International Journal of Science Education

ISSN: 0950-0693 (Print) 1464-5289 (Online) Journal homepage: http://www.tandfonline.com/loi/tsed20

Studying Gender Bias in Physics Grading: The role of teaching experience and country

Sarah I. Hofer

To cite this article: Sarah I. Hofer (2015) Studying Gender Bias in Physics Grading: The role of teaching experience and country, International Journal of Science Education, 37:17, 2879-2905, DOI: 10.1080/09500693.2015.1114190

To link to this article: http://dx.doi.org/10.1080/09500693.2015.1114190

International Journal of Science Education, 2015
Vol. 37, No. 17, 2879–2905, http://dx.doi.org/10.1080/09500693.2015.1114190

Studying Gender Bias in Physics Grading: The role of teaching experience and country

Sarah I. Hofer*
Institute for Research on Learning and Instruction, ETH Zurich, Zurich, Switzerland

The existence of gender-STEM (science, technology, engineering, and mathematics) stereotypes has been repeatedly documented. This article examines physics teachers' gender bias in grading and the influence of teaching experience in Switzerland, Austria, and Germany. In a 2 × 2 between-subjects design, with years of teaching experience included as moderating variable, physics teachers (N = 780) from Switzerland, Austria, and Germany graded a fictive student's answer to a physics test question. While the answer was exactly the same for each teacher, only the student's gender and specialization in languages vs. science were manipulated. Specialization was included to gauge the relative strength of potential gender bias effects. Multiple group regression analyses, with the grade that was awarded as the dependent variable, revealed only partial cross-border generalizability of the effect pattern. While the overall results in fact indicated the existence of a consistent and clear gender bias against girls in the first part of physics teachers' careers that disappeared with increasing teaching experience for Swiss teachers, Austrian teachers, and German female teachers, German male teachers showed no gender bias effects at all. The results are discussed regarding their relevance for educational practice and research.

Keywords: *Gender bias; Teaching experience; Physics instruction*

Introduction

Even today, a considerably smaller proportion of females than males opt for a career in engineering and science. In Switzerland 1.4% of all females work in one of these fields (compared to 6.6% of all males). In Austria 1.1% of all women (compared to 2.8% of all men) and in Germany 1.5% of all women (compared to 4.8% of all men) pursue a career in engineering or science (European Commission, 2013). Only 9% of the students at a big university in Switzerland who graduated with a Bachelor degree in

*Corresponding author. Institute for Research on Learning and Instruction, ETH Zurich, Clausiusstrasse 59, CH-8092 Zurich, Switzerland. Email: sarah.hofer@ifv.gess.ethz.ch

physics in 2013 were female (ETH Zurich Annual Report, 2013). Based on own data from $N = 396$ Swiss secondary school students that were gathered between 2012 and 2014, girls receive significantly lower physics grades than boys ($p < .001$). All these figures exemplarily illustrate the large gender gap that is still present in the STEM (science, technology, engineering, and mathematics) fields. Secondary school was identified as a crucial point in time to consolidate differences between girls and boys in terms of STEM performance, interest, and participation (Ceci, Williams, & Barnett, 2009). Among all of the explanations that were provided for the gender gap, the present study addressed the basic aspect of gender biased grading in secondary school physics. A number of studies examined accuracy and various biases in teachers' judgments of student performance (e.g. Dünnebier, Gräsel, & Krolak-Schwerdt, 2009; Glock & Krolak-Schwerdt, 2014; Südkamp, Kaiser, & Möller, 2012). There is no recent work, however, that explicitly investigated whether secondary school teachers' grading in physics indeed reveals a bias to the detriment of girls. The present study hence aimed to fill this gap and additionally shed light on the role of teaching experience. The generalizability of potential gender bias effects in secondary school physics was examined by comparing teachers' bias patterns across three German-speaking countries that are culturally closely related.

In the following sections, the literature addressing gender bias in teachers' judgments in STEM fields is outlined. First, mechanisms that underlie biased judgments and existing research on gender bias in academic judgments are considered. Then I turn to the potential influence of teaching experience. Finally, the cross-border generalizability of gender bias effects is briefly addressed, before the present study is introduced.

Gender Bias in Teachers' Judgments in STEM Fields

There are only a few studies that focus on gender bias in teachers' judgments in the specific domain of physics. Therefore, most of the findings and theoretical considerations that are summarized in the following sections relate to the broader category of STEM fields.

Underlying Mechanisms: Gender-STEM stereotypes

To be able to navigate through our highly demanding social environment, schemata are applied that efficiently categorize our perceptions (Bartlett, 1932). Schemata that refer to members of social groups are stereotypes. A stereotype associates a social group with one or a set of attributes (e.g. Greenwald et al., 2002). The most acknowledged models that were proposed to explain the influence of stereotypes on judgment processes are dual process models (see Brewer, 1988; Fiske & Neuberg, 1990) and parallel-constraint-satisfaction models (see Kunda & Thagard, 1996). Dual process models assume a serial processing of information about an individual. Information that refers to stereotypes is processed first (e.g. the person who likes action movies is uneducated). Information that refers to specific attributes of the

individual (e.g. the person is interested in history, enjoys literature, etc.) is processed afterwards, in case that the perceiver is able and motivated to engage in that kind of more controlled processing (see Brewer, 1988; Fiske & Neuberg, 1990). Parallel-constraint-satisfaction models represent all kinds of information including observed attributes, observed behavior, inferred attributes, and stereotypes as connected nodes in a network. Nodes that are associated with observed information about the individual (e.g. female, white coat) are activated or deactivated depending on the valence of the associations (see Kunda & Thagard, 1996). To give a simplified example, the observed information 'female' and 'white coat' and their associations with other information in the network might lead to a strong activation of 'pharmacist', 'helpful', and 'academic', for instance. The whole network that is activated in the specific judgment situation finally determines how all information is interpreted. All models arrive at very similar conclusions in terms of factors that are expected to affect the extent of a stereotype's influence on the judgment process. Accordingly, among others, cognitive business or limited cognitive resources in the judgment situation and ambiguous information can increase the probability that stereotypes take effect and dominate specific information about the individual (e.g. Chaiken & Maheswaran, 1994; Kunda & Spencer, 2003; Kunda & Thagard, 1996). Under these conditions, attributes that are part of a certain activated stereotype may influence a judgment while specific information about the individual is ignored.

In the present study, gender-STEM stereotypes were expected to potentially bias teachers' judgments. In general, gender stereotypes (how we picture a typical female or male) to some extent reflect but also contribute to existing gender differences in behavior (see Eagly & Wood, 2013). Perceived incongruity between gender stereotypes and stereotypic job roles (how we picture a typical hairdresser, construction worker, politician, or teacher) may lead to biased evaluations and prejudice against those females (or males) performing in a nontraditional domain (a female construction worker, for instance; e.g. Eagly & Koenig, 2008; Eagly, Wood, & Diekman, 2000). In line with this, existing research points to a commonly perceived mismatch between stereotypic views of women, on the one hand, and scientists, on the other hand (see Farenga & Joyce, 1999; Kessels, Rau, & Hannover, 2006; Nosek, Banaji, & Greenwald, 2002; Nosek et al., 2009). Accordingly, gender-STEM stereotypes can be defined as stronger associations between STEM-related content and males than females (see Miller, Eagly, & Linn, 2014; Nosek et al., 2002, 2007). Gender-physics stereotypes are one instantiation of gender-STEM stereotypes. A perceived mismatch between women and physics is repeatedly observed on the gender-science Implicit Association Test. This test measures the strength of the implicit association of male vs. female words with words that either represent science or liberal arts. Physics is presented as one instantiation of science words (e.g. Nosek et al., 2009). Gender stereotypes in the domain of physics may be even stronger than gender stereotypes in some of the other STEM domains. When comparing physics and math teachers' implicit theories about their students' achievement and ability in their respective fields, physics teachers' cognition tended to be even slightly more gender-

biased in favor of boys than math teachers' cognition (Heller, Finsterwald, & Ziegler, 2010).

To sum up, when physics teachers evaluate the performance of students, gender-STEM stereotypes may influence the judgment process, especially in judgment situations that are cognitively demanding and provide ambiguous information.

Existing Research: Disentangling bias and accuracy

There are two main approaches that dominate research on teachers' judgment biases. In the first approach, the characteristics that are expected to trigger biased evaluations in a particular judgment domain are manipulated, while the content that has to be judged stays the same in each condition. Focusing on potential gender-STEM bias effects, Moss-Racusin, Dovidio, Brescoll, Graham, and Handelsman (2012) applied this approach to the educational domain and found that science faculty staff derived significantly higher competence levels from identical application materials with a male name than those with a female name. By investigating secondary school science teachers' evaluations of the same essays that were either indicated to originate from a girl or a boy, Goddard Spear (1984a) also reported a rather consistent bias towards boys, with regard to grades, estimated competence, and the students' perceived inclination for science (see also Goddard Spear, 1984b). Although the author used a similar design, Baird (1998) did not find any gender bias in grading for A-level examinations in chemistry.

In the second, correlational approach, teachers' judgments of student performance are compared to objective performance measures to estimate judgment accuracy and biases. Judgments that are influenced by stereotypes are regarded as accurate or biased depending on the degree they reflect actual group differences (see, e.g. Jussim & Eccles, 1992; Madon et al., 1998). There is evidence that teachers tend to overestimate their male students' proficiency in math when actual performance is accounted for (Jussim & Eccles, 1992; Robinson-Cimpian, Lubienski, Ganley, & Copur-Gencturk, 2014). In keeping with Robinson-Cimpian et al. (2014), equally performing girls have to outmatch boys in terms of teachers' perceived effort, diligence, and manners to be rated as equally proficient in math.

In the present study, the highly controlled experimental approach was preferred to the correlational approach for two reasons. First, small but existent self-fulfilling prophecy effects (see Jussim & Harber, 2005) and stereotype threat effects (e.g. Nguyen & Ryan, 2008) may always be reflected in actual student performance that is assessed in the correlational approach. It may be simply not possible to measure teachers' judgment accuracy by comparing teachers' judgments to a standard that is unaffected by stereotypes (Walton & Spencer, 2009). The experimental approach does not need a standard that reflects real students' actual performance in order to assess gender bias effects. Second, no conclusions about teachers' judgment accuracy were intended. Only the correlational approach allows such conclusions. In the present study, 'bias' did not refer to a systematic deviation from an objective assessment of actual student performance. Bias was meant to indicate a systematic variation in

teachers' judgments as a function of the experimental variation of a stereotyped characteristic.

To sum up, there is some evidence for a bias against females in STEM fields. Most of the research up to now, however, focused on science in general or on math, but not on physics. How teaching experience may affect a bias against females in STEM fields is addressed in the following section.

The Role of Teaching Experience

Stereotypes are particularly influential in judgment situations that are cognitively demanding and provide ambiguous information. In the classroom, the information that is available to make a decision about a student's performance level usually is complex, ambiguous, and open to various interpretations. The accuracy of teachers' ratings of students' performance indeed seems to be lower for science and social studies than for reading, language arts, or mathematics (Hopkins, George, & Williams, 1985) and lower for conceptual questions than for computational questions (Coladarci, 1986), which inherently provide less strict evaluation criteria and more interpretative ambiguity.

Both the perceived ambiguity of information and a high demand for cognitive resources in the judgment situation can be expected to diminish with increasing teaching experience. There is evidence that expert teachers, in comparison to novices, are able to automatize parts of their work (see Carter, Sabers, Cushing, Pinnegar, & Berliner, 1987; Leinhardt & Greeno, 1986), and to quickly and correctly recognize more meaningful patterns as a function of their experience (see Berliner, 2001). Expert teachers, but not novices, seem to use elaborated schemata as frameworks to efficiently interpret and understand the often complex information that has to be processed (Carter et al., 1987). Although, in general, mere experience is not sufficient to determine expert teachers (see Palmer, Stough, Burdenski, & Gonzales, 2005), it is suggested that these skills develop with more experience of teaching.

Accordingly, the need to invoke stereotypes in grading may also decrease with increasing experience of teaching, which is supported by the following findings. Krolak-Schwerdt, Böhmer, and Gräsel (2009, 2012) instructed participants to read students' case reports and to either form an impression of the students' performance and personality or to predict future performance. The latter was stressed to be relevant for the students' future academic careers. The authors found that teachers with at least 10 years of teaching experience but not laymen were able to process information that referred to relevant attributes of the student when they had to predict the students' performance. When they only had to form an impression, they relied on information that referred to a stereotype. Laymen were unable to switch between these modes of processing. In a related study on the judgment of student performance (Dünnebier et al., 2009), student teachers were more influenced by prior information about the student than teachers with at least 8 years of experience. Finally, Babad (1985) found that elementary school teachers' grading in the context of text comprehension varied significantly as a function of the fictitious performance label (excellent vs. weak

student) in the group of the less experienced teachers (not more than 8 years of experience), but not in the group of the more experienced teachers.

To sum up, teachers with little experience of teaching may show a gender bias in their judgment of student performance. With increasing experience, teachers can be expected to develop the cognitive resources that are necessary to avoid the influence of stereotypes on the process of evaluation.

Cross-Border Generalizability of Gender-STEM Bias Effects

Overall, more than 70% of the participants in a study by Nosek et al. (2009) from 34 countries all over the world held implicit gender-STEM stereotypes. The degree of national stereotype endorsement turned out to predict nation-level gender achievement gaps in school science (Nosek et al., 2009). Also the proportion of women who participated in tertiary science education predicted the degree of nation-level gender-STEM stereotype endorsement (Miller et al., 2014). On a general level, the cultural context shapes the categories that are used to organize our perceptions and hence also influences the content of stereotypes (see, e.g. Fiske, Kitayama, Markus, & Nisbett, 1998). Gender-STEM bias effects in teachers' judgments may thus generalize over countries that are culturally closely related and that are comparable in terms of the nation-level representation of women in STEM fields and in terms of gender differences in science performance measures.

The Present Study

The present study applied the experimental approach to examine gender bias in physics teachers' judgments and the role of teaching experience. Secondary school physics teachers received a physics test question and the same written student answer, accompanied by the prompt to assign a grade. The question asked a fictive student for a written explanation about his or her conceptual understanding of Newtonian mechanics (detailed information on the judgment situation is provided in the section 'The Judgment Situation'). Two factors were manipulated in a short introductory text: student gender and specialization in languages vs. science. The second factor, specialization, was only included to gauge the relative strength of potential gender bias effects. Effects of gender on grading could then be compared to the effects of another category (students focusing on languages vs. students focusing on science) that was assumed to more clearly reflect actual group differences but to represent a less prevailing and less distinct social category.

Based on existing research, the present study expected physics teachers to show a gender bias in grading, to the detriment of girls. Student gender was assumed to more strongly influence grading than the less prominent social category student specialization. The study further aimed to investigate the potential moderating effect of teaching experience, which may reduce gender bias with increasing years of practice. In comparison to most other studies that contrasted groups of less and more experienced teachers, this study included teaching experience as continuous variable.

Because the three German-speaking countries Switzerland, Austria, and Germany are culturally closely related and comparable in terms of the nation-level representation of women in STEM fields (e.g. European Commission, 2013) and in terms of an existing advantage for boys in science performance measures (e.g. Organisation for Economic Co-operation and Development, 2011), a generally valid pattern of bias effects independent of German-speaking country was expected.

Method

Design

This study applied a 2 × 2 between-subjects factorial design. The two independent variables were student gender (female vs. male) and student specialization (languages vs. science). The grade that teachers assigned to the answer of the fictive student was the dependent variable. In real evaluation situations at school, a student's oral or written performance is generally evaluated by assigning a grade. Hence, asking the teachers to assign a grade ascertained ecological validity and allowed fast and intuitive processing of the survey. Teaching experience in years served as continuous moderating variable to investigate the influence of teaching experience on the effects of gender and specialization on the grade that was awarded. The effect pattern was compared between samples from Switzerland, Austria, and Germany to be able to examine its generalizability.

The study was run through the use of an online-survey tool, SoSciSurvey (Leiner, 2014), which could be accessed from every web-enabled device via a link. Physics teachers' associations and science education research institutions in Switzerland, Austria, and Germany were contacted and asked to distribute a request for participation that included the survey link to their mailing lists. The mailing lists explicitly addressed physics teachers. In the request for participation in the email and in the survey itself, it was emphasized that the study was exclusively aimed at physics teachers. Three country-specific links and surveys were prepared. Certain demographic and personal questions, as well as the grading system, were adapted to the countries' respective national standards. Both in the request for participation and in the introductory text in the survey itself, the overall objective of the study was described as investigating the process of performance evaluation in secondary school physics. The research interest in gender bias, however, was not made explicit in order to reduce social desirability biases and conscious efforts to avoid prejudice that could have, otherwise, distorted the findings. Hence, teachers were told that this research project particularly aimed to examine the correspondence between two approaches to assess a student's performance on a test. The teachers were informed that in the first approach, the test is split into the single test questions and each test question is evaluated by a different physics teaching expert, while in the second approach, one expert evaluates the complete test. According to this cover story, every participating teacher was assigned to an assessment situation that complied with one of these two approaches. In actual fact, all teachers had to evaluate the same answer to the same

single test question. This cover story, however, justified why the teachers were asked to evaluate a single test answer by assigning a grade. Because the teachers were told that they were evaluating a real student's test answers that had been provided by different schools, this study examined gender bias in an experimental design while maintaining good ecological validity.

Teacher Samples

A sample size of 20 physics teachers per experimental condition, which resulted in at least 80 teachers per country, was set as the lower limit. Country-specific data collection was finished after this limit was reached and when the survey was not accessed for at least four days. Following this procedure, 167 cases were initially registered from Switzerland, 178 from Austria, and 589 from Germany. In all of the three German-speaking countries, physics is more extensively instructed only in the higher tracks of secondary school. To arrive at comparable samples, only those participants who indicated that they taught at a higher level secondary school were considered. Participants whose age suggested that they had already retired were further excluded from the analyses. When no grade was awarded, the participant's data were eliminated. By checking IP-addresses and personal data, multiple completions of the survey were detected, and the respective data were deleted. Hence, the Swiss sample finally included $N = 116$ (14 women) physics teachers. On average, they were $M = 48.83$ (SD = 9.26) years old and had $M = 18.32$ (SD = 10.20) years of teaching experience. Due to the multilingualism in Switzerland, German language proficiency was additionally collected at the beginning of the survey in order to directly exclude teachers who did not have a German-speaking background. The Austrian sample included $N = 137$ (59 women) teachers, with a mean age of $M = 47.03$ (SD = 10.89) years and a mean length of teaching experience of $M = 19.58$ (SD = 12.40) years. The German sample included $N = 527$ (125 women) physics teachers, with a mean age of $M = 46.64$ (SD = 10.96) years and a mean length of teaching experience of $M = 17.17$ (SD = 11.84) years. The gender distribution in the three samples resembled country-specific statistical information on the gender distribution of physics teachers: The proportion of women among secondary school teachers in physics is about 16% in Switzerland (personal communication with ETH Zurich and University of Education Berne), about 45% in Austria (personal communication with the Austrian Federal Ministry for Education and Women), and about 37% in Germany (Destatis, 2013). The total sample included $N = 780$ German-speaking secondary school physics teachers.

Procedure

When accessing the online-survey, a brief introductory text informed the participants about the study's aim (according to the cover story) and the procedure. After the anonymous assessment of demographic and personal information, including years of teaching experience, participants were randomly forwarded to one of the four conditions.

In all of the four conditions, teachers received exactly the same information, with the exception of all of the terms that referred to the fictive student's gender and the student's specialization (languages vs. science), which were interchanged based on the condition. Following a short text that introduced the student, the teachers saw the physics test question, which asked the fictive student for a written explanation that targeted his or her conceptual understanding of Newton's mechanics, and the answer of the student. The teachers were asked to evaluate the student's answer by assigning a grade according to their respective country-specific school grading systems. Answers were graded by moving a continuous slider that instantaneously provided the corresponding number of the grade to one decimal point. Due to the randomization of the experimental conditions, systematic individual differences in the severity or leniency of the judgment could be neglected. For illustrative purposes, essential parts of the German online-survey were translated into English and are summarized in Figure 1.

The test questions, student answers, and brief descriptions of the context are directly provided by the participating schools. We only summarize the information that we receive. The content that is assessed in the test questions had always been taught in the lessons before.

In the following text, you will see a test question on Newtonian mechanics and a _____ (female/male student's) answer. _____ (She/He) is in _____ (her/his) Junior Year and takes Honors/AP courses. During _____ (her/his) school career, _____ (she/he) has focused on _____ (languages/the natural sciences), thus far. Please evaluate the _____ (female/male student's) answer.

_____ (She/He) was asked the following test question:
Two skateboarders who significantly differ in their masses each stand on a skateboard, face to face. They are connected by a tensioned rope. The left and lighter skateboarder actively pulls the rope, while the heavier right skateboarder only holds it. What happens? Explain your assumption in approximately five to six sentences. Friction is negligible.

The _____ (female/male student's) answer:
In general, force is composed of a person's mass and acceleration. The right skateboarder has to hold the rope as strongly as the left skateboarder pulls it. Both of the skateboarders are, thus, affected by forces of equal strengths, although only the left skateboarder pulls the rope. Consequently, nothing should happen because the two forces cancel out one another. Because the mass of the left skateboarder, however, is smaller than the mass of the right skateboarder, the left skateboarder has a higher acceleration than the right skateboarder. As a result, the left skateboarder most likely should at least move a small amount in the direction of the right skateboarder.

Please evaluate this _____ (female/male student's) answer by assigning a grade on a scale from A, with a corresponding grade point of 4.0, to F, with a corresponding grade point of 0.0.

In order to do this, please move the slider to the desired position.

Figure 1. English adaptation of the instructions and information that the teachers received. Terms that were interchanged in the four conditions are omitted, and variants are presented in parentheses. Note that in the German language, the student's gender is simply indicated by slightly changing the word's ending (female student = Schülerin, male student = Schüler)

The Judgment Situation

In this study, a student's answer to a conceptual question in Newtonian mechanics was used as the judgment situation for two reasons. First, compared to problems that require computation, conceptual questions turned out to be more difficult to evaluate accurately (Coladarci, 1986). Answers to conceptual questions can be expected to imply higher ambiguity and leeway in construal. Since stereotypes influence the judgment process particularly in judgment situations that are cognitively demanding and provide ambiguous information, gender bias effects may be especially pronounced for conceptual questions. Second, Newtonian mechanics is a topic that is most likely addressed in every physics classroom. Even though some secondary school students may have only a few physics lessons, almost all students deal with basic Newtonian mechanics. Consequently, the evaluation of questions on Newtonian mechanics is highly relevant, both for teachers and students.

While in standard grading situations grades are rarely assigned on the basis of one answer to one test question, there were reasons why only one item was used in the context of this study. First, the teachers' evaluation of a student's answer to an open conceptual question was considered to be a good proxy not only for the evaluation of written tests but also for the evaluation of a student's daily classroom contributions. If a student's answer to a teacher's question is evaluated differently as a function of the student's gender, this bias can be expected to show up in every evaluation of students' answers and accumulate over time. The judgment situation implemented in this study allowed assessing this kind of immediate evaluation that may reflect a very general bias in how physics teachers' process information about their students. Second, in this study, data were gathered using an online-survey which enabled the researcher to collect data from a large number of teachers from three different countries. In contrast to settings that imply personal contact, someone who has received a link to participate in an online-survey does not necessarily decide to participate. This decision can be expected to largely depend on the assumed costs (time and effort) of participation. The prompt to carefully evaluate a student's answers to several test questions could have restrained many teachers from participating. Some of those who would nevertheless have decided to participate might have lost interest over time and worked less conscientiously, reducing the quality of the data. Therefore, the use of a single answer was also the result of balancing the advantages of a broader assessment of teachers' evaluation behavior against the realistic risk of small sample sizes and data of low quality.

The conceptual test question that was used in this study was adapted from the Test of basic Mechanics Conceptual Understanding (bMCU; Hofer, Schumacher, & Rubin, 2015), a Rasch-scaled multiple-choice test on the conceptual understanding of Newton's mechanics. The 'skateboarder question' (see Figure 1) was chosen because it covered a problem that, in fact, frequently appeared in physics textbooks, exams, and classroom instruction in all of the three countries and required a complex answer that potentially included several correct and incorrect statements. In the process of the development of the bMCU Test, a variety of oral and written student answers to this conceptual question were recorded. These answers were

used to design three exemplary student answers. The aim was to arrive at an answer that represented average student performance and was neither completely wrong nor absolutely correct, in order to leave room for interpretation. The three answers were given to five informed physics teaching experts who were asked to assign a grade to each of them. The answer that most unequivocally reflected average performance was finally chosen and used in the study (see Figure 1).

Data Analysis

To investigate a potential gender bias in physics grading and the influence of teaching experience within and across the three countries, multiple group regression analyses were performed with Mplus Version 7.11 (Muthén & Muthén, 2012) with country as grouping variable. Grades were transformed into *z*-scores for each country in order to account for the different grading scales and, if necessary, recoded to create a grade scale where higher values indicated higher performance. This joint grade scale is referred to in the following sections and was used in the analyses.

Grades were regressed on student gender (0 = female, 1 = male) and specialization (0 = languages, 1 = science). Teaching experience, the interaction between gender and teaching experience, as well as the interaction between specialization and teaching experience were further included as predictors to be able to examine the potential moderating effect of teaching experience. Teaching experience, which was measured in years, was entered into the regression without being *z*-standardized to be able to examine the potential change in the gender bias effect with growing years of teaching experience. Consequently, the regression coefficient of gender reflected the influence of a fictive student's gender on the grades at the beginning of the teaching career—with zero experience of teaching.

To gain further insights into the meaning of potential interaction effects between the fictive student's gender and teaching experience in the empirical, and not linearly modeled, data, an additional analysis was performed. Grades were averaged within bins of 5 years of teaching experience, resulting in nine bins. Within each teaching experience bin, the mean grades that were awarded to a fictive female student were compared to the mean grades that were awarded to a fictive male student using *t*-tests.

Existing research suggests that the teacher's gender should have no influence on (gender) bias effects (see Moss-Racusin et al., 2012; Swim, Borgida, Maruyama, & Myers, 1989). Nevertheless, to rule out such influences, measurement invariance in terms of the regression model was investigated across the teachers' gender within each country separately. Only after the analysis of measurement invariance across the teachers' gender, which indicated whether female and male teachers from the same country could be reasonably considered together or had to be considered separately, the cross-border generalizability of the effect pattern was investigated.

The general strategy that was pursued in this study to analyze measurement invariance in terms of countries and the teachers' gender is outlined in what follows. One aim of this study was to examine the generalizability of potential gender bias effects. A generally valid pattern of effects and, thus, no differences between countries or

female and male physics teachers (i.e. measurement invariance) was expected. This research hypothesis had to be tested against the hypothesis that there was no generally valid pattern. Differences in the patterns (i.e. in the regression coefficients) between countries or female and male physics teachers would have disproved the notion of a generally valid pattern of bias effects. Models were constructed that defined how the regression was estimated in each country or for female and male teachers in each country, respectively (the specific models that were constructed are described in the 'Results'). In general, a model that assumed a universally valid pattern of effects according to the research hypothesis constrained the regression to be the same for all teachers (restrictive model). An alternative model that assumed differences in the patterns, by contrast, allowed the regression to vary between countries or female and male physics teachers and was hence less restrictive. The log-likelihood significance test that was used in this study to compare different models examined whether a more restrictive model described the data significantly worse than a less restrictive model. If this was the case, the more restrictive model had to be rejected. Using a significance level of $\alpha = .05$ meant that there was a 5% probability that the more restrictive model was rejected although it did not fit the data worse than the less restrictive model (a type I error). Increasing the significance level increased the power of the significance test to detect that the more restrictive model fitted the data worse than the less restrictive model. Increasing the significance level hence increased the chances of a type I error, but decreased the chances of a type II error. In this study, a type II error meant that the effect patterns were assumed to be identical across countries or female and male teachers when the effect patterns in fact differed. This study aimed at confirming the fit of the restrictive model to substantiate the research hypothesis that the effect patterns did not differ across countries and the teachers' gender. Assuming that the effect patterns did not differ between countries or female and male teachers when the effect patterns in fact did differ (a type II error) could thus be regarded as more problematic than assuming that the effect patterns differed when they did not differ (a type I error). Therefore, the significance level for all significance tests of invariance was set to $\alpha = .20$ to increase the rigor of the test of the research hypothesis.

Results

In the original regression model, grades were regressed on student gender and specialization, teaching experience, the interaction between gender and teaching experience, as well as the interaction between specialization and teaching experience. The specialization of the fictive student as well as the interaction between specialization and teaching experience, however, turned out to have no systematic influence on the grade that was awarded (all ps $\geq .12$), neither for female nor for male teachers in any of the countries. Therefore, specialization and the interaction between specialization and teaching experience were excluded from all analyses that are reported in the following sections.

Descriptive Statistics

Descriptive statistics that are related to the grade scale, without considering teaching experience, can be found in Table 1. Grade data are presented for each country separately organized according to the two experimentally manipulated variables.

Effects of the Teachers' Gender

The regression model, which now only included the three predictor variables gender, teaching experience, and the interaction between gender and teaching experience, was tested in terms of measurement invariance across the teachers' gender within each country separately (i.e. multiple group regression analyses with the teachers' gender as grouping variable). This meant that two models were compared within each of the three countries. A model that specified no restrictions in terms of the estimation of the regression coefficients was called the unrestrictive model. In the unrestrictive model, the regression was estimated independently for female and male physics teachers within each country. The regression coefficients of female and male physics teachers were hence allowed to differ. In the restrictive model, by contrast, the regression coefficients were constrained to be equal between female and male physics teachers of the same nationality. So this restrictive model assumed no differences in the effect patterns between female and male teachers (i.e. measurement invariance). The unrestrictive and the restrictive model only differed in the fact that regression coefficients were estimated freely or were constrained. Consequently, the two models were nested. Log-likelihood significance tests were carried out to compare the nested models (for

Table 1. Country- and condition-specific descriptive statistics for the grade scale and unstandardized grades

	Gender					
	Female			Male		
Specialization	*n*	*M*	SD	*n*	*M*	SD
CH						
Languages	28	0.07 (4.07)	1.14 (0.92)	32	0.02 (4.03)	1.04 (0.84)
Science	27	−0.16 (3.89)	0.95 (0.77)	29	0.05 (4.05)	0.89 (0.72)
AU						
Languages	33	−0.04 (3.11)	1.09 (1.11)	32	0.11 (2.95)	1.03 (1.05)
Science	35	−0.20 (3.26)	0.92 (0.93)	37	0.13 (2.93)	0.97 (0.99)
GE						
Languages	126	−0.03 (3.32)	1.02 (1.07)	143	0.06 (3.22)	1.00 (1.05)
Science	125	0.06 (3.22)	0.96 (1.01)	133	−0.10 (3.39)	1.02 (1.06)

Notes: The grade scale results from *z*-standardization within each country and recoding so that higher values indicate higher performance. Statistics for the unstandardized grades are in parentheses. In Switzerland (CH), grades range from 6 (best) to 1 (worst); in Austria (AU), grades range from 1 (best) to 5 (worst); and in Germany (GE), grades range from 1 (best) to 6 (worst).

detailed information on the test, see UCLA: Statistical Consulting Group, 2014). In the case of no significant discrepancies in model-fit ($p \geq .20$), the more restrictive model that suggested that regression coefficients did not differ between female and male teachers was preferred. In addition, the nested models were compared using information criteria that were inspected to gauge the fit of each model to the data. Two frequently used information criteria, the sample-size adjusted Bayesian information criterion (aBIC; Sclove, 1987) and the standard Bayesian information criterion (BIC; Schwarz, 1978), were applied in this study. Lower values on these criteria indicated better model-fit.

As regards Swiss female and male teachers, both the aBIC and BIC (restrictive: 333 and 355 vs. unrestrictive: 334 and 365) and the log-likelihood significance test ($p = .35$) indicated measurement invariance allowing a joint consideration of Swiss female and male teachers. Yet, based on a sample of only 14 Swiss female teachers, gender differences in the effect pattern cannot be ruled out definitely until further research confirms this finding. Also in the Austrian sample, the aBIC and BIC (restrictive: 392 and 414 vs. unrestrictive: 394 and 426) as well as the log-likelihood significance test ($p = .40$) revealed measurement invariance across the teachers' gender. In the German sample, however, the results suggested differences in the effect patterns of female and male physics teachers. Although the aBIC and BIC (restrictive: 1,502 and 1,524 vs. unrestrictive: 1,506 and 1,538) again favored the restrictive model, the log-likelihood significance test indicated a better fit of the unrestrictive model ($p < .20$). Prompted by the outcome of the log-likelihood significance test, the German sample was split to be able to take into account even small differences in the effect patterns of female and male German teachers.

Effects of Student Gender and Teaching Experience across Countries

Based on the analysis of effects of the teachers' gender, the German sample was divided into female and male physics teachers while in Switzerland and Austria female and male teachers were considered together. Now effect patterns could be compared across countries and the two German subsamples. Three models were constructed. The most restrictive model, Model 1, suggested similar effects across all of the three countries. This model hence represented the original research hypothesis that the bias effect pattern generalizes over all of the three countries and is generally valid. Model 3, by contrast, constituted the unrestrictive model that allowed for unique effect patterns within each country and the two German subsamples. If Model 3 proved to fit the data best, the effect patterns could be considered highly context-specific. An inspection of the Model 3 regression coefficients that were estimated independently within each country and the two German subsamples (see Table 2) suggested similar effects for Swiss teachers, Austrian teachers, and German female teachers but not for German male teachers. To be able to investigate the apparently divergent effect pattern of German male physics teachers, an additional model, Model 2, was constructed that consequently suggested similar effects across all of the three countries except for German male teachers. Model 2 thus represented a

Table 2. Comparison of three multiple group regression models that predict grades based on gender, teaching experience, and the interaction between gender and teaching experience (gender × exp)

	Model 1 (CH = AU = GE)					Model 2 (CH = AU = GE females)					Model 3 (unrestrictive)				
Variable in country	*b*	SE	*p*	ICs	LL *p*	*b*	SE	*p*	ICs	LL *p*	*b*	SE	*p*	ICs	LL *p*
				2,233/2,268	–				2,229/2,273	.00				2,248/2,311	.93
CH															
Gender[a]	0.28	0.13	.03			0.77	0.18	.00			0.84	0.37	.02		
Experience	0.00	0.00	.79			0.02	0.01	.03			0.02	0.02	.23		
Gender × exp	−0.02	0.01	.02			−0.03	0.01	.00			−0.04	0.02	.04		
AU															
Gender	0.28	0.13	.03			0.77	0.18	.00			0.86	0.28	.00		
Experience	0.00	0.00	.79			0.02	0.01	.03			0.02	0.01	.04		
Gender × exp	−0.02	0.01	.02			−0.03	0.01	.00			−0.03	0.01	.01		
GE females															
Gender	0.28	0.13	.03			0.77	0.18	.00			0.64	0.32	.05		
Experience	0.00	0.00	.79			0.02	0.01	.03			0.01	0.01	.54		
Gender × exp	−0.02	0.01	.02			−0.03	0.01	.00			−0.03	0.02	.09		
GE males															
Gender	0.28	0.13	.03			−0.10	0.18	.59			−0.10	0.18	.59		
Experience	0.00	0.00	.79			−0.01	0.01	.03			−0.01	0.01	.03		
Gender × exp	−0.02	0.01	.02			0.00	0.01	.78			0.00	0.01	.78		

Notes: CH = Switzerland, AU = Austria, GE = Germany; LL *p* = *p*-values that resulted from the log-likelihood significance tests.

[a]0 = female, 1 = male.

less strict version of the expected cross-border generalizability of gender-STEM bias effects (i.e. partial generalizability). No further models were constructed because these models sufficed to examine the generalizability of the effect pattern. Accordingly, regression analyses were run with the regression coefficients constrained to be equal across all of the three countries (Model 1), across all of the three countries with the exception of the German male physics teachers that were freed (Model 2), and with all of the coefficients estimated freely within each country and the two German subsamples (Model 3).

After the three models were estimated, their fit to the data was contrasted to find the model which best described the effect patterns across the countries including the two German subsamples. Hence, Model 1, Model 2, and Model 3 were compared, again using the aBIC and BIC as well as log-likelihood significance tests. The most restrictive Model 1 was accordingly compared to the less restrictive Model 2 as an alternative model. In a second step, Model 2 was compared to the unrestrictive Model 3 as an alternative model. In the case of no significant discrepancies in model-fit, always the more restrictive, more parsimonious model was chosen. In the case of significant differences, however, the more restrictive model fitted the data significantly worse than the less restrictive model and consequently the less restrictive model was chosen. The results of the multiple group regression analyses and the model comparisons are summarized in Table 2.

In regard to the log-likelihood significance tests, Model 2 fitted the data significantly better than the most restrictive Model 1 ($p < .01$). The least restrictive Model 3 did not fit the data significantly better than the more restrictive Model 2 ($p = .93$). The aBIC further indicated the superiority of Model 2. Although the BIC favored the most restrictive Model 1, the BIC of Model 2 only slightly exceeded the value calculated for Model 1. Hence, Model 2, which suggested similar effect patterns across all of the three countries with the exception of the German male physics teachers, turned out to best describe the effect patterns across countries and the two German subsamples and is interpreted in what follows.

According to Model 2 (see Table 2), the analysis revealed both a significant main effect of gender ($b_{gender} = 0.77$) and a clear moderating effect of teaching experience on the relationship between gender and grades ($b_{gender \times exp} = -0.03$) in the samples of Swiss, Austrian, and female German teachers. The gender effect that was reflected in an advantage of approximately 0.77 standard deviations on the grade scale for the fictive boy thus represented teachers' gender bias at the beginning of their career (without teaching experience). The negative interaction between gender and teaching experience indicated that the initial gender bias decreased with increasing years of teaching experience. The additional significant main effect of the continuous variable teaching experience ($b_{exp} = 0.02$) suggested that the fictive girl's grades improved by approximately 0.02 standard deviations per year of teaching experience. In the German male sample, only teaching experience ($b_{exp} = -0.01$) significantly influenced grading. Accordingly, with growing teaching experience, lower grades were awarded. While all of the other teachers showed a consistent bias pattern, the gender-neutral grading behavior of the German male teachers was exceptional. In the following

analyses and figures that were aimed at gaining further information on gender bias effects as a function of teaching experience, I hence focused on all Swiss, all Austrian, and the German female teachers. Yet, it is important to always keep in mind that all that is reported in the following does not apply to the whole teacher sample. The reported gender bias effects were not generally valid and showed only partial cross-border generalizability with the German male physics teachers demonstrating a divergent pattern of effects.

In Figure 2, an interaction plot for the equated samples of all Swiss, all Austrian, and the German female teachers based on Model 2 is depicted. To illustrate the moderating effect of teaching experience, grades were regressed on student gender and the mean teaching experience ($M = 17.76$ years, SD = 11.41 years) minus or plus one standard deviation.

Effects of Student Gender and Teaching Experience in the Binned Empirical Data

In keeping with Model 2, *t*-tests were applied on the data from all Swiss, all Austrian, and the German female physics teachers ($n = 378$). The comparisons between the

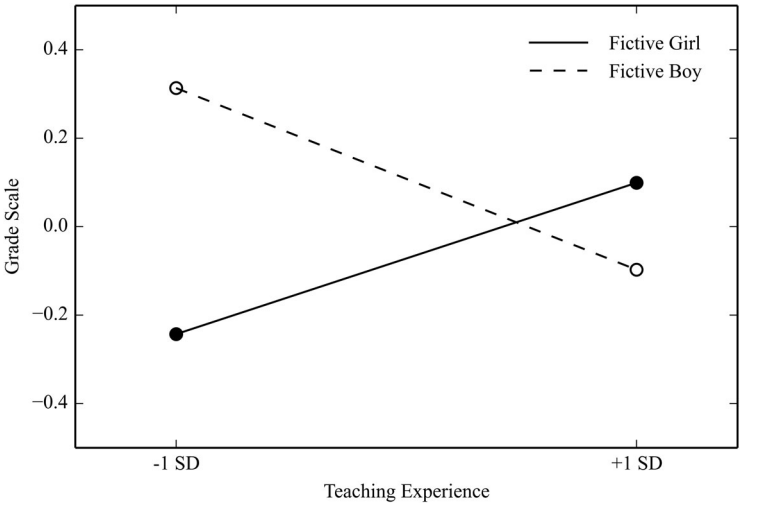

Figure 2. Interaction plot based on the equated samples of all Swiss, all Austrian, and the German female teachers. Teaching experience is depicted on the *x*-axis. The left tag on the *x*-axis marks the position on the teaching experience scale that is one standard deviation below the mean of teaching experience (approximately 6 years of teaching experience). The right tag on the *x*-axis marks the position on the teaching experience scale that is one standard deviation above the mean of teaching experience (approximately 29 years of teaching experience). The *y*-axis depicts the grades measured on the *z*-standardized grade scale that are predicted according to the regression in Model 2. The continuous line represents the grades that are predicted for the fictive girl (i.e. when student gender is set at female). The dashed line represents the grades that are predicted for the fictive boy (i.e. when student gender is set at male). The circles denote the grades that are predicted for the fictive female or male student when teaching experience is one standard deviation below or above the mean

mean grades that were awarded to a fictive female vs. male student in each of the nine 5-year bins of teaching experience revealed that after approximately 10 years of teaching experience, the gender-specific discrepancy in the grades was not significant any more. Hence, the mean grade difference was $M_{\Delta} = 0.87$ ($t(51) = 3.61$, $p < .001$) in the first bin and $M_{\Delta} = 0.67$ ($t(54) = 2.40$, $p < .05$) in the second bin, compared to $M_{\Delta} = 0.15$ ($t(59) = 0.65$, $p = .52$) in the third bin (in all of the six other bins, all ps $\geq .34$). The problem of multiple testing (i.e. the nine t-tests) was considered negligible here taking into account the severity of the problem of even small bias effects in grading. Expressed in the country-specific unstandardized grade scales, a difference of $M_{\Delta} = 0.87$ on the z-standardized grade scale corresponded to about 0.7 Swiss grades, to about 0.9 Austrian grades, and to about 0.9 German grades. Figure 3 visualizes the relationship between teaching experience and grading based on both individual data points and binned data (i.e. the interpolation line) as a function of the fictive student's

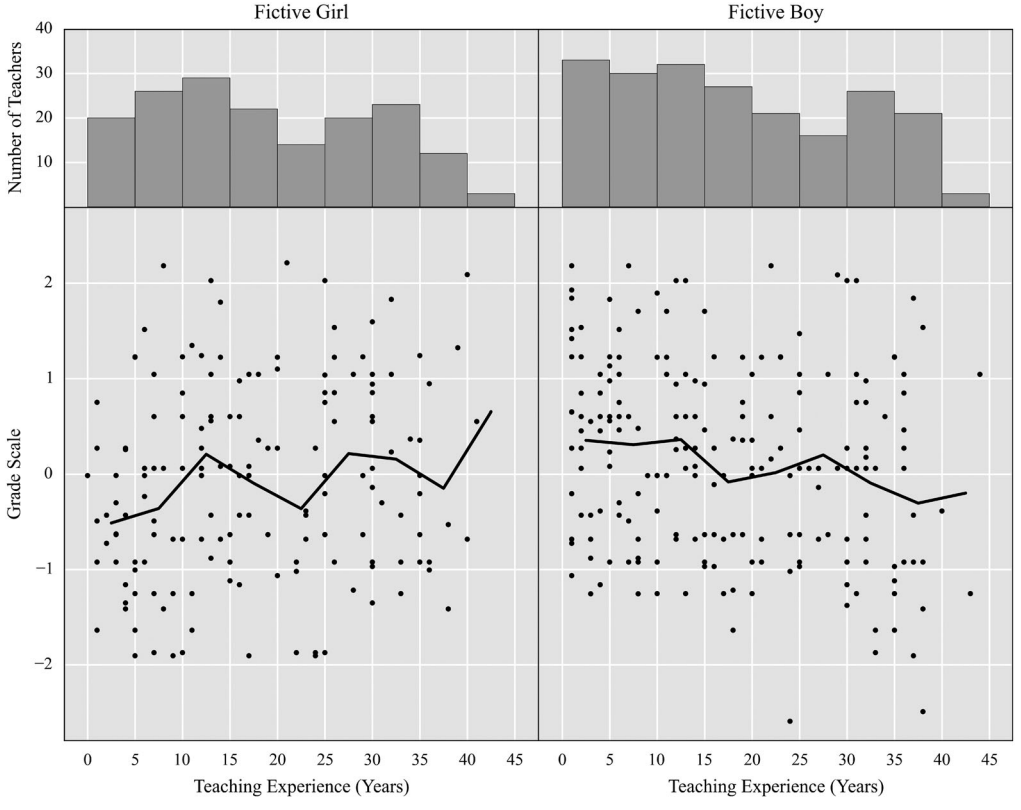

Figure 3. Scatter plots showing the relationship between teaching experience and grades for a fictive girl on the left side and a fictive boy on the right side, based on the samples of Swiss, Austrian, and German female physics teachers. The interpolation line connects the mean grades that were calculated within bins of 5 years of teaching experience. The histograms in the upper part of the figure visualize the number of teachers within each teaching experience bin

gender. Importantly, Figure 3 additionally provides information about the number of teachers within each bin of teaching experience in the form of histograms.

To conclude, the findings of this additional analysis suggested an interpretation of the interaction between gender and teaching experience in the sense that teaching experience removed the strong initial bias against girls but did not reverse it.

Discussion

Gender Bias and Teaching Experience: Effects Exist, but not for all Teachers

This experimental online-study investigated gender bias and the role of teaching experience in grading a fictive student's answer to a conceptual test question in secondary school physics. Contrary to prior expectations, the overall sample displayed no generally valid pattern of bias effects. This finding indicates that bias effects do not generalize easily across contexts. Nevertheless, this study revealed the existence of a consistent cross-border pattern of gender bias effects that applies to Swiss, Austrian, and female German teachers. Overall, the inspection of the data regarding this international group of teachers suggested a consistent clear gap between girls' and boys' grades that was significant for teachers with up to about 10 years of teaching experience and disappeared with increasing teaching experience. Yet, unexpectedly, German male physics teachers showed no gender bias effects at all. In the German male sample, performance of both female and male students was rated lower with years of teaching practice. The teacher samples from the three countries were highly comparable in terms of age and years of teaching experience. In general, the training program of German physics teachers and the gender distribution also closely resemble the situation in Austria (while the training program of Swiss teachers slightly differs from the other two countries and the gender distribution shows a more extreme preponderance of male teachers). There were no distortions in the sampling of the German male sample regarding teaching experience or age. In trying to find an explanation for the divergent pattern of the German male teachers, differences between the German male sample and the other samples were examined regarding the proportion of teachers teaching at rural vs. urban schools and regarding the time spent with the survey. However, also these analyses revealed no irregularities. To additionally factor in the considerably larger sample size of the German male physics teachers, random subsamples including approximately 25% (i.e. $n \sim 100$) of the overall German male teachers sample were drawn and analyzed. In none of the five subsamples analyzed gender bias effects emerged, suggesting that the pattern of bias effects in the samples of all Swiss, all Austrian, and the German female teachers was not merely an effect of distorted samples due to small sample sizes. Further research is required in order to detail the specifics of German male physics teachers that might explain their differing, gender-neutral grading behavior. It remains to be investigated how the patterns of female and male physics teachers in other countries compare with the two patterns revealed in this study, searching for regularities. Such research may help to understand in which contexts gender bias effects in

physics grading can be expected and when they do not appear, providing important information for remediating interventions.

Focusing now on the pattern of bias effects found in the samples of all Swiss, all Austrian, and the German female teachers, the moderating effect of teaching experience may partially explain the heterogeneity of existing findings on gender bias, where characteristics of the raters or judges were not taken into consideration (see Swim et al., 1989). Thus, the rater's experiences, with regard to the context of the judgment task, can play an important role in determining to what extent or whether or not a gender bias may arise. Future research that also closely examines the rater is needed to elucidate the process that underlies bias changing with experience (see Kunda & Spencer, 2003; Kunda & Thagard, 1996). On the one hand, there is good reason to assume that the need to invoke stereotypes decreases to the extent that the perceived ambiguity of information and a high demand for cognitive resources in the judgment situation diminish with increasing experience. On the other hand, experience could also reduce gender bias by changing the stereotype itself via repeated exposure to individuals (e.g. girls who are good at physics) that challenge formerly held beliefs (see Glock & Krolak-Schwerdt, 2013; Koenig & Eagly, 2014; Miller et al., 2014).

Student specialization in languages vs. science was not systematically considered in the grading process. This finding may indicate that student specialization, as compared to student gender, does not activate a shared social category that is used in the grading process examined in this study. In contrast to specialization, the fictive student's gender, however, seems to serve as a cue, activating cognitive structures that systematically influence some teachers' decision making during grading.

Relevance for Physics Classrooms

The study investigated teachers' evaluations of a student's answer to one conceptual test question. The distinct gender bias effects found for all Swiss, all Austrian, and the German female teachers who have been teaching for less than 10 years underpin the importance of straightforward assessment criteria—which mimic the more elaborated cognitive schemata of experienced teachers—whenever student performance is evaluated and especially when ill-defined conceptual problems are to be judged. By reducing the perceived ambiguity and cognitive overload of beginning teachers in the judgment situation, the need to draw on stereotypes and the resulting biases may be avoided. The use of standardized and maybe computerized testing could alleviate the problem of biased assessment. Standardized tests that can be evaluated by simply adding up correct answers enable highly objective assessment (e.g. the Test of basic Mechanics Conceptual Understanding; Hofer, Schumacher, & Rubin, 2015). Computerized procedures automatically evaluate a student's performance with respect to standards or performance criteria defined by the curriculum. Up to now, such standardized tests and computerized procedures are almost entirely restricted to test questions with predefined solution alternatives (e.g. multiple or single choice questions). This kind of narrow testing cannot capture all aspects of

physics literacy. To be able to assess creative thinking and flexible problem solving in physics, for instance, it may be adequate to ask questions that allow a wide variety of different solutions. Consequently, standardized procedures should be developed, evaluated, disseminated, and applied by practitioners whenever appropriate. In written assessments that require the problems in the test to be less restrictive, however, other strategies are available that could tackle the problem of gender biased grading. So students could use anonymized identification numbers and write their tests on the computer or comparable devices in case that the handwriting can reveal the identity of the particular student. Cooperation between teachers in the context of assessment may also reduce the risk of biased grading. Teachers could ask a colleague to evaluate the anonymized tests of their students according to clearly defined evaluation criteria (and vice versa).

The problem of gender-biased evaluations may not only be addressed by modifying the process of assessment itself but also by rethinking physics instruction on a more general level, as explained in what follows. The present study points to the possibility of more general effects of gender-STEM stereotypes on the teaching process. A student's answer to a conceptual question that could also be regarded as a good proxy for a student's oral classroom contributions was used as the judgment situation. The answer represented average student performance and was neither completely wrong nor absolutely correct, leaving room for interpretation. The student answer could hence be considered to reflect an intermediate state in the student's knowledge development process with some correct elements and some elements that still needed to be restructured or even abandoned. A bias favoring boys (or penalizing girls, respectively) may correspond to different ways of interpreting the student's answer. Hence, such an answer originating from a boy may be interpreted as indicating a promising step on the way to full understanding (in the sense of a benefit of the doubt) resulting in supportive instructional actions. On the contrary, such an answer originating from a girl may be interpreted as indicating profound misconceptions being hard to overcome resulting in teachers resigning and putting less effort into supportive instructional actions. To conclude, instruction that is explicitly designed to support all students' active knowledge construction could be expected to particularly help girls by providing support that might otherwise be less available. The cognitively activating physics instruction described by Hofer, Schumacher, Rubin, and Stern (2015) or interactive engagement methods (e.g. frequent feedback and group discussions) described by Lorenzo, Crouch, and Mazur (2006) are just two examples of the many instructional approaches that proved to be beneficial for all students but particularly for female students. Innovative instructional approaches could directly support female students but also contribute to reducing the teachers' gender-STEM stereotypes. To tackle biases in the evaluation of oral examinations or the students' daily classroom contributions requires teachers to revise their cognitive schemata concerning females and physics. If instruction allows and encourages female students to more strongly engage in physics, formerly held beliefs about girls and physics may slowly become untenable.

In addition to the promotion of promising instructional approaches, the present findings suggest that teacher education and teacher supervision should also focus more strongly on supporting beginning teachers in monitoring their (socio-)cognitive processes when student achievement is evaluated. Literature and information on bias effects may be provided. Teacher educators could communicate the importance of applying straightforward assessment criteria that structure the process of evaluation and discuss strategies like anonymized testing and the advantages and disadvantages of standardized tests and computerized procedures.

In real classroom situations teachers get to know their students after a while, and this knowledge base may at least reduce the application of stereotypes (see Kunda & Spencer, 2003). Nevertheless, a teacher's evaluations and grading at the beginning of the school year, which resemble the situation that was implemented in this study (i.e. little personal information), may lead to long-lasting self-fulfilling prophecy (e.g. De Boer, Bosker, & van der Werf, 2010; Jussim & Eccles, 1992) or stereotype threat effects (e.g. Marchand & Taasoobshirazi, 2013; Nguyen & Ryan, 2008).

Limitations and Suggestions for Future Research

In all of the three countries, the correlation between the teachers' years of teaching experience and the teachers' age was very high ($.86 \leq r \leq .90$). Consequently, with the cross-sectional design used in this study, it was difficult to determine whether teaching experience or the different socialization of the age cohorts influenced gender bias in grading. In trying to nevertheless estimate the relative impact of experience vs. age, an additional regression analysis was conducted for those Swiss, Austrian, and female German teachers with below average age and above average teaching experience ($n = 18$) and those teachers with above average age and below average teaching experience ($n = 30$). Albeit this analysis was based on very small sample sizes and the coefficients were not significant, a negative coefficient of the main effect of gender in the group of the younger but more experienced teachers and a positive coefficient in the group of the older but less experienced teachers suggested that not age but teaching experience determined the change in gender bias for Swiss, Austrian, and female German teachers. This conclusion, however, has to be underpinned by future research.

In this study, the application of gender-STEM stereotypes was not explicitly examined but deduced from the teachers' evaluation behavior and theoretical assumptions, since this study primarily aimed at describing physics teachers' gender bias in grading as a function of teaching experience. Now that there is evidence that gender bias effects in fact have to be considered in physics grading, further studies could add specific and detailed measures of (implicit) stereotype activation and application (see, e.g. Glock & Kovacs, 2013; Nosek et al., 2009) that also allow for a differentiation between general gender-STEM stereotypes, on the one hand, and more specific gender-physics stereotypes, on the other hand. Closely related, the domain specificity of the observed effect patterns was not addressed in this study. The generalizability of the observed patterns to other STEM fields (e.g. chemistry or mathematics) still has to be investigated.

To measure the teachers' evaluation of student performance, teachers had to grade a student's answer to a single test question. This judgment situation was considered appropriate to investigate whether gender bias effects in physics grading existed in Switzerland, Austria, and Germany. However, it did not perfectly match real grading situations. The present results consequently provide evidence that gender bias effects in physics teachers' evaluations exist, but it is not clear whether these effects would still be present in the evaluation of a whole test. More information about a student's performance that is provided by a number of answers to different test questions could reduce the need to draw on stereotypes. Nevertheless, also in real grading situations, each answer to a test question should be evaluated separately. So the information that is available in each judgment situation is a single answer to a particular test question, just as in this study. Even if the impression formed on the basis of one answer is used to inform the evaluation of other answers, biases that influenced the process of impression formation in the first place might be passed on, too. Although the judgment situation implemented in this study can hence be considered an appropriate probe to investigate bias effects, future work that is built upon the present results may apply more comprehensive instruments to assess the teachers' evaluation of student performance. In a different setting that involves direct contact to teachers, teachers could evaluate extensive materials that resemble standard physics exams. Studies may examine how the type of test question (e.g. conceptual questions vs. computational questions) influences gender bias effects. In the test question used in this study, Newtonian mechanics was applied to a problem involving the movement of two skateboarders. Further studies could investigate the impact of the specific problem context on gender bias effects. Maybe gender-STEM stereotypes would not be activated in contexts that are commonly perceived as gender-neutral or even female contexts (e.g. problems of health physics).

The teachers' familiarity with the particular physics problem used in this study, the 'skateboarder question', may be an alternative to teaching experience to explain the observed bias pattern. How often the teachers in the sample have come across a physics problem similar to the 'skateboarder question' can be expected to depend on the length of their professional experience. Familiarity with the problem might have helped teachers to interpret and evaluate the student answer by comparing it to mental representations of answers that different students have provided over the years. It can be argued, however, that familiarity is equivalent to the more and more efficient structuring of a physics problem's cognitive schema which is expected to proceed with teaching experience. Familiarity with single, frequently met problems (like the 'skateboarder question') could hence be expected to inevitably accompany growing teaching practice. Future studies could implement different and less familiar judgment situations to investigate familiarity with the test question as alternative explanation.

Conclusion

In the first decade of Swiss, Austrian, and female German physics teachers' careers, grading is affected by a gender bias that is in line with the common gender-STEM

stereotypes. Gender bias disappears with increasing years of teaching practice. German male teachers, by contrast, display gender-neutral grading behavior. It remains to be clarified why this group of teachers behaves differently.

Despite the only partial generalizability of gender bias effects, even today gender bias in grading seems to represent a real problem in at least some physics classes. Since gender bias effects in grading should not appear at all, this finding has to be taken seriously. Ultimately, some girls' underperformance in physics may be an inevitable consequence of the social learning environment, while the existence of gender-STEM stereotypes may be an inevitable consequence of the girls' underperformance. We can try to break this vicious circle by implementing standardized tests and assessment techniques that allow anonymous evaluation, by sensitizing student teachers and novice physics teachers to the problem of gender bias in grading, by emphasizing straightforward criteria to assess student performance, or by modifying physics instruction to tackle gender-STEM stereotypes on a more general level. Such efforts have the potential to alleviate the gender gap in physics.

Acknowledgements

I wish to thank David Wintgens, Andreas Vaterlaus, Martin Hopf, Knut Neumann, Karsten Reckleben, Gerhard Röhner, Franz Kranzinger, and Silke Eckstein for their help in contacting physics teachers in Switzerland, Austria, and Germany; Elsbeth Stern, Andreas Lichtenberger, Clemens Wagner, Bahar Behzadi, André van der Graaff, Rolf Strassfeld, and Sebastian Seehars for their support; and, in particular, all of the physics teachers for their participation. I would also like to thank Bruno Rütsche and Peter Edelsbrunner for their help with technical and software issues.

Disclosure statement

No potential conflict of interest was reported by the author.

References

Babad, E. Y. (1985). Some correlates of teachers' expectancy bias. *American Educational Research Journal*, *22*(2), 175–183. doi:10.3102/00028312022002175

Baird, J. (1998). What's in a name? Experiments with blind marking in A-level examinations. *Educational Research*, *40*(2), 191–202. doi:1080/0013188980400207

Bartlett, S. F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge: Cambridge University Press.

Berliner, D. C. (2001). Learning about and learning from expert teachers. *International Journal of Educational Research*, *35*(5), 463–482. doi:10.1080/0013188980400207

Brewer, M. B. (1988). A dual process model of impression formation. In T. K. Srull & R. S. Wyer, Jr. (Eds.), *Advances in social cognition* (pp. 1–36). Hillsdale, NJ: Lawrence Erlbaum Associates.

Carter, K., Sabers, D., Cushing, K., Pinnegar, S., & Berliner, D. C. (1987). Processing and using information about students: A study of expert, novice, and postulant teachers. *Teaching and Teacher Education*, *3*(2), 147–157. doi:10.1016/S0883-0355(02)00004-6

Ceci, S. J., Williams, W. M., & Barnett, S. M. (2009). Women's underrepresentation in science: Sociocultural and biological considerations. *Psychological Bulletin*, *135*(2), 218–261. doi:10.1037/a0014412 218

Chaiken, S., & Maheswaran, D. (1994). Heuristic processing can bias systematic processing: Effects of source credibility, argument ambiguity, and task importance on attitude judgment. *Journal of Personality and Social Psychology*, *66*(3), 460–473.

Coladarci, T. (1986). Accuracy of teacher judgments of student responses to standardized test items. *Journal of Educational Psychology*, *78*(2), 141–146.

De Boer, H., Bosker, R. J., & van der Werf, M. P. C. (2010). Sustainability of teacher expectation bias effects on long-term student performance. *Journal of Educational Psychology*, *102*(1), 168–179. doi:10.1037/a0017289

Destatis. (2013). *Prüfungen an Hochschulen* [Exams at universities]. Wiesbaden: Federal Statistical Office.

Dünnebier, K., Gräsel, C., & Krolak-Schwerdt, S. (2009). Urteilsverzerrungen in der schulischen Leistungsbeurteilung: Eine experimentelle Studie zu Ankereffekten. *Zeitschrift für Pädagogische Psychologie*, *23*(34), 187–195. doi:10.1024/1010-0652.23.34.187

Eagly, A. H., & Koenig, A. M. (2008). Gender prejudice: On the risks of occupying incongruent roles. In E. Borgida & S. T. Fiske (Eds.), *Beyond common sense: Psychological science in the courtroom* (pp. 63–81). Malden, MA: Blackwell Publishing.

Eagly, A. H., & Wood, W. (2013). The nature–nurture debates: 25 years of challenges in understanding the psychology of gender. *Perspectives on Psychological Science*, *8*(3), 340–357. doi:10.1177/1745691613484767

Eagly, A. H., Wood, W., & Diekman, A. B. (2000). Social role theory of sex differences and similarities: A current appraisal. In T. Eckes & H. M. Trautner (Eds.), *The developmental social psychology of gender* (pp. 123–174). Mahwah, NJ: Lawrence Erlbaum Associates.

ETH Zurich Annual Report. (2013). Retrieved from https://www.ethz.ch/content/dam/ethz/common/docs/publications/annual-reports/2013/ETH_Annualreport_2013_updated.pdf

European Commission. (2013). *She figures 2012: Gender in research and innovation*. Luxembourg: Publications Office of the European Union. Retrieved from http://ec.europa.eu/research/science-society/document_library/pdf_06/she-figures-2012_en.pdf

Farenga, S. J., & Joyce, B. A. (1999). Intentions of young students to enroll in science courses in the future: An examination of gender differences. *Science Education*, *83*(1), 55–75.

Fiske, A. P., Kitayama, S., Markus, H. R., & Nisbett, R. E. (1998). The cultural matrix of social psychology. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology, Vols. 1 and 2* (4th ed.) (pp. 915–981). New York, NY: McGraw-Hill.

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.

Glock, S., & Kovacs, C. (2013). Educational psychology: Using insights from implicit attitude measures. *Educational Psychology Review*, *25*(4), 503–522. doi:10.1007/s10648-013-9241-3

Glock, S., & Krolak-Schwerdt, S. (2013). Does nationality matter? The impact of stereotypical expectations on student teachers' judgments. *Social Psychology of Education*, *16*(1), 111–127. doi:0.1007/s11218-012-9197-z

Glock, S., & Krolak-Schwerdt, S. (2014). Stereotype activation versus application: How teachers process and judge information about students from ethnic minorities and with low socioeconomic background. *Social Psychology of Education*, *17*(4), 589–607. doi:10.1007/s11218-014-9266-6

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

Goddard Spear, M. (1984b). The biasing influence of pupil sex in a science marking exercise. *Research in Science & Technological Education*, *2*(1), 55–60. doi:10.1080/0263514840020107

Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., & Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review*, *109*(1), 3–25. doi:10.1037//0033-295X.109.1.3

Heller, K. A., Finsterwald, M., & Ziegler, A. (2010). Implicit theories of mathematics and physics teachers on gender-specific giftedness and motivation. In K. A. Heller (Ed.), *Munich studies of giftedness* (pp. 239–252). Berlin: LIT.

Hofer, S. I., Schumacher, R., & Rubin, H. (2015). *The Test of basic Mechanics Conceptual Understanding (bMCU): Using Rasch analysis to develop and evaluate an efficient multiple-choice test on Newton's mechanics.* Manuscript in preparation.

Hofer, S. I., Schumacher, R., Rubin, H., & Stern, E. (2015). *Fostering physics learning with cognitively activating instruction: A classroom intervention study*. Manuscript in preparation.

Hopkins, K. D., George, C. A., & Williams, D. D. (1985). The concurrent validity of standardized achievement tests by content area using teachers' ratings as criteria. *Journal of Educational Measurement*, *22*(3), 177–82.

Jussim, L., & Eccles, J. S. (1992). Teacher expectations II: Construction and reflection of student achievement. *Journal of Personality and Social Psychology*, *63*(6), 947–961.

Jussim, L., & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review*, *9*(2), 131–155.

Kessels, U., Rau, M., & Hannover, B. (2006). What goes well with physics? Measuring and altering the image of science. *British Journal of Educational Psychology*, *76*(4), 761–780. doi:10.1348/000709905X59961

Koenig, A. M., & Eagly, A. H. (2014). Evidence for the social role theory of stereotype content: Observations of groups' roles shape stereotypes. *Journal of Personality and Social Psychology*, *107*(3), 371–392. doi:10.1037/a0037215

Krolak-Schwerdt, S., Böhmer, M., & Gräsel, C. (2009). Verarbeitung von schülerbezogener Information als zielgeleiteter Prozess: Der Lehrer als «flexibler Denker». *Zeitschrift für Pädagogische Psychologie*, *23*(34), 175–186. doi:10.1024/1010-0652.23.34.175

Krolak-Schwerdt, S., Böhmer, M., & Gräsel, C. (2012). Leistungsbeurteilungen von Schulkindern: Welche Rolle spielen Ziele und Expertise der Lehrkraft? *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *44*(3), 111–122. doi:10.1026/0049-8637/a000062

Kunda, Z., & Spencer, S. J. (2003). When do stereotypes come to mind and when do they color judgment? A goal-based theoretical framework for stereotype activation and application. *Psychological Bulletin*, *129*(4), 522–544. doi:10.1037/0033-2909.129.4.522

Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological Review*, *103*(2), 284–308.

Leiner, D. J. (2014). *SoSciSurvey—oFb—der onlineFragebogen*. Retrieved from https://www.soscisurvey.de

Leinhardt, G., & Greeno, J. G. (1986). The cognitive skill of teaching. *Journal of Educational Psychology*, *78*(2), 75–95. doi:10.1037/0022-0663.78.2.75

Lorenzo, M., Crouch, C. H., & Mazur, E. (2006). Reducing the gender gap in the physics classroom. *American Journal of Physics*, *74*(2), 118–122. doi:10.1119/1.2162549

Madon, S., Jussim, L., Keiper, S., Eccles, J., Smith, A., & Palumbo, P. (1998). The accuracy and power of sex, social class, and ethnic stereotypes: A naturalistic study in person perception. *Personality and Social Psychology Bulletin*, *24*(12), 1304–1318. doi:10.1177/01461672982412005

Marchand, G. C., & Taasoobshirazi, G. (2013). Stereotype threat and women's performance in physics. *International Journal of Science Education*, *35*(18), 3050–3061. doi:10.1080/09500693.2012.683461

Miller, D. I., Eagly, A. H., & Linn, M. C. (2014). Women's representation in science predicts national gender-science stereotypes: Evidence from 66 nations. *Journal of Educational Psychology*, No Pagination Specified. doi:10.1037/edu0000005

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, *109*(41), 16474–16479.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide (*7th ed.*)*. Los Angeles, CA: Muthén & Muthén. Retrieved from http://www.statmodel.com/download/usersguide/Mplus%20user%20guide%20Ver_7_r3_web.pdf

Nguyen, H.-H. D., & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology*, *93*(6), 1314–1334.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Math = male, me = female, therefore math ≠ me. *Journal of Personality and Social Psychology*, *83*(1), 44–59. doi:10.1037/0022-3514.83.1.44

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., … Banaji, M. R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, *18*(1), 36–88. doi:10.1080/10463280701489053

Nosek, B. A., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, A., … Greenwald, A. G. (2009). National differences in gender–science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences*, *106*(26), 10593–10597. doi:10.1073/pnas.0809921106

Organisation for Economic Co-operation and Development. (2011). *PISA 2009 results: What students know and can do: Student performance in reading, mathematics and science* (Vol. 1). Paris: OECD Publications.

Palmer, D. J., Stough, L. M., Burdenski, Jr., T. K., & Gonzales, M. (2005). Identifying teacher expertise: An examination of researchers' decision making. *Educational Psychologist*, *40*(1), 13–25. doi:10.1207/s15326985ep4001_2

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

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, *52*(3), 333–343.

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

Swim, J. K., Borgida, E., Maruyama, G., & Myers, D. G. (1989). Joan McKay versus John McKay: Do gender stereotypes bias evaluations? *Psychological Bulletin*, *105*(3), 409–429.

UCLA: Statistical Consulting Group. (2014, July 18). *Mplus FAQ. How can I compute a chi-square test for nested models with the MLR or MLM estimators*? Retrieved July 18, 2014, from http://www.ats.ucla.edu/stat/mplus/faq/s_b_chi2.htm

Walton, G. M., & Spencer, S. J. (2009). Latent ability: Grades and test scores systematically underestimate the intellectual ability of negatively stereotyped students. *Psychological Science*, *20*(9), 1132–1139. doi:10.1111/j.1467-9280.2009.02417.x