European Journal of Psychology of Education (2019) 34:825–846
https://doi.org/10.1007/s10212-018-0405-x

Preservice teachers' evaluations of students' achievement development in the context of school-track recommendations

Florian Klapproth[1] · Birthe Doreen Fischer[1]

Received: 19 October 2016 / Revised: 31 July 2018 / Accepted: 13 September 2018 /
Published online: 3 October 2018

Abstract
We experimentally examined whether preservice teachers considered both the gender of primary school students and their development of GPAs, indicated by two successive school reports, when making secondary school-track recommendations. We used student vignettes to mimic real students and orthogonally varied student gender, their GPA development (positive vs. negative), and the grand mean of grades. We found that students who improved were more than twice as likely to be recommended for the highest track than those who deteriorated. Moreover, we could show that with deteriorating students, male participants preferred boys relative to girls. These results are discussed regarding possible rules of extrapolating current achievements and concerning the gender of students and participants, and conclusions are drawn with respect to preservice teacher education.

Keywords School placement recommendation · Development of achievement · Preservice teachers · Student gender · Participant gender · Experiment

Introduction

How do preservice teachers evaluate the improvement in students' achievement, and how the deterioration in achievement? Is there a gender bias in the evaluation of student performance, with girls preferred to boys? The study at hand addresses these questions in the context of school placement recommendations that are made in Germany (and some other European countries) at the end of primary school in order to suggest a student for one of the different school tracks in secondary school.

Secondary school in Germany is hierarchically structured, consisting of higher and lower tracks, where "high" and "low" correspond to both the necessary achievement level students should have to be eligible for being taught in a specific track, and the level of instruction that is delivered in the tracks. When students finish primary school, which is—dependent on the federal

✉ Florian Klapproth
florian.klapproth@medicalschool-berlin.de

[1] Medical School Berlin, Calandrellistrasse 1-9, 12247 Berlin, Germany

state—at the end of school year 4 (e.g., in Bavaria) or school year 6 (e.g., in Berlin), teachers decide on the future track students should attend in secondary school. Traditionally, three distinctive tracks constitute the German secondary school. The lowest track ("Hauptschule") is dedicated to students with major learning difficulties and below-average achievement profiles. Students who attend this track can acquire qualifications for rather limited vocational areas. The intermediate track ("Realschule") provides students with general education and vocational-training courses. The highest track ("Gymnasium") offers students with above-average achievement profiles the qualification for university entrance when the students successfully accomplished this track.

According to the German transition regulations, teachers should value the student's achievements as the major factor when making a school placement recommendation (KMK 2010). The students' achievements that are considered for placement recommendations are mainly represented by their school grades, given in the last but one school report in primary school. Moreover, working habits and social behavior, related to achievement and mentioned in the school report, do also play a role for track recommendations. School grades in Germany vary between 1 and 6, with 1 meaning "very good" and 6 meaning "insufficient," that is, lower scores on the grade scale represent higher achievements. School grades are based on predetermined education standards representing the knowledge and skills that should be mastered at each stage of the school system. Whereas school grades seem to be criterion-referenced at first glance, they are actually affected by the school or even the class where the teacher examines the students. For example, the meaning of "very good" performance might vary between classes of different achievement levels (Phillips 1991). Hence, the likelihood for students to get lower grades increases with the ability level of the class (Preckel et al. 2008; Zeidner and Schleyer 1999). School grades are therefore not directly comparable between schools.

In Berlin, which is both a federal state and the capital of Germany, teachers are given two school reports instead of one to reach a decision about a school-track recommendation. In Berlin, where secondary school starts in grade 7, the school reports serving as the decision base are from the last semester of school year 5 (5/2) and the first semester of school year 6 (6/1). The teachers are advised to take both school reports into consideration when opting for one of the different tracks. If the grand mean of the grades of both school reports equals 2.2 or less, the recommendation for the highest track is advised. If the grand mean is 2.8 or larger, students are recommended for the lower track. If, however, the grand mean is between 2.2 and 2.8, additional information regarding the student's skills and achievements (e.g., her or his reflection of the learning process) should be gathered (Senatsverwaltung für Bildung, Jugend und Wissenschaft 2015). One possible advantage of using two reports rather than one might be that they provide a more stable and reliable estimate of the "true" achievements of a student. However, since it is likely that both school reports differ from one another in their grade point averages (GPA), teachers are faced with both a grand mean across the grades obtained in semester 5/2 and semester 6/1 and the difference between the GPAs of the two school reports. According to official regulations and school policies, only the grand mean should be used for placement decisions (Senatsverwaltung für Bildung, Jugend und Wissenschaft 2015). However, information about how to handle changes between both reports is not given.

Factors affecting school-track recommendations

Although it is clearly stated in official regulations that achievement be the only factor that should affect the teachers' track recommendations (KMK 2010), factors that are not related to achievement have also been shown to affect track recommendations. Recent research has brought ample evidence that the main predictors of school placement decisions are the school grades (Bos et al. 2004; Stubbe and Bos 2008), scores from standardized achievement tests (Bos et al. 2004; Stahl

2007), achievement motivation (Bos et al. 2004; Neugebauer 2011), the socioeconomic status of the students (Stubbe 2009), the immigration status of the students (Kristen 2006), parental aspirations (Braun and Mehringer 2009), the composite achievement level of the school class (Trautwein and Baeriswyl 2007), and students' gender (Arnold et al. 2007).

Gender

Several studies show that girls were more likely than boys to be recommended for the highest track in German secondary school (Arnold et al. 2007; Jürges and Schneider 2011; Lehmann and Peek 1997; Milek et al. 2009; Schmitt 2008; Schneider 2011; Schulze et al. 2009). This effect suggests that boys are, compared to girls, disadvantaged in the German school system. Although the predictive power of gender for school-track recommendations dropped when the students' achievement was controlled for, the effect of gender was still significant in some studies (Arnold et al. 2007; Jürges and Schneider 2011; Lehmann and Peek 1997; Schulze et al. 2009). An explanation as to why teachers were prone to recommend girls more often to the highest track than boys despite equal achievements might be based on different expectations in regard to the students' cognitive and personal development in secondary school. These expectations might result from teachers' knowledge about students' maturation and its effect on school performance. Boys lag behind girls by 2 years in physical maturation (Marshall and Tanner 1970). There is also evidence that girls are emotionally more mature than boys (Carrothers et al. 2000) and outperform boys in regard to self-efficacy (Kumar and Lal 2006). With respect to intelligence, boys and girls mature at different rates. Girls mature on average faster at the age of about 9 years and remain in advance of boys until the age of 14 or 15 years (Colom and Lynn 2004; Lynn 1999). Similarly, girls have been shown to mature earlier than boys in personality. For instance, girls exhibit higher levels of agreeableness (Branje et al. 2007; Klimstra et al. 2009) and conscientiousness (Klimstra et al. 2009) in early and middle adolescence. Both personality traits contribute to positive interpersonal relationships, which in turn teachers regard as valuable for learning and classroom behavior (e.g., Lane et al. 2004).

Teacher expectations might also refer to gender-related differences in student attitudes to school (e.g., Chesterfield and Enge 1998; Parks and Kennedy 2007), which may also result from girls' earlier maturation (Darom and Rich 1988). For instance, girls show on average a more positive attitude to school than boys (OECD 2004), they usually enjoy going to school more than boys (Segeritz et al. 2010; Van Ophuysen 2008), and they tend to show more positive approaches to learning such as attentiveness and task persistence than boys (Ready et al. 2005). All these attitudes and behaviors might alleviate the subjective burden that comes along with the transition from primary to secondary school and, therefore, might be considered when teachers make their school-track recommendations. Likewise, results of a recent study (Timmermans et al. 2016) suggest that teachers' expectations for the future academic performance of their students during the final grade of primary school were related to several teacher perceptions of student attributes. In general, teachers had higher expectations for a student if they perceived the student as having positive working habits. Positive working habits, however, are generally associated with female students rather than with male students (Reyna 2000; Siegle and Reis 1994).

Teacher expectations may also be driven by implicit gender stereotypes (Chalabaev et al. 2009). Much research has focused on stereotypes that teachers have concerning girls' ability in mathematics and science (e.g., Li 2005). Teachers tend to stereotype mathematics as a male domain and attribute boys' successes and failures to ability, whereas they tend to attribute girls'

successes and failures to effort (Fennema et al. 1990; Tiedemann 2002). Even if female students are high achievers, they are seen as less logical, less independent in mathematics, and liking mathematics less compared to equally achieving male students (Fennema et al. 1990). Gender stereotypes favoring male students are discussed as a factor influencing educational choices, resulting—for example—in an unequal distribution of males and females in STEM (science, technology, engineering, and mathematics) fields of studies (Glass and Minnotte 2010). If teachers tend to underestimate the mathematical ability of girls relative to boys (Frome and Eccles 1998), girls may be placed at a disadvantage.

However, teachers' stereotypical perceptions of students are moderated by the students' performance level. Teachers attribute more developmental resources in mathematics to male than to female primary school students if they are low or average achievers, but not if they are high achievers (Tiedemann 2002). Since high-achieving students (male or female) are likely to be recommended for the highest track in secondary school, teachers' stereotyped perceptions may not necessarily result in favoring boys in placement recommendations. Moreover, whereas boys are usually stereotyped as having stronger mathematical abilities than girls, girls are stereotyped as having stronger verbal abilities than boys (Plante et al. 2013). Primary school teachers, however, weight mathematical abilities to a smaller degree than verbal abilities, when it comes to placement recommendations (e.g., Bos et al. 2004; Klapproth et al. 2013). Hence, high-track recommendations may be more likely for girls than for boys.

Teacher-student gender interaction

Student assessment might not only depend on students' gender, but also on their teachers' gender. According to the gender-stereotypic model (Martin and Marsh 2005), boys achieve higher scores in classes taught by males, and girls are better when instructed by female teachers. Therefore, some policy makers (e.g., in Great Britain) attempted to increase the number of male teachers in primary schools (Francis et al. 2008), although studies on the effect of teacher gender on the achievement, attitudes, or behaviors of their male and female students are quite rare (Driessen 2007). Teachers might have preferences over students of their own gender, and hence, female teachers might assess girls better than boys, whereas male teachers might prefer boys to girls (Holmlund and Sund 2005). Indeed, some studies corroborate this hypothesis. For instance, Lavy (2004) could show that girls got on average higher scores than boys in the main school subjects, but in some school subjects (e.g., biology and chemistry), the difference between boys and girls was larger with female teachers than with male teachers. Likewise, Dee (2007) revealed that in secondary school, boys and girls were evaluated more positively when they were taught by a same-gender teacher rather than by a teacher of the opposite gender. Contrarily, however, Hopf and Hatzichristou (1999) found an interaction between student and teacher gender, such that male primary school teachers judged boys to show more problematic behavior than did female primary school teachers. Other studies found that female teachers generally evaluated both boys and girls more positively than male teachers (Ehrenberg et al. 1995), or even did not find a teacher-gender bias in assessing male and female students (e.g., Driessen 2007). Particularly in Germany, studies concerning the effect of teacher gender on student assessment have focused on school placement decisions. For instance, Helbig (2010) as well as Neugebauer (2011) could show that there was no significant interaction between teacher gender and student gender. One reason for not finding teacher-student gender interactions might be that both male and female teachers simply follow the rules given by their administrations or schools when judging students. Since these rules often

entail students' learning motivation (Neugebauer 2011), girls are on average preferred over boys, as girls show on average higher degrees of learning motivation than boys (Ready et al. 2005).

Development of achievement

When teachers in Germany are urged to opt for one of the different tracks in secondary school, they mainly resort to grades as indicators of the students' achievement (Arnold et al. 2007), which are given in the school reports. In the federal state (Berlin) where our study was conducted, teachers are presented with two school reports instead of one when it comes to placement recommendations. Although successive school reports from a single student may be very similar, they are hardly exactly the same.

How are changes of achievement perceived or evaluated by teachers, when it comes to school placement recommendations? When recommending a student for the highest track, the teacher in charge is likely to expect that the student will perform adequately in this track. According to Jussim et al. (1998), students' past performance predicts teacher expectations about students' future performance. A student who performed well in the past is expected to perform well in the future, whereas a student who performed poorly in the past is expected to perform poorly in the future. When teachers expect students to continue to perform according to previously established patterns, resulting expectation effects have been categorized as "sustaining expectation effects" (Cooper 1985; Good and Brophy 2003). Evidence for sustaining expectation effects comes, for example, from a study conducted by Cooper et al. (1976) where college students imagined themselves as primary school teachers who had to predict the performance of a child whose report card reflected either an increasing or decreasing performance pattern. Results indicated that expectations were higher in the increasing rather than decreasing condition. Rolison and Medway (1985) reported similar results with teachers as participants.

Although in Germany the development of academic achievement is not considered a factor for determining the suitable track, Caro et al. (2009) demonstrated that students growing more rapidly in their mathematics skills, measured by standardized achievement tests, were more likely to get a high-track recommendation than students with a smaller degree of growth. Caro et al. (2009) concluded that teachers value the growth rate of students for their placement recommendations. However, when teachers in Germany make placement recommendations, they usually do not have access to achievement test data, nor are they legally allowed to use these data. Therefore, in the Caro et al. study, teachers may have used indicators of achievement growth that were not explicitly given by test scores, but instead by the information that was actually present in the school reports. In particular, the teachers possibly gauged the development of achievement from the two school reports that were given for each student, even if the differences between them were rather small.

Research questions and hypotheses

The aim of the present study was to examine whether preservice teachers, who will become primary school teachers after successful completion of their teacher study program, and who therefore will eventually be in charge of making school placement decisions, are biased by two factors when making these decisions: the gender of the students and whether the students improved or deteriorated within one semester at the end of primary school. To the authors' knowledge, no study so far has investigated this research question experimentally, although student achievement in general is the major determinant of school placement decisions.

The rationale of the present study was as follows. Since teachers (or even preservice teachers) may evaluate girls to be more mature than boys, they would presumably predict higher achievements in secondary school for girls than for boys, even if both currently show equal achievements. Moreover, since teachers are likely to predict students' future achievements on the basis of their previous development of achievement, an increasing pattern of achievement would result in the prediction of higher achievements in secondary school than a decreasing pattern of achievement.

We therefore hypothesized that female students would be more likely to be recommended for the highest track than male students, despite their grades being equal. Furthermore, we assumed that showing preservice teachers two different school reports as the basis for deciding whether the students are eligible for being taught in the highest track or not, the preservice teachers will opt more frequently for the highest track when the GPA of the second report was lower (i.e., "better") than that of the first report, than when the GPA of second report was higher (i.e., "worse") than that of the first report. Hence, we expected that an improvement of the students in terms of GPA would make the recommendation for the highest track more likely, whereas with the deterioration of the GPA, the likelihood for the high-track recommendation would decrease.

In addition to the change of the GPA, we postulated that—as it is the standard official rule for teachers in Germany—the grand mean of the grades (i.e., the average over all grades of both school reports) would also be predictive for the recommendation of the participants. In particular, we expected that the lower the grand mean was, the more likely it would be that the participants recommend students for the highest track.

Finally, we examined whether the gender of the participants contributed differentially to the recommendations for male and female students, that is, whether there is an interaction between participants' gender and students' gender with respect to school-track recommendations. However, based on previous research, a straightforward hypothesis was hardly derivable, so that we treated this issue as an open research question.

Method

Participants

In total, 260 preservice teachers took part at the study. Of these participants, 172 (66.2%) completed the study fully, whereas 83 participants (31.9%) had missing values on more than 3%, but less than 10% of the responses. The remaining participants (1.9%) omitted 50% or more of the possible responses and were excluded from further analyses. In the sample used ($N = 255$), 176 (69.0%) participants were female and 79 (31.0%) were male. The distribution of gender in our sample matched pretty well the distribution of teachers' gender in German primary schools (Neugebauer and Gerth 2013). The participants' age varied between 18 and 40 years, with a mean age of 22.1 years ($SD = 3.6$).

For the 83 participants with less than 10% missing values, we conducted multiple imputation on the dependent variables by using the respective tool offered by SPSS, Version 23, since missing data pose a problem on data interpretation when not appropriately handled (Peugh and Enders 2004). Imputation was conducted five times, and the five imputations were finally aggregated to a single data set. Prior to multiple imputation, the pattern of missing values was analyzed. Little's MCAR (missing completely at random) test (Little 1988) yielded a χ^2 $(364) = 384.31$, $p = 0.101$, indicating that the values were missing randomly.

All participants were enrolled in a primary school teacher education program at the time of the study. In teacher education programs, the assessment of students and the training of diagnostic competences is a major part (e.g., Abs 2006). All participants had previous teaching experiences as student teachers or student observers in the classroom within the framework of their study program (the mean practice duration was 19.6 weeks (SD = 19.7)), and they had studied on average for 4.5 semesters (SD = 2.3). The study was open for 4 weeks.

Materials

Each participant received all 24 student vignettes, which were displayed online via the internet platform "soscisurvey.de" and which were accessible on every computer device connected to the internet. The vignettes mimicked two school reports that teachers in Berlin primary schools receive to make their track recommendations for secondary school. The vignettes were developed according to guidelines presented by Evans et al. (2015). Prior to the study, the vignettes were pretested in a small sample ($N = 6$) of students and teachers who were asked to rate the content of the vignettes with respect to plausibility and comprehensibility, and whether they appear similar to real-life student reports. One exemplar of the vignettes is shown in the Appendix.

Each vignette contained six grades varying between 1 ("very good") and 4 ("sufficient"), with each grade being related to one school subject. However, the school subjects were not specified (e.g., subject A: "2," subject B: "2," subject C: "4," subject D: "1," subject E: "2," subject F: "3"), so that the participants were to rely only on the amount of the grades and therefore could not apply subjective weighting of the school subjects. The rationale behind leaving the school subjects unspecified was as follows: To control for subjective weighting of school subjects, a design would have been necessary that would allow every single school subject to be crossed with all remaining factors, rendering the design virtually unfeasible. To reach a single GPA (e.g., 2.17), the combination of grades (e.g., 1; 2; 2; 2; 2; 4) was always the same.

The realized grand means of the grades displayed in the vignettes were $M = \{2.33; 2.50; 2.67; 2.83; 3.00; 3.17\}$, with higher means representing lower achievements. However, the grand means emerged from two school reports showing either improvement or deterioration in grades. In case of improvement, the GPAs of the first school report (representing grades obtained in semester 5/2) were 2.50, 2.67, 2.83, 3.00, 3.17, or 3.33, and the corresponding GPAs of the second school report (representing grades obtained in semester 6/1) were 2.17, 2.33, 2.50, 2.67, 2.83, or 3.00, respectively, so that the magnitude of improvement was always the same between two school reports. Accordingly in case of deterioration, the GPAs of the first school report were smaller than those of the second school report. Note that the change of GPAs was always due to the change of grades in one school subject by an amount of 2.0. For instance, when a student improved in her GPA (e.g., from 2.50 to 2.17, yielding a grand mean of 2.33), she realized this improvement by the change of grades in a single school subject from 4 ("sufficient") to 2 ("good"). The grand means of the GPAs were unrelated to both the students' gender and whether there was improvement or deterioration in grades.

The gender of the students was manipulated by the names that were assigned to the students. The names were common either for male or female German students.

Each vignette was supplemented with information regarding the students' working habits and social behavior in order to make the vignettes more similar to real-world school reports. This information was delivered by two rather short sentences, which were derived from standardized sentences used for appraisal of the working habits and social behavior in school (Niedersächsisches Kultusministerium 2010). All sentences used in the vignettes displayed

behavior that is regarded in school as "meeting the expectations." Thus, all vignettes showed student behavior that was evaluated in quite the same way.

The three factors (the students' grand mean of GPAs, their gender, and whether they improved or deteriorated) were varied orthogonally, resulting in a 2 (gender: male vs. female) × 2 (change of GPA: positive vs. negative) × 6 (grand mean 2.33, 2.50, 2.67, 2.83, 3.00, 3.17) within-subjects factorial design. The dependent variable was the decision of the participants, which was either in favor of or against placement in the highest school track. We additionally collected data about the participants' sociodemographic background.

Procedure

The participants were instructed to imagine that they were a teacher of a class in the last grade of primary school and were to make a decision about every student of this class on her or his future track in secondary school. They were given the options "in favor of the highest track" or "not in favor of the highest track" at the end of each student description. The participants were instructed to make use of the information that was presented to them for each of the 24 students. After the general instruction, an example task followed which should make the participants get acquainted with the procedure. After that, the student vignettes followed in random order. In case a participant did not make a judgment (and instead clicked on the "next" button), a prompt popped up which reminded the participant to make a judgment, or otherwise to proceed the experiment without making a judgment. A new vignette was shown on the screen after the preceding vignette was closed by the decision of the participant. After the participants had made decisions for all 24 students, they were asked to give some information about their sociodemographic background. Finally, they were all thanked, debriefed about the purpose of the study, and were given the opportunity to take part at a lottery with the prospect of a 20 Euro prize to win.

Data analyses

We used multilevel logistic regression analysis to test our hypotheses. In this analysis, the judgments of the participants are nested within the participants. Hence, the level-1 unit of the analysis consists of the repeated measures for each participant, and the level-2 unit is the participant. The predictors in the regression model were the grand mean of grades as a metric covariate (with the values 2.33, 2.50, 2.67, 2.83, 3.00, and 3.17) and student gender (female = 0, male = 1) as well as the change of GPA (negative = 0, positive = 1) as binary factors. Since in our hypotheses we predicted three main effects to occur, we firstly estimated a regression model that contained only main effects (model 1). However, since we could not exclude the possibility of interactions between the predictors, we also estimated a model that specified interaction terms (model 2). We additionally examined whether the gender of the participants affected their placement recommendations (model 3). Finally, in order to examine whether the results obtained in this study can be generalized to potentially all "admissible" preservice teachers, we conducted a generalizability study.

Results

The results were reported according to guidelines suggested by Jaccard (2001). Table 1 shows the mean proportions as well as the respective standard deviations for each condition.

Apparently, the proportions of high-track recommendations were dependent on the grand mean of grades, with a smaller grand mean resulting in higher proportions. Moreover, Table 1 also indicates that the change of grades contributed to the proportions of high-track recommendations, as a positive change (meaning that students improved) yielded higher proportions compared to a negative change. To examine whether these apparent effects were statistically significant, we conducted multilevel logistic regression analysis, of which the results are depicted in Table 2.

The resulting logistic regression equation for model 1, which contained only main effects, reads as follows:

(I) Predicted logit of high-track recommendation = 6.09 + 0.94 * Change of GPA + 0.05 * Student Gender − 2.69 * Grand Mean.

Except for student gender, all predictors were shown to be significant. For change of GPA, holding student gender and the grand mean constant, the logit increased by $B = 0.94$, when the change was positive. This effect translates to an odds ratio of 2.55, meaning that the odds of getting a recommendation for the highest track increased by factor 2.55 when the change in GPA was positive rather than negative. For the grand mean, lower values corresponded significantly with higher probabilities for high-track recommendations. An increase of one unit on the German grade scale (which ranges from 1 ("very good") to 6 ("insufficient")) corresponded with roughly a 15 times lower chance for a high-track recommendation.

In model 2, we added four interaction terms to the main effects, resulting in the following logistic regression model:

(II) Predicted logit of high-track recommendation = 3.49 + 4.04 * Change of GPA + 0.55 * Student Gender − 1.72 * Grand Mean + 2.80 * Change of GPA × Student Gender

Table 1 Means and standard deviations of the proportions of high-track recommendations as a function of students' grand mean in grades, their gender, and whether their grades improved or declined

Student gender	Change of grades	Grand mean in grades	*M*	SD
Male	Positive	2.33	0.69	0.46
		2.50	0.75	0.43
		2.67	0.54	0.50
		2.83	0.32	0.47
		3.00	0.18	0.39
		3.17	0.09	0.28
	Negative	2.33	0.44	0.50
		2.50	0.34	0.47
		2.67	0.23	0.42
		2.83	0.20	0.40
		3.00	0.19	0.39
		3.17	0.13	0.34
Female	Positive	2.33	0.71	0.45
		2.50	0.59	0.49
		2.67	0.43	0.50
		2.83	0.38	0.49
		3.00	0.28	0.45
		3.17	0.16	0.37
	Negative	2.33	0.39	0.49
		2.50	0.31	0.46
		2.67	0.25	0.43
		2.83	0.18	0.39
		3.00	0.13	0.34
		3.17	0.16	0.37

Table 2 Results of multilevel logistic regression analyses

Parameter	Model 1				Model 2				Model 3			
	B(SE)	Exp(*B*)	Wald χ^2 (*df*)	*p*	*B*(SE)	Exp(*B*)	Wald χ^2 (*df*)	*p*	*B*(SE)	Exp(*B*)	Wald χ^2 (*df*)	*p*
Intercept	6.09 (0.47)	442.36	168.98 (1)	< 0.001	3.49 (0.64)	32.63	29.66 (1)	< 0.001	9.49 (0.87)	13,169.58	119.15 (1)	< 0.001
Change	0.94 (0.08)	2.55	124.40 (1)	< 0.001	4.04 (0.68)	56.77	35.57 (1)	< 0.001	1.22 (0.94)	3.39	1.70 (1)	0.192
SG	0.05 (0.05)	1.05	1.20 (1)	0.274	0.55 (0.61)	1.74	0.81 (1)	0.367	− 0.20 (0.91)	0.82	0.05 (1)	0.822
GM	− 2.69 (0.18)	0.068	216.33 (1)	< 0.001	− 1.72 (0.24)	0.18	51.60 (1)	< 0.001	− 4.19 (0.34)	0.02	148.39 (1)	< 0.001
PG									− 10.97 (1.22)	171,600	81.33 (1)	< 0.001
Change × SG					2.80 (0.94)	16.49	8.94 (1)	0.003	5.52 (1.23)	248.70	20.05 (1)	< 0.001
Change × GM					− 1.15 (0.25)	0.32	20.45 (1)	< 0.001	0.16 (0.36)	1.17	0.19 (1)	0.663
SG × GM					− 0.17 (0.23)	0.84	0.57 (1)	0.450	0.06 (0.38)	1.06	0.03 (1)	0.864
SG × PG									2.39 (1.52)	10.88	2.48 (1)	0.116
Change × PG									2.77 (1.45)	15.89	3.64 (1)	0.057
GM × PG									4.56 (0.46)	95.33	99.17 (1)	< 0.001
Change × SG × GM					− 1.06 (0.35)	0.35	9.29	.002	− 2.09 (0.47)	0.13	19.38 (1)	< 0.001
Change × SG × PG									− 5.96 (2.16)	0.01	7.62 (1)	0.006
Change × GM × PG									− 1.55 (0.55)	0.21	8.07 (1)	0.005
SG × GM × PG									− 0.78 (0.56)	0.46	1.95 (1)	0.460
Change × SG × GM × PG									2.24 (0.80)	9.42	7.89 (1)	0.005
QIC	7058.71				6990.56				6321.22			

Note. QIC means quasi-likelihood under the independence criterion. Coding of the binary predictors was as follows. Change: 0 = negative, 1 = positive; SG: 0 = female, 1 = male, PG: 0 = female, 1 = male

Change change of grade point average, *SG* student gender, *GM* grand mean (mean of all grades obtained in two successive school reports), *PG* participant gender

− 1.15 * Change of GPA × Grand Mean − 0.17 * Gender × Grand Mean − 1.06 * Change of GPA × Student Gender × Grand Mean.

As compared to model 1, the goodness of fit, indicated by the quasi-likelihood under the independence model criterion (QIC), was smaller in model 2 and, hence, indicated better fit to the data. As in model 1, the same main effects were significant, which were the effect due to the change of GPA and the effect due to the grand mean. However, the effect of the change of GPA increased (from $B = 0.94$ in model 1 to $B = 4.04$ in model 2), whereas the effect of the grand mean decreased (from $B = -2.69$ in model 1 to $B = -1.72$ in model 2). Note that when interaction terms are included in a logistic regression equation, the coefficients for the main effects no longer represent main effects in the traditional sense, but instead odds ratios comparing the odds of one predictor of the interaction term when the other predictor of that interaction term is set to zero. Model 2 revealed three significant interaction terms: the Change of GPA × Student Gender interaction, the Change of GPA × Grand mean interaction, and the three-way interaction. However, the Student Gender × Grand mean interaction was not significant.

The Change × Grand Mean interaction means that when holding the value of student gender constant (e.g., when the students were all female or all male), the slopes of the functions were nonparallel, with steeper slopes for students who improved their grades than for students whose grades deteriorated. The Change of GPA × Student Gender interaction means that male students were more likely to get a high-track recommendation than female students, when they improved their grades rather than deteriorated.

The interpretation of the three-way interaction obtained from model 2 necessitates a look at the differences in logits between female-negative change and male-negative change students on the one hand, and female-positive change and male-positive change students on the other hand. At the smallest grand mean (2.33), the difference was Diff = 0.48 for positive-change students and Diff = 0.16 for negative-change students (with a higher probability of high-track recommendations for male than for female students). Hence, at this grand mean, gender of the students was of lower importance for judging negative-change students than for judging positive-change students. At the largest grand mean (3.17), the difference was Diff = − 0.56 for positive-change students and Diff = 0.01 for negative-change students, which indicates that at this achievement level gender played again a stronger role for the judgments of positive-change students than for the judgment of negative-change students. However, at this grand mean, female students were favored over male students, when they improved their achievements. Obviously, when students showed deterioration, the different grand means were considered to a lesser degree than when they improved, and the effect of the grand mean on the probability of a high-track recommendation was even stronger when the improving students were male rather than female.

In model 3, the participants' gender was included both as a main effect and as part of interaction terms. The resulting model was as follows:

(III) Predicted logit of high-track recommendations: 9.49 + 1.22 Change of GPA − 0.20 * Student Gender − 4.19 * Grand Mean − 10.97 * Participant Gender + 5.52 * Change of GPA × Student Gender + 0.16 * Change of GPA × Grand Mean + 0.06 * Student Gender × Grand Mean + 2.39 * Student Gender × Participant Gender + 2.77 * Change of GPA × Participant Gender + 4.56 * Grand Mean × Participant Gender − 2.09 * Change of GPA × Student Gender × Grand Mean − 5.96 * Change of GPA × Student Gender × Participant Gender − 1.55 * Change of GPA × Grand Mean × Participant Gender −

0.78 * Student Gender × Grand Mean × Participant Gender + 2.24 * Change of GPA × Student Gender × Grand Mean × Participant Gender.

Compared to model 1 and model 2, the QIC score was smaller and therefore indicated a better fit. Notably, the four-way interaction effect was significant, which we will illustrate in the following.

In Fig. 1, the predicted logits obtained from model 3, which corresponded to all combinations of factors realized in our study, were depicted. In the upper panels of Fig. 1, logits obtained from students with a positive change in GPA, and in the lower panels, logits obtained from students with a negative change in GPA are shown. The figure will help to interpret the four-way interaction of model 3.

To interpret the four-way interaction obtained from model 3, it is useful first to consider the effects for positively and negatively changing students separately. For students with positively changing grades, the slopes of the functions were steeper with female ($B = -5.04$) than with male ($B = -1.31$) participants, meaning that the grand mean of grades affected the participants' decisions to a larger degree when the participants were female rather than male. Moreover, the relationship between the grand mean of grades and student gender was dependent on the participants' gender. The difference in slopes between male and female students was larger with female (Diff = 2.03) than with male (Diff = 0.56) participants. In addition, whereas female participants judged male and female students differentially depending on their grand mean of grades, male participants preferred boys relative to girls at almost all values of the grand mean of grades.

The pattern of results was quite different for students with negatively changing grades. With female participants, the slopes of the functions were still steeper ($B = -4.16$) than with male ($B = 0.01$) participants. However, compared to positively changing students, female

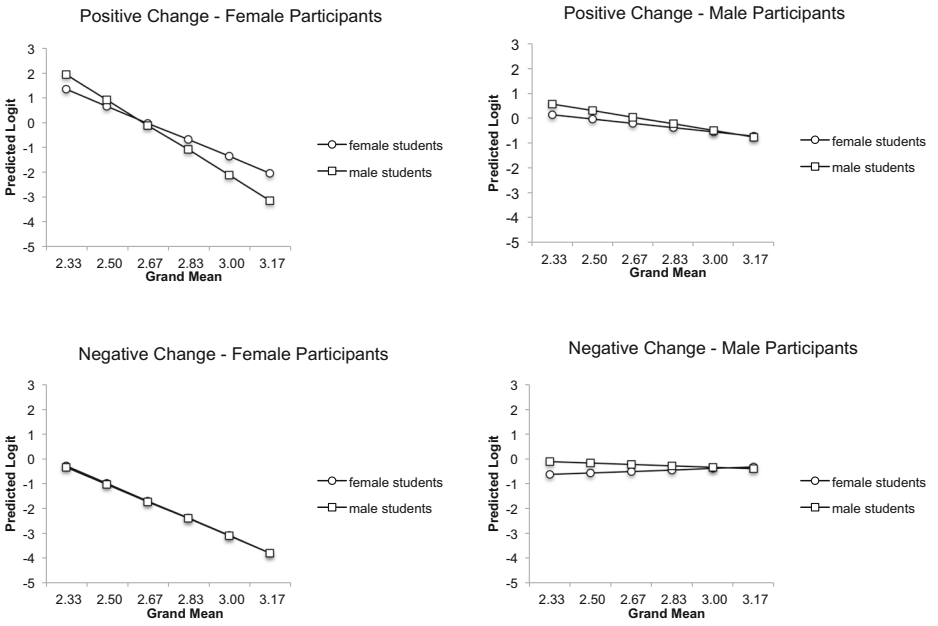

Fig. 1 Predicted logits of the probability of high-track recommendations, obtained from the different conditions in the experiment, depending on the grand mean of all grades. Upper panels: predicted logits for students who increased in grades. Lower panels: predicted logits for students who decreased in grades. Left panels: predicted logits for female participants. Right panels: predicted logits for male participants

participants did not make any visible difference when recommending boys or girls, since the slopes for boys and girls were virtually the same (Diff = 0.06). Male participants, however, appeared to devalue high-performing girls more than low-performing girls, since the slope for the girls was positive ($B = 0.37$), whereas the slope for the boys was negative ($B = -0.35$).

Finally, we conducted a generalizability study in order to examine whether the results obtained in this study can be generalized to all admissible preservice teachers. Generalizability (G) theory is a statistical framework that allows for the identification and estimation of different sources of measurement error (Shavelson and Webb 2006) in order to determine the limits of generalizability of the results obtained from the measurement made (Cronbach et al. 1972). These different sources of error (e.g., item, occasion, participants, test form) are called facets of a measurement. In a simple G study, we attempted to isolate and estimate the measurement error attributed by the participants. In the G study we conducted, participants were considered a facet, whereas the independent variables of the study (grand mean, change, gender) were considered the objects of measurement. Table 3 shows the variance components for the main effects and interactions.

The largest variance component was obtained for participants (12.8%) and for the four-way interaction (44.7%). The variance component for participants shows that—averaging over all levels of the independent variables—the participants in the sample differed systematically in their judgments. The large variance component for the four-way interaction reflects that the participants' judgments were dependent on the independent variables—a result that was predicted in advance and confirmed by logistic regression analysis. In order to give an estimate about the degree to which the results can be generalized to (similar) persons not part of the sample, we estimated the dependability index (Brennan 2001; see also Shavelson and Webb 2006). The dependability index is analogous to the reliability coefficient in classical test theory, and its formula is as follows (cf. Shavelson and Webb 2006):

$$\Phi = \sigma^2{}_p / \left(\sigma^2{}_p + \sigma^2{}_\Delta\right)$$

with $\sigma^2{}_p$ equals the variance component for participants, and $\sigma^2{}_\Delta$ equals the error variance. Due to the large number of participants in the study ($N = 255$), the error variance is quite small ($\sigma^2{}_\Delta = 0.00073$), yielding a dependability index of $\Phi = 0.976$.

Table 3 Variance components for the main effects and interactions

Source	Variance	Percent of total variability
Participant (p)	0.030	12.8
Grand mean (gm)	0.023	9.8
Change (c)	0.015	6.4
Gender (g)	0.000	0
p × gm	0.013	5.5
p × c	0.012	5.1
p × g	0.000	0
gm × c	0.008	3.4
gm × g	0.000	0
c × g	0.001	0.4
p × gm × c	0.022	9.4
p × gm × g	0.001	0.4
p × c × g	0.002	0.8
gm × c × g	0.003	1.3
p × gm × c × g	0.105	44.7

Discussion

Comparison between the hypotheses and the results obtained

With the present study, we aimed at examining whether preservice teachers valued both the gender of primary school students as well as their development of GPAs, indicated by two successive school reports, when making recommendations for the students track in secondary school. The results obtained seem to be highly reliable, as the G study showed, so that the results might be generalizable to similar participants. We predicted that—according to legal regulations—the grand mean of the grades should affect the probability of a high-track recommendation, with lower grand means (i.e., higher achievements) resulting in higher probabilities. This hypothesis could be confirmed. When the grand mean was reduced by one unit on the German grade scale, the probability for a high-track recommendation increased by approximately factor 15. This result clearly shows that the preservice teachers acknowledged the overall achievement indicated by the grades of two school reports as a basis for their decision. Hence, what legal regulations envisage was actually adopted by our participants.

However, we additionally assumed that when students improved their GPA within one-half year of schooling, they would be more likely to get a high-track recommendation than students who deteriorated within the same time period, even if their grand mean was the same. This hypothesis was also confirmed. Actually, students who improved were two and a half times more likely to get a high-track recommendation than students who deteriorated. This result clearly contradicts the official regulations provided by the authorities, since teachers (and therefore preservice teachers) are allowed only to take into consideration the grand mean of the GPA of both school reports, but not their change. Also note that this effect occurred due to a rather moderate change in GPA. The difference between the successive school reports was an increase or a decrease of one single grade (out of six) by the amount of 2 units on the German grade scale. That is, the GPAs between both school reports differed by 1/3 unit, with one unit being roughly equivalent to a letter grade in the US grading system.

Finally, we predicted that preservice teachers would account for student gender by preferring girls over boys in their high-track recommendations. This hypothesis was not confirmed since the assumed main effect of gender was not significant, which means that on average the participants did not make a difference between male and female students when judging their suitability for the highest track.

However, there were some unexpected significant interactions that qualified the obtained main effects. First, the significant Change of GPA × Grand mean interaction showed that when students were rather low in achievement, the probabilities of receiving a high-track recommendation were quite similar between improving and deteriorating students. However, with high-performing students (indicated by a low grand mean), positively developing students were much better off than negatively developing students.

A possible explanation refers to the students' achievements that were indicated by the second school report, as the GPA of the second report might have served as a predictor of future achievement. When the grand mean of all grades was rather high (e.g., 3.00), the student's improvement yielded a second-report GPA of 2.83, and when the student deteriorated, the second-report GPA was 3.17. Although both second-report GPAs were produced by a change of 0.33, both the improvement and the deterioration would hardly justify a recommendation for the highest track. That is, at this grand mean, a change of the GPA in either direction would probably be of no consequence for the participants' decision. Consider now

the grand mean of 2.33, indicating rather high achievements. When students improved, their GPA of the second school report was 2.17, whereas when they deteriorated, the GPA of the second report was 2.50. If the participants valued the second report as a predictor of future achievement, a student who improved would certainly be recommended for the highest track, whereas a student who deteriorated would presumably judged as being not eligible for the highest track by quite a high number of participants. Thus, the change between both school reports appears to be valued differently depending on the second-report GPA. Since the GPA of the second report was directly related to the grand mean, the change effect was dependent on the grand mean, thus resulting in a Change of GPA × Grand mean interaction.

Second, the significant Change of GPA × Gender interaction means that the difference in the probabilities of getting a high-track recommendation between boys and girls was larger when the students grew in achievement than when they were downgraded. Contrary to our prediction, male students were preferred over female students, at least when they improved rather than declined. A possible explanation refers to maturational differences between boys and girls, which the participants presumably had assumed. If the participants recognized improvement with boys, they might have seen a latent potential in male students that eventually could result in high performance in secondary school.

However, the significant three-way interaction qualified all the significant two-way interactions. Only with rather high-performing students, positively developing boys were preferred over girls. Yet, when the students were rather low performers, the reverse was the case, meaning that girls were favored over boys. Since female primary school students are on average more mature than their male counterparts (Colom and Lynn 2004; Lynn 1999) and show more positive attitudes to school (OECD 2004), more attentiveness and task persistence (Ready et al. 2005), and more positive working habits (Reyna 2000) than boys, the participants might have evaluated highly performing and positively developing male students as exceptionally good. As such, the participants could have attributed more positive characteristics to these students and hence might have expected a more positive development in secondary school compared to their female counterparts, who in contrast were usually expected to be good performers in school. Indeed, studies on teacher biases in identifying talented students have revealed that teachers who were asked to nominate students for gifted programs based on hypothetical student profiles were more likely to select profiles where the students' behavior did not match the expected gender stereotype (e.g., Bianco et al. 2011; Powell and Siegle 2000). For example, Powell and Siegle (2000) could show that teachers who generally believed that female students were better at reading than male students rated the profile of a male student who was a very good reader higher than a female student with the same skills. Similar results were obtained by Bianco et al. (2011) who demonstrated that teachers were less willing to refer a female student to a gifted and talented program than a male student, who actually were identically described regarding their characteristics.

How did the participants judge the change of students' achievement?

This study brought evidence that preservice teachers regarded the development of grades as an indicator of future success in school, since students who improved their GPA were more than twice as likely to be recommended for the highest track than those who deteriorated. Despite this intriguing effect, the mechanism according to which this effect occurred is still unclear. We assumed that the participants would extrapolate the students' development in achievement (indicated by their successive GPAs) in line with their sustaining expectations (Cooper 1985;

Good and Brophy 2003) according to which improvement would be followed by further improvement, and impairment would be followed by further impairment. When applying this rule of development to the student vignettes of our study, an improving student would certainly be judged as being more successful in secondary school than a deteriorating student, since the former is likely to improve further, whereas the latter is likely to get worse. Even if we assume a less strict rule, for instance by proposing that an increase might be followed by further increase *or* by maintaining the level of achievement, high-track recommendations would be still more likely compared to students who fall behind their initial achievements. When the participants adopted a growth rule like this, they ignored that the change of GPA might have been the result of pure randomness. When students increase or decrease in their grades, this could happen by a variety of factors, which might be a "true" change in achievement or skills, but also—and perhaps equally likely—factors like a change of motivation, a change of teachers and a corresponding change of grading rules, or changes of the learning environment at home.

How did the participants' gender affect their judgments?

Our final analysis has shown that participants' gender significantly contributed to the likelihood of their high-track recommendations. On average, male participants recommended students more frequently for the highest track than female participants. This "leniency" of male participants in recommending students for the highest track is also mirrored in the results from studies investigating teachers' characteristics as predictors of their ratings of students. For instance, Taylor et al. (2001) could show that female participants (inservice as well as preservice teachers) rated the degree of learning and behavioral problems of videotaped students on average higher than male participants. Moreover, the female participants in our study considered the grand mean of grades more strongly than did the male participants for their high-track recommendations for both improving and deteriorating students, and even more strongly for boys than for girls when the students were improving in grades. Similarly, research has shown that female teachers were more accurate than male teachers in identifying students' behavioral problems (Ritter 1989) or learning difficulties (Hopf and Hatzichristou 1999). Hence, it seems that the female participants of our study evaluated the students' grades with more caution and precision, particularly the boys, and were less optimistic than the male participants. This effect might also be due to a proneness of men to be more optimistic than women. This gender-dependent "optimistic bias" (Weinstein 1989) has been found in several areas, such as marriage (Lin and Raghubir 2005), self-evaluation (Beyer and Bowden 1997), and the accuracy of grade expectancies (Beyer 1999). With deteriorating students, however, female participants did not differentiate between boys and girls, whereas male participants preferred boys relative to girls and even devalued high-performing girls relative to low-performing girls. The male participants' preference of boys indicates a same-gender bias, which has also been found in some previous investigations (e.g., Dee 2007; Lavy 2004).

Limitations

Some limitations should be mentioned that were inherent in the study. First, we did not realize baseline conditions wherein no change between successive school reports occurred. Second, this study was experimental in nature. While we therefore could expect the realization of a high level of internal validity, field investigations are nevertheless needed in order to show the effects obtained in a natural environment. Third, at the end of the experiment, we did not ask the participants about their preferences with respect to the placement decisions they made, for instance, whether or not

they favored male over female students. If we had asked them for subjective reasons for their preferences, we might have got more insight into the participants' judgment process. Fourth, grades were not associated with specific school subjects. This might have been confusing for some participants, as in practice grades are always related to school subjects. However, if we had assigned grades to specific school subjects, it is likely that students would have been judged based on the grades in distinct school subjects. To make the design of the study feasible, we abstained from specifying school subjects. For the same reason, we did not incorporate personality traits of students, which might have had an effect on the participants' judgments, into the vignettes. Fifth and finally, since we examined decisions made by preservice teachers, we were not able to securely infer from the study's results to inservice teachers' decision-making. Recent research has identified some differences in decision-making between (rather inexperienced) preservice teachers or students and experienced inservice teachers. For example, Krolak-Schwerdt et al. (2009) could show that teachers (i.e., experts) were more flexible than university students of natural sciences (i.e., laymen) to switch between different modes of information processing, when the task was to evaluate characteristics of primary school students. In addition, teachers seem to be not only more flexible in choosing the appropriate information processing strategy, but they also use more information than rather inexperienced teacher students (e.g., Sabers et al. 1991). There is also evidence that when decisions are at high stakes (as it is certainly the case with real school-track recommendations), teachers decide more carefully than if they make decisions in a rather artificial experimental context (cf. Glock et al. 2012). However, like preservice teachers, inservice teachers are prone to be affected by their implicit attitudes toward students (e.g., Glock and Karbach 2015; Mertler 2004). Hence, preservice teachers should be made aware of their proneness to judge students differently according to factors not related to the predetermined educational standards.

Follow-up studies would shed more light into the complex decision-making processes when preservice or even inservice teachers evaluate students regarding their appropriateness for a secondary school track. For instance, establishing a baseline condition within a similar experimental design would allow for examining whether or not participants weight improvement and deterioration of grades equally or differentially. Furthermore, the provision of instructions that explicitly state how to handle information from both school reports could reduce the bias in placement recommendations. Finally, in order to validate the experimental studies, case studies could be applied in which participants would be presented with more elaborative student descriptions, and participants' responses would be coded qualitatively.

Conclusions

In Germany (and some other European countries), students leave primary school with a certification of their teachers that recommends them for one of the school tracks in secondary education. The tracks in secondary school are hierarchically ordered, with only the highest track ("Gymnasium") allowing for university entrance after being successfully accomplished. Hence, it is of great importance to know for students, preservice teachers, teachers, and policy makers, which factors have an effect on these recommendations. This study is the first one that showed that the change of grades between two successive school semesters has a large impact on preservice teachers' judgments as to the eligibility of students for the highest track in secondary school. Moreover, this effect was dependent on the gender of the students, the gender of the participants, and the students' overall achievement indicated by the grand mean of the grades of both school reports. Since the effect of the change of grades on the preservice

teachers' judgments is not envisaged by school authorities, and presumably not known by preservice teachers or educational personnel involved in teacher education, we deemed it of utmost importance to provide this knowledge for teacher education programs. Further studies should examine whether the effects obtained in this study generalize to real school settings. If data provide evidence that students are partially evaluated by their grades' development, the tracking policy would have to be reconsidered. Moreover, even in school systems where students are not separated into different school tracks, the results of this study should be considered as a caveat since the development of school marks, measured at two occasions, might be the result of coincidence and does not necessarily reflect real change.

Appendix

Name of student: Klaus

School Report Semester 5/2

Subject A	2	Subject D	2
Subject B	4	Subject E	2
Subject C	1	Subject F	1

School Report Semester 6/1

Subject A	2	Subject D	2
Subject B	4	Subject E	2
Subject C	3	Subject F	1

Notes on social behavior and working habits:
Klaus' working habits meet the expectations of his teachers. He frequently takes part at common social activities.

Would you recommend Klaus for the highest track (Gymnasium) in secondary school?

YES O NO O

References

Abs, H. J. (2006). Zur Bildung diagnostischer Kompetenz in der zweiten Phase der Lehrerbildung. In C. Allemann-Ghionda & E. Terhart (Eds.), *Kompetenzen und Kompetenzentwicklung von Lehrerinnen und Lehrern* (pp. 217–234). Weinheim: Beltz.

Arnold, K.-H., Bos, W., Richert, P., & Stubbe, T. C. (2007). Schullaufbahnpräferenzen am Ende der vierten Klassenstufe. In W. Bos, S. Hornberg, K.-H. Arnold, G. Faust, L. Fried, E.-M. Lankes, K. Schwippert, & R. Valtin (Eds.), *IGLU 2006. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (pp. 271–297). Waxmann: Münster.

Beyer, S. (1999). Gender differences in the accuracy of grade expectancies and evaluations. *Sex Roles, 41*(3/4), 279–296.

Beyer, S., & Bowden, E. M. (1997). Gender differences in self-perceptions: convergent evidence from three measures of accuracy and bias. *Personality and Social Psychology Bulletin, 23*(2), 157–172.

Bianco, M., Harris, B., Garrison-Wade, D., & Leech, N. (2011). Gifted girls: gender bias in gifted referrals. *Roeper Review, 33*(3), 170–181.

Bos, W., Voss, A., Lankes, E.-M., Schwippert, K., Thiel, O., & Valtin, R. (2004). Schullaufbahnempfehlungen von Lehrkräften für Kinder am Ende der vierten Jahrgangsstufe. In W. Bos, E.-M. Lankes, M. Prenzel, K. Schwippert, R. Valtin, & G. Walther (Eds.), *IGLU - Einige Länder der Bundesrepublik Deutschland im nationalen und internationalen Vergleich* (pp. 191–228). Münster: Waxmann.

Branje, S. J. T., van Lieshout, C. F. M., & Gerris, J. R. M. (2007). Big five personality development in adolescence and adulthood. *European Journal of Personality, 21*(1), 45–62.

Braun, C., & Mehringer, V. (2009). Familialer Hintergrund, Übertrittsempfehlungen und Schulerfolg bei Kindern mit und ohne Migrationshintergrund. In J. Hagedorn, V. Schurt, C. Steber, & W. Waburg (Eds.), *Ethnizität, Geschlecht, Familie und Schule: Heterogenität als erziehungswissenschaftliche Herausforderung* (pp. 55–79). Wiesbaden: Springer.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Caro, D. H., Lenkeit, J., Lehmann, R., & Schwippert, K. (2009). The role of academic achievement growth in school track recommendation. *Studies in Educational Evaluation, 35*(4), 183–192.

Carrothers, R. M., Gregory, S. W., & Gallagher, T. J. (2000). Measuring emotional intelligence of medical school applicants. *Academic Medicine, 75*(5), 456–463.

Chalabaev, A., Sarrazin, P., Trouilloud, D., & Jussim, L. (2009). Can sex-undifferentiated teacher expectations mask an influence of sex stereotypes? *Journal of Applied Social Psychology, 39*(10), 2469–2498.

Chesterfield, R., & Enge, K. (1998). Gender, cognitive categorization, and classroom interaction patterns of Guatemalan teachers. *Human Organization, 57*(1), 108–116.

Colom, R., & Lynn, R. (2004). Testing the developmental theory of sex differences in intelligence on 12-18 year olds. *Personality and Individual Differences, 36*(1), 75–82.

Cooper, H. (1985). Models of teacher expectation communication. In J. B. Dusek (Ed.), *Teacher expectancies* (pp. 135–158). Hillsdale, NJ: Erlbaum.

Cooper, H., Lowe, C., & Baron, R. (1976). Pattern of past performance and expected future performance: a reversal of the unexpected primacy effect. *Journal of Applied Social Psychology, 6*(1), 31–39.

Cronbach, L., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: theory of generalizability for scores and profiles*. New York: Wiley.

Darom, E., & Rich, Y. (1988). Sex differences in attitudes toward school: student self-reports and teacher perceptions. *British Journal of Educational Psychology, 58*(3), 350–355.

Dee, T. S. (2007). Teachers and the gender gaps in student achievement. *Journal of Human Resources, 42*, 528–554.

Driessen, G. (2007). The feminization of primary education: effects of teachers' sex on pupil achievement, attitudes and behaviour. *Review of Education, 53*, 183–203.

Ehrenberg, R., Goldhaber, D., & Brewer, D. (1995). Do teachers' race, gender, and ethnicity matter? Evidence from the National Educational Longitudinal Study of 1988. *Industrial and Labor Relations Review, 48*(3), 547–560.

Evans, S. C., Roberts, M. C., Keeley, J. W., Blossom, J. B., Amaro, C. M., Garcia, A. M., Odar Stough, C., Canter, K. S., Robles, R., & Reed, G. M. (2015). Vignette methodologies for studying clinicians' decision-making: validity, utility, and application in ICD-11 field studies. *International Journal of Clinical and Health Psychology, 15*(2), 160–170.

Fennema, E., Peterson, P. L., Carpenter, T. P., & Lubinski, C. A. (1990). Teachers' attributions and beliefs about girls, boys and mathematics. *Educational Studies in Mathematics, 21*(1), 55–69.

Francis, B., Skelton, C., Carrington, B., Hutchings, M., Read, B., & Hall, I. (2008). A perfect match? Pupils' and teachers' views of the impact of matching educators and learners by gender. *Research Papers in Education, 23*(1), 21–36.

Frome, P. M., & Eccles, J. S. (1998). Parents' influence on children's achievement-related perceptions. *Journal of Personality and Social Psychology, 74*(2), 435–452.
Glass, C., & Minnotte, K. L. (2010). Recruiting and hiring women in STEM fields. *Journal of Diversity in Higher Education, 3*(4), 218–229.
Glock, S., & Karbach, J. (2015). Preservice teachers' implicit attitudes toward racial minority students: Evidence from three implicit measures. *Studies in Educational Evaluation*, *45*, 55–61.
Glock, S., Krolak-Schwerdt, S., Klapproth, F., & Böhmer, M. (2012). Improving terachers' judgments: accountability affects teachers' tracking decision. *International Journal of Technology and Inclusive Education, 1*, 89–98.
Good, T. L., & Brophy, J. E. (2003). *Looking in classrooms* (9th ed.). Boston, MA: Allyn and Bacon.
Helbig, M. (2010). Sind Lehrerinnen für den geringeren Schulerfolg von Jungen verantwortlich? *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 62*(1), 93–111.
Holmlund, H., & Sund, K. (2005). *Is the gender gap in school performance affected by the sex of the teacher? Working Paper 5/2005*. Stockholm: Swedish Institute for Social Research.
Hopf, D., & Hatzichristou, C. (1999). Teacher gender-related influences in Greek schools. *British Journal of Educational Psychology, 68*, 1–18.
Jaccard, J. (2001). Interaction effects in logistic regression. In *Sage University papers series on quantitative applications in the social sciences, 07-135*. Thousand Oaks: Sage.
Jürges, H., & Schneider, K. (2011). Why young boys stumble: early tracking, age and gender bias in the German school system. *German Economic Review, 12*(4), 371–394.
Jussim, L., Smith, A., Madon, S., & Palumbo, P. (1998). Teacher expectations. In J. Brophy (Ed.), *Advances in research on teaching: expectations in the classroom* (p. 148). Greenwich, CT: JAI Press.
Klapproth, F., Glock, S., Krolak-Schwertdt, S., Martin, R., & Böhmer, M. (2013). Prädiktoren der Sekundarschulempfehlung in Luxemburg. *Ergebnisse einer Large-Scale-Untersuchung. Zeitschrift für Erziehungswissenschaft, 16*(2), 355–379.
Klimstra, T. A., Hale, W. W., Raaijmaakers, Q. A. W., Branje, S. J. T., & Meeus, W. H. J. (2009). Maturation of personality in adolescence. *Journal of Personality and Social Psychology, 96*(4), 898–912.
KMK – Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2010). *Übergang von der Grundschule in Schulen des Sekundarbereichs I*. Retrieved from http://www.kmk.org/ fileadmin/ veroeffentlichungen_beschluesse/ 2010/2010_10_18-Uebergang-Grundschule-S_eI1-Orientierungsstufe.pdf.
Kristen, C. (2006). Ethnische Diskriminierung in der Grundschule? Die Vergabe von Noten und Bildungsempfehlungen. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 58*(1), 79–97.
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.
Kumar, R., & Lal, R. (2006). The role of self-efficacy and gender differences among the adolescents. *Journal of the Indian Academy of Applied Psychology, 32*, 249–254.
Lane, K. L., Givner, C. C., & Pierson, M. R. (2004). Teacher expectations of student behaviour: social skills necessary for success in elementary school classrooms. *The Journal of Special Education, 38*(2), 104–110.
Lavy, V. (2004). Do gender stereotypes reduce girls' human capital outcomes? Evidence from a natural experiment. *NBER Working Paper, 10678*.
Lehmann, R. H., & Peek, R. (1997). *Aspekte der Lernausgangslage von Schülerinnen und Schülern der fünften Jahrgangsstufe an Hamburger Schulen. Bericht über die Untersuchung im* September 1996. Hamburg, Germany: Behörde für Schule, Jugend und Berufsausbildung.
Li, J. (2005). Mind or virtue: Western and Chinese beliefs about learning. *Current Directions in Psychological Science, 14*(4), 190–194.
Lin, Y. C., & Raghubir, P. (2005). Gender differences in unrealistic optimism about marriage and divorce: are men more optimistic and women or realistic? *Personality and Social Psychology Bulletin, 31*(2), 198–207.
Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83*(404), 1198–1202.
Lynn, R. (1999). Sex differences in intelligence and brain size: a developmental theory. *Intelligence, 27*(1), 1–12.
Marshall, W. A., & Tanner, J. M. (1970). Variations in the patterns of prepubertal changes in boys. *Archives of Disease in Childhood, 45*(239), 13–23.
Martin, A., & Marsh, H. (2005). Motivating boys and motivating girls: does teacher gender really make a difference? *Australian Journal of Education, 49*(3), 320–334.
Mertler, C. A. (2004). Secondary teachers' assessment literacy: does classroom experience make a difference? *American Secondary Education, 33*, 49–64.
Milek, A., Lüdtke, O., Trautwein, U., Maaz, K., & Stubbe, T. C. (2009). Wie konsistent sind Referenzgruppeneffekte bei der Vergabe von Schulformempfehlungen? Bundeslandspezifische Analysen mit Daten der IGLU-Studie. *Zeitschrift für Erziehungswissenschaft, 12*, 282–301.

Neugebauer, M. (2011). Werden Jungen von Lehrerinnen bei den Übergangsempfehlungen für das Gymnasium benachteiligt? Eine Analyse auf Basis der IGLU-Daten. In A. Hadjar (Ed.), *Geschlechtsspezifische Bildungsungleichheiten* (pp. 235–260). Wiesbaden: Springer.

Neugebauer, M., & Gerth, M. (2013). Weiblicher Schulkontext und Schulerfolg von Jungen. In R. Becker & A. Schulze (Eds.), *Bildungskontexte* (pp. 431–455). Wiesbaden: Springer.

Niedersächsisches Kultusministerium. (2010). *Bewertung des Arbeits- und Sozialverhaltens. Runderlass des Kultusministeriums*. Hannover: Niedersächsisches Kultusministerium.

OECD. (2004). *Learning for tomorrow's world—first results from PISA 2003*. Paris: OECD.

Parks, F. R., & Kennedy, J. H. (2007). The impact of race, physical attractiveness, and gender on education majors' and teachers' perceptions of students. *Journal of Black Studies, 37*(6), 936–943.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: a review of reporting practices and suggestions for improvement. *Review of Educational Research, 74*(4), 525–556.

Phillips, D. (1991). Assessment in German schools. *Journal of Curriculum Studies, 23*(6), 544–548.

Plante, I., de la Sablonnière, R., Aronson, J. M., & Théorêt, M. (2013). Gender stereotype endorsement and achievement-related outcomes: the role of competence beliefs and task values. *Contemporary Educational Psychology, 38*(3), 225–235.

Preckel, F., Zeidner, M., Goetz, T., & Schleyer, E. J. (2008). Female "big fish" swimming against the tide: the "big-fish-little-pond effect" and gender-ratio in special gifted classes. *Contemporary Educational Psychology, 33*(1), 78–96.

Powell, T., & Siegle, D. (2000). Teacher bias in identifying gifted and talented students. *The National Research Center on the Gifted and Talented Newsletter, Spring, 2000*, 13–15.

Ready, D., LoGerfo, L., Burkham, D. T., & Lee, V. E. (2005). Explaining girls' advantage in kindergarten literacy learning: do classroom behaviors make a difference? *Elementary School Journal, 106*(1), 21–38.

Reyna, C. (2000). Lazy, dump, or industrious: when stereotypes convey attribution information in the classroom. *Educational Psychology Review, 12*(1), 85–110.

Ritter, D. (1989). Teachers' perceptions of problem behaviour in general and special education. *Exceptional Children, 55*(6), 559–564.

Rolison, M. A., & Medway, F. J. (1985). Teachers' expectations and attributions for student achievement: effects of label, performance pattern, and special education intervention. *American Educational Research Journal, 22*(4), 561–573.

Sabers, D. S., Cushing, K. S., & Berliner, D. C. (1991). Differences among teachers in a task characterized by simultaneity, multidimensional, and immediacy. *American Education Research Journal, 28*, 63–88.

Schmitt, M. (2008). Die Bedeutung von sozialer Herkunft und bundeslandspezifischen Übergangsregelungen für die Grundschulempfehlung. In E.-M. Lankes (Ed.), *Pädagogosche Professionalität als Gegenstand empirischer Forschung* (pp. 111–121). Münster: Waxmann.

Schneider, T. (2011). Die Bedeutung der sozialen Herkunft und des Migrationshintergrundes für Lehrerurteile am Beispiel der Grundschulempfehlung. *Zeitschrift für Erziehungswissenschaft, 14*(3), 371–396.

Schulze, A., Wolter, F., & Unger, R. (2009). Bildungschancen von Grundschülern: Die Bedeutung des Klassen- und Schulkontextes am Übergang auf die Sekundarstufe I. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 61*(3), 411–435.

Segeritz, M., Stanat, P. & Walter, O. (2010). Muster des schulischen Erfolgs von Mädchen und Jungen mit Migrationshintergrund. In C. P. Allemann-Ghionda, Stanat, K. Göbel, & C. Röhner (Eds.), *Migration, Identität, Sprache und Bildungserfolg. 55. Beiheft der Zeitschrift für Pädagogik*, 165–186.

Senatsverwaltung für Bildung, Jugend und Wissenschaft. (2015). Berliner Schulwegweiser. In *Wohin nach der Grundschule?* Berlin: SenBJW.

Shavelson, R. J., & Webb, N. M. (2006). Generalizability theory. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 309–322). Hillsdale: Lawrence Erlbaum.

Siegle, D., & Reis, S. M. (1994). Gender differences in teacher and student perceptions of student ability and effort. *The Journal of Secondary Gifted Education, 6*, 86–92.

Stahl, N. (2007). Schülerwahrnehmung und -beurteilung durch Lehrkrafte. In H. Ditton (Ed.), *Kompetenzaufbau und Laufbahnen im Schulsystem. Ergebnisse einer Längsschnittuntersuchung an Grundschulen* (pp. 171–198). Waxmann: Münster.

Stubbe, T. C. (2009). *Bildungsentscheidungen und sekundäre Herkunftseffekte: Soziale Disparitäten bei Hamburger Schülerinnen und Schülern der Sekundarstufe I*. Münster: Waxmann.

Stubbe, T. C., & Bos, W. (2008). Schullaufbahnempfehlungen von Lehrkräften und Schullaufbahnentscheidungen von Eltern am Ende der vierten Jahrgangsstufe. *Empirische Pädagogik, 22*, 49–63.

Taylor, P. B., Gunter, P. L., & Slate, J. R. (2001). Teachers' predictions of inappropriate student behavior as a function of teachers' and students' gender and ethnic background. *Behavioral Disorders, 26*(2), 146–151.

Tiedemann, J. (2002). Teachers' gender stereotypes as determinants of teachers perceptions in elementary school mathematics. *Educational Studies in Mathematics, 50*(1), 49–62.
Timmermans, A. C., de Boer, H., & van der Werf, M. P. C. (2016). An investigation of the relationship between teachers' expectations and teachers' perceptions of student attributes. *Social Psychology of Education, 19*(2), 217–240.
Trautwein, U., & Baeriswyl, F. (2007). Wenn leistungsstarke Klassenkameraden ein Nachteil sind: Referenzgruppeneffekte bei Übertrittsentscheidungen. *Zeitschrift für Pädagogische Psychologie, 21*(2), 119–133.
Van Ophuysen, S. (2008). Zur Veränderung der Schulfreude von Klasse 4 bis 7. Eine Längsschnittanalyse schulformspezifischer Effekte von Ferien und Grundschulübergang. *Zeitschrift für Pädagogische Psychologie, 22*(34), 293–306.
Weinstein, N. D. (1989). Optimistic biases about personal risks. *Science, 246*(4935), 1232–1233.
Zeidner, M., & Schleyer, E. J. (1999). The big-fish-little-pond effect for academic self-concept, test anxiety, and school grades in gifted children. *Contemporary Educational Psychology, 24*(4), 305–329.

Florian Klapproth. Medical School Berlin, Calandrellistrasse 1-9, 12247 Berlin, Germany. Email: florian.klapproth@medicalschool-berlin.de

Current themes of research:

Predictors and consequences of grade retention. Predictive validity of tracking decisions. Prediction of school failure.

Most relevant publications in the field of Psychology of Education:

Klapproth, F., Schaltz, P., Brunner, M., Keller, U., Fischbach, A., Ugen, S., & Martin, R. (2016). Short-term and medium-term effects of grade retention in secondary school on academic achievement and psychosocial outcome variables. *Learning and Individual Differences, 50*, 182–194.
Klapproth, F. (2015). Do algorithms homogenize students' achievements in secondary school better than teachers' tracking decisions? *Education Policy Analysis Archives, 23*, 1–18.
Klapproth, F., & Schaltz, P. (2014). Who is retained in school, and when? Survival analysis of predictors of grade retention in Luxembourgish secondary school. *European Journal of Psychology of Education, 30*, 119–136.
Klapproth, F., & Schaltz, P. (2013). Identifying students at risk of school failure in Luxembourgish secondary school. *International Journal of Higher Education, 2*, 191–204.

Birthe Doreen Fischer. Medical School Berlin, Calandrellistrasse 1-9, 12247 Berlin, Germany

Current themes of research:

In the field of Social and Educational Psychology focusing on stereotypes and stereotyping and she does not have a previous publication.