Well done (for someone of your gender)! Experimental evidence of teachers' stereotype-based shifting standards for test grading and elaborated feedback

Schuster, Carolin; Narciss, Susanne; Bilz, Jessica

Published in:
Social Psychology of Education

DOI:
10.1007/s11218-021-09633-y

Publication date:
2021

Document Version
Publisher's PDF, also known as Version of record

Social Psychology of Education (2021) 24:809–834
https://doi.org/10.1007/s11218-021-09633-y

Well done (for someone of your gender)! Experimental evidence of teachers' stereotype-based shifting standards for test grading and elaborated feedback

Carolin Schuster[1] · **Susanne Narciss**[2] · **Jessica Bilz**[3]

Received: 12 May 2020 / Accepted: 1 April 2021 / Published online: 23 April 2021

Abstract
In three experiments (*N*s = 327/137/210), we investigated whether test grades and elaborated feedback in a stereotypically male (Math) and a stereotypically female subject (German) are biased by the student's gender. For this purpose, pre-service teachers graded and provided written feedback on tests which were allegedly from boys or girls. In addition, participants' belief in stereotypes was measured in Study 1 and 2 and manipulated in Study 3 to test its moderating role. A meta-analysis across the three studies confirmed the following pattern: a small to moderate stereotype-contrasting grading bias, if the evaluators endorsed stereotypes, but no bias if they did not. Tests from the gender that, according to the stereotype, is weaker in the domain, were graded better. Study 1 and 3 further showed that the supposedly weaker gender received more elaborated feedback. The results are discussed in terms of shifting standards and previous findings in gender bias in school.

Keywords Gender stereotypes · Performance evaluation · Grades · Marks · Feedback · Teacher · Gender bias · Stereotype endorsement · Shifting standards

1 Introduction

The grades students receive in school serve as a criterion by which they are selected for further educational tracks, awarded scholarships, or offered jobs. If grading was biased by social categories and stereotypes about them, it might substantially

✉ Carolin Schuster
carolin.schuster@leuphana.de

[1] Institute of Psychology, Leuphana University Lüneburg, Universitätsallee 1, 21335 Lüneburg, Germany

[2] Technische Universität Dresden, Dresden, Germany

[3] University of Passau, Passau, Germany

contribute to inequality. The present research examines whether such a bias exist for the social category of gender.

Grades do not only function as a selection criterion, they also have a feedback function for the students throughout their academic paths. Specifically, grades are supposed to provide information to the students about their current skill level relative to expectations, norms and standards. In addition, grades are often accompanied by more or less *elaborated* verbal feedback, in written or oral form. This feedback can consist of positive or critical *evaluative* components (i.e., information about what is correct or not, well or poorly done) as well as *formative* components (i.e., information intended to foster the learning process). If the grades are biased, then the elaborated feedback might be as well. In fact, biases might even be more pronounced because providing high quality and unbiased elaborated feedback is less often a topic of teacher education than preventing bias in performance diagnostics, at least in Germany. In addition, differential feedback to boys and girls can be expected to have a considerable impact on differential skill development, given the role of feedback for learning (Hattie and Timperley, 2007; Narciss, 2004; Shute, 2008). Therefore, the aim of the present research is to examine gender biases in grading as well as in elaborated feedback provided to the student.

1.1 Gender bias in performance evaluations

Gender stereotypes comprise that males have greater ability in mathematics and science and are generally more competent intellectually, whereas females are more diligent, communicative, and have better skills in the verbal domain (e.g., Heyder and Kessels, 2015; Retelsdorf et al., 2015; Tiedemann, 2002). It seems intuitively plausible that evaluations of the performance of males and females may be biased consistent with these stereotypes, such that each gender is evaluated better than the other in the domains ascribed to them. We refer to this type of effect as a *stereotype-consistent bias*. A meta-analysis of experimental studies on this hypothesis has, accordingly, shown (slightly) more negative evaluations of females in stereotypically male or neutral domains, but no differences in stereotypically female domains (Swim et al., 1989). However, only few studies focused specifically on gender bias in teachers' grading (Duval, 1980; Hofer, 2015; Spear, 1984a, 1984b). They compared science or mathematics evaluations of allegedly male and female secondary school students and mostly found, in line with the meta-analysis, more negative evaluations of females in these stereotypically male domains (for an exception, see Duval, 1980). Further support for a stereotype-consistent bias comes from a study where pre-service teachers evaluated the math performance levels of 12 primary school students (Bonefeld et al., 2020). In contrast to the previous studies, they were provided not only with the students' test answers but also their final scores. The results show that students with randomly assigned girls' names were evaluated worse than those assigned boys' names.

However, there is a surprising lack of research on the role of the evaluators' stereotypic beliefs in gender bias even though the bias is argued to be rooted in gender

stereotypes (Hofer, 2015). Another gap in the present research is that, to our knowledge, no experimental studies on gender bias in grading in stereotypically female domains (e.g., verbal skills) exist so far. The present work addresses this gap by investigating gender bias in both the mathematical and the verbal domain and considering the role of the evaluators' belief in stereotypes.

Whereas much of the previous experimental literature seems to show a bias to the disadvantage of females, another line of research found gender biases in the opposite direction of stereotypes (i.e., a *stereotype-contrasting bias*): Several analyses of large-scale educational assessment data have shown that girls, more than boys, are evaluated better than predicted by their standardized test performance in the domain (Cornwell et al., 2013; Falch and Naper, 2013; Kuhl and Hannover, 2012; Lavy and Sand, 2015; Lindahl, 2016; Marcenaro-Gutierrez and Vignoles, 2015; Robinson and Lubienski, 2011; Robinson-Cimpian et al., 2014b). This could point towards a gender bias favoring girls over boys. However, such a bias may not result from the belief that girls are actually more talented in the domain, but from factoring girls' presumably better learning behavior into mathematics performance evaluations (Kuhl and Hannover, 2012). Despite apparently favoring girls, such a bias can actually exacerbate gender gaps in math performance on the long run (Robinson-Cimpian et al., 2014a, 2014b). This type of empirical educational research has, of course, its own limitations for understanding psychological processes involved in gender bias in school. Most of all, the evaluations by the teachers did not refer to the same performances as the standardized tests and thus could be influenced by many different factors besides gender.

From a theoretical perspective, both a stereotype-consistent bias (e.g., grading girls worse than boys in math) and a stereotype-contrasting bias (e.g., grading girls better than boys in math) could be rooted in the same stereotypical associations of the gender and the domain (e.g., mathematics as a male domain). According to the shifting standards model, contrast effects occur when the individual is evaluated by a group-specific standard (or, in other words, frame of reference), which may be lower in the context of negative stereotypes (Biernat, 2012; Biernat and Manis, 1994). For example, the same moderate math proficiency could be evaluated as good (for a girl) but not so good (for a boy). Hence, even though the evaluation seems to be equally good or better for the negatively stereotyped gender, the stereotype-contrasting bias reflects the assumption that the genders play in different leagues–as it is the case in many sports. Indirect evidence for shifting standards in school comes from a study where pre-service teachers estimated the original test results of a female or male student on either an objective standardized math test or a portfolio task focused on individual progress, which had happened earlier. They were only informed that, based on this original test result, it had been concluded that the student has average mathematic ability. Participants estimated similar test results for the male and female student in the portfolio, but better results of the male than female student on the standardized test (Holder and Kessels, 2017, Study 1). This shows pre-service teachers' stereotype-consistent assumptions that a girl described as having average math ability has lower objective math test scores than a boy. It also shows their assumption that such gender differences would not show in a portfolio that allows the use of different reference standards.

In summary, based on different lines of research, both a stereotype-consistent as well as a stereotype-contrasting bias in the grading of female and male student in stereotypically male domains could be predicted. So far, it remains unclear whether a bias in any direction is mirrored in the grading of male students in stereotypically female domains. Furthermore, there is a gap in addressing the role of teachers' stereotypic beliefs, especially since the gender roles have been changing in the last decades. Even though stereotypic beliefs seem to be a precondition of stereotypic bias, they have not been reported in previous experimental studies on gender bias in grading.

1.2 Research on gender bias in teacher's feedback

Not only quantitative performance evaluations like grades may be susceptible to bias in the context of stereotypes. In fact, gender bias may even be more pronounced in more qualitative evaluations such as in the context of performance feedback to students. Nevertheless, there is a large gap in research on whether teachers provide elaborated evaluative and formative feedback on test performances differentially depending on student's gender. This is surprising given the crucial role feedback plays in the learning process (Evans, 2013; Narciss, 2008, 2017; Shute, 2008).

To our knowledge, there are no studies yet that experimentally compare the feedback on a test given to male and female students. However, a few studies looked at feedback based on whether the (higher education) students were European American or of a minority ethnicity. They showed that ethnic minority members received more positive or less negative feedback by white American evaluators (Croft and Schmader, 2012; Harber, 1998; Harber et al., 2012). It is unclear whether this generalizes to gender groups, however, as it may be driven by concerns of majority members to not appear racist (Croft and Schmader, 2012).

In addition, several observational studies of classroom interactions exist that provide some information on gender-biased teacher behavior (Heller and Parsons, 1981; Jones and Dindia, 2004; Parsons et al., 1982). A meta-analysis concluded that teachers interact more with boys than girls, but mostly more negatively (e.g., more reprimands; Jones and Dindia, 2004). It has to be noted that this mostly reflects reactions to classroom behavior rather than performance. Nevertheless, some observational studies that focused on evaluative feedback on performance in mathematics (Heller and Parsons, 1981; Parsons et al., 1982) found gender differences such that boys received more criticism than girls. Praise seemed to be dependent on expectations as well, such that for girls for whom the teacher had high expectancy fewer interactions were praised than for high and low expectancy boys, and low expectancy girls. In addition, one study found gender specific patterns of the content of feedback, which reflects attributions of competence for boys' performance and of effort for girls' performance (Dweck et al., 1978; see also Tiedemann, 2002). While these observational studies suggest that gender might affect how teachers give feedback, their findings are not completely consistent. The lack of experimental studies on gender bias in elaborated feedback leaves the question of causality open.

Given the wide gaps in the research, we will take an exploratory approach. Specifically, we are interested in whether there are differences in the absolute or relative amount of positive and critical formative feedback statements given to boys and girls in the verbal and mathematical domain.

1.3 The present research

The present research aimed to make three main contributions: The first goal was to examine the existence and direction of gender bias in grading in a stereotypically male domain, mathematics, and a stereotypically female domain, German. The second goal was to examine the role of the teachers' stereotype endorsement on gender bias in grading for the first time. The third goal was to explore gender biases in the elaborated feedback given to the student.

With regard to gender biases in grading, previous experimental research points to an stereotype-consistent effect (i.e., evaluations differ in the direction of the relevant ability stereotypes; also called *assimilation* effect) for girls in mathematics and science (e.g., Hofer, 2015). However, there are also studies that indicate more lenient evaluations of girls in mathematics that may be caused by the same negative stereotypes (Holder and Kessels, 2017; Kuhl and Hannover, 2012). For boys in the verbal domain, there are no experimental studies on gender bias in grading to this day. Based on this mixed evidence, we predict that a gender bias exists and it might be either in the form of a stereotype-consistent or a stereotype-contrasting effect. In addition, as this hypothesis rests on the assumed impact of stereotypes we predict that this bias appears mostly if the evaluating teacher believes that gender stereotypes reflect actual differences in boys' and girls' abilities.

H1 In domains with prevalent gender stereotypes (e.g., mathematics, German), tests of boys and girls will be graded differently.

H2 Gender bias in grading is moderated by teachers' stereotype endorsement, such that it appears only or stronger if the teacher believes in stereotypes.

With regard to the exploratory research question of gender bias in feedback, we test gender-based differences in the total and relative amount of positive and critical formative (i.e., directed at improving things the student cannot yet perform well) feedback.

We present three studies to examine these research questions. Study 1 tests the effect the alleged gender of the student on grades and feedback on identical Math or German tests in a between-subjects design and explores the moderating role of teacher's stereotype endorsement. In Study 2, the students' genders were experimentally varied within subjects: each participant graded four test versions of either a Math or a German test. In a pre-registered Study 3, the students' genders were also manipulated within participants, who graded and gave feedback to a German test. In addition, we manipulated the belief in stereotypes experimentally between subjects to test the prediction of its moderating role on gender bias.

2 Study 1: Between-participants effect of student gender

The first aim of Study 1 was to test whether the student's gender causally affects the grade on a test, to examine the direction of this potential difference, and whether this direction was symmetrical in Mathematics (i.e., a stereotypically masculine subject) and German (i.e., a stereotypically feminine subject). The second aim was to examine, for the first time, whether elaborated feedback on the test differed by gender and subject. In addition, we aimed to consider the role of the teacher's stereotype endorsement for gender bias.

2.1 Method

2.1.1 Design

Study 1 represents the integrative data analysis (Curran and Hussong, 2009) of two separate experiments, henceforth referred to as Sample A and B. Both experiments had the same 2 (student gender: boy or girl) by 2 (test subject: Math or German) between-subject design and measured the points, grade, and feedback on the test as dependent variables. In addition, both experiments included a measure of the exploratory moderator, stereotype endorsement.

2.1.2 Samples and recruitment

Sample A ($n=204$) stems from an online experiment for which pre-service teachers were recruited by contacting universities that train teachers across Germany and having them send the study link to their mailing lists. Sample B ($n=123$) stems from a paper and pencil experiment conducted in a social psychology lecture for pre-service teachers at a university in the state of Lower Saxony in Germany. Both samples consist of pre-service teachers of different school types and school subjects. Sample A had been planned to find a medium-sized 2*2 interaction of gender and subject with 95% power. Sample B was limited by the number of attendants of the lecture who were willing to participate, which resulted in a low sensitivity for this analysis in this sample ($f=0.32$). Given that the tendency to endorse stereotypes (yes/no) was explored as an additional, not evenly distributed moderator, both samples separately were underpowered. To increase power, the samples were thus combined and the type-I-error inflation resulting from post-hoc sample size augmentation was accounted for in the analyses as recommended by Sagarin, Ambler, and Lee (2014). The biggest difference between the samples is how they participated: online and probably in privacy at their computer (A) or on paper in a full lecture hall, in which a social psychology class was going to take place (B). Further differences in the procedure of the experiments that could lead to differences between the samples will be disclosed in the procedure section and analyses for each sample separately are reported in the Supplementary file (Tables 4, 5, 6).

The complete sample ($N=327$) consists of 230 women, 91 men, and 6 persons of other or undisclosed gender. Their mean age was 23.59 years ($SD=3.81$). They

were studying education in universities across 13 out of 16 federal states in Germany (which have different education systems). It was not a requirement for participation to specialize in German or Mathematics as a school subject or to specialize in primary education. Based on the data available on specialization (partly unavailable in Sample B) all school forms and tracks are represented in the sample.

2.1.3 Procedure and manipulation

Both experimental samples were invited to participate in a study on grading and giving feedback on school tests without further information about the hypothesis. After giving informed consent for this voluntary and anonymous study, participants were told they would be given the test of a 10-year old third grader (either a German or a Math test). The gender of the student was manipulated in Sample 1 by referring to the student as a boy or girl in the beginning and using the male or female word for student (Schüler/Schülerin) throughout the test. As this manipulation led to a high number of participants failing the manipulation check (see Supplementary file, Table 2), student gender was manipulated with a common female (Sarah) or male (David) name in Sample B.

The test material and the answer of the alleged student were identical in both experimental samples and consisted of eight problems in the respective subject. Participants had to award points on each problem and then assign a grade to the test in total. Then participants were asked to give direct feedback to the student. In Sample A they did so in three separate text fields: with regard to what he/she already masters well in the subject (positive feedback), what he/she does not master sufficiently yet (critical feedback), and with suggestions on how to improve (formative feedback). Based on participant comments that the latter two categories were impossible to differentiate they were integrated for analyses, and in Sample B, only two text fields were offered for positive and critical-formative feedback. The following parts of the questionnaires included different exploratory items in both samples, which are not relevant for the present study: In Sample A these were rating and open-ended questions on the student's performance in the respective subject. In Sample B there were questions on the perception of the performance relative to the age group, on the goals for giving feedback, a scale on the motivation to judge prejudice-free (Banse and Gawronski, 2003), and two questions on the valence of associations with the names Sarah and David. These measures will not be discussed further but are fully listed in the Supplementary file (Tables 1 and 2). Finally, the questionnaires in both samples included a measure of stereotype endorsement and demographic questions.

2.1.4 Test materials and performance evaluations

In contrast to previous studies with the Goldberg Paradigm (Swim et al., 1989), we did not ask for evaluations of essays, but less ambiguous performance indicators: Both the Math and the German test consisted mainly of example questions from VERA-3 (a standardized test for third graders to be administered by teachers), which can be retrieved from the homepage of the 'Institut zur Qualitätsentwicklung im

Bildungswesen' [Institute for Quality Development in Education] (IQB, 2013). For each of the eight test problems, the maximum number of points was indicated. The test answers were constructed at a good to medium performance level and written in infantile handwriting. An example problem of the German test was: "Find the opposite. long/short; heavy/_____; strong/_____; hard/____ [3 pt"]. The answer could translate into something like "light as a feather, weak, zoft", including the spelling error, thus leaving some room for interpretation. An example problem of the Math test was: "Chances: You throw a dice and win if it is 6. You throw a coin and win if it is head up. Give a reason for at which game you have higher chances to win [3 pt]". The answer was "The coin only has two sides". The maximum number of points in total was 25.5 on the German and 26 on the Math test. The percentage of points given thus represents a relatively objective, criterion-oriented form of performance evaluation. In addition, participants graded the performance on the usual German 6-point grading scale (i.e., 1 'excellent' to 6 'insufficient'). A specific standard for transforming the points into a grade was not provided; therefore, the grade reflects a more subjective performance evaluation.

2.1.5 Feedback

We measure positive and critical-formative feedback in terms of the number of words written down in each of these categories. We suggest that more/longer feedback could be cautiously interpreted as desirable, given that it may contain more useful and specific information. Given the online experimental setting in Sample A and the setting in a lecture in Sample B it is unlikely that length of feedback reaches an extent at which it becomes undesirably complex. However, given that the settings differ and the time spent to write feedback may be more normative in the group setting of Sample B, dataset differences need to be accounted for especially for this dependent variable. Values of zero words of feedback will be coded as missing values.

2.1.6 Stereotype endorsement

In order to make the measure of stereotype endorsement more resistant against social desirability, we told participants that some attributes may be attributed to genders in a society. They were first asked to indicate whether they thought verbal/mathematical abilities were regarded as typical for boys, for girls, or neither in the German society, independent of their own beliefs. They were then told that men and women actually differ on average in some attributes, for example physical strength, and that some other assumptions about differences are just prejudice. Participants then indicated whether they thought the following statements about men and women were actually true. Then they rated several items (four in Sample A, two items were accidently missing, Cronbach's $\alpha = 0.80$; six in Sample B, $\alpha = 0.87$) on a Likert scale ranging from 1 (not at all true) to 6 (completely true). All items were statements of stereotypes about boys' and girls' gender specific characteristics and differences (e.g. *Boys don´t handle language as fast and fluently as girls*). We assumed that if stereotype endorsement were a moderator in biased evaluations it would be based on

Table 1 Results of the ANOVA of grades in Study 1

Dependent variable	Grade		Grade adj. for points	
Factors	$F(1, 319)$	p	$F(1, 319)$	p
Gender	0.007	0.933	0.279	0.598
Subject	1.057	0.305	0.009	0.923
StE	1.973	0.161	0.120	0.729
Gender*subject	3.951	**0.048**	3.292	0.071
Gender*StE	1.360	0.244	1.009	0.316
Subject*StE	0.022	0.882	0.004	0.948
Gender*subject*StE	7.132	**0.008**	9.639	**0.002**

Gender: test was allegedly written by a boy or a girl. Subject: Mathematics or German. StE (stereotype endorsement): participants' belief that relevant gender stereotypes are true or not. Grade adjusted for points: standardized residuals of regressing each gender's grade on their points percentage; effects thus refer to differences in the grades given for the same amount of points. Statistically significant effects are highlighted in bold letters

whether or not a person tends to believe that stereotypes are true at all, rather than a linear effect of the extent to which these beliefs are held. Therefore, this variable was recoded by splitting it at the mid-point of the scale, into participants who tend to disagree with stereotypes (<3.5, $n=200$) and those who tend to agree (≥ 3.5, $n=127$). Thus, the specific items are less of a concern for the moderator and the differences in the measurement in Sample A and B seem neglectable.

2.2 Analyses and results

2.2.1 Methods for integrative data analysis

As the Sample B was added post-hoc to the initial Sample A, the repeated analysis inflates the alpha-error. Sagarin and colleagues (2014) propose a p-augmented statistic that reports the alpha-error inflation of post-hoc sample augmentation. It displays a range of what the actual alpha error might be (i.e., it is always >0.05) in the augmented sample: from the case that the decision to collect further data was based on the exact p value found in the first analysis to the case that further data would have been collected even at $p=1$. We will report p_{aug} for the hypothesized effects to help readers interpret the p value in the augmented sample.

2.2.2 Grades

Hypothesis 1 and 2 were tested with a three-factorial univariate ANOVA with the student's gender, the test subject, and participants' stereotype endorsement as factors. The results in the left column of Table 1 shows the predicted subject*gender interaction (H1) with $p_{aug}=[0.072;\ 0.077]$, and a significant moderation of this

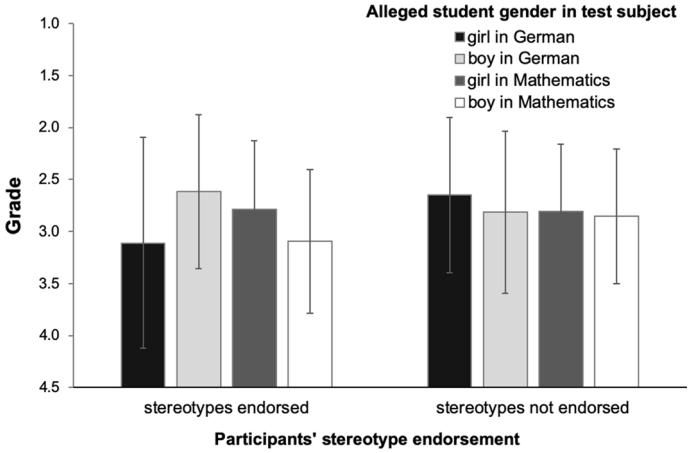

Fig. 1 Means and standard deviations of grades by condition in Study 1. The grades are coded in the way of the German school system, ranging from 6 (insufficient) to 1 (excellent)

effect by stereotype endorsement (H2) with $p_{aug} = [0.050; 0.052]$.[1] The group means and standard deviations are displayed in Fig. 1 (the exact numbers are reported in the Supplementary file, Table 15), which shows the hypothesized gender bias in grading by participants who endorsed stereotypes. This bias manifests in better grades for the girl in the math test and better grades of the boy in the German test.

To examine the nature of this moderated contrast effect more closely, we tested whether this effect would hold when controlling for the percentage of points given on the test. As can be seen in Table 1, the three-way interaction effect remains significant when analyzing the residuals of the grade regressed on the percentage of points.[2] This means that participants awarded different grades for the same amount of points in the different conditions, thereby applying different standards of evaluation.

2.2.3 Feedback

We conducted a repeated measures ANOVA with feedback as the dependent variable,[3] type (positive /critical formative) as a within-subject factor and student gender, subject, and stereotype endorsement as between-subject factors. There was an unpredicted significant main effect of subject, $F(1,299) = 5.887$, $p = 0.016$, $\eta_p^2 = 0.021$, as well as a subject* type interaction, $F(1,277) = 8.407$, $p = 0.004$, $\eta_p^2 = 0.029$.

[1] The three-way interaction effect remains significant when including dataset as a covariate (Online Resource, Table 7).

[2] The three-way interaction effect remains significant when including dataset as a covariate. The results are reported in the Online Resource, Table 7.

[3] The distributions of positive as well as critical formative feedback were severely right-skewed in all student gender and subject conditions. To achieve a less-skewed distribution, the number of words was log-transformed and the analyses conducted again with the transformed DV. Given the results did not lead to different conclusions (Online Resource, Table 8). We report the analysis of the easier-to-interpret original analysis.

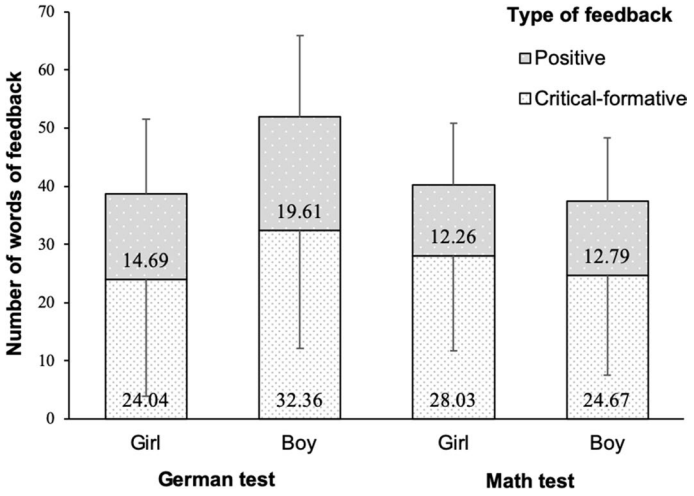

Fig. 2 Number of words of feedback by alleged student gender and test subject in Study 1. The numbers displayed in the stacked bars are the means for the respective feedback type

In support of the existence of a gender bias, there was also a gender*subject interaction effect, $F(1,277) = 6.905$, $p = 0.009$, $\eta_p^2 = 0.024$, and a gender*subject*type interaction, $F(1,277) = 4.652$, $p = 0.032$, $\eta_p^2 = 0.017$. This effect represents a pattern of means where boys ($M = 1.15/1.45$, $SE = 0.04/0.04$ for positive/critical formative) receive more feedback than girls ($M = 1.08/1.31$, $SE = 0.04/0.04$) on the German test, and girls ($M = 1.00/1.40$, $SE = 0.03/0.03$) more feedback than boys ($M = 0.97/1.29$, $SE = 0.04/0.04$) on the Mathematics test, and this is more pronounced for critical formative than positive feedback (Fig. 2).

2.3 Discussion

Study 1 showed initial support for the existence of a gender bias in grading primary school students' German and Mathematics tests among pre-service teachers who endorsed gender stereotypes. In contrast to previous experimental evidence on bias in the evaluations of secondary school students' STEM essays (Hofer, 2015; Spear, 1984a, 1984b), we found a stereotype-contrasting effect: The gender stereotyped as less competent in the subject (i.e. girls in Mathematics, boys in German) was graded more favorably. Participants who did not endorse stereotypes did not significantly show this bias. In addition, Study 1 indicated, for the first time, a bias in the amount of feedback given to boys and girls, which was not moderated by stereotype endorsement. The gender stereotyped as less competent received more feedback, especially more critically formative feedback. Both good grades as well as much feedback could be considered benevolent behavior. Hence, both findings suggest that participants might feel inclined to give more support or make it easier for the gender that they think has less talent and thus has to work harder. The exploratory finding that the bias on grades appeared when statistically controlling the raw points given on

the test implies that the contrast effect is a matter of different reference standards applied to boys and girls for the same perceived performance.

The findings of Study 1 have to be interpreted tentatively. The effects are small and found in an integrative analysis of two separate experiments with different settings of data collection. The subsamples have mean differences in the dependent variables, for which we control statistically. Given that both samples separately have too low power to examine small three-way interaction effects, the integrative analysis is more informative than separate analyses (reported in Supplementary file, Table 4, 5, 6), which show non-significant effects in both, but consistent patterns of means. Replications of the findings are needed to draw reliable conclusions.

3 Study 2: Within-participant effect of student gender

To increase power in Study 2, we used a design where student gender was manipulated within subjects. Specifically, each participant evaluated the performance of four alleged students named Sarah, Lisa, Tobias, and David. This design also reflects the situation of teachers in the classroom better, who typically have to correct tests of several students and evaluate them one after the other. The goal of Study 2 was to confirm the finding from Study 1 that stereotype endorsement would moderate the gender bias effect on grades. Specifically, we hypothesized that stereotype endorsers will show a stereotype-contrasting gender bias in both subjects, but stereotype non-endorsers will not show such a bias. In addition, we attempted to replicate the finding that the negatively stereotyped gender receives more feedback.

3.1 Method

3.1.1 Design and sample

The design was a two (student gender: boys/girls) by two (subject: Maths/German) by two (stereotype endorsement: yes/no) design. Student gender was manipulated within participants, subject between participants, and stereotype endorsement was a measured moderator. The sample was collected in psychology lectures for education students, hence the sample size in this study was not a-priori planned. The sample consisted of 137 students of education (pre-service teachers), 113 of them female, from two German universities. Most of them were in their second or third year of studies, and they were in study programs preparing for various school types (55% primary school). For a 2*2*2 within-between interaction effect, this sample has a sensitivity to find a small effect of $f = 0.183$ with a power of 0.90, given a correlation of $r = 0.139$ between the repeated measure of grades.

3.1.2 Procedure and measures

Students of the psychology lectures were invited to participate voluntarily and anonymously in this study, which took about 45 min. Questionnaires were only

administered after consent was given and could be withdrawn privately anytime by not completing it. The goals of the study were communicated correctly but vaguely as "a study about how tests are evaluated and how feedback is given" to not make participants aware of the relevance of gender stereotypes. The questionnaire contained four completed Math or German tests. Each test was introduced by the information that it was written by 10-year old Sarah (or Lisa, Tobias, David, respectively—all names are popular, non-stigmatized German names). The respective name was repeated on the test and the following evaluation sheet to increase the salience of gender. Participants had to correct it and give up to 10 points for each problem, grade the test, and provide written feedback about what the student already mastered (positive) and about what he/she did not master yet, and how to improve it (critical formative). After the four tests, a general questionnaire followed, starting with an open-ended question about what participants thought to be the most important functions of feedback. Then followed six items, $\alpha = .74$, to exploratively measure motivation to evaluate without prejudice (Items 2, 5, 7, 10,12, 16, from Banse and Gawronski, 2003); we will not further report on these results. Then the measure of stereotype endorsement followed. Finally, participants answered some demographical questions.

3.1.3 Test materials and randomization

To achieve an acceptable duration of the study, shorter tests with three problems per test were used. An example problem of the German test was to rewrite a short text in a different tense. The Math problems were text-based and required, for example, calculating the required number of roles of wallpaper à 6 m^2 given the height of the wall, the size of a wardrobe, and the space beside the wardrobe. The four answer versions of the tests were constructed to contain a similar number and types of mistakes of a medium performance level and presented in a similarly orderly handwriting. The test versions were then systematically combined with the names resulting in 24 versions, each of which was administered at least twice. By mistake, the order of names was not implemented randomly, resulting in about two thirds of the sample being ordered girl-girl-boy-boy and one third boy-boy-girl-girl. Due to this error, effects of order position cannot be fully controlled but only estimated for the specific order versions that were used.

3.1.4 Dependent variables

For each problem, participants could award up to 10 points. The grades were given as German school grades ranging from 1 'excellent' to six 'insufficient'. The total point percentages and grades of the two girls and the two boys were combined to mean scales for each gender. As in Study 1, feedback was measured by the number of words written as positive and critical formative feedback. For these as well, mean scales for girls and boys were computed.

Table 2 Results of the mixed-model ANOVA of grades in Study 2

Dependent variable	Grade		Grade adj. for points	
	$F(1,128)$	p	$F(1,128)$	p
Within-subject factors				
Gender	0.691	.407	0.029	.866
Gender*subject	3.205	.076	0.728	.395
Gender*StE	0.039	.845	0.163	.687
Gender*subject*StE	4.225	**.042**	4.659	**.033**
Between subject factors				
Subject	3.978	.048	3.745	.055
StE	0.031	.860	0.025	.874
Subject*StE	0.002	.967	0.540	.464

StE (stereotype endorsement). Grade adjusted for points: standardized residuals of regressing each gender's grade on their points percentage. Statistically significant effects are highlighted in bold letters

3.1.5 Stereotype endorsement

As in Study 1, the stereotype endorsement (StE) measure was preceded by two questions on whether other people in the society might believe mathematical and verbal ability to be more typical for one gender or the other. Then, after an introductory note that some assumptions about gender differences are true and some are just prejudice, participants were asked how much they personally agreed with six items on relevant gender stereotypes (*Boys have higher mathematical abilities than girls, Girls have more difficulties solving mathematical problems than boys;* 6-item $\alpha = 0.83$). As in Study 1, participants were split in two groups that either tended to agree (≥ 3.5, $n = 39$) or disagree (< 3.5, $n = 97$) with the stereotypes.

3.2 Analyses and results

3.2.1 Controlling for order effects

For each analysis, we tested whether including order (i.e., girls first or boys first) as a covariate showed significant order effects or order*gender interaction effects. If this was not the case, the analyses without including order are reported.

3.2.2 Grades

A mixed model ANOVA was calculated with grade as a dependent variable, student gender as within-subject factor and test subject and StE as between-subject factors.[4] The results in Table 2 show no significant two-way interaction of gender * subject, which was predicted by H1. Confirming H2, there is a significant three-way

[4] Controlling for order did not change the results substantially, nor was it significant as a covariate (Online Resource, Table 11a).

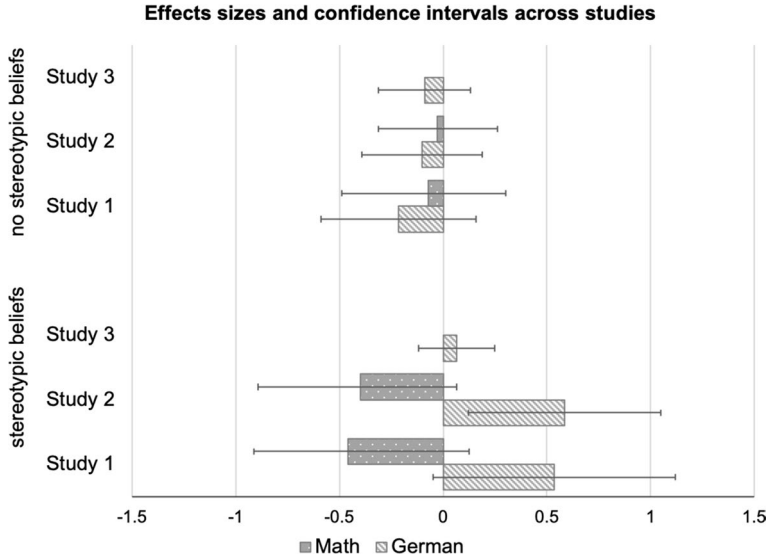

Fig. 3 Effect size of gender difference in grades by condition across Study 1–3. Positive effect sizes indicate that boys are graded better than girls and negative effect sizes indicate girls are graded better than boys

interaction of gender * subject * StE. This interaction means that, in line with Study 1, better grading of girls in Math and of boys in German only among teachers who endorsed stereotypes, but not among those who did not endorse stereotypes (see Fig. 3 for the effect size of gender bias by condition). The right column of Table 2 shows this interaction also remains significant when controlling for the students' point percentage by analyzing the residuals of the grades regressed on the point percentage. The means and standard deviations by group are reported in Supplementary file (Table 15).

Exploratively, we tested whether the school type for which participants trained played a role in the results. Given that the test was primary school level, primary school teachers might be more motivated or better able to evaluate the student. Including school type (primary / other) as a factor in the analyses showed a significant main effect on grade, $F(1,124) = 5.151$, $p = 0.025$, such that primary school teachers were estimated to give better grades, EMM = 2.129, SE = 0.060, than non-primary school teachers, EMM = 2.384, SE = 0.095. Besides that, there were significant gender*subject, $F(1,124) = 5.493$, $p = 0.21$, and gender*subject*StE interactions, $F(1,124) = 6.241$. $p = 0.014$ in the predicted direction.

3.2.3 Feedback

A mixed model ANOVA with gender and type of feedback (positive or critical formative) as within subject factors, and subject and StE as between subject factors, as well as order as covariate, showed a significant main effect of gender, $F(1,93) = 50.571$, $p < 0.001$, but also a gender*order interaction, $F(1,93) = 37.93$,

$p < 0.001$ (full results in the Supplementary file, Table 11). This interaction effect reflects a tiring effect, where the gender that was given feedback first was given more feedback. Given that the order was not correctly randomized, we refrain from interpreting the gender effect.

3.3 Discussion

In summary, Study 2 confirmed the pattern of bias in grading indicated by Study 1: Pre-service teachers who endorse stereotypes seemed to be more lenient on the gender negatively associated with a subject. As in Study 1, the effect was there when controlling for raw points, indicating the evaluation by shifting group-specific standards. This is in line with the theoretical assumption that negative stereotypes can lower the standards by which individuals are judged (Biernat and Manis, 1994). Interestingly, the shifting standards literature predicts that for grades a common, objective standard would be used. The effect is even more surprising as it has been found in a within-subject design, meaning that the same person applied different grading standards to the performances of alleged boys and girls.

Study 2 has two major limitations. First, due to a procedural mistake the order in which the genders were represented was only partially varied and about two thirds of the sample received both boys' tests before the girls' tests. The order did not affect the grades; thus, we can interpret the effects on them despite this unbalanced ordering. However, participants seemed to tire from writing feedback throughout the study; thus, we cannot interpret gender effects on amount of feedback, which are partially confounded with order. A second limitation was that the subgroup (stereotype endorsers) in which significant biases were found was small (39 ppts.), thus limiting power despite the within-subject design. In addition, there is a need to test the causal role of stereotypic beliefs as a moderator of gender bias. The moderation by measured stereotype endorsement is in line with that interpretation but it needs experimental corroboration.

4 Study 3: Pre-registered test of the effect of student gender

Study 3 was conducted to address the limitations of the previous studies. We aimed to test causally whether the belief in stereotypic gender differences in ability would trigger the contrast effects in grades. Although stereotype endorsement is typically thought of as a belief that persists across situations, the use of stereotypes in a specific situation is a heuristic that will most likely be qualified by more concrete information about the target person or group (Fiske and Neuberg, 1990). Therefore, we hypothesize that providing specific information about the (non-)existence of relevant gender differences will have the same effect as stereotype endorsement in previous studies. We will refer to the manipulated variable as stereotype accuracy belief (SAB). To increase power in Study 3, we focused only on one subject, German.

We preregistered our hypotheses (see https://osf.io/25stq) to find a grading bias to the advantage of boys (H1), which we hypothesized to be moderated by stereotype

accuracy beliefs (H2). Based on the previous findings, we expect to find that in the true stereotype condition, boys would be graded more favorably than girls, whereas in the untrue stereotype condition, there would be no or an opposite bias. We also expected to replicate the effect from Study 1 that boys receive more feedback in the stereotypically female subject than girls (H3).

4.1 Method

4.1.1 Sample and design

The Study had a 2 (student gender: boy/girl) × 2(stereotype accuracy: true/untrue) mixed experimental design with student gender as within-subject factor. To have high power to detect even a small effect, we planned to collect data of 208 participants who were required to pass a manipulation check and to take at least 10 min to complete the study. To facilitate recruitment, participation was open to student teachers of all school types even though the test was primary school level. The final sample ($N = 210$) included 91 prospective primary school teachers and 119 of other school types. The majority were female (164; 22 male, 24 unknown gender) and their mean age was 23.19 years ($SD = 3.69$).

4.1.2 Procedure, manipulations, and measures

After completing the informed consent confirmation on the first page, participants were told that they would now grade two German tests of third graders named Sarah and David and give them written feedback on it. Stereotype accuracy beliefs (SAB) were manipulated with the following instruction: *It is important to not let your judgement be guided by prejudice. The abilities assessed by tests like the following can sometimes differ between groups. However, there are proven to be no differences in the average verbal abilities between boys and girls* (in the untrue condition)/*For example, there are proven to be differences in the average verbal abilities of boys and girls* (in the true condition). The instruction thus leaves open in which direction ability differences would be. Two of the four test material versions from Study 3 were used and presented to participants in counterbalanced combinations and orders of version and gender. After grading and giving feedback, participants answered some exploratory questions on their own and the sample's average diagnostic accuracy and on what they thought might cause bias (handwriting, order, gender). Finally, they answered the items on stereotypes and stereotype endorsement from Study 2, some demographic questions, and an item on the valence of their associations with the names Sarah and David. Then they could sign up for a survey to win a 20 € voucher for online shopping and were debriefed.

Table 3 Results of the mixed-model ANOVA of grades in Study 3

Dependent variable	Grade		Grade adj. for points	
	$F(1,199)$	p	$F(1,191)$	p
Within-subject factors				
gender	0.003	.958	0.187	.666
gender*SAB	0.434	.511	5.698	.018
Between subject factors				
SAB	1.161	.282	2.497	.116

SAB (stereotype accuracy beliefs). Grade adjusted for points: standardized residuals of regressing each gender's grade on their points percentage

4.2 Analyses and results

4.2.1 Preregistered and exploratory procedures

The preregistration specifies mixed-model ANOVAs of the grade and feedback with SAB as between factor and gender—and with regard to feedback, type of feedback—as within-subject factors. We will use the same procedure as in Study 2 to control for order effects.[5] Analyses that are not preregistered will be explicitly referred to as exploratory.

4.2.2 Grades

The mixed-model ANOVA of grades on gender and subject did not result in the expected main or moderated gender effect, as Table 3 shows.[6] Thus, H1 and H2 cannot be confirmed by the preregistered analysis. However, the exploratory analysis of grades adjusted for points in the right column of Table 3, as in the previous studies, indicates an interaction effect of gender and SAB. This interaction means that the effect sizes of the gender difference in the grade given for a certain percentage of points differ dependent on whether the grader believes in stereotypes: participants who were made to believe that gender stereotypes were accurate tended to grade boys, $M = 0.00$, $SD = 0.98$; better than girls with the same points, $M = 0.16$, $SD = 1.10$, whereas those who were made to believe that stereotypes are not true tended to grade girls, $M = -0.23$, $SD = 0.79$, better than boys with the same points, $M = 0.00$, $SD = 1.03$.

Exploratively, we also tested the role of school type for participants' grading by including it as a factor in the ANOVA of grades. The only significant effect is a gender*SAB*school type interaction, $F(1,197) = 4.764$, $p = 0.030$, all other Fs < 1.5,

[5] There was no procedure preregistered in case of order effects.

[6] Including order as a covariate did not point to order effects and did not change the results (Online Resource, Table 14a).

$ps > 0.138$. The interaction effect indicates that the pattern predicted by H2 is found among pre-service primary school teachers, but not among other teachers (full results in Supplementary file, Table 13).

4.2.3 Feedback

As in Study 2, there was a significant order*gender interaction on the number of words of feedback in Study 3, $F(1,207) = 19.180$, $p < 0.001$, besides a significant main effect of gender, $F(1,207) = 13.066$, $p < 0.001$, and a gender*SAB interaction, $F(1,207) = 4.109$, $p = 0.044$.

This reflects a pattern of means where the male student receives moderately more feedback than the female when given feedback first, $dz = -- 0.374$, $CI[0.572, 0.175]$, but only slightly less when given feedback second, $dz = 0.197$ [0.005, 0.390]. This seems to be particularly driven by the participants who were made to believe stereotypes are inaccurate: among those the difference between the female and the male student was substantial when the male was first, $dz = 0.472$ [0.803, 0.141], and non-significant if the female was first, $dz = 0.038$, [0.242, 0.318]. In summary, these findings support H3 that boys receive more feedback in a stereotypically female subject.

4.3 Discussion

The results of Study 3 did not confirm the hypotheses about gender-biased grading based on the preregistered analyses: In the complete sample, the gender bias in the context of belief in the accuracy of stereotypes on grades was not found. This may be due to an unpredicted effect of the school type in which the participants where aspiring to teach. Those aspiring to teach in primary school, i.e. the level where the test scenario stems from, showed exactly the predicted pattern. In addition, we confirmed the bias in the complete sample when adjusting for the points given on the test, which may be most indicative of a gender bias in the sense of shifting reference standards. In addition, we find support for the hypothesis that the negatively stereotyped gender receives more feedback when considering the place in order, confirming the findings of Study 1.

5 Synthesis of results across studies

Figure 3 shows the effect sizes of the gender difference in grades across studies. The hypotheses were further tested meta-analytically as suggested by McShane and Böckenholt (2017). For this purpose, the subject conditions (German test or Math test) in Study 1 and 2 were treated as separate samples and the student's gender was recoded in all studies with regard to the stereotypical fit: girls were thus coded as "fit" in the German and "no fit" in the Math test condition, and boys as "fit" in the Math and "no fit" in the German test condition. As a second factor, the stereotype endorsement measure in Study 1 and 2 and the stereotype accuracy manipulation in Study 3 were included. The meta-analysis hence has a 2 stereotypical fit (fit/no

fit) by 2 stereotypic belief (true/untrue) design. Hypothesis 1 is thus represented by a main effect of the stereotypical fit factor (i.e., contrast coefficients 1 1 1 1), and Hypothesis 2 by an interaction effect (i.e., contrast coefficients 1 1 0 0), on the grade as a dependent variable. Overall, the $I^2 = 90.86$ indicates considerable non-random heterogeneity. The estimate for the first contrast was non-significant, 0.128, *CI*[0.282, 0.026] but the estimate for the second contrast was significant, 0.188, *CI*[0.294, 0.082]. This supports H2, showing a grading bias favoring the gender that has no stereotypical fit with the test subject among teachers that think of relevant gender stereotypes as true. Similarly, Contrast 2 is significant with regard to the grades accounted for points (see supplemental material). Given the low interpretability due to order effects in Studies 2 and 3, we refrain from meta-analyzing the feedback.

6 General discussion

Across three studies, a small but consistent pattern was found with regard to a gender bias on primary school grade evaluations. Specifically, pre-service teachers who believed that gender stereotypes about subject specific abilities are true graded boys in German and girls in Math more generously. For pre-service teachers who did not believe in gender stereotypes, there was no gender bias found for either subject. This pattern was significant in Study 1 and 2, where the teachers' belief in stereotypes was measured. In the preregistered Study 3, where the belief in stereotypes was experimentally manipulated, this pattern was unexpectedly moderated by the school type of the pre-service teachers in the sample, and only appeared among future primary school teachers. Overall, the predicted bias on grading in the form of a contrast effect was confirmed meta-analytically across all studies. In summary, these findings support the hypothesis of a gender-biased evaluation that is moderated by personal beliefs, but not an overall gender bias.

In addition, exploratory analyses show that when the grading was controlled for the raw points given to the test answers, the same pattern of results, a stereotype-contrasting effect, appeared in all three studies. This finding shows that even at the same level of raw performance perception (i.e., reflected in the points), the teachers evaluated this performance as more or less good (i.e., reflected in the grades) depending on the child's gender and its fit with the subject stereotype.

The present research also examined the amount of feedback given to boys and girls in each subject. In Study 1 and 3, we find support for a tendency to provide more feedback (positive and critical-formative) to the gender that has less stereotypical fit with the subject, that is, to girls in mathematics and boys in German. In Study 2, there is no evidence of this. However, this also is hardly interpretable as gender and order were seriously confounded. The effects on feedback were not moderated by the belief in stereotypes.

The present research is the first to experimentally show a stereotype-contrasting bias on the grading of school tests. It thereby provides evidence of a phenomenon previously overlooked in the experimental research on gender bias in school. In addition, our work shows the importance of the teacher's belief in stereotypes for

the influence of the bias on their evaluations. People who believe less in group differences seem to be more likely to evaluate the performance of boys and girls by the same standard–at least with regard to grades. A third major contribution is the initial finding of gender bias in elaborated instructive feedback.

6.1 Limitations

One limitation of this work is that the participants were not yet active as teachers. Their lack of experience may contribute to the variability in grades overall, as well as to the gender bias. In previous research, a gender bias in the opposite direction (i.e., grading girls worse than boys in Physics) was found to be stronger among inexperienced teachers (Hofer, 2015).

Second, it has to be noted that even though we used realistic test materials, the grading process in the experiments differs somewhat from conditions in the field. Teachers might follow the recommendation to formulate clear criteria for performance evaluations in advance, which was not possible in the experiments. They also may use a fixed formula for all students to transform raw points on a test into grades (though Ingenkamp (1971) provides evidence that they not always do). In addition, the effects of gender were small compared to the generally large range of grades for the same performance. When considering the practical implications of this works, this needs to be considered.

Finally, the findings of gender bias in feedback are not yet conclusive and have to be considered a first step in experimental research on this important question. The strong order effects in Study 2 and 3 point to a factor that might limit teachers' written feedback in the field similarly or even more, given that they will have to give feedback to many students. Nevertheless, the findings from Study 1 and 2 that more feedback was given to the gender that was negatively stereotyped in the domain show that more studies are needed to examine how stereotypes and social cognition affect elaborated feedback. This is highly relevant as feedback is one of the most important factors to affect learning (Hattie and Timperley, 2007; Narciss, 2008, 2017). Feedback, in the present studies, was analyzed with regard to the number of words in the categories positive and critical formative, and in total. This is a very objective measure. But to better understand its validity, it would be useful to code how constructive the content of feedback is. Due to the various relevant dimensions of this constructiveness, these analyses would be beyond the scope of this article.

6.2 Implication for practice and further research

The present experimental evidence of gender bias, particularly the awarding of different grades for the same raw points, does not necessarily mean that teachers in the field use shifting standards similarly when they grade the test of a complete class; Hopefully and most likely, they will define a common transformation rule of points to grades for the whole class. Nevertheless, teachers have considerable margins of discretion in evaluations, the use of which may be subtly affected by reference standards based on gender stereotypes. Stereotype-consistent biases may, in practice, be

more pronounced in settings with larger margins of discretion than our test material out of standardized test questions. In this regard, our findings are in line with previous work that shows that teachers expect that larger margins of discretion, as for instance in the evaluation of portfolio work, would lead to the disappearance of stereotype-consistent gender difference which would show in standardized performance tests (Holder and Kessels, 2017). In the field, effects like the ones found in the present work may appear, for instance, in the grading and feedback of oral presentations or in evaluations that include perceptions of classroom behavior (e.g., Kuhl and Hannover, 2012).

The tendency to be more lenient to the supposedly cognitively disadvantaged might even be rooted in benevolent motives. In the study by Holder and Kessels (2017) this is reflected in ideological differences used to introduce the different forms of tests to participants: Standardized tests on the grounds of high objective education standards and portfolio work on the grounds of inclusion. In the present research, the benevolence of motives might reflect in the tendency to give more positive as well as critical-formative elaborated feedback to the supposedly disadvantaged gender. In addition, the idea that contrast effects of stereotypes appear based on benevolent motives to help the student may also explain why we found a stereotype-contrasting bias, whereas previous experiments with a similar design have found a stereotype-consistent bias (Hofer, 2015; Spear, 1984a, 1984b). The previous studies were set in a secondary school context, whereas ours was set in a primary school context, which arguably aims much more at motivating students of all ability levels to learn and acquire basic skills and less at preparing for selection into different careers. This also may explain why in Study 3, the stereotype-consistent bias on grades was only found among primary school pre-service teachers. Future research could examine the teacher's educational philosophy and goals as a predictor for biases, besides their belief in stereotypes.

6.3 Conclusion

Synthesizing three experiments, we found a significant gender bias on grades among stereotype endorsing teachers, such that they graded girls better than boys on a Math test and boys better than girls on a German test. In two of the experiments, a novel matching pattern was also found among all teachers with regard to providing more written tutorial feedback. This is the first evidence that grades and feedback given to boys, as well as girls, can be biased by subject-specific gender stereotypes in similar ways. It is also the first direct experimental evidence of the shifting of reference standards to more leniency towards the negatively stereotyped gender, which may be specific to a presumably student-focused primary school context. In summary, the findings support assumptions of shifting standards theory but point to the importance of considering motivational processes as well as individuals' stereotype endorsement for making predictions about the direction of biases. In addition, they suggest that in practice, negative stereotypes and counteracting, but similarly biased motivational tendencies may conceal each other and thus be underestimated.

It remains an important question for future research how such subtle biases affect students.

7 Declarations

This research was conducted without external funding. The authors declare that they have no conflicts of interest with regard to the studies. For low-risk studies like the ones described in this manuscript, there is no ethics approval required in Germany. Participants gave informed consent according to the guidelines of the American Psychological Association and were debriefed about the purpose of the study after completion. The original study materials, the data of all studies (for anonymization excluding open-ended demographic data and comments), and the preregistration of Study 3 are available in the project on Open Science Framework (https://osf.io/auwdk/). The first author contributed to this article by elaborating the study ideas; collecting the data for Study 1, partly Study 2, and Study 3; and by writing the manuscript draft. The second author contributed by providing expertise on the topics of feedback and performance diagnostics, by partially collecting data for Study 2, and by revising the manuscript. The third author gave the initial impulse for the project and prepared Study 1 before she passed away untimely. This research project is dedicated to her.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11218-021-09633-y.

Acknowledgements The authors would like to note that this line of research was initiated by our dear friend and colleague Jessica Bilz, who unexpectedly passed away before the first study was completed. We are continuing this research with her on our minds and in our hearts. In addition, we thank Lara Moen, Janina Höpfner, and Morzal Goia for their help with preparation and data collection.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Banse, R., & Gawronski, B. (2003). Die Skala Motivation zu vorurteilsfreiem Verhalten: Psychometrische Eigenschaften und Validität. *Diagnostica, 49*(1), 4–13. https://doi.org/10.1026//0012-1924.49.1.4

Biernat, M. (2012). Chapter one—stereotypes and shifting standards: Forming, communicating, and translating person impressions. In P. Devine & A. Plant (Eds.), *Advances in experimental social psychology.* (pp. 1–59). Academic Press. https://doi.org/10.1016/B978-0-12-394286-9.00001-9

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

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

Cornwell, C., Mustard, D. B., & Van Parys, J. (2013). Noncognitive skills and the gender disparities in test scores and teacher assessments: Evidence from primary school. *Journal of Human Resources, 48*(1), 236–264. https://doi.org/10.3368/jhr.48.1.236

Croft, A., & Schmader, T. (2012). The feedback withholding bias: Minority students do not receive critical feedback from evaluators concerned about appearing racist. *Journal of Experimental Social Psychology, 48*(5), 1139–1144. https://doi.org/10.1016/j.jesp.2012.04.010

Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychological Methods, 14*(2), 81–100. https://doi.org/10.1037/a0015914 PubMed.

Duval, C. M. (1980). Differential teacher grading behavior toward female students of mathematics. *Journal for Research in Mathematics Education, 11*(3), 202–214. https://doi.org/10.2307/748941

Dweck, C. S., Davidson, W., Nelson, S., & Enna, B. (1978). Sex differences in learned helplessness: II. The contingencies of evaluative feedback in the classroom and III. An experimental analysis. *Developmental Psychology, 14*(3), 268–276. https://doi.org/10.1037/0012-1649.14.3.268

Evans, C. (2013). Making sense of assessment feedback in higher education. *Review of Educational Research, 83*(1), 70–120. https://doi.org/10.3102/0034654312474350

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

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. *Advances in Experimental Social Psychology, 23*, 1–74. https://doi.org/10.1016/S0065-2601(08)60317-2

Harber, K. D. (1998). Feedback to minorities: Evidence of a positive bias. *Journal of Personality and Social Psychology, 74*(3), 622–628. https://doi.org/10.1037/0022-3514.74.3.622

Harber, K. D., Gorman, J. L., Gengaro, F. P., Butisingh, S., Tsang, W., & Ouellette, R. (2012). Students' race and teachers' social support affect the positive feedback bias in public schools. *Journal of Educational Psychology, 104*(4), 1149–1161. https://doi.org/10.1037/a0028110

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Heller, K. A., & Parsons, J. E. (1981). Sex differences in teachers' evaluative feedback and students' expectancies for success in mathematics. *Child Development, 52*(3), 1015. https://doi.org/10.2307/1129106.

Heyder, A., & Kessels, U. (2015). Do teachers equate male and masculine with lower academic engagement? How students' gender enactment triggers gender stereotypes at school. *Social Psychology of Education, 18*(3), 467–485. https://doi.org/10.1007/s11218-015-9303-0

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

Ingenkamp, K. (1971). *Die Fragwürdigkeit der Zensurengebung [The dubiousness of school grades].* Beltz.

IQB. (2012). *IQB—Beispielaufgaben Deutsch Primarstufe*. https://www.iqb.hu-berlin.de/vera/aufgaben/dep

IQB. (2013). *IQB—Beispielaufgaben Mathematik Primarstufe*. https://www.iqb.hu-berlin.de/vera/aufgaben/map.

Jones, S. M., & Dindia, K. (2004). A meta-analytic perspective on sex equity in the classroom. *Review of Educational Research, 74*(4), 443–471. https://doi.org/10.3102/00346543074004443

Kuhl, P., & Hannover, B. (2012). Differenzielle Benotungen von Mädchen und Jungen. *Zeitschrift Für Entwicklungspsychologie Und Pädagogische Psychologie, 44*(3), 153–162. https://doi.org/10.1026/0049-8637/a000066

Lavy, V., & Sand, E. (2015). *On The Origins of Gender Human Capital Gaps: Short and Long Term Consequences of Teachers' Stereotypical Biases*. University of Warwick, Department of Economics. https://EconPapers.repec.org/RePEc:wrk:warwec:1085

Lindahl, E. (2016). Are teacher assessments biased?—evidence from Sweden. *Education Economics, 24*(2), 224–238. https://doi.org/10.1080/09645292.2015.1014882

Marcenaro-Gutierrez, O., & Vignoles, A. (2015). A comparison of teacher and test-based assessment for Spanish primary and secondary students. *Educational Research, 57*(1), 1–21. https://doi.org/10.1080/00131881.2014.983720

McShane, B. B., & Böckenholt, U. (2017). Single-paper meta-analysis: Benefits for study summary, theory testing, and replicability. *Journal of Consumer Research, 43*(6), 1048–1063. https://doi.org/10.1093/jcr/ucw085

Narciss, S. (2004). The impact of informative tutoring feedback and self-efficacy on motivation and achievement in concept learning. *Experimental Psychology, 51*(3), 214–228. https://doi.org/10.1027/1618-3169.51.3.214 pdh.

Narciss, S. (2008). Feedback strategies for interactive learning tasks. In J. M. Spector, M. D. Merrill, J. J. G. van Merrienboer, & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology, 3.* (pp. 125–144). Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203880869.ch11

Narciss, S. (2017). Conditions and effects of feedback viewed through the lens of the interactive tutoring feedback model. In D. Carless, S. M. Bridges, C. K. Y. Chan, & R. Glofcheski (Eds.), *Scaling up assessment for learning in higher education.* (pp. 173–189). Springer. https://doi.org/10.1007/978-981-10-3045-1_12

Parsons, J. E., Kaczala, C. M., & Meece, J. L. (1982). Socialization of achievement attitudes and beliefs: Classroom influences. *Child Development, 53*(2), 322–339. https://doi.org/10.2307/1128974

Retelsdorf, J., Schwartz, K., & Asbrock, F. (2015). "Michael can't read!" Teachers' gender stereotypes and boys' reading self-concept. *Journal of Educational Psychology, 107*(1), 186–194. https://doi.org/10.1037/a0037107 pdh.

Robinson, J. P., & Lubienski, S. (2011). The development of gender achievement gaps in mathematics and reading during elementary and middle school: Examining direct cognitive assessments and teacher ratings. *American Educational Research Journal, 48*(2), 268–302. https://doi.org/10.3102/0002831210372249

Robinson-Cimpian, J. P., Lubienski, S. T., Ganley, C. M., & Copur-Gencturk, Y. (2014a). Teachers' perceptions of students' mathematics proficiency may exacerbate early gender gaps in achievement. *Developmental Psychology, 50*(4), 1262–1281. https://doi.org/10.1037/a0035073

Robinson-Cimpian, J. P., Lubienski, S. T., Ganley, C. M., & Copur-Gencturk, Y. (2014b). Are schools shortchanging boys or girls? The answer rests on methods and assumptions: Reply to Card (2014) and Penner (2014). *Developmental Psychology, 50*(6), 1840–1844. https://doi.org/10.1037/a0036693

Sagarin, B. J., Ambler, J. K., & Lee, E. M. (2014). An ethical approach to peeking at data. *Perspectives on Psychological Science, 9*(3), 293–304. https://doi.org/10.1177/1745691614528214

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Spear, M. G. (1984a). The biasing influence of pupil sex in a science marking exercise. *Research in Science and Technological Education, 2*(1), 55–60

Spear, M. G. (1984b). Sex bias in science teachers' ratings of work and pupil characteristics. *European Journal of Science Education, 6*(4), 369–377. https://doi.org/10.1080/0140528840060407

Swim, J., Borgida, E., Maruyama, G., & Myers, D. G. (1989). Joan McKay versus John McKay: Do gender stereotypes bias evaluations? *Psychological Bulletin, 105*(3), 409–429. https://doi.org/10.1037/0033-2909.105.3.409

Tiedemann, J. (2002). Teachers' gender stereotypes as determinants of teacher perceptions in elementary school mathematics. *Educational Studies in Mathematics, 50*(1), 49–62. https://doi.org/10.1023/A:1020518104346

Carolin Schuster Carolin Schuster is an assistant professor for Applied Social Psychology at the Leuphana University Lüneburg. One of her major research interests is focused on the consequences of gender stereotypes and gender roles in educational, professional, and domestic contexts, aiming at a better understanding of how equality and freedom of all genders can be fostered. In addition, she examines how personal values guide people's behaviors in conflicts and other social contexts.

Susanne Narciss Susanne Narciss is a full professor and head of the research team "Psychology of Learning and Instruction (PsyLI)" at the Faculty of Psychology at the Technische Universität Dresden, Germany. The PsyLI-team is conducting theory-driven and design-based psychological research on issues related to life-long learning and instruction within socio-technical systems. Her current interests include research on (a) promoting self-regulated learning, (b) the role of motivation and metacognition in instructional contexts, (c) conditions and effects of interactive learning tasks, and (d) conditions and effects of formative feedback strategies.

Jessica Bilz Jessica Bilz had been a very talented and dedicated young researcher in Susanne Narciss's research team at the University Passau from April 2014 until her death in May 2015. She graduated from the teacher education program of the Technische Universität Dresden in Summer 2013. Her research interests focused on issues of how to train teachers in using formative feedback strategies that promote student learning and motivation.