011)	0.000	0.043
Total Mediated Effect	0.007 (0.045)	-0.085	0.093	0.031 (0.025)	-0.014	0.084
Mediation through IA	0.041 (0.035)	-0.029	0.107	0.031 (0.024)	-0.013	0.083
Mediation through IT	-0.047 (0.029)	-0.111	0.004	0.004 (0.008)	-0.010	0.023
Double mediation	0.013 (0.017)	-0.017	0.052	-0.003 (0.004)	-0.012	0.003

Note: Standard Deviations in parentheses. IA = Implicit Gender-typed Associations, IE = Implicit Gender-typed Expectations, GGB = Gender Grading Bias, ** p<0.01, * p<0.05

Table: Gender Grading Bias Against Female students.

Double Mediation of Implicit Gender-typed Associations and Gender-typed Expectations in the Relation Between Teachers' Gender and Gender Grading Bias against Female Students

	Dutch Language (N = 131)			Math (N = 227)		
	(1) IA	(2) IE	(3) GGB	(4) IA	(5) IE	(6) GGB
Constant	0.099 (0.066)	0.076 (0.089)	-0.218 (0.044)	0.539 (0.041)	-0.032 (0.064)	-0.335 (0.059)
Gender Teacher	0.509** (0.075)	-0.084 (0.118)	-0.008 (0.058)	-0.270** (0.055)	-0.065 (0.067)	0.044 (0.062)
Implicit Associations		0.052 (0.119)	0.012 (0.058)		-0.155* (0.077)	0.098 (0.072)
Expectations Talent			0.425** (0.043)			0.342** (0.062)
R^2	0.262**	0.004	0.437**	0.097**	0.018	0.123**
Constant	0.099 (0.066)	0.210 (0.090)	-0.226 (0.056)	0.539 (0.041)	0.194 (0.069)	-0.370 (0.063)
Gender Teacher	0.509** (0.075)	-0.192 (0.119)	-0.007 (0.074)	-0.270** (0.055)	0.048 (0.073)	0.015 (0.066)
Implicit Associations		0.104 (0.120)	0.014 (0.073)		0.163 (0.084)	0.025 (0.076)
Expectations Effort			0.193** (0.054)			0.122* (0.060)
R^2	0.262**	0.020	0.094**	0.097**	0.017	0.020

Note: Standard Deviations in parentheses. IA = Implicit Gender-typed Associations, IE = Implicit Gender-typed Expectations, GGB = Gender Grading Bias, ** p<0.01, * p<0.05

Table: Gender Grading Bias against Females

Indirect Effects of Gender on Gender Grading Bias against Female Students

	Dutch Language			Math		
	Effect Size (SE)	LLCI	ULCI	Effect Size (SE)	LLCI	ULCI
Total Mediated Effect	-0.018 (0.055)	-0.124	0.090	-0.035 (0.030)	-0.092	0.022
Mediation through IA	0.006 (0.031)	-0.055	0.069	-0.027 (0.022)	-0.072	0.016
Mediation through IT	-0.036 (0.046)	-0.128	0.055	-0.022 (0.023)	-0.066	0.026
Double mediation	0.011 (0.023)	-0.033	0.059	0.014 (0.008)	0.000	0.032
Total Mediated Effect	-0.020 (0.052)	-0.126	0.082	-0.006 (0.025)	-0.057	0.042
Mediation through IA	0.007 (0.044)	-0.080	0.093	-0.007 (0.023)	-0.054	0.038
Mediation through IT	-0.037 (0.024)	-0.088	0.005	0.006 (0.011)	-0.012	0.031
Double mediation	0.010 (0.015)	-0.016	0.046	-0.005 (0.004)	-0.016	0.000

Note: Standard Deviations in parentheses. IA = Implicit Gender-typed Associations, IE = Implicit Gender-typed Expectations, GGB = Gender Grading Bias, ** p<0.01, * p<0.05