Abstract
Learning progress monitoring (LPM) is an effective tool for teachers to improve students’ performance by systematically and quickly responding to achievement data. However, studies show that in-service and preservice teachers often have difficulties using LPM because of lacking graph literacy, especially with high data ambiguity. The present study examines whether (a) preservice teachers are biased by gender stereotypes when predicting students’ performance based on progress data, (b) the preservice teachers’ gender affected their predictions differentially depending on student gender, and (c) the insertion of a trend line or lowered data variability diminishes the gender bias in predictions. N = 134 preservice teachers received 16 experimental student vignettes online via the internet in random order which depicted the learning progress of boys and girls in oral reading fluency assessment over a period of 11 weeks. Half of the participants were presented with progress data accompanied by a trend line, the other half received progress data only. Results evidenced that preservice teachers were prone to a gender bias favoring girls. The gender bias was attenuated when a trend line was presented or when data variability was low, with male participants benefiting more from the trend line, and female participants benefiting more from low data variability. The adaptation of international training programs to enhance graph literacy and to diminish gender stereotyping in German teachers is recommendable.
Die Prädiktion der Leseleistung von Schülerinnen und Schülern im Rahmen von Lernverlaufsdiagnostik: Evidenz für einen Geschlechter-Bias

Zusammenfassung

Schlagworte
Lernverlaufsdiagnostik, graph literacy, Lesekompetenz, Datenambiguität, Geschlechter-Bias, Stereotype

1. Introduction

Learning progress monitoring (LPM) is an increasingly popular method teachers can use to track how students are progressing in basic academic areas such as math, reading, spelling, or writing. It usually entails using quick, frequently ad-
ministered standardized measures to assess students’ progress towards a long-term goal (Deno, 1985). Although LPM offers educators a sound foundation for making evidence-based decisions about whether students need help, instructions have to be modified, or teaching goals have to be changed (Hosp et al., 2007), the effectiveness of LPM for improving student achievement appears to be mixed (Ardoin et al., 2013; Christ et al., 2012; Stecker et al., 2005). One reason why LPM alone does not lead to better teaching, which in turn could improve student achievement, is that teachers may have difficulties using progress data to make reasonable predictions about students’ future achievement (Van den Bosch et al., 2017). Prediction of student achievement within LPM may be meaningful when teachers want to identify students at risk (Strathmann et al., 2010), or when ineffective interventions should be identified before the end of the school semester (Van Norman & Parker, 2018). Actually, the interpretation of progress data is affected by several factors, for instance data variability (Klapproth, 2018), or the presence versus absence of a trend line (Van Norman et al., 2013). Although several studies showed that girls are expected to have higher competence in reading than boys although performance of both is actually the same (e.g., Plante et al., 2013), little is known about whether teachers are also biased by gender stereotypes when interpreting students’ learning progress. In the present study in Germany, we sought to examine whether pre-service teachers are prone to a gender bias when predicting students’ achievement based on their learning progress. Additionally, we investigated whether the pre-service teachers’ gender affected their predictions differentially depending on student gender, and we finally assessed two factors that might lower a gender bias in predictions.

2. Theoretical and Empirical Background

2.1 Learning Progress Monitoring

Learning progress monitoring (LPM) is a broad term for progress monitoring systems designed to track the learning achievement progress of students within an academic domain. The measurement of the academic progress of an individual student or a group of students (e.g., a school class) usually serves the purpose of the evaluation of instructions for those students in order to judge whether students have reached a learning goal or instead need additional support (Deno, 2003). LPM involves the frequent administration of short measures of performance in an academic area of interest (e.g., reading, writing, or mathematics). The student’s achievement usually is visualized as a graph representing the achievement trajectory of the student over a predetermined period of time. Teachers may inspect the graph to gauge whether the instructional program is effective, whether a student has mastered a topic, or whether a student is expected to perform according to prespecified teaching goals. The latter judgment needs the extrapolation of future
achievements of students from their current LPM progress graphs. This is necessary when teachers want to decide whether or not they should change instruction (Good & Shinn, 1990) to achieve long-term goals in education (Shinn et al., 1989), or when they want to identify students at risk (Hosp et al., 2007; Strathmann et al., 2010). By systematically responding to achievement data with instructional adaptations, LPM could be an effective tool for teachers to improve students’ performance. However, teachers often have difficulties to improve their instruction when using LPM (Ardoin et al., 2013; Christ et al., 2012; Stecker et al., 2005). One possible reason is their lack of skills to correctly read and interpret data (Van den Bosch et al., 2017, 2019). Even the use of computer software aiming to help teachers in interpreting the graphs by providing statistics like, for example, the linear trend of the graph, does not result in adequate understanding of the progress data. Teachers often do not use these statistics, even if they have gained experiences with LPM (Espin et al., 2017). Instead, they more frequently rely on their visual inspection of the data (Van Norman et al., 2013). Visual inspection, however, is prone to error (Klapproth, 2006), and, consequently, teachers make mistakes when interpreting visualized progress data.

2.2 Understanding Progress Data

The capability of reading and interpreting progress data depicted as a graph has been called graph literacy (Friel et al., 2001). Graph literacy is defined as an individual’s ability to derive meaning from a graph and includes three components: the ability to extract data from the graph (“read the data”), the ability to interpret visualized data (“read between the data”), and the ability to evaluate data (“read beyond the data”). This approach has also been applied to data obtained from LPM (e.g., Van den Bosch et al., 2017). In this context, the ability to read data means that teachers are able to correctly describe the scores and the growth rate of a graph. Reading between data corresponds with interpreting the relation between the actual growth rate and the expected growth rate, whereas reading beyond data means linking data to the instructional context.

All these components of graph literacy matter when teachers use LPM data to adjust their instruction. However, even if teachers can read beyond data, effects on student achievement might still be absent if teachers do not implement modifications in their instructional programs (Stecker et al., 2005).

Graph literacy can be fostered by providing visual aids in the graphical depiction of progress data, which in turn would make judgments in LPM more accurate. A common tool in LPM to facilitate visual judgments of progress data is the presentation of a trend line. This is calculated by ordinary least squares regression such that it fits best through all data points (Van Norman & Christ, 2016). However, whereas visual analysis supplemented with trend lines tends to outperform pure visual analysis, it is still prone to incorrect interpretations (Nelson et al., 2017; Van Norman & Christ, 2016).
2.3 Biases in Interpreting and Predicting Progress Data

Several sources of bias have been identified when teachers use LPM data to interpret or predict student achievement. For instance, LPM data patterns can be ambiguous (Deno, 2013). Ambiguity can occur when features of the graph do not consistently speak in favor or against the existence of a positive or negative trend, for instance, when data cyclicity is present (Brossart et al., 2006). It may also be increased when achievement scores are contaminated by construct-irrelevant variance (Christ et al., 2012), which may make it difficult to visually identify a trend in the data. With LPM data in particular, teachers have difficulty to accurately estimate the rate of improvement when progress data are highly variable (Klapproth, 2018; Nelson et al., 2017; Tindal et al., 1983) or include extreme values (Klapproth, 2018; Nelson et al., 2017). However, if variability in the data is beyond what is generally expected, the confidence that the data allow for judgment and prediction may decrease (Horner & Odom, 2014).

2.4 Is There a Gender Bias in LPM?

Reading skills are essential for individuals to gain an understanding across subject domains in school and hence are an important predictor of their future socioeconomic status (e.g., Ritchie & Bates, 2013). Oral reading fluency directly measures phonological segmentation and recoding as well as fast word recognition and serves as an indicator of overall reading competence (Fuchs et al., 2001). It is one of the skills that is predominantly assessed by LPM (Tindal, 2013). Although boys have consistently been found to obtain lower scores in reading competence than girls across different countries and languages (e.g., Chiu & McBridge-Chang, 2006; McElvany et al., 2017; Mullis et al., 2017), little is known about gender differences in LPM oral reading fluency. Some studies, however, indicated that girls obtained higher scores in curriculum-based measurement of reading fluency than boys (e.g., Kranzler et al., 1999; Yeo et al., 2011).

Boys’ lower attainments in reading competence are discussed as being partly a result of a bias due to teachers’ gender stereotypes according to which reading is suited more for girls rather than for boys (e.g., Lorenz et al., 2016). Teachers’ gender stereotypes adhere to assumptions about student motivation and student working habits (Glock & Kleen, 2017; Jussim & Eccles, 1992).

Stereotypes can be defined as “shared [...] beliefs about traits that are characteristic of members of a social category” (Greenwald & Banaji, 1995, p. 14). Thus, they result from a categorization of individuals into groups according to presumed commonalities. On the one hand, stereotypes may serve as schemas that facilitate social interactions with unknown individuals. On the other hand, stereotypes can also serve as a social norm affecting expectations and behavior toward members of a particular social group (e.g., Schneider, 2004). When a target is categorized
as belonging to a given group, these beliefs are activated (e.g., Macrae et al., 1994; Van Knippenberg et al., 1999).

2.5 Stereotypes Might Reduce Uncertainty in Judgments

According to dual process theories of social judgment (e.g., Fiske et al., 1999; Fiske & Neuberg, 1990), people’s judgments of other people occur on a continuum of two concurrent processes. On one end of the continuum, judgments are automatic, quick, effortless, and follow social categories (“girl”, “boy”, “poor”, “rich”, etc.); on the other end of the continuum, a process is assumed that is slow, effortful, and voluntarily initiated to overcome and enrich the automatic process by integrating all available and relevant information about the to-be-judged person. However, the use of stereotypic categories may be facilitated if there are salient attributes of a person that fit a certain stereotype (Fiske & Neuberg, 1990), if the motivation or accountability to correctly judge the person is low (Tetlock & Kim, 1987), or if a judging person is uncertain about the correct interpretation of someone’s behavior (Campbell, 1967). For instance, if a teacher wants to judge the reading fluency progress of a student who shows ups and downs with no clear linear trend visible in the data, the teacher might be uncertain about the “true” achievement development of the student. Social stereotypical knowledge can act as an interpretive frame when evaluating members of a social group (Hicklin & Wedell, 2005; Hilton & von Hippel, 1996). As a consequence, evaluations of individual group members might be biased in the direction of the knowledge we have about the group as a whole (Spieß & Bekkering, 2020). The more complex or ambiguous a task is, the more likely is the application of stereotypes (Darley & Gross, 1983; Kunda & Sherman-Williams, 1993). Hence, when teachers also know about the student’s gender, they might judge a girl performing higher or a boy performing lower on reading skills, respectively. Thus, the gender of the student could provide information, which in turn might reduce uncertainty and ambiguity in judgments (Kossak & Johnson, 2001).

Evidence that coping with ambiguity might result in the use of stereotypes stems, for example, from studies in the medical context where researchers found that students who had a tendency to perceive ambiguous situations as threatening were more likely to make use of stereotypes about patients (Geller et al., 1990) or were more accepting the use of stereotyping in everyday situations (Valutis, 2015) than those who were, comparatively, highly tolerant of ambiguity. Further evidence comes from studies in perception. For instance, both Bar (2003) and Correll et al. (2015) have shown that stereotypes even guided visual processing when the perceived object was ambiguous.
2.6 Teacher-Student Gender Interaction

Student assessment might also depend on teachers’ gender. According to the gender-stereotypic model (Martin & Marsh, 2005), boys achieve higher scores in classes taught by males, and girls are better when instructed by female teachers. Teachers might favor students of their own gender (Holmlund & Sund, 2008). Some studies support this hypothesis. For instance, 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. Other studies found that female teachers generally evaluated both boys and girls more positively than male teachers (Ehrenberg et al., 1995), or did not find a teacher-gender bias in assessing male and female students (e.g., Driessen, 2007). In a large-scale study conducted by Neugebauer et al. (2011) on data from IGLU-E (an expansion of the Progress in International Reading Literacy Study, PIRLS, in Germany), no same-gender effect on student outcomes was obtained. Boys did not benefit from male teachers and girls did not benefit from female teachers. However, in a recent experimental study, Klapproth and Fischer (2019) found that female participants evaluated students with more caution and precision and were less optimistic than male participants when recommending them to the tracks of secondary school. Male participants, however, showed a preference for boys, indicating a same-gender bias.

2.7 Research Questions and Hypotheses

To the authors’ knowledge, this is the first experimental study that examined effects of social stereotypes on the interpretation of progress data within the frame of LPM. The rationale of the present study was as follows. Since teachers (or preservice teachers) may have stereotypical beliefs about girls and boys (Muntoni & Retelsdorf, 2018), they would presumably expect higher achievements in oral reading fluency for girls than for boys. Additionally, since ambiguity in achievement data should raise the likelihood of the use of stereotypes (Correll et al., 2015; Kossak & Johnson, 2001), we investigated whether data ambiguity would affect achievement prediction, and whether a gender bias in achievement prediction would be smaller when data ambiguity was reduced. Thus, we aimed at testing the following hypotheses:

1) We hypothesized that within LPM predicted achievements in oral reading fluency would be higher for girls than for boys, even when girls and boys show the same progress data.

2) We assumed that when data ambiguity is low, students’ gender should affect predictions of achievement to a lesser degree. We sought to realize reduction of data ambiguity by (a) the insertion of trend lines in the graphical visualization of progress data, and (b) lowering data variability. Hence, we tested two interac-
tion hypotheses: the trend line × student gender interaction and the data variability × student gender interaction.

Finally, we examined whether the gender of the participants contributed differentially to oral reading fluency predictions for male and female students, that is, whether there is an interaction between participants’ gender and students’ gender with respect to their reading fluency predictions. However, with respect to the mixed results reported in previous research we abstained from stating a specific hypothesis.

3. Method
3.1 Participants

Based on previous investigations (cf. Klapproth, 2018, for the effect of data variability, and Klapproth & Fischer, 2019, for the effect of student and participant gender), we expected an average medium to large effect ($f = .30$, which translates to $\eta^2 = .08$) of the independent variables on the participants’ predictions. We conducted an a priori power analysis for an analysis of variance (ANOVA) with repeated measures and four groups, using G*Power 3.1 (Faul et al., 2009). When prespecifying $f = .30$, $\alpha = .05$, $1-\beta = .90$, and the correlation among repeated measures with $r = .50$, power analysis yielded a total sample size of $N = 44$, which we deemed to be the minimum sample size.

Announcements on social media platforms were used to recruit a sample of preservice teachers enrolled in a primary or secondary school teacher education program at various German universities. A sample of 170 preservice teachers had been recruited for participation in the experiment. However, we excluded 31 participants from subsequent analyses because they dropped out from the study before they were presented with student vignettes or after they received the first vignette but did not continue to participate. Furthermore, five participants were not considered for analyses because they left the study during ongoing presentation of the vignettes (most of them dropped out before having received half of the vignettes). In total, $n = 36$ participants were dropouts and were excluded from analyses. From the remaining participants, data were complete.

Thus, a total of $N = 134$ preservice teachers (57.5% female, $M_{age} = 24.4$ years, $SD = 2.5$) participated in the study, which were much more than recommended by power analysis. However, we chose the larger sample because estimates of parameters are more precise with larger than with smaller samples. Most of the participants conducted their study program in universities located in North-Rhine Westphalia (35.8%) or Bavaria (32.1%), followed by participants from Berlin (11.9%), Hesse (3.7%), and remaining states (16.5%). They had studied on average for 7.5 semesters ($SD = 3.5$). No participant reported previous experiences with LPM. Table 1 shows the frequency of the participants’ gender across the teacher education
programs, Table 2 shows participants’ background information across the experimental groups (trend line versus no trend line).

### Table 1: Frequency of the Participants’ Gender Across Teacher Education Programs

<table>
<thead>
<tr>
<th>Participant gender</th>
<th>Teacher education program</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Primary</td>
</tr>
<tr>
<td>Male</td>
<td>19</td>
</tr>
<tr>
<td>Female</td>
<td>34</td>
</tr>
<tr>
<td>Sum</td>
<td>53</td>
</tr>
</tbody>
</table>

### Table 2: Participants’ Background Information Across the Groups

<table>
<thead>
<tr>
<th>Group</th>
<th>Participant gender</th>
<th>Teacher education program</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male</td>
<td>Female</td>
</tr>
<tr>
<td>Trend line</td>
<td>37</td>
<td>30</td>
</tr>
<tr>
<td>No trend line</td>
<td>20</td>
<td>47</td>
</tr>
<tr>
<td>Sum</td>
<td>57</td>
<td>77</td>
</tr>
</tbody>
</table>

Whereas the participants’ gender and their teacher education program were not significantly interrelated, $\chi^2(1) = 1.61$, $p = .205$, there was a significant relationship between the participants’ assignments to either experimental group (trend line versus no trend line) and both their gender, $\chi^2(1) = 8.82$, $p = .003$, and their education, $\chi^2(1) = 5.28$, $p = .022$. Male participants and participants enrolled in secondary education programs were more frequently assigned to the trend line group than female participants and participants enrolled in primary education programs.

### 3.2 Materials and Procedure

The experiment was conducted online on www.soscisurvey.de. The participants could perform the tasks of the experiment on a computer or any other device that was connected to the internet. The study was open for 14 days.

At the beginning of the experiment, the participants were welcomed and were given a brief introduction into learning progress monitoring based on oral reading fluency that had been obtained over a period of time and is depicted visually as a learning development graph. After that, the participants received 16 experimental student vignettes in random order. The vignettes depicted the learning progress of boys and girls in oral reading fluency assessment over a period of 11 weeks. The experimental vignettes were supplemented by the presentation of six distractor vignettes, which were created arbitrarily and served to camouflage the independent variables of the study, because knowledge of the independent variables may af-
fect the participants’ responses (Klein, et al., 2012). They mimicked results from curriculum-based oral reading fluency (R-CBM) tests, one of the most common in LPM, which is frequently applied in schools (Ardoin et al., 2013) and which has often been the focus of research (Reschly et al., 2009). Each vignette displayed a graph of student progress data. In line with the R-CBM, the y-axis represented the number of words read correctly (WRC), which could range from zero to 140 (Hosp et al., 2007). The x-axis represented the school week that the test was given and ranged from Week 1 to Week 17, thus spanning a typical school semester. WRC scores were given only for the first 11 weeks. The participants were told that they were to examine the development of oral reading fluency scores over a period of 11 weeks in order to make a prediction of the oral reading fluency score presumably obtained at Week 17. Half of the participants were presented with progress data accompanied by a trend line aiming at facilitating interpretation of the data, whereas the other half of participants were provided with progress data only. The trend line, depicted as a thin (1 mm) dotted black line from Week 1 to Week 11, reflected the linear trend that was estimated by ordinary least squares linear regression analysis. Hence, the occurrence of a trend line was a between-subjects variable. The participants were randomly assigned to both groups.

Independent within-subjects variables were the students’ gender and the amount of data variability. Each participant received two sets of different vignettes: eight vignettes, differing in data variability and labeled with a male name, and the same eight vignettes, labeled with a female name. The names used for this study were common for either male or female German students. Common names were chosen to prevent the activation of concepts like a certain socioeconomic background by rarely used names that are especially common in certain social and economic milieus (Gerhards, 2010).

Data variability was either low or high. In the literature, variability of learning progress graphs is often quantified as the standard error of the estimate (SEE), which is defined as the average magnitude of residuals around the trend line obtained from linear regression. According to the literature (e.g., Ardoin & Christ, 2009; Van Norman & Christ, 2016), SEE values usually vary between 5 and 20, where 5 means very low, and 20 very high variability. In our study, the SEE of progress data with high variability was 10.0, whereas in the low variability condition SEE was 5.0.

We used a mixture between a within-subjects design and a between-subjects design because a complete crossing of all variables (meaning that each participant would have received both vignettes with trend lines and without trend lines) would have lacked ecological validity, as in reality teachers are seldom presented with different types of graphs in LPM.

Both low data variability and the presence of a trend line were supposed to reduce the ambiguity of the data. The dependent variable was the participants’ prediction of the WRC score in Week 17.

Figure 1 shows four experimental vignettes, either with a trend line or without a trend line for both low and high data variability.
3.3 Data Analyses

The constant error (i.e., the difference between participants’ predictions and the prediction of a linear regression equation) served as a measure of bias (cf. Makridakis et al., 1998). To calculate the constant errors, ordinary least squares linear regression analysis of the given data was used to predict the WRC score at Week 17 in each of the experimental vignettes. The result of the linear regression analysis was then subtracted from participants’ prediction. A positive constant error indicates that participants predicted higher achievement relative to the linear regression equation (i.e., positive bias), while a negative constant error indicates that participants predicted lower achievement relative to the linear regression equation (i.e., negative bias). A repeated measures ANOVA with trend line as between-subject factor, and student gender and data variability as within-subject factors was then conducted. Additionally, the participants’ gender was included in our analysis as a further independent variable. With that analysis we were able to estimate the extent to which student gender, participant gender, data variability, and the presence or absence of a trend line affected the constant error of prediction. In order to assess whether the participants’ teacher education program affected their predictions, we additionally ran correlational analyses between the teacher educa-
tion program (primary vs. secondary education) and the constant error of predictions obtained from all realized conditions.

4. Results

Table 3 displays the means, standard errors, and the 95%-confidence intervals of the constant errors of the predicted WRC scores of each condition. In Table 4, these descriptive statistics are given for each independent variable.

Table 3: Means, Standard Errors (in Parentheses), and 95%-Confidence Intervals (in Brackets) of the Constant Errors of Predicted WRC Scores

<table>
<thead>
<tr>
<th>Trend line</th>
<th>Participant gender</th>
<th>Variability</th>
<th>Student gender</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>No</td>
<td>Male</td>
<td>−0.38 (3.05)</td>
<td>−6.45 (2.32)</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>2.36 (1.76)</td>
<td>−5.85 (1.52)</td>
</tr>
<tr>
<td>Yes</td>
<td>Male</td>
<td>−0.97 (1.25)</td>
<td>−3.26 (0.77)</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>2.88 (2.70)</td>
<td>−3.64 (1.78)</td>
</tr>
</tbody>
</table>

Table 4: Means, Standard Errors, and 95%-Confidence Intervals of the Constant Errors of Predicted WRC Scores per Independent Variable

<table>
<thead>
<tr>
<th>IV</th>
<th>M</th>
<th>SE</th>
<th>CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trend line</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>1.87</td>
<td>1.38</td>
<td>−0.85, 4.59</td>
</tr>
<tr>
<td>Yes</td>
<td>0.16</td>
<td>1.27</td>
<td>−2.35, 2.66</td>
</tr>
<tr>
<td>Participant gender</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>−0.07</td>
<td>1.43</td>
<td>−2.89, 2.76</td>
</tr>
<tr>
<td>Female</td>
<td>2.09</td>
<td>1.20</td>
<td>−0.29, 4.47</td>
</tr>
<tr>
<td>Student gender</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>−1.91</td>
<td>0.87</td>
<td>−3.64, −0.19</td>
</tr>
<tr>
<td>Female</td>
<td>3.94</td>
<td>1.06</td>
<td>1.84, 6.04</td>
</tr>
<tr>
<td>Data variability</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>1.62</td>
<td>1.02</td>
<td>−0.40, 3.64</td>
</tr>
<tr>
<td>High</td>
<td>0.41</td>
<td>0.91</td>
<td>−1.39, 2.20</td>
</tr>
</tbody>
</table>

The results of the ANOVA yielded two significant main effects and four significant interaction effects. All results are displayed in Table 5.
First, there was a significant main effect of student gender, $F(1, 130) = 117.41, p < .001, \eta^2 = .48$. Participants overestimated on average future reading fluency scores for female students ($M = 3.94, SD = 12.27$) compared to male students ($M = –1.91, SD = 10.07$), $d = 0.52$. Moreover, there was a significant main effect of data variability, $F(1, 130) = 6.25, p = .014, \eta^2 = .05$. When data variability was low, constant error was on average larger ($M = 1.62, SD = 11.81$) than with high data variability ($M = 0.41, SD = 10.53$), $d = 0.11$.

In addition to these main effects, the ANOVA revealed a significant student gender $\times$ trend line interaction, $F(1, 130) = 31.84, p < .001, \eta^2 = .20$. The difference in constant error between boys and girls was dependent on the presence or absence of a trend line. Simple effects test showed that when a trend line was present, the difference between boys and girls was small, $M_{\text{Diff}} = 2.80, d = 0.33$, yet significant, $F(1, 130) = 14.70, p < .001, \eta^2 = .10$, but when no trend line was shown, the difference was quite large, $M_{\text{Diff}} = 8.90, d = 0.97, F(1, 130) = 125.38, p < .001, \eta^2 = .49$. Note, that $p$ values were already adjusted according to Bonferroni.

There was also a significant student gender $\times$ data variability interaction, $F(1, 130) = 70.00, p < .001, \eta^2 = .35$. This interaction effect means that the difference in constant error between boys and girls was affected by data variability. When data variability was high, the difference between boys and girls was larger, $M_{\text{Diff}} = 10.40, d = 0.98, F(1, 130) = 128.51, p < .001, \eta^2 = .50$, than when data variability was low, $M_{\text{Diff}} = 1.30, d = 0.19, F(1, 130) = 5.06, p = .026, \eta^2 = .35$. However, there was no significant student gender $\times$ participant gender interaction, $F(1, 130) = 0.71, p = .401, \eta^2 = .01$. 

### Table 5: Results of the Analysis of Variance

<table>
<thead>
<tr>
<th>Effect</th>
<th>$F$ ratio</th>
<th>df</th>
<th>$p$</th>
<th>$\eta^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student gender</td>
<td>117.41</td>
<td>1, 130</td>
<td>&lt; .001</td>
<td>.48</td>
</tr>
<tr>
<td>Variability</td>
<td>6.25</td>
<td>1, 130</td>
<td>.014</td>
<td>.05</td>
</tr>
<tr>
<td>Trend line</td>
<td>0.84</td>
<td>1, 130</td>
<td>.360</td>
<td>.01</td>
</tr>
<tr>
<td>Participant gender</td>
<td>1.33</td>
<td>1, 130</td>
<td>.251</td>
<td>.01</td>
</tr>
<tr>
<td>Student gender $\times$ trend line</td>
<td>31.84</td>
<td>1, 130</td>
<td>&lt; .001</td>
<td>.20</td>
</tr>
<tr>
<td>Student gender $\times$ variability</td>
<td>70.00</td>
<td>1, 130</td>
<td>&lt; .001</td>
<td>.35</td>
</tr>
<tr>
<td>Student gender $\times$ participant gender</td>
<td>0.71</td>
<td>1, 130</td>
<td>.401</td>
<td>.01</td>
</tr>
<tr>
<td>Variability $\times$ trend line</td>
<td>1.03</td>
<td>1, 130</td>
<td>.313</td>
<td>.01</td>
</tr>
<tr>
<td>Variability $\times$ participant gender</td>
<td>0.53</td>
<td>1, 130</td>
<td>.470</td>
<td>.00</td>
</tr>
<tr>
<td>Trend line $\times$ participant gender</td>
<td>0.39</td>
<td>1, 130</td>
<td>.536</td>
<td>.00</td>
</tr>
<tr>
<td>Student gender $\times$ trend line $\times$ participant gender</td>
<td>4.36</td>
<td>1, 130</td>
<td>.039</td>
<td>.03</td>
</tr>
<tr>
<td>Student gender $\times$ variability $\times$ trend line</td>
<td>2.60</td>
<td>1, 130</td>
<td>.110</td>
<td>.02</td>
</tr>
<tr>
<td>Student gender $\times$ variability $\times$ participant gender</td>
<td>5.18</td>
<td>1, 130</td>
<td>.025</td>
<td>.04</td>
</tr>
<tr>
<td>Variability $\times$ trend line $\times$ participant gender</td>
<td>0.57</td>
<td>1, 130</td>
<td>.453</td>
<td>.00</td>
</tr>
<tr>
<td>Student gender $\times$ variability $\times$ trend line $\times$ participant gender</td>
<td>0.08</td>
<td>1, 130</td>
<td>.777</td>
<td>.00</td>
</tr>
</tbody>
</table>
In addition to these two-way interactions, the ANOVA yielded two significant three-way interactions, which are displayed in Figures 2 and 3. The first is the student gender × participant gender × trend line interaction, $F(1, 130) = 4.36$, $p = .039$, $\eta^2 = .03$. This interaction means that the two-way student gender × trend line interaction was dependent on the participant gender. As can be seen in Figure 2, the presence of a trend line attenuated the gender bias to a larger degree, when the participants were male rather than female. Simple interaction tests revealed that when participants were female, the student gender × trend line interaction was absent, $F(1, 75) = 0.03$, $p = .865$, $\eta^2 = .00$. However, with male participants, the student gender × trend line interaction was present, but not significant, $F(1, 55) = 3.00$, $p = .089$, $\eta^2 = .05$. To get more insight into the interaction, we conducted simple effects tests. With male participants, the difference between boys and girls was still significant when a trend line was present, $F(1, 36) = 7.53$, $p = .009$, $\eta^2 = .17$, yet without a trend line, the effect was larger, $F(1, 36) = 40.50$, $p < .001$, $\eta^2 = .68$.

The second three-way interaction is the student gender × participant gender × data variability interaction, $F(1, 130) = 5.18$, $p = .025$, $\eta^2 = .04$. According to Figure 3, this interaction revealed that the effect of data variability on reducing differences in constant error between girls and boys was larger for female than for male participants. Simple interaction tests showed that when participants were female, the student gender × variability interaction was significant, $F(1, 75) = 48.05$, $p < .001$, $\eta^2 = .39$. With male participants, the student gender × variance interaction was also significant, $F(1, 75) = 37.13$, $p < .001$, $\eta^2 = .40$. Simple effects tests addition-
ally shed more light on the interactions and confirmed the visual impression. With male participants, the difference between boys and girls was significant when data variability was low, $F(1, 55) = 8.35, p = .006, \eta^2 = .13$, and high, $F(1, 55) = 88.26, p < .001, \eta^2 = .62$. With female participants, however, the difference between boys and girls was only significant when data variability was high, $F(1, 75) = 74.06, p < .001, \eta^2 = .50$.

Figure 3: Illustration of the Student Gender × Participant Gender × Variability Interaction.

Note. The y-axis depicts the constant error, the x-axis depicts whether data variability was low or high.

In order to assess whether the participants’ teacher education program affected the predictions of the participants, we additionally conducted correlational analyses between the teacher education program (primary vs. secondary education) and the constant error of predictions obtained from all realized combinations of variables that were used to construct the student vignettes. We obtained positive, but weak and insignificant associations, with all $r_s < .18$ and all $p_s > .154$.

5. Discussion

This study evidences that within LPM preservice teachers were prone to a gender bias. When presented with a graph depicting the development of reading fluency of both boys and girls over a period of 11 weeks, the predictions of reading fluency in Week 17 were clearly higher for girls than for boys. In particular, preservice teachers predicted on average a higher score for girls compared to boys, which was 0.52 times the standard deviation of the distribution of constant errors obtained in this study. Therefore, we could find support for our hypothesis that preservice teachers...
were stereotyping boys and girls when making judgments about their reading fluency development.

Moreover, as predicted, the gender bias was attenuated when a trend line was presented or when data variability was low. We assumed that when we would reduce ambiguity of the data and hence increased preservice teachers’ subjective certainty about the trend inherent in the data, the participants would be less prone to use stereotypical beliefs about reading skills of boys and girls for making judgments. Both the student gender × trend line interaction and the student gender × data variability interaction supported our hypothesis. Notably, the sizes of the interaction effects were quite large ($\eta^2 = .20$, and $\eta^2 = .35$, respectively), which indicates that the reduction of the gender bias was quite effective.

Furthermore, we could not find a same-gender bias in male participants in our data. Both male and female participants favored girls over boys with respect to predicted reading fluency scores. However, the participants’ preference for girls reduced when ambiguity of the data was decreased or when a trend line was present. These effects, albeit small, were affected by the participants’ gender. Male participants benefitted more from a trend line than female participants, as males made similar predictions for both boys and girls and with only small constant errors. When data ambiguity was lowered by reducing data variability, female participants seemed to benefit more than males, since low data variability let the gender bias almost completely disappear in female participants.

Why did a gender bias in LPM occur, why was it so strong, and why did different ways of decreasing ambiguity result in different reductions of gender bias between male and female participants? In LPM, teachers or preservice teachers use the graphical depiction of a student’s achievement development to make several decisions. They could use the achievement trajectory to judge whether lessons had been effective, whether a student needs additional support, or whether their instruction has to be adjusted in order to better fit the needs of the students. Whatever the decisions to be made are, teachers have to understand the data depicted as a kind of a learning curve. Several studies (e.g., Espin et al., 2017; Van den Bosch et al., 2017; Van Norman et al., 2013) have called into question that teachers have adequate skills to fully understand what is presented in achievement trajectories, especially when relations between parameters of the curve have to be determined (e.g., how well students achieved their achievement goals) or when conclusions have to be drawn with respect to educational interventions. Since most of the learning progress data are variable (Van Norman & Christ, 2016), teachers necessarily would encounter difficulties in detecting the trend of the data, which in turn would make it difficult for them to both “read between the data” and “read beyond the data” (Friel et al., 2001).

Data ambiguity might encourage teachers to look for additional information that would facilitate their judgements. However, when there is no source of additional information, or when decisions have to be reached quickly, stereotypical beliefs are likely to guide expectations about those students (Kunda & Sherman-Williams, 1993). This might be even more the case if teachers have to predict stu-
students’ achievements based on what they see on a chart showing a student’s learning progress. Given current knowledge, predictions are always uncertain since there are multiple possible future states of nature (Stewart, 2000). If uncertainty cannot be reduced by the integration of additional information, heuristic and category-based processing of information is likely to occur (Fiske & Neuberg, 1990). Moreover, when teachers (or preservice teachers) are to make a prediction about a student’s future achievement, they are likely to use expectations they have about that student to guide their predictions. Compared to judgments of current states, predictions of future states might be even more susceptible for stereotypical expectation, like, for instance, the expectation that on average girls would show better achievements than boys.

The amount of gender bias obtained in this study might arise from two sources, which are the prediction task, and the absence of additional prediction-relevant information about the students. In prediction tasks, uncertainty is usually higher than in tasks where the judgment concerns actual instead of future events, since predictions involve the future, and the future is unknown. In addition, the prediction task in this study can be characterized as being complex, since it required the participants to both differentiate the information presented with regard to the different interpretations they allowed, and to integrate the information to a single judgment (cf. Tetlock & Kim, 1987). Complex tasks would elicit the use of stereotypes more likely than simple tasks would do (Bodenhausen & Lichtenstein, 1987). Therefore, the prediction task itself might have facilitated the use of stereotypical beliefs. Moreover, the participants were forced to make predictions without having access to information relevant for preservice teachers’ predictions besides what was presented in the graph. A lack of individuating information, however, corresponds with an increased use of stereotypes (Fiske et al., 1999; Macrae et al., 1993).

Apparently, male and female participants processed data ambiguity and its reduction in different ways. When presented with a trend line, male participants were less prone to a gender bias compared to male participants who received vignettes without a trend line. With female participants, however, the effect of the trend line on gender bias was smaller than for male participants, and they made more accurate predictions when data variability was low. The differences between male and female participants’ predictions might be caused by the differences between the realized ways of reducing data ambiguity. When data variability was low rather than high, reduction of ambiguity was produced by the students themselves. That is, students showing constant progress of achievement with only minor fluctuations might be regarded as stable and therefore easy to predict. Predictions based on the trend line, however, could neglect the ups and downs in the growth curve, such that predictions for both stable and unstable students could be made with similar certainty, provided that the trend line depicted the “real” achievement development.

It appears that female participants seemed to be more skeptical against the trend line, whereas male participants did more than female participants rely on external statistical aids.
6. Limitations

People use their stereotypical beliefs about social groups as a basis for judgment and prediction whenever they lack the desire or the ability to engage in more extensive processing of information (Bodenhausen, 1993). Such a lack of desire or ability may be common under most everyday life circumstances (Fiske & Neuberg, 1990), but not always in educational decisions made by teachers. Therefore, it might be a limitation of this study that the participants made predictions about virtual students, not real students. Hence, because of the artificial character of the experiment, the participants might not have felt highly accountable for their predictions (Tetlock & Kim, 1987) and therefore engaged less in effortful information processing than they would do in real school settings.

In addition, teachers in real classrooms have access to a variety of information that could bolster their decisions and predictions about students’ achievements. We could therefore expect that predictions of achievements with real students might be less affected by stereotypes and category-based information processing than in this study.

Moreover, the participants were preservice instead of in-service teachers. Although there is ample evidence that both make comparable judgments (Glock & Karbach, 2015; Mertler, 2004), preservice teachers usually lack the experience in-service teachers have when students are to be judged, and may use information about students differently (e.g., Sabers et al., 1991).

Furthermore, effects of the independent variables used in this study could be affected by the fact that the participants were inexperienced with LPM. It is possible that more experienced participants would have made less errors in prediction.

A further limitation might be that we did not establish a baseline condition where the gender of the students was not announced. This baseline condition would have shown how precise predictions are when the participants were unaffected by student gender.

Finally, we did not apply some sort of manipulation check. For instance, we could have checked for whether the participants actually had attended to the student names that were attached to the graphs by asking them at the end of the study. Manipulation checks may help figure out if participants noticed the treatment at all. However, making the manipulation check at the end of the study may compromise its validity, since the participants may not remember what they were feeling or thinking during the study (Hauser et al., 2018).

7. Conclusions

The results indicate that preservice teachers were prone to apply gender stereotypes when interpreting and predicting learning progress data. The likelihood of the occurrence of stereotyping was raised when the ambiguity of progress data was
high. Therefore, it can be concluded that caution is warranted when (preservice)
teachers base their interventions on learning progress data that is highly ambigu-
ous and not supported by visual or statistical aids.

Based on the results obtained, we recommend training of graph literacy (Friel
et al., 2001) for both preservice and in-service teachers. There are a few studies on
training programs for teachers in regard to the interpretation of learning progress
data, showing that training of the interpretation of visualized progress data is pos-
sible and successful (Kennedy et al., 2015; Van den Bosch et al., 2019). These train-
ing programs should be adapted to teacher education programs in Germany.

References

Standard errors associated with progress monitoring outcomes from DIBELS,
AIMSweb, and an experimental passage set. School Psychology Review, 38(2),

systematic review and summarization of the recommendations and research sur-
rounding Curriculum-Based Measurement of oral reading fluency (CBM-R) deci-
jsp.2012.09.004

Bar, M. (2003). A cortical mechanism for triggering top-down facilitation in visual ob-
org/10.1162/089892903321662976

Bodenhausen, G. V. (1993). Emotions, arousal, and stereotypic judgements: A heuristic
model of affect and stereotyping. In D. M. Mackie & D. L. Hamilton (Eds.), Affect,
B978-0-08-088579-7.50006-5

Bodenhausen, G. V., & Lichtenstein, M. (1987). Social stereotypes and information-
processing strategies: The impact of task complexity. Journal of Personality and

Brossart, D. F., Parker, R. I., Olson, E. A., & Mahadevan, L. (2006). The relation-
ship between visual analysis and five statistical analyses in a simple AB sing-
org/10.1177/0145445503261167

Psychologist, 22(10), 817–829. https://doi.org/10.1037/h0025079

doi.org/10.1207/s1532799xssr1004_1

based measurement of oral reading: Quality of progress monitoring outcomes.
800306

Correll, J., Wittenbrink, B., Crawford, M. T., & Sadler, M. S. (2015). Stereotypic vision:
How stereotypes disambiguate visual stimuli. Journal of Personality and Social

Darley, J. M., & Gross, P. G. (1983). A hypothesis-confirming bias in labeling ef-
org/10.1037/0022-3514.44.1.20


Schneider, D. J. (2004). *The psychology of stereotyping.* Guilford.


