education sciences

MDPI

Article

A Gender Bias in Curriculum-Based Measurement across Content Domains: Insights from a German Study

Florian Klapproth * and Holger von der Lippe

Department of Psychology, Medical School Berlin, 14197 Berlin, Germany;
holger.vonderlippe@medicalschool-berlin.de
* Correspondence: florian.klapproth@medicalschool-berlin.de

Abstract: By immediately responding to achievement progress data, teachers can improve students' performance by using curriculum-based measurement. However, there are studies showing that teachers are prone to make biased judgments about the students providing the data. The present investigation experimentally examined whether pre-service teachers in Germany were biased by the use of gender stereotypes when judging students' achievement derived from progress data. $N = 100$ pre-service teachers received graphs that depicted the development of either oral reading fluency or math achievement of girls and boys over a time interval of 11 weeks. The results obtained confirmed the hypotheses partially. The participants did not favor girls over boys on average. However, they judged achievement in reading to be higher for girls than for boys, and math achievement to be higher for boys than for girls. The results suggest that gender stereotypes (boys are good at math, girls are good at reading) are still prevalent in pre-service teachers.

Keywords: curriculum-based measurement; gender bias; gender stereotypes; experiment; pre-service teachers; Germany

Citation: Klapproth, F.; Lippe, H.v.d. A Gender Bias in Curriculum-Based Measurement across Content Domains: Insights from a German Study. *Educ. Sci.* **2024**, *14*, 76. https://doi.org/10.3390/educsci14010076

Academic Editors: Gavin T. L. Brown and Ian Hay

Received: 16 August 2023
Revised: 10 December 2023
Accepted: 20 December 2023
Published: 9 January 2024

1. Introduction

Teachers are increasingly turning to curriculum-based measurement (CBM) as a tool for monitoring student development in fundamental academic domains like arithmetic, reading, spelling, and writing. In most cases, it comprises employing brief, routinely administered standardized tests to gauge pupils' advancement toward a long-term objective [1]. However, the effectiveness of CBM for raising student achievement appears to be mixed [2–5], despite the fact that it provides teachers with a strong framework for making judgments based on evidence regarding whether students need help, instructions need to be revised, or teaching objectives need to be adjusted [6]. Research has shown that teachers have difficulty using progress data to inform and guide their instruction [7,8]. The interpretation of progress data is impacted by a number of factors, which is one reason why CBM alone does not result in better teaching [9–12]. These factors may be related to the lack of attention teachers devote to relevant aspects of the graph [13], may be connected with characteristics of the progress data itself or its presentation, or may belong to characteristics of the to-be-judged students. For example, refs. [9,11,12] found that high data variability results in relative overestimation of current trends. Ref. [14] could show that the presence of a trend line, which visually depicts the linear component of progress data points, reduces judgment errors. Moreover, ref. [11] demonstrated that pre-service teachers judged progress data of reading fluency obtained from girls more positively than the same data obtained from boys.

The aim of the present study was to contribute to the growing literature showing that in-service and pre-service teachers have difficulty interpreting and making adequate decisions on progress data. In particular, the gender bias obtained by [11] was sought to be replicated with a different task. Furthermore, it was investigated whether a gender bias

would depend on the content domain wherein progress data were obtained. Specifically, it was assumed that although pre-service teachers should on average favor girls over boys, boys should benefit from a math domain, whereas girls should be at an advantage in a reading domain.

1.1. Curriculum-Based Measurement

Curriculum-based measurement is a general term for assessment systems that track a student's progress in learning within a particular academic subject. In order to determine whether students have achieved a learning goal or instead require extra support, it is necessary to frequently evaluate their abilities [15].

CBM entails the frequent administration of brief measures of performance in a chosen academic domain of interest (such as reading, writing, or mathematics). A graph showing the student's learning trajectory over a set time period is typically used to depict the student's achievement. The graph can be used by teachers to assess the efficacy of a lesson plan, a student's mastery of a subject, or whether a student is expected to perform in accordance with pre-set learning objectives. CBM can be a useful tool for teachers to raise students' performance by systematically responding to accomplishment data with instructional adjustments. However, when employing CBM, teachers frequently struggle to enhance their education [2–5,8]. Although CBM graphs are constructed to facilitate teachers' understanding of their students' progress, their comprehension appears to be challenging [8,16]. One probable explanation for teachers' lack of adequate response to the presentation of progress data is their inability to read and understand data accurately [17,18]. Even using computer software designed to help teachers analyze graphs by presenting statistics like the graph's linear trend does not produce an adequate grasp of the progress data. Moreover, teachers frequently do not use these statistics [16]. Instead, they rely on visual assessment of the data more frequently [14]. Visual inspection, on the other hand, is prone to inaccuracy [19,20], and as a result, teachers make errors when evaluating visible progress data. For instance, ref. [12] presented teachers with CBM graphs and assessed their ability to grasp information from them. They found that teachers were prone to ignore the relevant information and to focus on rather marginal details. Similar results were revealed by [13] who examined teachers' eye movements when judging CBM graphs. The quality of data interpretation seems also to depend on the intensive support of teachers by researchers. Without such support, teachers are likely to use CBM data inconsistently and inappropriately [21]. According to [22], merely providing teachers with students' data will not necessarily result in their using it, as long as they do not believe in the importance of these data.

1.2. Origins of CBM

CBM was invented in the 1970s by Deno and Mirkin [23]. It was developed within the field of special education with the aim of allowing teachers to formatively evaluate their instructional programs by successively using probes to test their students' basic academic skills, so that they were able to describe the students' growth in academic achievement. The invention of CBM was followed by a 5-year program of research conducted at the Institute for Research on Learning Disabilities at the University of Minnesota [24]. Since CBM offers a system for monitoring students' attainment of academic goals and evaluating instructional programs, its use has been formalized among school districts in the United States. Today, U.S.-based norms and data management are available on internet platforms, e.g., www.aimsweb.com (accessed on 10 December 2023).

1.3. CBM in Germany

Whereas in Germany, where the present study was run, formative evaluation of the achievement of students has been debated and applied since the 1970s, the particular use of CBM started after the seminal paper from [25] about the history and development of CBM. Since then, research in Germany on CBM flourished, leading to internet platforms that

provide teachers with diagnostic instruments to monitor students' learning progress [26]. However, the relevant tests must be obtained by the respective school or the teacher via providers; they are therefore not made available to the teacher as part of the teaching material that is available anyway. Instruction in diagnostic practice is usually provided via videos or handouts available online. Training programs for teachers do not seem to exist so far. Hence, some authors e.g., [5,8] call for teacher training programs that offer working with assessment data and using these data to modify instruction. Similar to the U.S.—but established there for a long time—there are also approaches to the standardization of learning progress data in Germany (e.g., [27]).

1.4. Different Content Domains

CBM was developed initially to help teachers at the primary school level increase the achievement of students struggling to learn basic skills in reading, writing, and arithmetic [15]. Since many students struggle with reading [28], one of the first domains where CBM was applied was reading. Reading CBM often consists of oral reading fluency [6], where students read aloud from a passage for a limited time (e.g., 1 min). To do this, students need to use a variety of different literacy skills, for instance, decoding, vocabulary, and comprehension [29]. Teachers score reading CBM by first counting the total amount of words attempted in 1 min, then counting the total number of errors, and finally subtracting the total number of errors from the total number of words, yielding the words read correctly (WRC) score [6].

As with reading, math is a skill that is essential for success in life. In parallel to reading CBM, math CBM has been developed to assess computation e.g., [30], with the majority of research and development focused on the primary school level [31]. Math CBM is conducted by having students answer computational problems for a certain amount of time (e.g., 2 min; [32]). When scoring math CBM, usually the number of correct digits is used [6].

1.5. Biases in Interpreting Progress Data

When teachers use CBM data to judge student achievement, several causes of bias have been discovered. For instance, when progress data are highly variable, teachers find it challenging to predict the rate of progress accurately [9,12,33]. Ref. [33] could show that teachers tended to overestimate student progress when data variability was high. Similar results were obtained by [9,12]. They could show that pre-service teachers tended to overestimate the current trend. This result could be explained by the participants' proclivity to identify trends in random patterns. Peaks in achievement in progress data with a high level of random variability may imply that those children will perform better than students with the same trend but a lower amount of random variability and hence lower peaks.

1.6. A Gender Bias in CBM

Girls typically outperform boys in reading competency across countries and languages e.g., [34–38]. In math, gender differences are also likely to occur. Boys continue to outperform girls in math, with a wider disparity among the highest achievers, despite gender gaps in job market involvement and educational attainment narrowing [39,40].

Despite the relative advantage of boys in math, several studies in different countries have shown that, on average and across domains, girls outperform boys e.g., [40]. Compared to boys, girls are more likely to display high-achieving developmental patterns e.g., [41].

Differences between boys and girls in achievement are usually reflected in differences in teachers' assessment of their achievement [42]. However, gender-related differences in assessment might also arise from bias that is not based on achievement or skills. Boys' lower reading proficiency levels and their relatively higher math successes are discussed as being partially the product of a bias due to teachers' gender stereotypes, which hold that

reading is more appropriate for girls than for boys e.g., [43], whereas math is better suited for boys than for girls [38,44]. Gender stereotypes among teachers conform to stereotypes about student motivation and working habits [45,46].

1.7. Gender Stereotypes as a Source of Gender Bias

Stereotypes can be defined as "shared [. . .] beliefs about traits that are characteristic of members of a social category" [47] (p. 14). Thus, they are the result of categorizing individuals into groups based on supposed commonality. Stereotypes can serve as norms, affecting expectations and behavior toward members of a particular social group, and as schemas, enhancing social interactions with strangers [48]. These expectations are activated when a target is classified as belonging to a specific group [49,50].

According to dual process theories of social judgment e.g., [51], people's evaluations of other people take place along a continuum of two simultaneous processes. On one end of the spectrum, judgments are, quick, effortless, automatic, and based on social categories (e.g., "girl", "boy", "immigrant"); on the other end of the spectrum, it is assumed that a slow, laborious, voluntarily initiated process will outweigh and enrich the automatic process by incorporating all pertinent information about the subject of the judgment.

If a person exhibits salient characteristics that are consistent with a certain stereotype, or if the judging person is unsure about the proper interpretation of the other person's behavior, the use of stereotypical categories becomes more likely [52,53].

Gender stereotypes in particular cause female students to be seen as less talented than male students in all areas of science, whereas male students are considered inferior to female students in the domain of languages [54].

1.8. Pre-Service Teacher Education in Germany

The German teacher education system has traditionally been divided into two different phases: After graduating from a university with a first state examination (or a master's degree), the second phase of teacher education is the traineeship, which might take up to two years. During this time, the student teacher observes lessons, takes classes on general topics related to teaching in schools, and works as a teacher in one or various schools [55]. There is ample evidence that even in pre-service teachers' stereotypical beliefs about students regarding their ethnicity or gender exist [56,57]. Once established, these stereotypes are likely to affect the judgment of students later during their teaching in school e.g., [58].

1.9. Research Questions and Hypotheses

The following was the justification for the current study. Because pre-service teachers may have different stereotypical expectations about achievement development of boys and girls in school cf. [53], they would presumably judge the achievement trajectories of girls to be higher than those of boys, even when they are actually the same. Beyond that general gender achievement bias, pre-service teachers should in particular overestimate girls' oral reading fluency compared to that of boys, whereas the opposite should be true in math. Consequently, the following four hypotheses were tested:

1. It was assumed that estimates of achievement progress, depicted as CBM graphs, should in general be higher for girls than for boys, irrespective of the content domain.
2. In addition, it was hypothesized that estimates of achievement progress in oral reading fluency would be higher for girls than for boys, even when both exhibit identical learning progress.
3. On the contrary, it was supposed that estimates of achievement progress in math would be higher for boys than for girls, even when both exhibit the same learning progress.
4. Finally, it was assumed that the participants would estimate achievement progress of both girls and boys to be higher when the linear trend of the data is steep rather than flat, and when data variability is high rather than low.

2. Materials and Methods

2.1. Participants

Based on previous investigations [11], an average medium-sized effect (f = 0.30, corresponding to η^2 = 0.08) of the independent variables on the participants' judgments was assumed. An a priori power analysis for ANOVA with repeated measures and two groups was conducted by applying G*Power 3.1 [59]. When prespecifying f = 0.30, α = 0.05, $1 - \beta$ = 0.90, and the correlation among repeated measures with r = 0.50, we determined that N = 22 was the minimum sample size based on the results of the power analysis.

Pre-service teachers enrolled in primary or secondary school teacher education programs at several German colleges were recruited via social media announcements. Students of primary and secondary teacher study programs were recruited because in both levels of schooling CBM is applied in Germany. A total of 128 pre-service teachers volunteered to participate in the experiment. However, 28 individuals were omitted from further analyses because they left the study either before receiving student vignettes or after receiving the first vignette but did not continue to take part. From the remaining participants, data were complete.

Thus, a total of N = 100 pre-service teachers (M_{age} = 24.9 years, SD = 2.0) participated in the study. 85 participants were female, 15 were male. The majority of the students were in the third or fourth year of their study program, and had, on average, studied teaching for 7.1 semesters (SD = 3.3). The majority of the participants (n = 87) reported to have no previous experiences with CBM. The remaining participants stated to have heard about CBM in lectures or seminars, but without ever practicing it.

2.2. Materials and Procedure

The experiment was run via www.soscisurvey.de (accessed on 12 April 2022). The participants could complete the experiment's tasks on a computer or any other electronic device that was linked to the Internet. The study was accessible for 22 days.

The participants were welcomed and requested to consent to participate in the experiment before the experiment began. After that, they were given a brief introduction to curriculum-based measurement. In the first part of the introduction, the participants received a short text (264 words) containing information about the aim of CBM and how teachers can benefit from it. In the second part of the introduction, the participants learned how learning progress data is visually represented (including explaining the different parts of the graphs, i.e., the learning curve, the trend line, and the goal line). Moreover, they were provided with the trend line rule, according to which the student is making adequate progress, if the trend line and the goal line are similar [6]. The introduction was finished with the presentation of an example of the graphical representation of a primary school student's learning progress. The participants were told to judge the student's learning progress by estimating whether the student should receive further support and whether the goal line should be raised.

After the introduction, the participants were randomly assigned in equal numbers to one of two conditions. However, due to a programming error, in one condition (Condition 1), the number of participants was n = 51, whereas in Condition 2, 49 participants were enrolled.

In Condition 1, the participants were presented with progress data obtained from an oral reading fluency assessment. In Condition 2, the participants were presented with progress data obtained from arithmetic tasks. In either condition, each participant received eight experimental student vignettes in random order. In these vignettes, the learning progress of four boys and four girls over a time period of 11 weeks was illustrated. Six arbitrary distractor vignettes were presented in addition to the experimental vignettes to mask the study's independent factors because knowledge of the independent variables could influence the participants' responses [60].

Each vignette showed a graph of the progress data of each student. In Condition 1, the y-axis represented the number of words read correctly (WRC), ranging from 0 to

140 cf. [6]. In Condition 2, the *y*-axis represented the number of correct digits cf. [6], also ranging from 0 to 140. In both conditions, the *x*-axis represented the school week that the test was administrated, which ranged from week 1 to week 11. The progress data were accompanied by a trend line and a goal line. The linear trend was estimated by ordinary least squares linear regression analysis and was represented by a thin (1 mm) dotted black line from week 1 to week 11. The goal line was displayed as a continuous thin (1 mm) black line from week 1 to week 11, which served as an aid to facilitate judgments about the progress each student made. The steepness of the goal line was equal for all student vignettes and was given by $y = 3x + 25$, with x being the week of assessment.

The participants were asked to examine the development of students' achievements over a period of 11 weeks to make a judgment of each student's progress. In particular, they were instructed to rate how strongly they would agree to the following statements: (1) "The student needs further assistance". (2) "The student's goal line should be raised". Each rating was done by using a six-point Likert scale ranging from 1 to 6, with 1 meaning "totally disagree" and 6 meaning "totally agree". Note that agreement on the first statement indicated that the participants judged the student's progress to be rather low, whereas agreement on the second statement reflected the participant's opinion that the student was doing quite well. Two items instead of one single item were developed to cover different facets of the assessment of students' achievement progress.

The independent within-subjects variables included the slope of the linear trend of the data, the amount of data variability, and the students' gender. Students' gender was indicated by their names. The names chosen for this study were typical of German children, both boys and girls. Typical names were used in order to prevent the activation of stereotypes associated with certain social and economic milieus, which can be triggered by rarely used names that are frequently chosen in these milieus [61].

The slope of the linear trend of the data was either low or high. The linear trend in both conditions was given by the following function: $y = bx + 9$, with b representing the rate of improvement and x representing the school week. A steep linear trend was operationalized as $b = 5$, whereas a flat trend was given by $b = 2$.

Data variability was either low or high. The standard error of the estimate (*SEE*), which is defined as the average magnitude of residuals around the trend line derived from linear regression, is frequently used in the literature to quantify the variability of learning progress graphs. *SEE* values often vary between 5 and 20, with 5 meaning a very low and 20 a very high variability e.g., [62,63]. In the present study, the *SEE* of high-variable progress data was 10.0 and 5.0 with low-variable data.

The between-subjects independent variable was the content domain of the progress data, which represented learning progress either in reading or in math.

Figure 1 shows the experimental vignettes of boys of the reading condition for illustration purposes.

2.3. Data Analyses

A repeated-measures ANOVA, with content domain (reading vs. math) as the between-subjects factor, and student gender, slope, and data variability as the within-subjects factors, was run for each dependent variable.

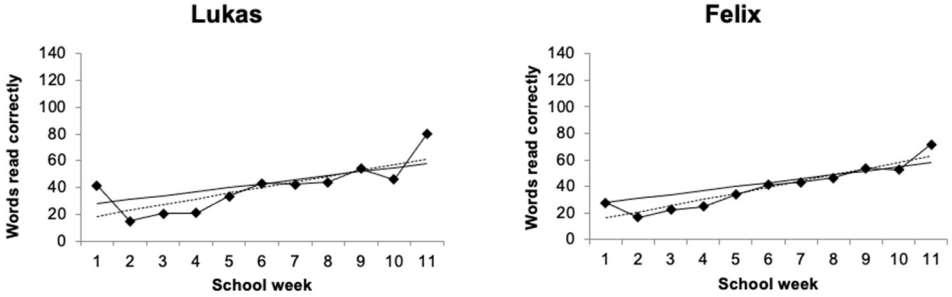

Figure 1. *Cont.*

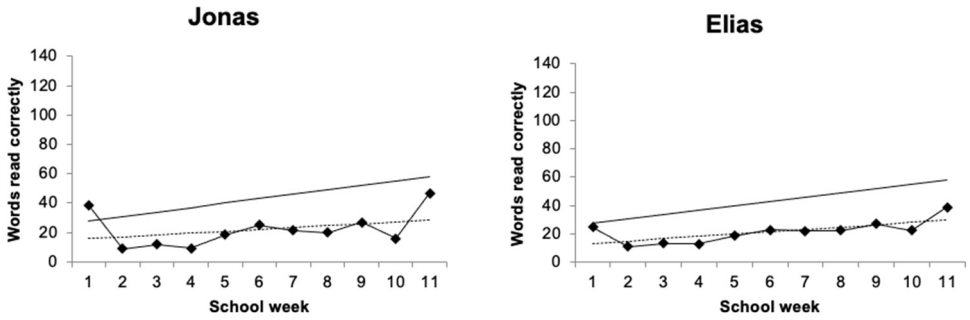

Figure 1. Experimental vignettes of boys of the reading condition. *Note*. Upper panel: Steep linear trend of the data; lower panel: flat linear trend of the data; left panel: high variability of the data; right panel: low variability of the data. The names of the students were shown on the vignettes to indicate the gender of the student. The meaning of the three different lines (solid: goal line; dotted: trend line; solid with data points: student progress data) was explained to the participants. The original labeling of the axes was in German.

3. Results

Tables 1 and 2 display the means, standard deviations, and 95% confidence intervals of the dependent variables of each condition.

Table 1. Means, standard deviations (in parentheses), and 95% confidence intervals [in brackets] of the dependent variable "need for assistance".

Domain	Slope	Student Gender			
		Male		Female	
		Variability			
		Low	High	Low	High
Math	Flat	2.33 (1.65) [1.85, 2.81]	3.27 (1.56) [2.90, 3.64]	5.33 (0.83) [4.94, 5.71]	4.08 (1.53) [3.62, 4.54]
	Steep	1.51 (0.79) [1.14, 1.88]	2.63 (1.51) [2.21, 3.05]	4.71 (1.50) [4.29, 5.14]	3.06 (1.68) [2.67, 3.46]
Reading	Flat	4.04 (1.72) [3.57, 4.51]	5.02 (0.99) [4.66, 5.38]	2.96 (1.73) [2.58, 3.34]	2.37 (1.71) [1.92, 2.82]
	Steep	3.90 (1.68) [3.54, 4.27]	4.45 (1.45) [4.04, 4.86]	2.55 (1.47) [2.14, 2.96]	1.80 (1.04) [1.42, 2.19]

Table 2. Means, standard deviations (in parentheses), and 95% confidence intervals [in brackets] of the dependent variable "raising goal".

Domain	Slope	Student Gender			
		Male		Female	
		Variability			
		Low	High	Low	High
Math	Flat	4.53 (1.50) [4.09, 4.98]	3.63 (1.56) [3.26, 4.01]	1.61 (0.84) [1.27, 1.96]	3.00 (1.57) [2.55, 3.45]
	Steep	5.35 (0.95) [4.96, 5.73]	4.24 (1.65) [3.84, 4.65]	2.37 (1.45) [1.95, 2.78]	4.14 (1.55) [3.78, 4.50]
Reading	Flat	2.80 (1.64) [2.37, 3.24]	1.76 (1.01) [1.40, 2.13]	3.82 (1.51) [3.48, 4.16]	4.61 (1.60) [4.17, 5.05]
	Steep	3.02 (1.67) [2.64, 3.40]	2.06 (1.21) [1.66, 2.46]	4.33 (1.48) [3.93, 4.74]	5.37 (0.92) [5.02, 5.73]

Regarding the dependent variable "need for assistance", the results of the analysis of variance yielded one significant main effect and three significant interaction effects. All results are shown in Table 3.

Table 3. Results of the analysis of variance for the dependent variable "need for assistance".

Effect	*F* Ratio	*df*	*p*	η^2
Slope	23.25	1, 98	<0.001	0.19
Variability	1.57	1, 98	0.213	0.02
Student Gender	0.06	1, 98	0.804	0.00
Condition	0.15	1, 98	0.701	0.00
Slope × Condition	1.99	1, 98	0.161	0.02
Variability × Condition	4.08	1, 98	0.046	0.04
Student Gender × Condition	186.59	1, 98	<0.001	0.66
Variability × Slope	2.79	1, 98	0.098	0.03
Variability × Student Gender	29.46	1, 98	<0.001	0.23
Slope × Student Gender	0.92	1, 98	0.340	0.01
Slope × Variability × Condition	0.56	1, 98	0.457	0.01
Slope × Student Gender × Condition	0.04	1, 98	0.850	0.00
Variability × Student Gender × Condition	2.12	1, 98	0.149	0.02
Slope × Variability × Student Gender	0.42	1, 98	0.519	0.00
Slope × Variability × Student Gender × Condition	3.13	1, 98	0.080	0.03

First, there was a significant main effect of the slope of the linear trend. The participants estimated the need for further assistance to be higher when the slope was flat (M = 3.67, SD = 0.30) rather than high (M = 3.08, SD = 0.44).

In addition to this main effect, ANOVA yielded a significant data variability × content domain interaction. In the math domain, participants judged students showing low variability (M = 3.47, SD = 0.40) to need more assistance than students showing high variability (M = 3.26, SD = 0.47), whereas in the reading domain, students showing low variability (M = 3.36, SD = 0.40) were judged to need less support than students with high variability (M = 3.41, SD = 0.46). The difference in judgments between low and high variability students was significant only in the math domain, $F(1, 98) = 5.25$, $p = 0.024$, but not in the reading domain, $F(1, 98) = 0.30$, $p = 0.585$. All simple effects were Bonferroni-adjusted.

Moreover, there was a significant student gender × content domain interaction. In the math domain, students were judged to need more support if they were girls (M = 4.30, SD = 0.74) rather than boys (M = 2.43, SD = 0.78), whereas in the reading domain, the participants judged the needed support for boys (M = 4.35, SD = 0.77) higher than for girls (M = 2.42, SD = 0.72). In both domains, the differences between boys and girls were significant (math: $F(1, 98) = 88.16$, $p < 0.001$; reading: $F(1, 98) = 98.70$, $p < 0.001$).

Finally, ANOVA produced a significant data variability × student gender interaction. With boys, the participants judged the need for assistance to be higher when the variability of the data was high (M = 3.84, SD = 0.85) rather than low (M = 2.94, SD = 0.91), $F(1, 98) = 21.11$, $p < 0.001$. However, with girls, the participants did the reverse and judged the need for assistance to be higher when the variability of the data was low (M = 3.89, SD = 0.83) rather than high (M = 2.83, SD = 0.86), $F(1, 98) = 32.06$, $p < 0.001$.

Concerning the dependent variable "raising goal", ANOVA produced three significant main effects and four significant interactions. The results are shown in Table 4.

As with the dependent variable "need for assistance", there was a significant main effect of slope, meaning that a high slope resulted in higher ratings (M = 3.86, SD = 0.45) than a low slope (M = 3.22, SD = 0.48). Furthermore, there was a significant main effect of data variability. The participants preferred raising the goal for students showing high data variability (M = 3.60, SD = 0.31) over students showing low data variability (M = 3.48, SD = 0.28). In addition, a main effect of the content domain was obtained. The ratings for raising the goal were higher in the math domain (M = 3.61, SD = 0.30) than in the reading domain (M = 3.47, SD = 0.29).

In addition to the main effects, ANOVA yielded a significant data variability × content domain interaction. In the math domain, ratings were higher for students showing high variability (M = 3.76, SD = 0.45) than for students showing low variability (M = 3.46, SD = 0.41), $F(1, 98) = 11.40$, $p = 0.001$, whereas in the reading domain, ratings were higher

for students showing low variability (M = 3.50, SD = 0.40) than for students showing high variability (M = 3.45, SD = 0.43), but this difference was not significant, $F(1, 98) = 0.27$, $p = 0.602$.

Table 4. Results of the analysis of variance for the dependent variable "raising goal".

Effect	*F* Ratio	*df*	*p*	η^2
Slope	30.04	1, 98	<0.001	0.24
Variability	4.18	1, 98	0.044	0.04
Student Gender	2.75	1, 98	0.101	0.03
Condition	5.40	1, 98	0.022	0.05
Slope × Condition	2.74	1, 98	0.101	0.03
Variability × Condition	7.71	1, 98	0.007	0.07
Student Gender × Condition	181.96	1, 98	<0.001	0.65
Variability × Slope	1.29	1, 98	0.258	0.01
Variability × Student Gender	41.96	1, 98	<0.001	0.30
Slope × Student Gender	10.56	1, 98	0.002	0.10
Slope × Variability × Condition	0.11	1, 98	0.743	0.00
Slope × Student Gender × Condition	0.60	1, 98	0.439	0.01
Variability × Student Gender × Condition	0.93	1, 98	0.337	0.01
Slope × Variability × Student Gender	2.45	1, 98	0.121	0.02
Slope × Variability × Student Gender × Condition	0.72	1, 98	0.400	0.01

Moreover, a significant student gender × content domain was obtained. In the math domain, boys (M = 4.44, SD = 0.78) were rated higher than girls (M = 2.78, SD = 0.75), $F(1, 98) = 68.63$, $p < 0.001$, whereas in the reading domain, the opposite was the case, with girls (M = 4.53; SD = 0.74) rated higher than boys (M = 2.41, SD = 0.77), $F(1, 98) = 117.04$, $p < 0.001$.

Additionally, there was a significant student gender × slope interaction. With flat linear trends, there was no significant difference between boys (M = 3.18, SD = 0.74) and girls, (M = 3.26, SD = 0.69), $F(1, 98) = 0.27$, $p = 0.603$, whereas with steep linear trends, the difference between boys (M = 3.67, SD = 0.69) and girls (M = 4.05, SD = 0.68) was significant, $F(1, 98) = 6.94$, $p = 0.010$.

Finally, there was a significant data variability × student gender interaction. With boys, the participants' ratings were higher when data variability was low (M = 3.93, SD = 0.90) rather than high (M = 2.93, SD = 0.83), $F(1, 98) = 15.24$, $p < 0.001$, whereas with girls, ratings were higher when data variability was high (M = 4.28, SD = 0.83) rather than low (M = 3.03, SD = 0.81), $F(1, 98) = 38.83$, $p < 0.001$.

We also conducted correlational analyses between the teacher gender (male vs. female) and each dependent variable obtained from all realized combinations of variables that were used to construct the student vignettes in order to determine whether the gender of the participants had an impact on the ratings of the participants. With all $|r| < 0.16$ and all ps > 0.127, only weak and negligible relationships were obtained.

4. Discussion

This study demonstrated that pre-service teachers' judgments of students' learning progress were biased by the gender of the students. The judgments of reading fluency were higher for girls than for boys. In particular, the participants judged the need for assistance to be lower for girls than for boys, and they opted more strongly for raising the goal line for girls than for boys. The contrary was the case when the participants were to judge a graph presenting the trajectory of math achievement. With math, boys were judged to be superior to girls on both dependent measures. Strikingly, the difference between boys and girls was large for either content domain. Pre-service teachers rated on average a score for girls compared to boys that was 2.7 times the standard deviation of the distribution of the dependent variables used in this study. A similar result was obtained for boys in the math domain (d = 2.32). These effects were comparable, albeit somewhat larger, to those obtained

from [11]. Therefore, the hypothesis was confirmed that pre-service teachers stereotype boys and girls when judging the progress of their reading fluency and math achievement.

We also expected the participants to judge the achievement development of girls on average to be higher than that of boys. The rationale behind this assumption was the evidence obtained from several studies [40,41] showing that girls outperform boys on average in both reading and math. In correspondence to this achievement difference, teachers should expect girls to perform better and develop faster than boys [53]. However, the overall difference between the participants' judgments of boys and girls was not significant in this study; hence, the hypothesis had to be rejected. One reason for the absence of the student gender main effect was the disordinal interaction between student gender and content domain, which impressively showed that the advantage of one gender over another was strongly dependent on the content domain.

It was further assumed that steep achievement trajectories would correspond with a higher likelihood for high judgments of achievement (i.e., low ratings of needed support and high ratings of raising the goal line) compared to flat achievement trajectories. This hypothesis was confirmed. With both dependent variables, a significant main effect of steepness occurred, which accounted for 21.5% of the variance on average. Although this result might seem odd, as it appears to reflect just an expectable tendency of judging fast development higher than slow development, it demonstrates that the participants actually perceived and responded to the manipulation of the graphs, indicating that the results obtained in this study were internally valid.

In the final hypothesis, it was assumed that high-variable trajectories were more likely to indicate faster development than trajectories with low data variability. This hypothesis was informed by empirical results, showing that high levels of variability in progress data often correspond with a higher probability of detecting a trend in the data than low levels of variability [9,33]. Moreover, theoretical assumptions guided this hypothesis, as high data variability produces higher peaks in progress data than low data variability, and high peaks could be perceived as a zone of achievement a student is potentially capable of cf. [64]. In line with this hypothesis, the participants judged high-variable progress data as indicating faster progress than low-variable progress data, but only with respect to raising the goal line.

The latter result shows that findings obtained from the dependent variable "need for assistance" did not exactly mirror the results obtained from the dependent variable "raising the goal line". With the latter, the independent variables produced more significant effects. On a descriptive level, all effects that were significant with "raising the goal line" were also observed with "need for support", but only some of them reached statistical significance. Obviously, "raising the goal line" was more sensitive to the manipulation of the graph characteristics and the gender of the students.

The judged superiority of highly variable developing students was dependent on gender and content domain. Concerning student gender, data variability affected the participants' judgments in opposite directions depending on student gender. With boys, high-variable achievement trajectories resulted in lower ratings of performance, whereas with girls, the reverse was the case, with high-variable trajectories yielding higher ratings of performance. In the math domain, the participants judged the achievement trajectories to be higher when data variability was high than when it was low, whereas in the reading domain, data variability did not significantly affect the participants' judgments. Both interaction effects cannot be explained by the visual characteristics of the achievement trajectory, but have to be ascribed rather to characteristics that relate to the students to be judged, and to the content where the achievement had been demonstrated.

When data variability was substantial, higher ratings for girls and lower ratings for boys may have resulted from the participants' impressions about the students derived through studying each trajectory. Since the source of data variability is difficult to interpret, it offers educators in general and the participants of the present study, in particular, a bunch of possible reasons according to which the data might vary. When straightforward

interpretations of given data are not possible to apply, the judgment itself will become uncertain [65]. Uncertainty in judgments has been shown to produce judgments that are prone to rely on stereotypical assumptions about the individuals to be judged [52,66,67]. Hence, high variability in progress data could have elicited more stereotype-based judgments than low variability because the latter appeared to be easier to attribute to plain causes than the former. Consequently, the likelihood that the participants used their stereotypes about boys and girls in their judgments was higher with high than with low variable data. In accordance with gender stereotypes, the achievement of boys should be lower than that of girls. This assumption fits well the data obtained and replicates the gender × data variability effect shown by [11].

There is evidence that variability in reading time is a predictor of reading comprehension [68], as variability has been shown to be an important component of response preparation with executive control functions [69]. Thus, regarding the interaction between data variability and content domain, it is possible that the participants interpreted high variability in reading as a clue for a rather low level of achievement, whereas in math high variability seems not to be associated with low performance.

This study also yielded a significant student gender × slope interaction. Only with steep trends, there was a significant difference between boys and girls, with participants associating lower progress with boys rather than with girls, irrespective of the content domain. Again, gender stereotypes held by participants may account for this result. Since stereotypes about girls correspond to high achievement, whereas those about boys adhere to rather low achievement [45], steep achievement trajectories conform to the stereotype about a girl rather than to that about a boy. Since stereotypes are more likely to be activated when they match the description of individuals rather than when they are in contradiction to them [70], we deem it likely that steep trajectories were automatically associated with girls rather than with boys.

5. Limitations

Every time they lack the motivation or capacity to participate in a more thorough analysis of information, people rely on their stereotyped ideas about social groupings as a basis for their judgments [71]. Although this kind of lack of desire or aptitude may be typical in most real-world situations [51], it is not always the case when teachers make decisions regarding students' education. Therefore, the fact that the participants in this study made judgments about virtual rather than actual students may be considered a study weakness. As a result, participants may not have felt strongly responsible for their judgments [72] and therefore may have exerted less effort to analyze information than they would do in school.

Additionally, teachers in classrooms can access various data that may support their judgments of their students' performance. Hence, it is likely that judgments of student accomplishment in the classroom may be less influenced by stereotypes than in this study.

The participants' lack of CBM expertise may also have had an impact on the results of the independent factors employed in this study. Actually, little is known about the strategies teachers use to interpret learning progress data presented by graphs [73], and there is evidence that teachers have difficulty interpreting CBM data [17]. Training teachers in graph literacy might therefore improve their understanding of learning progress and even help improve their judgments of learning progress graphs e.g., [12]. However, to the authors' knowledge, there is no study showing that a lack of expertise in interpreting CBM data is related to more biased judgments. According to the literature, teachers' expertise is only weakly correlated with their judgment accuracy [74,75]. Further studies should use teachers' expertise as an independent variable in the investigation of sources of biased judgments.

Even the unequal distribution of participant gender might have biased the results, as women and men might differ regarding their judgments of students' achievements in

reading and math [76]. However, the distribution of participants' gender in the present study matched the current distribution of teachers' gender in Germany [77].

6. Conclusions

The obvious prevalence of gender stereotypes in pre-service teachers' judgments of CBM graphs coincides with an increase in actual gender-related differences in students' achievements in math and reading in some countries, particularly in Germany, where the present study was conducted. Compared to a decade ago, girls lag behind boys in math, but outperform boys in reading, to an even larger degree [34,78]. The mere existence of gender stereotypes may contribute to differences in achievement between boys and girls. According to the theory of stereotype threat [79], girls' math performance is decreased because girls feel threatened by the possibility that their performance will confirm the negative stereotype associated with their social group. Stereotype threat may also explain the underperformance of boys in reading [80]. A vicious cycle of the interrelatedness of stereotypes and achievement may manifest once the gender-related differences have been established and appear to be stable. As a consequence, the gender gap in reading and math may even increase in the long term. To disrupt this cycle, we recommend teacher training on two paths. First, focusing on curriculum-based measurement, in-service and pre-service teachers should be trained in graph literacy [81], which is the competence to read and understand visualized progress data, and which is apparently lacking in many pre-service and in-service teachers e.g., [12,82]. Second, the sensitivity to gender stereotypes should be increased, both within teacher study programs and in in-service teacher training programs. One possibility to do this is by increasing the awareness of teachers' own stereotypes and biases (see for example [83] for an intervention).

Author Contributions: Conceptualization, F.K. and H.v.d.L.; methodology, F.K.; formal analysis, F.K.; data curation, H.v.d.L.; writing—original draft preparation, F.K.; writing—review and editing, F.K. and H.v.d.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of the Medical School Berlin (protocol code 2021-74, date of approval 9 November 2021).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Deno, S.L. Curriculum-based measurement: The emerging alternative. *Except. Child.* **1985**, *52*, 219–232. [CrossRef] [PubMed]
2. Ardoin, S.P.; Christ, T.J.; Morena, L.S.; Cormier, D.C.; Klingbeil, D.A. A systematic review and summarization of the recommendations and research surrounding curriculum-based measurement of oral reading fluency (CBM-R) decision rules. *J. Sch. Psychol.* **2013**, *51*, 1–18. [CrossRef] [PubMed]
3. Christ, T.J.; Zopluoglu, C.; Long, J.D.; Monaghen, B.D. Curriculum-based measurement of oral reading: Quality of progress monitoring outcomes. *Except. Child.* **2012**, *78*, 356–373. [CrossRef]
4. Espin, C.A.; van den Bosch, R.M.; van der Liende, M.; Rippe, R.C.A.; Beutick, M.; Langa, A.; Mol, S.E. A systematic review of CBM professional development materials: Are teachers receiving sufficient instruction in data-based decision-making? *J. Learn. Disabil.* **2021**, *54*, 256–268. [CrossRef] [PubMed]
5. Peters, M.T.; Förster, N.; Hebbecker, K.; Forthmann, B.; Souvignier, E. Effects of data-based decision-making on low-performing readers in general education classrooms: Cumulative evidence from six intervention studies. *J. Learn. Disabil.* **2021**, *54*, 334–348. [CrossRef]
6. Hosp, M.K.; Hosp, J.L.; Howell, K.W. *The ABCs of CBM. A Practical Guide to Curriculum-Based Measurement*; Guilford Press: New York, NY, USA, 2007.
7. Raffe, C.P.; Loughland, T. "We're not data analysts": Teachers' perspectives on factors impacting their use of student assessment data. *Issues Educ. Res.* **2021**, *31*, 224–240.

8. Zeuch, N.; Förster, N.; Souvignier, E. Assessing teachers' competencies to read and interpret graphs from learning progress assessment: Results from tests and interviews. *Learn. Disabil. Res. Pract.* **2017**, *32*, 61–70. [CrossRef]
9. Klapproth, F. Biased predictions of students' future achievement: An experimental study on pre-service teachers' interpretation of curriculum-based measurement graphs. *Stud. Educ. Eval.* **2018**, *59*, 67–75. [CrossRef]
10. Klapproth, F. Stereotype in der Lernverlaufsdiagnostik. In *Stereotype in der Schule II*; Glock, S., Ed.; Springer: Berlin, Germany, 2022; pp. 49–88.
11. Klapproth, F.; Holzhüter, L.; Jungmann, T. Prediction of students' reading outcomes in learning progress monitoring. Evidence for the effect of a gender bias. *J. Educ. Res. Online* **2022**, *14*, 16–38. [CrossRef]
12. Jungjohann, J.; Gebhardt, M.; Scheer, D. Understanding and improving teachers' graph literacy for data-based decision-making via video intervention. *Front. Educ.* **2022**, *7*, 919152. [CrossRef]
13. Van den Bosch, R.M.; Espin, C.A.; Sikkema-de Jong, M.T.; Chung, S.; Boender, P.D.M.; Saab, N. Teachers' visual inspection of curriculum-based measurement progress graphs: An exploratory, descriptive eye-tracking study. *Front. Educ.* **2022**, *7*, 921319. [CrossRef]
14. Van Norman, E.R.; Nelson, P.M.; Shin, J.-E.; Christ, T.J. An evaluation of the effects of graphic aids in improving decision accuracy in a continuous treatment design. *J. Behav. Educ.* **2013**, *22*, 283–301. [CrossRef]
15. Deno, S.L. Developments in curriculum-based measurement. *J. Spec. Educ.* **2003**, *37*, 184–192. [CrossRef]
16. Espin, C.A.; Waymann, M.M.; Deno, S.L.; McMaster, K.L. Data-based decision making: Developing a method for capturing teachers' understanding of CBM graphs. *Learn. Disabil. Res. Pract.* **2017**, *32*, 8–21. [CrossRef]
17. Van den Bosch, R.M.; Espin, C.A.; Chung, S.; Saab, N. Data-based decision making: Teachers' comprehension of curriculum-based measurement progress-monitoring graphs. *Learn. Disabil. Res. Pract.* **2017**, *32*, 46–60. [CrossRef]
18. Van den Bosch, R.M.; Espin, C.A.; Pat-El, R.J.; Saab, N. Improving teachers' comprehension of curriculum-based measurement progress monitoring graphs. *J. Learn. Disabil.* **2019**, *52*, 413–427. [CrossRef] [PubMed]
19. Wilbert, J.; Bosch, J.; Lüke, T. Validity and judgment bias in visual analysis of single-case data. *Int. J. Res. Learn. Disabil.* **2021**, *5*, 13–24. [CrossRef]
20. Klapproth, F. Mental models of growth. In *Culture and Development in Japan and Germany*; Helfrich, H., Zillekens, M., Hölter, E., Eds.; Daedalus: Münster, Germany, 2006; pp. 141–153.
21. Gesel, S.A.; LeJeune, L.M.; Chow, J.C.; Sinclair, A.C.; Lemons, C.J. A meta-analysis of the impact of professional development on teachers' knowledge, skill, and self-efficacy in data-based decision-making. *J. Learn. Disabil.* **2021**, *54*, 269–283. [CrossRef]
22. Lai, M.K.; Schildkamp, K. Inservice teacher professional learning: Use of assessment in data-based decision-making. In *Handbook of Human and Social Conditions in Assessment*; Brown, G.T.L., Harris, L.R., Eds.; Routledge: Oxfordshire, UK, 2016; pp. 77–94.
23. Deno, S.L.; Mirkin, P. *Data Based Program Modification: A Manual*; Leadership Training Institute for Special Education: Minneapolis, MN, USA, 1977.
24. Tindal, G. Curriculum-based measurement: A brief history of nearly everything from the 1970s to the present. *ISRN Educ.* **2013**, *2013*, 958530. [CrossRef]
25. Klauer, K.J. Erfassung des Lernfortschritts durch curriculumbasierte Messung. *Heilpädagogische Forsch.* **2006**, *32*, 16–26.
26. Blumenthal, S.; Gebhardt, M.; Förster, N.; Souvignier, E. Internetplattformen zur Diagnostik von Lernverläufen von Schülerinnen und Schülern in Deutschland. Ein Vergleich der Plattformen Lernlinie, Levumi und quop. *Z. Für Heilpädagogik* **2022**, *73*, 153–167.
27. Förster, N.; Kuhn, J.-T.; Souvignier, E. Normierung von Verfahren zur Lernverlaufsdiagnostik. *Empirische Sonderpädagogik* **2017**, *9*, 116–122.
28. Mullis, I.V.S.; von Davier, M.; Foy, P.; Fishbein, B.; Reynolds, K.A.; Wry, E. *PIRLS 2021. International Results in Reading*; Boston College: Chestnut Hill, MA, USA, 2023.
29. Fuchs, L.S.; Fuchs, D.; Hosp, M.K. Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Sci. Stud. Read.* **2001**, *5*, 239–256. [CrossRef]
30. Fuchs, L.S.; Fuchs, D.; Compton, D.L.; Bryant, J.D.; Hamlett, C.L.; Seethaler, P.M. Mathematics screening and progress monitoring at first grade: Implications for responsiveness to intervention. *Except. Child.* **2007**, *73*, 311–330. [CrossRef]
31. Nelson, G.; Kiss, A.J.; Codding, R.S.; McKevett, N.M.; Schmitt, J.F.; Park, S.; Romero, M.E.; Hwang, J. Review of curriculum-based measurement in mathematics: An update and extension of the literature. *J. Sch. Psychol.* **2023**, *97*, 1–42. [CrossRef] [PubMed]
32. Christ, T.J.; Scullin, S.; Tolbize, A.; Jiban, C.L. Implications of Recent Research: Curriculum-Based Measurement of Math Computation. *Assess. Eff. Interv.* **2008**, *33*, 198–205. [CrossRef]
33. Nelson, P.M.; Van Norman, E.R.; Christ, T.J. Visual analysis among novices: Training and trend lines as graphic aids. *Contemp. Sch. Psychol.* **2017**, *21*, 93–102. [CrossRef]
34. McElvany, N.; Lorenz, R.; Frey, A.; Goldhammer, F.; Schilcher, A.; Stubbe, T.C. *IGLU 2021. Lesekompetenzen von Grundschulkindern im Internationalen Vergleich und im Trend Über 20 Jahre*; Waxmann: Münster, Germany, 2023.
35. Mullis, I.V.S.; Martin, M.O.; Foy, P.; Hooper, M. *PIRLS 2016: International Results in Reading*; TIMSS & PIRLS International Study Center; Lynch School of Education; Boston College International Association for the Evaluation of Educational Achievement (IEA): Chestnut Hill, IL, USA, 2017.
36. Manu, M.; Torppa, M.; Vasalampi, K.; Lerkkanen, M.-K.; Poikkeus, A.-M.; Niemi, P. Reading development from kindergarten to age 18: The role of gender and parental education. *Read. Res. Q.* **2023**, *58*, 505–538. [CrossRef]

37. Meissel, K.; Meyer, F.; Yao, E.S.; Rubie-Davies, C.M. Subjectivity of teacher judgments: Exploring student characteristics that influence teacher judgments of student ability. *Teach. Teach. Educ.* **2017**, *65*, 48–60. [CrossRef]
38. Carlana, M. Implicit stereotypes: Evidence from teachers' gender bias. *Q. J. Econ.* **2019**, *134*, 1163–1224. [CrossRef]
39. OECD. Are Boys and Girls Equally Prepared for Life? 2014. Available online: https://www.oecd.org/pisa/pisaproducts/PIF-2014-gender-international-version.pdf (accessed on 14 April 2023).
40. Tian, L.; Li, X.; Chen, X.; Huebner, E.S. Gender-specific trajectories of academic achievement in Chinese elementary school students: Relations with life satisfaction trajectories and suicidal ideation trajectories. *Learn. Instr.* **2023**, *85*, 101751. [CrossRef]
41. Fu, R.; Chen, X.; Wang, L.; Yang, F. Developmental trajectories of academic achievement in Chinese children: Contributions of early social-behavioral functioning. *J. Educ. Psychol.* **2016**, *108*, 1001. [CrossRef]
42. Hoge, R.D.; Coladarci, T. Teacher-based judgments of academic achievement: A review of literature. *Rev. Educ. Res.* **1989**, *59*, 297–313. [CrossRef]
43. Lorenz, G.; Gentrup, S.; Kristen, C.; Stanat, P.; Kogan, I. Stereotype bei Lehrkräften? Eine Untersuchung systematisch verzerrter Lehrererwartungen. *Kölner Z. Für Soziologie Und Sozialpsychologie* **2016**, *68*, 89–111. [CrossRef]
44. Cvencek, D.; Kapur, M.; Meltzoff, A.N. Math achievement, stereotypes, and math self-concepts among elementary-school students in Singapore. *Learn. Instr.* **2015**, *39*, 1–10. [CrossRef]
45. Glock, S.; Kleen, H. Gender and student misbehavior: Evidence from implicit and explicit measures. *Teach. Teach. Educ.* **2017**, *67*, 93–103. [CrossRef]
46. Jussim, L.; Eccles, J. Teacher expectations: II. Construction and reflection of student achievement. *J. Personal. Soc. Psychol.* **1992**, *63*, 947–961. [CrossRef]
47. Greenwald, A.G.; Banaji, M.R. Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychol. Rev.* **1995**, *102*, 4–27. [CrossRef] [PubMed]
48. Schneider, D.J. *The Psychology of Stereotyping*; Guilford Press: New York, NY, USA, 2004.
49. Macrae, C.N.; Milne, A.B.; Bodenhausen, G.V. Stereotypes as energy-saving devices: A peek inside the cognitive toolbox. *J. Personal. Soc. Psychol.* **1994**, *66*, 37–47. [CrossRef]
50. Van Knippenberg, A.; Dijksterhuis, A.; Vermeulen, D. Judgement and memory of a criminal act: The effects of stereotypes and cognitive load. *Eur. J. Soc. Psychol.* **1999**, *29*, 191–201. [CrossRef]
51. Fiske, S.T.; Neuberg, S.L. A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. *Adv. Exp. Soc. Psychol.* **1990**, *23*, 1–74. [CrossRef]
52. Campbell, D.T. Stereotypes and the perception of group differences. *Am. Psychol.* **1967**, *22*, 817–829. [CrossRef]
53. Muntoni, F.; Retelsdorf, J. Gender-specific teacher expectations in reading—The role of teachers' gender stereotypes. *Contemp. Educ. Psychol.* **2018**, *54*, 212–220. [CrossRef]
54. Ellmers, N. Gender stereotypes. *Annu. Rev. Psychol.* **2018**, *69*, 275–298. [CrossRef]
55. Kertz-Welzel, A. Bildung and the master teacher: Issues in preservice teacher education in Germany. In Proceedings of the 18th International Seminar of the ISME Commission on Music Policy: Culture, Education, and Mass Media, Birmingham, UK, 20–22 July 2016; p. 498.
56. Acar-Erdol, T.; Bostancioglu, A.; Gözütok, F.D. Gender equality perceptions of preservice teachers: Are they ready to teach it? *Soc. Psychol. Educ.* **2022**, *25*, 793–818. [CrossRef]
57. Frühauf, M.; Hildebrandt, J.; Mros, T.; Zander, L.; McElvany, N.; Hannover, B. Does an immigrant teacher help immigrant students cope with negative stereotypes? Preservice teachers' and school students' perceptions of teacher bias ad motivational support, as well as stereotype threat effects on immigrant students' learning. *Soc. Psychol. Educ.* **2023**, 1–41. [CrossRef]
58. Yendell, O.; Claus, C.; Bonefeld, M.; Karst, K. "I wish I could say, 'Yeah, both the same'": Cultural stereotypes and individual differentiations of preservice teachers about different low socioeconomic origins. *Soc. Psychol. Educ.* **2023**, 1–36. [CrossRef]
59. Faul, F.; Erdfelder, E.; Buchner, A.; Lang, A.-G. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behav. Res. Methods* **2009**, *41*, 1149–1160. [CrossRef]
60. Klein, O.; Doyen, S.; Leys, C.; da Gama, P.A.; Miller, S.; Questienne, L.; Cleeremans, A. Low hopes, high expectations: Expectancy effects and the replicability of behavioral experiments. *Perspect. Psychol. Sci.* **2012**, *7*, 572–584. [CrossRef] [PubMed]
61. Gerhards, J. *Die Moderne und Ihre Vornamen: Eine Einladung in die Kultursoziologie*, 2nd ed.; VS Verlag: Wiesbaden, Germany, 2010.
62. Ardoin, S.P.; Christ, T.J. Curriculum-based measurement of oral reading: Standard errors associated with progress monitoring outcomes from DIBELS, AIMSweb, and an experimental passage set. *Sch. Psychol. Rev.* **2009**, *38*, 266–283. [CrossRef]
63. Van Norman, E.R.; Christ, T.J. How accurate are interpretations of curriculum-based measurement progress monitoring data? Visual analysis versus decision rules. *J. Sch. Psychol.* **2016**, *58*, 41–55. [CrossRef] [PubMed]
64. Vygotsky, L.S. *Mind in Society: The Development of Higher Psychological Processes*; Harvard University Press: Cambridge, MA, USA, 1978.
65. Jungermann, H. Reasons for uncertainty: From frequencies to stories. *Psychol. Beiträge* **1997**, *39*, 126–139.
66. Darley, J.M.; Gross, P.G. A hypothesis-confirming bias in labeling effects. *J. Personal. Soc. Psychol.* **1983**, *44*, 20–33. [CrossRef]
67. Kunda, Z.; Sherman-Williams, B. Stereotypes and the construal of individuating information. *Personal. Soc. Psychol. Bull.* **1993**, *19*, 90–99. [CrossRef]
68. Li, J.J.; Cutting, L.E.; Ryan, M.; Zilioli, M.; Denckla, M.B.; Mahone, E.M. Response variability in rapid automatized naming predicts reading comprehension. *J. Clin. Exp. Neuropsychol.* **2009**, *31*, 877–888. [CrossRef] [PubMed]

69. Clare Kelly, A.M.; Uddin, L.Q.; Biswal, B.B.; Castellanos, F.X.; Milham, M.P. Competition between functional brain networks mediates behavioral variability. *NeuroImage* **2008**, *39*, 527–537. [CrossRef]
70. Casper, C.; Rothermund, K.; Wentura, D. Automatic stereotype activation is context dependent. *Soc. Psychol.* **2010**, *41*, 131–136. [CrossRef]
71. Bodenhausen, G.V. Emotions, arousal, and stereotypic judgements: A heuristic model of affect and stereotyping. In *Affect, Cognition, and Stereotyping*; Mackie, D.M., Hamilton, D.L., Eds.; Academic Press: San Diego, CA, USA, 1993; pp. 13–37. [CrossRef]
72. Tetlock, P.E.; Kim, J.I. Accountability and judgement processes in a personality prediction task. *J. Personal. Soc. Psychol.* **1987**, *52*, 700–709. [CrossRef]
73. Blumenthal, S.; Blumenthal, Y.; Lembke, E.S.; Powell, S.R.; Schultze-Petzold, P.; Thoma, E.R. Educator perspectives on data-based decision making in Germany and the United States. *J. Learn. Disabil.* **2021**, *54*, 284–299. [CrossRef]
74. Jansen, T.; Vögelin, C.; Machts, N.; Keller, S.; Köller, O.; Möller, J. Judgment accuracy in experienced versus student teachers: Assessing essays in English as a foreign language. *Teach. Teach. Educ.* **2021**, *97*, 103216. [CrossRef]
75. McElvany, N.; Schroeder, S.; Hachfeld, A.; Baumert, J.; Richter, T.; Schnotz, W.; Horz, H.; Ullrich, M. Teachers' diagnostic skills to assess student abilities and task difficulty of learning materials incorporating instructional pictures. *Ger. J. Educ. Psychol.* **2009**, *23*, 223–235. [CrossRef]
76. Kleen, H.; Baumann, T.; Glock, S. Der demografische Match zwischen Schüler*innen und Lehrer*innenmerkmalen: Geschlecht, sozialer Status, Migrationshintergrund—Wer profitiert am meisten? In *Stereotype in der Schule*; Glock, S., Ed.; Springer VS: Berlin/Heidelberg, Germany, 2022; pp. 379–400.
77. Statista. Anteil der Weiblichen Lehrkräfte an Allgemeinbildenden Schulen in Deutschland im Schuljahr 2022/2023 Nach Schulart. 2023. Available online: https://de.statista.com/statitik/daten/studie/1129852/umfrage/frauenanteil-unter-den-lehrkraeften-in-deutschalnd-nach-schulart/ (accessed on 23 August 2023).
78. IPN. *MINT-Nachwuchsbarometer 2023*; Joachim-Herz-Stiftung: Hamburg, Germany, 2023.
79. Steele, C.M.; Aronson, J. Stereotype threat and the intellectual test performance of African Americans. *J. Personal. Soc. Psychol.* **1995**, *69*, 797–811. [CrossRef]
80. Pansu, P.; Régner, I.; Max, S.; Colé, P.; Nezlek, J.B.; Huguet, P. A burden for the boys: Evidence of stereotype threat in boys' reading performance. *J. Exp. Soc. Psychol.* **2016**, *65*, 26–30. [CrossRef]
81. Friel, S.N.; Curcio, F.R.; Bright, G.W. Making sense of graphs: Critical factors influencing comprehension and instructional implications. *J. Res. Math. Educ.* **2001**, *32*, 124–158. [CrossRef]
82. Mandinach, E.B.; Gummer, E.S. What does it mean for teachers to be data literate: Laying out the skills, knowledge, and dispositions. *Teach. Teach. Educ.* **2016**, *60*, 366–376. [CrossRef]
83. Bonefeld, M. Reflexion eigener Stereotype als Motor zur nachhaltigen Stereotypreduktion bei angehenden Lehrkräften. In *Stereotype in der Schule*; Glock, S., Ed.; Springer VS: Berlin/Heidelberg, Germany, 2022; pp. 341–378.