Social Science Research 116 (2023) 102938

Contents lists available at ScienceDirect

Social Science Research

journal homepage: www.elsevier.com/locate/ssresearch

# Teacher judgements and gender achievement gaps in primary education in England, Germany, and the US

Melanie Olczyk [a,*], Sarah Gentrup [b], Thorsten Schneider [c], Anna Volodina [d,e], Valentina Perinetti Casoni [f], Elizabeth Washbrook [f], Sarah Jiyoon Kwon [g], Jane Waldfogel [h]

[a] Martin-Luther-University Halle-Wittenberg, Paracelsusstr. 22, 06114 Halle (Saale), Germany
[b] Humboldt-Universität zu Berlin, Unter den Linden 6, 10099, Berlin, Germany
[c] Leipzig University, Beethovenstraße 15, 04107, Leipzig, Germany
[d] Institute for Educational Quality Improvement at the Humboldt-Universität zu Berlin, Luisenstraße 56, 10117, Berlin, Germany
[e] University of Bamberg, Augustenstraße 6, 96047, Bamberg, Germany
[f] University of Bristol, 35 Berkeley Square Clifton, Bristol, United Kingdom
[g] The University of Chicago, 969 E. 60th St., Chicago, IL, 60637, USA
[h] Columbia University, 1255 Amsterdam Avenue, New York, NY, 10027, USA

A B S T R A C T

We examined whether inaccurate teacher judgements of primary school student achievement correlate with students' gender and whether such bias contributes to gender achievement gaps in language and mathematics. Our study used ex-post harmonised longitudinal data from England, Germany, and the US. We observed domain-specific teacher judgement bias with a positive bias for girls in the language domain and for boys in mathematics. Furthermore, biased teacher judgements partly mediated the effect of gender on later achievement. Despite these common findings, cross-country differences emerged in the extent of teacher judgement bias as well as its mediation of gender achievement gaps. We conclude that this is a topic of relevance across national contexts and where the institutional and societal setting needs more attention in future research.

## 1. Introduction

Findings from international large-scale assessment studies in primary and lower secondary school, such as the Progress in International Reading Literacy Study (PIRLS), the Trends in International Mathematics and Science Study (TIMSS), and the Programme for International Student Assessment (PISA), consistently show gender achievement gaps across countries taking part in these studies: While girls perform significantly better in reading than boys, there is a male advantage in mathematics, albeit less consistent and less pronounced (see, e.g., McElvany et al., 2017; Mullis et al., 2017a, p. 36; Wendt et al., 2016, for primary education; see, e.g., OECD, 2016, pp. 168–169, 196–197, for secondary education).

Researchers propose multiple possible explanations for gender achievement gaps which range from focussing on biologically based sex differences to psycho-sociological perspectives stressing the importance of gender-based socialisation, education, and teaching (see, e.g., Halpern, 2011; Hyde, 2014, for an overview). According to the psycho-sociological perspective, gender differences in academic interests and behaviours can, for example, emerge as a child observes and imitates the (gender-based) behaviours of people or

* Corresponding author. Martin-Luther-University Halle-Wittenberg, Paracelsusstr. 22, 06114, Halle (Saale), Germany.
*E-mail addresses:* melanie.olczyk@soziologie.uni-halle.de (M. Olczyk), sarah.gentrup@hu-berlin.de (S. Gentrup), thorsten.schneider@uni-leipzig.de (T. Schneider), anna.volodina@iqb.hu-berlin.de (A. Volodina), valentina.perinetticasoni@bristol.ac.uk (V. Perinetti Casoni), Liz.Washbrook@bristol.ac.uk (E. Washbrook), sarahjiyoonkwon@uchicago.edu (S.J. Kwon), j.waldfogel@columbia.edu (J. Waldfogel).

https://doi.org/10.1016/j.ssresearch.2023.102938
Received 12 December 2022; Received in revised form 8 September 2023; Accepted 30 September 2023
Available online 21 October 2023
0049-089X/© 2023 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

when a child's behaviour gets confirmed or invalidated by significant others, such as parents, peers, or teachers, based on their gender-specific beliefs (Bussey and Bandura, 1999). When children enter school, teachers and their expectations and judgements become important to them (e.g., Gentrup and Rjosk, 2018). Even though teachers' evaluations of student characteristics are generally accurate (e.g., Südkamp et al., 2012), there is emerging research showing that teachers' stereotypes coinciding with student gender (e.g., Carlana, 2019; Dersch et al., 2022) can lead to gender bias in teacher judgements (Jussim et al., 1996; Tenenbaum and Ruck, 2007). Gender stereotypes are defined as generalised beliefs about the characteristics, attributes, and behaviours female and male persons (allegedly) have or should have (Eagly, 1987; Hannover and Wolter, 2019). Such stereotypical beliefs can influence the perceptions a person forms about individual group members (Lorenz, 2018; Mast and Krings, 2008), such as teacher judgements and expectations of the academic performance of a specific boy or girl. Biased teacher judgements and expectations can, in turn, trigger different verbal (e.g., less warm and supportive, low-quality feedback; Gentrup et al., 2020; Rubie-Davies, 2007) and non-verbal teacher behaviours (e.g., reduced eye contact; Babad, 1990, 1993), and, eventually, result in self-fulfilling prophecies confirming the initially biased expectations (Wang et al., 2018). In the case of self-fulfilling effects of teacher judgements and expectations on student achievement, such processes could exacerbate gender achievement gaps.

Current findings on the existence and extent of gender bias in teacher judgements are mixed. While there is research that did not find bias related to student gender (e.g., Jussim et al., 1996; Karing et al., 2011; McKown and Weinstein, 2008), some studies identified gender differences in teacher judgements even after controlling for actual student achievements. The latter points to domain-specific results with boys' achievement often being overestimated in mathematics (e.g., Gentrup and Rjosk, 2018; Lee and Newton, 2021; Riegle-Crumb and Humphries, 2012), while girls' skills are overestimated in the language domain (e.g., Campbell, 2015; Lorenz et al., 2016; Ready and Wright, 2011). Further, research on the effects of biased judgements and expectations on girls' and boys' achievements is also rather sparse. The few existing studies mainly suggest that biased teacher expectations may contribute to gender achievement gaps (e.g., Muntoni and Retelsdorf, 2018; Robinson-Cimpian et al., 2014).

In our study, we want to provide new evidence on these topics. First, we investigate whether teacher judgements of students' skills at the beginning of primary school are biased by student gender. Second, we examine whether these (potentially biased) teacher judgements contribute to gender gaps in student achievement at the end of primary education (age 10 to 11). Third, by considering three national contexts, namely England, Germany, and the US, we explore whether patterns are robust. Fourth, as our study relies on data from three different countries, we tentatively address the question of whether systematic cross-country variation might exist.[1] Linked to this, we refer to two types of features which might contribute to cross-country variation: gender (in)equality within a country which might be a constituent element of stereotypes and specific characteristics of the education system, such as ability grouping, that have been discussed in studies on the moderating role of the (educational) context (e.g., Geven et al., 2021).

All analyses are based on ex-post harmonised data from three large-scale longitudinal surveys, the Millennium Cohort Study (MCS) for England, the Starting Cohort 2 of the National Educational Panel Study (NEPS-SC2) for Germany, and the Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) for the US.

## 2. Theoretical considerations and empirical findings

### 2.1. Teacher judgements and judgement bias

Conceptually, teacher judgements are the result of cognitive information processing and diagnostic thinking (see Loibl et al., 2020, for a conceptional framework). According to dual-process theories, the degree to which judgements involve the deliberate integration of target information or are based on stereotypes and generalisations varies (e.g., Fiske et al., 1999). Whether target information is comprehensively considered in forming judgements or not depends, amongst other factors, on teachers' personal characteristics (e.g., attitudes, knowledge, or mindsets) and situational characteristics such as time pressure or judgement goals (see, e.g., Loibl et al., 2020, for an overview). Furthermore, teacher judgements are more likely to be based on categorisation and generalisations when information is ambiguous (Gawronski et al., 2003).

Research shows that, in general, teachers estimate students' achievements in different domains relatively accurately (Hoge and Coladarci, 1989; Südkamp et al., 2012). However, systematic gender differences in teacher evaluations also have been observed pointing in part to domain-specific gender biases. In the language domain, teachers usually rate girls' achievement higher than the achievement of boys, even after controlling for achievement test scores (for England see, e.g., Campbell, 2015; Lee and Newton, 2021; Plewis, 1997; for Germany see, e.g., Gentrup and Rjosk, 2018; Lorenz et al., 2016; Stahl, 2007; Hinnant et al., 2009; Ready and Wright, 2011). In mathematics, on the contrary, girls' achievement is often underestimated compared to equally performing boys (for England see, e.g., Campbell, 2015; Lee and Newton, 2021; for Germany see, e.g., Gentrup and Rjosk, 2018; Tiedemann, 2000, 2002; for the US see, e.g., Jussim and Eccles, 1992; Riegle-Crumb and Humphries, 2012). As the aforementioned studies controlled for students' scores on standardised performance tests, the observed differences in teacher judgements can be interpreted as gender bias. However, there are also studies that did not report differences in teacher judgements of girls' and boys' achievements after controlling for students' test scores and, therefore, did not identify gender bias in the language domain (for Germany see, e.g., Karing et al., 2011; for the US see, e.g., Doherty and Conolly, 1985; Hoge and Butcher, 1984; McKown and Weinstein, 2008) or in mathematics (for Germany see, e.g.,

[1] In a previous paper, the authors used the same data sets to address teacher bias by student's social origin in a cross-country perspective (Olczyk et al., 2022). Therefore, the data and operationalisation section (Section 3), as well as the analytical approach section (e.g., Table 3) have a high degree of overlap with this previous work.

Schrader and Helmke, 1990; for the US see, e.g., Hinnant et al., 2009; Jussim et al., 1996; Jussim and Eccles, 1995; Madon et al., 1998; McKown and Weinstein, 2008).

### 2.2. Teacher judgements and achievement development

Teacher judgements of current student performance are strongly related to teachers' expectations about students' future performance and can affect learning and achievement development similarly. Referring to the concept of the self-fulfilling prophecy (Merton, 1948), Rosenthal (1973) proposed four dimensions of teacher behaviour that may mediate such effects of teacher expectations on student development: (1) input that the children receive by the teacher, (2) student's opportunities for output such as calling on students, (3) the feedback given by the teacher, and (4) the climate of the teacher-student relationships. Empirical evidence on these four potential processes and their (relative) relevance for transmitting biased judgements is scarce (see, e.g., Harris and Rosenthal, 1985; Urhahne and Wijnia, 2021). Because of this, our review on the relevance of teacher judgement and expectations for achievement development focuses on research investigating associations instead of testing specific mechanisms.

For the US, Robinson-Cimpian and colleagues (2014) examined to what extent gender gaps in mathematics are mediated by prior and current teacher perceptions of primary school students' math proficiency. They reported that teacher perceptions accounted for varying shares of the gender gap in mathematics ranging from 42% up to 85% (see Models 5 to 7 in Tables 4 and 5 in Robinson-Cimpian et al., 2014). For Germany, Muntoni and Retelsdorf (2018) showed for secondary school students from Grades 5 and 6 that teacher expectations mediated the link between student gender and reading achievement. For the first year of primary education (Grade 1), Gentrup and Rjosk (2018) found no substantial contribution to gender gaps in reading and mathematics by teacher expectations in Germany. However, they reported that strongly underestimated girls were more vulnerable to negative teacher expectation effects in mathematics than boys (Gentrup and Rjosk, 2018). Other studies also showed that girls are more sensitive to lower teacher expectations in mathematics than were boys (Jussim et al., 1996; McKown and Weinstein, 2002). For England or the UK, to our knowledge, no studies until now have investigated the contribution of self-fulfilling teacher judgements to gender achievement gaps. Thus, even though a few earlier studies investigated whether self-fulfilling teacher judgements and expectations contribute to gender gaps in student achievement, empirical evidence on this issue is still scarce.

### 2.3. Relevance of national contexts for teacher judgement bias and its impact on achievement

In Sections 2.1 and 2.2, we introduced general theoretical considerations. As information processing and stereotypes are embedded in and could be affected by the broader institutional context, the extent of judgement bias should differ across countries (see, e.g., Geven et al., 2021; Olczyk et al., 2022; Weinstein, 2002, for a similar argumentation). Besides the conditions and regulations of schools and teacher training, the broader institutional context also refers to norms, values, as well as cultural-cognitive beliefs that frame and guide social action (see Geven et al., 2021, for a similar argumentation). Further, the institutional setting may also alter the extent to which teacher judgements and expectations may result in self-fulfilling prophecies and, eventually, contribute to gender achievement gaps. Teachers shape educational careers not only by teaching students but also by grading and sorting them. Hence, it might be the case that in some contexts, teachers might have a greater impact on the educational careers of students than in others, for example, with the presence/absence of ability grouping (e.g., Finnigan and Gross, 2007; Geven et al., 2021; Kelly and Carbonaro, 2012; Lee et al., 2014; Lerner and Tetlock, 1999; Pit-ten Cate et al., 2020).

In our study, we examine the robustness of gender bias in teacher judgements, how bias predicts later student achievement, and how it contributes to gender achievement gaps in three countries. Thus, we also explore differences in the effects between countries. Although the data sets do not allow consideration of specific features of the institutional and societal setting potentially causing any cross-country variations, we outline and provide some examples of how the national context might shape the extent of gendered judgement bias and its effects on later achievement as a starting point for future research.

The extent of gender bias in teacher judgements might be associated with *gender (in)equality* in a specific country. As gender (in) equality in education and the labour market are important sources of gendered beliefs of persons (Hilton and von Hippel, 1996), gender stereotypes mirror to some extent actual differences in girls' and boys' educational success as well as women's and men's position in the labour market. Therefore, one might argue that an overall higher degree of (vertical) gender equality in a country should be accompanied with less prevalent gender stereotypes. However, some recent studies have observed that gender differences in diverse societal areas tend to be larger in wealthier countries that have reached a higher degree of formal gender equality in the sense of equal economic opportunities and legal rights. This unintuitive pattern of higher (vertical) gender equality and stronger (horizontal) gender gaps, the so-called "gender equality paradox" (Stoet and Geary, 2018), has, for example, been observed for self-reported personality traits (Fors Connolly et al., 2020), self-esteem (Zuckerman et al., 2016), math-related values (Stoet et al., 2016) or adolescents' occupational aspirations (Stoet and Geary, 2022). To explain this phenomenon, some scholars rely on the assumption that genetically based differences in the preferences of women and men come more easily into play in wealthier countries in which women and men share similar rights (Schmitt et al., 2017). From a sociological and socio-psychological perspectives, it seems more likely that the higher degree of horizontal gender segregation is rooted in a reshaping of gender-related beliefs. Recent research supports the latter

and shows that gender stereotypes that assume female and male persons will succeed in different areas (e.g., the "math is for males" stereotype) are *more pronounced* in wealthier countries that have reached a higher degree of formal gender equality (Breda et al., 2020; Napp and Breda, 2022). In combination with structural changes in the labour market, such as more job opportunities in the service sector, gendered choices and job pressures can reinforce horizontal gender segregation even in formally more gender-equal countries (Barone, 2011). Hence, in line with these results, we would expect more pronounced gender bias in accordance with domain-specific gender stereotypes in countries with higher formal gender equality, but pronounced segregation of the genders in different societal eras. That is, gender bias in mathematics in favour of boys and gender bias in favour of girls in the language domain should be more prevalent in countries with a higher degree of horizontal gender inequality.

In terms of the institutional setting, particularly a country's education system, *ability grouping and tracking* might be associated with teachers' judgement formation as well as its impact on student achievement. As Krolak-Schwerdt and colleagues (2018) argue, teachers could be encouraged to more thoroughly consider their students' performance and learning in educational systems where forms of ability grouping are common. Therefore, we would expect teacher judgements to be more precise in these systems (see, e.g., Olczyk et al., 2022, for a similar argumentation). This should be particularly the case for teacher judgements preceding the separation of students into tracks and streams and, hence, in systems where grouping is more influential for the later educational career (see Krolak-Schwerdt et al., 2018, for evidence from experimental research). Furthermore, if course levels persist over multiple years – in the case of path dependencies – then the degree to which teachers feel responsible for their decisions may also increase. Consequently, teacher judgement bias might be even lower. On the contrary, attendance at a certain course level could also act as a signal or label and, in turn, shape future judgements by the teachers (e.g., Macqueen, 2013). In other words, initial groupings may function as an "anchor" for later judgements. With respect to the impact of teacher judgement bias on subsequent achievement, ability grouping might be relevant as well (see Ready and Chu, 2015, for a similar argumentation). Students whose ability is underestimated by their teachers will be assigned to less-demanding courses with lower quantity and slower pace of instruction (e.g., Gamoran, 1986; Pallas et al., 1994) and, therefore, will receive educational instruction that does not support them to achieve more highly. Furthermore, such 'inadequate' placements might de-motivate students. Both processes – less demanding input and student demotivation – can lead to lower performance than would have been possible under other conditions. If teacher judgements correlate with students' gender net of abilities and skills, 'inadequate' placements resulting from biased judgements would then contribute to the persistence and even exacerbation of gender achievement gaps. These effects should be strongest when students remain in the same ability group for a substantial period of time. In consequence, we would expect a stronger association between teacher judgement bias and later achievement in countries where ability grouping is common compared to countries where ability grouping is relatively uncommon during primary education.

*School accountability* might be another institutional factor which could be relevant for bias in teacher judgements and expectations as well as its impact on gender achievement gaps. With respect to teacher judgements and expectations, accountability could frame and affect teachers' diagnostic thinking and information processing (e.g., Krolak-Schwerdt et al., 2013, 2018), for example by setting specific governmental goals and evaluating the performance of schools according to these goals (Finnigan and Gross, 2007). Accountability could be accompanied by a stronger feeling of responsibility for the decisions made by teachers. An increasing responsibility, in turn, could increase the motivation to get an accurate sense of the situation and to avoid stereotypical judgements and expectations (Bodenhausen et al., 1999; Geven et al., 2018; Krolak-Schwerdt et al., 2013, 2018). Hence, teachers could be expected to have more incentive to judge student achievement accurately in systems in which they are held accountable for their work.[2]

*Standardisation* that often accompanies accountability could be another factor that influences the probability that teacher judgements and expectations affect student achievement and which could lead to cross-country differences (see also Olczyk et al., 2022, for a similar argumentation): the more input factors such as curricular goals, but also teaching and test materials are predetermined, the less "room" will exist for biased teacher judgements to impact teachers' behaviours and, in consequence, students' achievement development (Klenowski and Wyatt-Smith, 2010). This argument mainly addresses the input mechanism of self-fulfilling prophecies, but may also shape student's opportunities for output and the feedback given by the teacher (for more information on the dimension of teacher behaviour relevant for self-fulfilling prophecies, see Harris and Rosenthal, 1985). From a cross-country perspective, in line with this argumentation, we would expect a weaker association between teacher judgement bias and later achievement for countries with a comparatively high degree of standardisation.

### 2.4. Gender (in)equality and features of the education system in England, Germany, and the US

The three countries show a comparable and high degree of vertical gender equality when compared to gender equality worldwide – at least according to common indices such as the Gender Gap Index (GGI; Germany: 0.787, the UK: 0.767, the US: 0.724; World Economic Forum, 2019) or the gender wage gap (female to male disadvantage in employees' median earnings; Germany: 14%, the UK: 14%, the US: 17%; OECD, 2023). With respect to horizontal gender segregation, all three counties show substantial inequality. However, the rank order between the countries depends on the specific indicator used. For example, with regard to career aspirations, gender segregation seems to be more pronounced in the US than in Germany and the UK. In 2018 in the US, for every three 15-year old boys seeing themselves working as a science or engineering professional in the future there was only one girl sharing the same career aspiration (17% males and 6% females had the aspiration to work as a science or engineering professional in the future; OECD, 2019:

[2] Even if teachers undermined these processes – for example, by manipulating the data (e.g., Espeland and Sauder, 2007), they should have a more informed view of which students are especially low performers. The expected positive effects on teacher's motivation and effort, however, could be offset if accountability policies increased teacher's stress (e.g., Berryhill et al., 2009; Jerrim and Sims, 2022; Perryman and Calvert, 2020).

Table II.B1.8.20). In Germany and the UK, the ratio was around two boys to one girl (Germany: 12% males and 7% females, the UK: 19% males and 9% females). In tertiary education, in contrast, the share of female graduates from STEM programs was lowest in Germany (28%, 2017), followed by the US (34%, 2016) and the UK (38%, 2016; WorldBank, 2023).

With respect to the educational system, the three national contexts show more variation. In England and the US, there is a well-established comprehensive school system, although within-school ability grouping is relatively common and flexible with low-threshold opportunities for changes over time (for the UK see Boliver and Capsada-Munsech, 2021; Hallam and Parsons, 2013; for the US see Condron, 2008; Lleras and Rangel, 2009; Loveless, 2013). In contrast, ability grouping during primary education is relatively uncommon in Germany (Ammermueller and Pischke, 2009; Eckhardt, 2019). However, in Germany, at the end of primary education (usually at the end of Grade 4 but at the end of Grade 6 in a few federal states), teachers recommend the school track the student should attend in lower secondary education based on, amongst other things, students' school performance in primary education. This is an important recommendation, which is also binding in some federal states, and which has an impact on students' educational careers (i.e., attending a vocational or academic track in secondary education). It remains open whether and how teachers' knowledge that they will have to recommend a school track for each student at the end of primary education already affects teachers' judgements in Grade 1.

Furthermore, each of the three countries developed accountability policies, whereby accountability policies in England and the US may be more consequential for school organisation; here, results from standardised tests are used to monitor school quality and are related to further policies such as allocation of teaching staff or funding of schools (Hartong, 2014; Muench and Wieczorek, 2022). In particular, in England, primary education is divided into Key Stage (KS) 1, which covers ages 5 to 7 (Years 1 and 2), and Key Stage 2, which covers ages 7 to 11 (Years 3–6). At the end of Key Stage 2 (age 11), national standardised exams take place. Key Stage results are published as aggregate statistics and determine ranking in public school league tables, which then affects school desirability to parents, enrolment numbers, and the money schools receive (Burgess and Greaves, 2009; Hall and Ozerk, 2010). As a consequence, schools have an interest to score high in Key Stage exams. These policies are complemented by agencies ensuring that teachers meet set performance standards (Muench and Wieczorek, 2022). In addition, school quality is monitored through self-evaluations by school management and independent school inspectorates (Muench and Wieczorek, 2022). In the US, since the *No Child Left Behind Act* of 2001, there also have been state-wide standardised tests (at least once between Grades 3 and 5) in public schools (e.g., Abernathy, 2007; Figlio and Loeb, 2011; Hanushek and Rivkin, 2010). Results from standardised tests in public schools are used to identify whether schools have met the required progress for a specific year (*Adequate Yearly Progress*, AYP; Hartong, 2014) defined by each state. The AYP holds schools, school districts, and states accountable for the performance of (groups of) students as it forms the basis for educational governance and financial allocations and, hence, is linked to a series of school policy rewards and punishments (Abernathy, 2007; Hartong, 2014). In Germany, in contrast, results from standardised tests in Grade 3, which assess the level of students' competencies compared to the binding cross-states educational standards (the so-called VERA 3; see, e.g., KMK, 2015) are not linked to, for example, resource allocation. Teachers receive feedback on VERA 3 test results at class and student level, each supplemented by descriptions of proficiency levels and the national reference values.

Linked to this, in all three countries, there is a certain degree of standardisation in the form of defined learning goals and curricula (see, e.g., Burgess and Greaves, 2009, 2013; Hall and Ozerk, 2010, for more information on England; see, e.g., Eckhardt, 2019; Muench and Wieczorek, 2022, for Germany; see, e.g., Hartong, 2014, for the US), whereby schools have some autonomy over the learning material used in lessons (Mullis et al., 2016, 2017b). However, especially England shows a high degree of standardisation accompanied by standardised testing, while Germany and the US are characterised by more decentralisation.

Overall, the theoretical considerations from Section 2.3 refer to general processes and applying them to specific contexts might be over-simplified due to the fact that within-country variation could lead to a weakening or strengthening of patterns. In our study, this should especially apply to Germany and the US. In Germany, the federal states are responsible, for example, for the curricula (Eckhardt, 2019, p. 110). In the US, the educational system is also characterised by decentralisation (McGuinn and Manna, 2013): Each state has its own standards and subject-specific accountability requirements (e.g., Figlio and Loeb, 2011; Hanushek and Rivkin, 2010). Further, curricula and funding are determined by school districts (Yanushevsky, 2011) and there are large differences between schools in different districts with respect to curriculum, school resources etc. In addition, the extent of and the use of students' ability as a criterion for grouping varied within the US (e.g., Loveless, 2013, pp. 15–17). For example, for 4th grade reading instruction, Loveless (2013) showed that in 2009, although ability was the criterion for most groupings, in some cases students' interest or aspects of diversity were considered.

## 3. Data and operationalizations

### 3.1. Data

In the present study, we analysed data from three large-scale longitudinal studies conducted in England, Germany, and the US. For the purposes of our study, we only used information from the period of primary education (see Table 1 for further information).

The Millennium Cohort Study (MCS) is an ongoing observational cohort study that began in 2000–2001 (Joshi and Fitzsimons, 2016; University College London, UCL Institute of Education, Centre for Longitudinal Studies, & Department for Education, 2021;

**Table 1**
Survey and data information by country.

| | England | Germany | US |
|---|---|---|---|
| Survey | MCS | NEPS-SC2 | ECLS-K:2011 |
| Primary sampling units | electoral wards | schools | schools |
| Birth cohorts | 2000/1 | 2005/6 | 2004/5 |
| T1: begin of primary education | Y2: age 7 | Grade 1: age 6-7 | Grade 1: age 6-7 |
| T2: end of primary education | Y6: age 11 | Grade 4: age 9-10 | Grade 5: age 10-11 |

Note. MCS: Millennium Cohort Study; NEPS-SC2: National Educational Panel Study Starting Cohort 2; ECLS-K: Early Childhood Longitudinal Study kindergarten.
Source. Compilation according to Olczyk et al. (2022).

University of London, Institute of Education, Centre for Longitudinal Studies, 2021). In its first wave, the MCS drew a representative sample of 18,552 families from across the UK. For data availability reasons, we focus on children resident in England only (i.e., one of the four countries in the UK; $N = 8{,}883$ at T1). Information on Key Stage test results at the end of primary school (i.e., our measures of language and mathematics skills at T2) were matched to the MCS from administrative National Pupil Database (NPD) and were administered only to children in state schools. Matches were achieved for 7,625 children (86%) with valid Key Stage mathematics scores at T2 (7,602 children with matched language scores). Of the approximately 1,250 cases without matched Key Stage scores, around 450 (5% of the baseline sample) are estimated to be children educated in private schools, while the remaining 9% are predominantly due to lack of parental consent for NPD data linkage (see Rihal and Gomes, 2021, for further information on the MCS-NPD data linkage). Another major source of data loss in the analytical sample was lack of response on the teacher postal questionnaire, the source of the teacher assessments of student ability at T1 (38.1% of missing cases from the initial sample). Around 10% of the missing teacher questionnaires were due to lack of parental consent, or insufficient provision of teacher contact details, to send out the postal questionnaire; the remaining missing responses were overwhelmingly due to the failure of teachers to return the postal questionnaires, even after two reminders (Johnson et al., 2011). At the end, the analytical samples for England included 4,721 (language skills) resp. 4,717 (mathematical skills) students.

The German National Educational Panel Study (NEPS) is an ongoing national longitudinal study aimed at providing data on competence development, educational decisions, and returns to education throughout the lifespan using a multi-cohort sequence design (Blossfeld and Roβbach, 2019). In the present study, we used data from the Starting Cohort 2 (NEPS-SC2; https://doi.org/10.5157/NEPS:SC2:9.0.0; NEPS Network, 2020), which is assumed to be representative for children in Grade 1 in the school year 2012/13.[3] 6,733 students from 374 schools were tested in Grade 1, spring 2013, and 5,636 parents were interviewed by telephone. In general, participation in the NEPS is voluntarily. This does not only lead to missing teacher assessment in Grade 1, but also to panel drop-outs at later waves (see Zinn et al., 2020: 194f., Table 9, for more information on selectivity in panel drop-outs in the NEPS). The analytical samples for Germany comprised 3,870 (language skills) resp. 3,738 (mathematical skills) students.

The Early Childhood Longitudinal Study, Kindergarten Class 2010–11 (ECLS-K:2011) is a longitudinal study which followed a nationally representative sample of students across the US from kindergarten through the fifth grade (Tourangeau et al., 2015).[4] The baseline sample size is approximately 18,170 children enrolled in about 970 schools. In Grade 1, about 15,110 students took part (sample sizes are rounded to nearest 10, as required by the National Center for Education Statistics). At T2, when students attend Grade 5, approximately 20% of the baseline sample became ineligible for the survey due to various reasons, including moving out of the country, transferring to schools that fall outside the scope of the study, or passing away. The analytical samples for the US included information on 8,420 (language skills) resp. 8,410 (mathematical skills) students.

### 3.2. Instruments

#### 3.2 1. Teacher judgements at the beginning of primary education (T1)

In England, teachers rated each students' reading, writing, and speaking ability in relation to all children of the same age (1 = *well below average* to 5 = *well above average*).[5] For our analyses, we used the mean score of these three ratings ($\alpha = 0.91$). In Germany, too, a mean score of teacher rating (three items; $\alpha = 0.91$) was calculated. Here, teachers assessed students' general language skills (e.g., vocabulary, sentence construction) and students' written language abilities (e.g., ability to understand and write texts) in relation to those of same-aged children (1 = *much worse* to 5 = *much better*). Furthermore, teachers rated the statement "The child has very good language skills" (1 = *does not apply* to 4 = *does apply*). In the US, a single item was used: teachers reported how they would rate the language and literacy skills of a student in comparison to other children of the same grade level (1 = *far below average* to 5 = *far above average*).

In addition, in all three studies, teachers were asked to rate mathematical skills of each student on a 5-point-scale (1 = *much worse* to

[3] The NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi, Germany) in cooperation with a nationwide network.
[4] All reported ECLS-K:2011 sample sizes are rounded to the nearest 10 in accordance with National Center for Education Statistics (NCES) regulations.
[5] In England, the original scale goes form *well above* to *well below average* so we reversed it.

5 = *much better* in Germany; 1 = *well/far below average* to 5 = *well/far above average* in England and in the US). While teachers in the US compared the child to other children of the same grade level, teachers in England and Germany were asked to compare the student to same-aged children.

### 3.2 2. *Tests on language skills, mathematical skills, and cognitive abilities*

*Language skills.* In England, language skills at the beginning of primary education were captured by the British Ability Score II (BAS II) Word Reading test (Elliott et al., 1996). The BAS II assessed students word decoding ability such as the recognition and oral reading of single words, as well as vocabulary knowledge. We used IRT-scaled ability scores to scale the results, accounting for the set of responses the child was presented with (adaptive routing led to variation in the difficulty of the items with which children were presented). At the end of primary school, we relied on the Key Stage 2 reading test score marks.

In Germany, we used the information on students' receptive vocabulary (as measured by a modified Peabody Picture Vocabulary Test; PPVT) tested in Grade 1 (Berendes et al., 2013) and receptive grammar skills (as measured by the Test for Reception of Grammar [TROG] in Grade 1; Lorenz et al., 2017). At Grade 4, we relied on students' performance in reading comprehension. The NEPS developed a test of reading comprehension (see Gehrer et al., 2013; Weinert et al., 2011, for the description of the theoretical framework) which aims to assess three cognitive requirements (i.e., finding information in texts, drawing text-related conclusions, and reflecting and assessing). The test consists of five texts that each represent one text type or text function (i.e., information, commenting or arguing, literary, instruction and advertising) and five item sets referring to these texts. All test results are IRT-scaled by the NEPS data centre.

In the US, students language skills assessment in Grades 1 and 5 in the ECLS-K:2011 were largely based on the National Assessment of Educational Progress (NAEP) Reading Frameworks for 2009 and 2011, respectively (Najarian et al., 2018, 2020). The test included items on basic literacy skills (e.g., print familiarity, word recognition), vocabulary knowledge, and reading comprehension (Tourangeau et al., 2019). We used an IRT-scaled score provided in the ECLS-K:2011, which measures a child's latent ability in each domain (i.e., language and mathematics).

**Table 2**
Descriptive statistics by country.

| | | England | | Germany | | US | |
|---|---|---|---|---|---|---|---|
| | Time | *M*/% | *SD* | *M*/% | *SD* | *M*/% | *SD* |
| **Language skills** | | | | | | | |
| Teacher assessment: lang. (std.) | T1 | 0 | 1 | 0 | 1 | 0 | 1 |
| Lang. achievement (std.) | T1 | 0 | 1 | 0 | 1 | 0 | 1 |
| Lang. achievement, grammar (std.) | T1 | *n.a.* | | 0 | 1 | *n.a.* | |
| Lang. achievement (std.) | T2 | 0 | 1 | 0 | 1 | 0 | 1 |
| Cognitive abilities (std.) | T1 | 0 | 1 | 0 | 1 | 0 | 1 |
| Late assessment at T1 | T1 | 59.7 | | 36.4 | | 61.5 | |
| Age-in-months at T1 testing | T1 | 86.75 | 2.92 | 84.85 | 4.67 | 85.57 | 4.38 |
| Time span testing T2-T1 (in months) | T2-T1 | 48.47 | 1.96 | 32.13 | 1.56 | 48.10 | 1.08 |
| Student female | T1 | 50.5 | | 50.9 | | 49.3 | |
| Highest parental education | T1 | | | | | | |
| High | | 32.9 | | 37.8 | | 42.7 | |
| Medium | | 27.4 | | 52.0 | | 29.3 | |
| Low | | 39.6 | | 10.1 | | 28.1 | |
| Student of immigrant descent | T1 | 19.3 | | 22.0 | | 31.5 | |
| *N* | | 4,721 | | 3,870 | | 8,420[a] | |
| **Mathematics** | | | | | | | |
| Teacher assessment: math. (std.) | T1 | 0 | 1 | 0 | 1 | 0 | 1 |
| Math. achievement (std.) | T1 | 0 | 1 | 0 | 1 | 0 | 1 |
| Math. achievement (std.) | T2 | 0 | 1 | 0 | 1 | 0 | 1 |
| Cognitive abilities (std.) | T1 | 0 | 1 | 0 | 1 | 0 | 1 |
| Late assessment at T1 | T1 | 59.6 | | 35.8 | | 61.5 | |
| Age-in-months at T1 testing | T1 | 86.75 | 2.91 | 84.83 | 4.66 | 85.57 | 4.38 |
| Time span testing T2-T1 (in months) | T2-T1 | 48.46 | 1.96 | 32.13 | 1.57 | 48.10 | 1.08 |
| Student female | T1 | 50.2 | | 51.1 | | 49.3 | |
| Highest parental education | T1 | | | | | | |
| High | | 32.7 | | 38.0 | | 42.7 | |
| Medium | | 27.4 | | 51.7 | | 29.3 | |
| Low | | 39.9 | | 10.4 | | 28.1 | |
| Student of immigrant descent | T1 | 19.3 | | 22.1 | | 31.5 | |
| *N* | | 4,717 | | 3,738 | | 8,410[a] | |

*Notes*. n.a.: not applicable. std.: z-standardised.
Time refers to measurement time with T1 indicating Grade 1 (Germany, US) or Year 2 (England) and T2 indicating Grade 4 (Germany), Grade 5 (US) or Year 6 (England).
[a] For the US, sample sizes are rounded to nearest 10, as required by the National Center for Education Statistics.
Sources. Own calculations based on MCS, NEPS-SC2, and ECLS-K:2011 (see also Olczyk et al., 2022).

*Mathematical skills.* In England, mathematical skills were captured by an adapted version of the National Foundation for Educational Research (NFER) Progress in Maths test (which assessed numbers, space, measurement, and data handling) when children were 7 years old. We used the NFER Progress in Maths test ability score. At the child age of 11 years (end of Year 6), we relied on Key Stage 2 mathematics test score marks (i.e., a component of the compulsory standardised assessment) from the National Pupil Database.

In Germany, we used results from mathematical tests constructed by the NEPS and assessed in Grade 1 and Grade 4. The tests cover content-related (i.e., quantity, space and shape, change and relationships, data and change) and process-related components (i.e., applying technical skills, representing, modelling, communicating, problem solving; Schnittjer et al., 2020). Again, test results are IRT-scaled by the NEPS data centre and available in the Scientific Use Files.

In the US, mathematical skills were assessed by mathematical tests conducted in Grades 1 and 5. The ECLS-K:2011 mathematics framework for Grade 1 is primarily based on the mathematics assessment from the older cohort of the ECLS-K. The ECLS-K framework was developed based on the NAEP 1996 framework. The ECLS-K:2011 mathematics framework was updated to reflect recommendations in the NAEP 2005, the Principles and Standards for School Mathematics guidelines of the National Council of Teachers of Mathematics, and state standards (California, New Jersey, Tennessee, Texas, and Virginia; Najarian et al., 2018). The mathematics framework for Grade 5 was largely based on the framework for Grade 1 and also integrated recommendations in the NAEP 2011 framework. The framework aims to measure skills in number properties and operations, measurement, geometry, data analysis and probability, and algebra (Najarian et al., 2020). We used an IRT-scaled score to measure mathematical skills of children in the US.

*Nonverbal cognitive abilities (T1).* In England, nonverbal cognitive abilities were measured by the British Ability Score II Pattern Construction test (Elliott et al., 1996; Jones and Schoon, 2008). Children's task was to replicate patterns presented to them using solid plastic cubes. In the present study, we used the ability score, as the number of items administered depended on students' performance during the assessment.

In Germany, we used results from the NEPS-MAT test (sum score; Lang et al., 2014) administered at Grade 2 to assess nonverbal cognitive abilities. The test included horizontally and vertically arranged fields with different geometrical elements. Children's task was to select the right complement for one free field from several offered solutions on the basis of deduced logical rules which underlie the patterns of the geometrical elements.

In the US, working memory was measured by the Numbers Reversed task (Blackwell, 2001). Children's task was to recall an orally presented sequence of numbers and repeat the sequence in reverse order. In our analyses, we used the age-standardised score (the W score), representing both a child's ability and the task difficulty (Tourangeau et al., 2015).

We z-standardised all indicators for student's achievement and teacher judgements to allow for cross-country comparisons.

#### *3.2 3. Time aspects*

In all three studies, the *data collection* spanned over several months in the case of teacher judgement. As in earlier work (Olczyk et al., 2022), we accounted for this variation in the present study by creating a dummy variable (0 = *a teacher rated the students early*; 1 = *a teacher rated the students later in the data collection process*). In England, where the school year runs from September to July, students who were assessed by teachers between September and January were assigned to early, and those assessed between February and July to late. In Germany, we considered early assessments to take place between February and March and late ones between April to June. In the US, we defined assessments between January and March as early ones and those between April and June as late. Besides the time of assessment, we also considered the *age in months at the time of testing* at T1. Furthermore, examining the effects of biased assessment on later achievement, the *time span* between testing at T2, towards the end of primary education, and testing at T1, at the beginning of primary education (in months), were controlled for.

#### *3.2 4. Student characteristics*

A central variable in our analyses was *student gender* (0 = *male*; 1 = *female*).

Furthermore, we considered the *highest education* (i.e., low, middle, and high) of parents living together with the child at the beginning of primary education. Hence, where there was only one co-resident parent, the family was categorised based on her or his level of education, whilst in case a child lived with two resident parents, the family was categorised based on the more highly educated of the two.

As an indicator of *immigration status*, we considered information provided by parents on whether a student and/or at least one parent was foreign born (0 = *no immigration*; 1 = *foreign-born student and/or at least one parent*).

In general, in all three countries information provided by parents was used; if missing, in the case of Germany, we relied on further information sources from teachers.

Descriptive statistics of the study variables are displayed in Table 2.

### *3.3. Analytic strategy*

To identify the proportion of bias in teacher judgements, we followed the residual approach suggested by Madon and colleagues (1997; see, e.g., Gentrup et al., 2020; Hinnant et al., 2009, for other studies applying this approach). Therefore, we conducted multiple regression analyses predicting teacher judgements in language and mathematics skills from students' results in language or mathematical achievement tests as well as students' general cognitive abilities. We used cluster robust standard errors, as students are clustered by teachers (see also our earlier work, Olczyk et al., 2022). The residuals of these regressions show the variance which could not be explained by the abovementioned student characteristics and, hence, could be interpreted as teacher judgement bias. We z-standardised the residuals. A positive residual score represents positively biased judgements (i.e., teacher overestimation), while a

negative one reflects a negatively biased judgement (i.e., teacher underestimation). Values close to zero indicated unbiased judgements, which represent an accurate prediction of student achievement (Madon et al., 1997). Aiming to investigate whether teacher judgement bias varied with student gender, we compared the mean residual scores for boys and for girls in the language domain and in mathematics, respectively.

To examine whether teacher judgement bias was related to gender achievement gaps, we separately estimated linear regression models for language and mathematical skills towards the end of primary education and applied a stepwise approach (see also our earlier work, Olczyk et al., 2022). In particular, in Model 1, we studied whether there are gender gaps in achievement development by using information on test results from Grade 1 (Germany and US) or Year 2 (England), as well as further controls (i.e., parental education and immigration status); again, we estimated clustered standard errors. In Model 2, the residual score was added with the aim of investigating the question of whether gender gaps are mediated by teacher judgement bias.

All results are based on complete case analysis. Inspection of the distribution of variables available for the initial samples revealed only very minor differences compared to the distributions in the analysed samples with only complete cases (see Tables S1a–S1c in the supplementary material, for information on sample characteristics including the share of missing values based on the initial samples).

## 4. Results

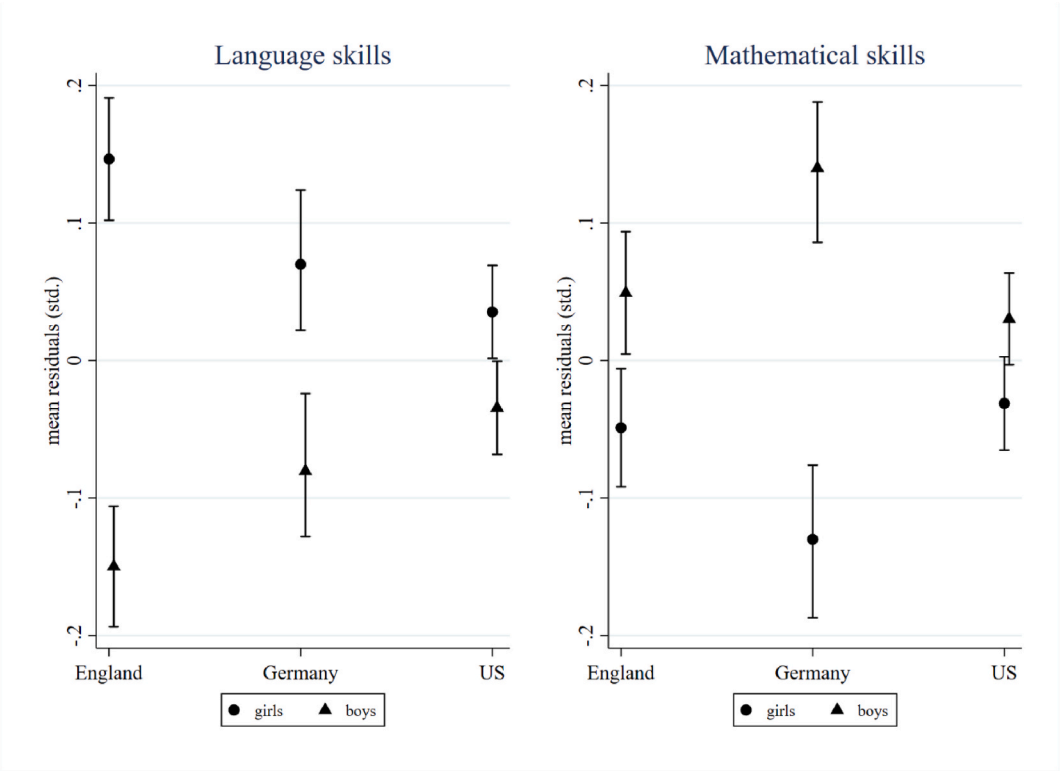
### 4.1. Teacher judgement bias and students' gender

Table 3 displays results from linear regressions with teacher judgement of student's mathematical and language skills as outcomes. Overall, the share of explained variance was highest for teacher judgements in the language domain in England (54% versus 37% for mathematics) and in the US (51% versus 40% for mathematics), and much lower in Germany (25% for both domains).

**Table 3**
Results of regression models for teacher judgement (z-standardised) by country.

| | England | Germany | US |
|---|---|---|---|
| | β (*SE*) | β (*SE*) | β (*SE*) |
| **Language skills** | | | |
| T1 lang. achievement (std.) | .64* | .19* | .68* |
| | (.02) | (.02) | (.03) |
| Late assessment at T1 (*ref. early*) | .15* | .01 | −.04* |
| | (.02) | (.04) | (.02) |
| Interaction between late assessment at T1 and T1 lang. achievement (std.) | .03 | −.00 | .01 |
| | (.02) | (.04) | (.03) |
| T1 lang. achievement, grammar (std.) | *n.a.* | .30* | *n.a.* |
| | | (.02) | |
| Interaction between late assessment at T1 and T1 grammar (std.) | *n.a.* | −.01 | *n.a.* |
| | | (.04) | |
| T1 cognitive abilities (std.) | .16* | .16* | .05* |
| | (.01) | (.01) | (.01) |
| Age-in-months at T1 testing | .01* | −.01* | −.00 |
| | (.00) | (.00) | (.00) |
| Constant | −0.94* | 1.07* | 0.05 |
| | (.31) | (.26) | (.17) |
| $R^2$ | .539 | .253 | .506 |
| *N* | 4,721 | 3,870 | 8,420[a] |
| **Mathematics** | | | |
| T1 math. achievement (std.) | .48* | .43* | .57* |
| | (.02) | (.02) | (.02) |
| Late assessment at T1 (*ref. early*) | .22* | −.12* | −.01 |
| | (.03) | (.04) | (.02) |
| Interaction between late assessment at T1 and T1 math. achievement (std.) | .00 | .04 | .00 |
| | (.02) | (.03) | (.02) |
| T1 cognitive abilities (std.) | .18* | .15* | .10* |
| | (.01) | (.02) | (.01) |
| Age-in-months at T1 testing | .02* | −.01 | −.00 |
| | (.00) | (.00) | (.00) |
| Constant | −1.66* | 0.44 | 0.20 |
| | (.36) | (.28) | (.18) |
| $R^2$ | .365 | .255 | .401 |
| *N* | 4,717 | 3,738 | 8,410[a] |

*Notes*. *$p < .05$. n.a.: not applicable. std.: z-standardised.
[a] Sample sizes are rounded to nearest 10, as required by the National Center for Education Statistics.
Sources: Own calculations based on MCS, NEPS-SC2, and ECLS-K:2011 (see also our earlier work, Olczyk et al., 2022).




*Note.* Displayed are the means and 95% confidence intervals for girls and boys (values are in Table S2a in the supplementary material).

**Fig. 1.** Teacher judgement bias (mean residuals) for students' achievement by student gender and country
*Note.* Displayed are the means and 95% confidence intervals for girls and boys (values are in Table S2a in the supplementary material).
Sources. Own calculations based on MCS, NEPS-SC2, and ECLS-K:2011.

Next, we examined whether the standardised residuals from these regressions and, hence, the degree and direction of inaccuracy varied systematically by students' gender (see Fig. 1, for the mean residuals by gender). In the language domain, teacher judgements of female students showed on average a positive bias while male students on average were underestimated by their teachers. In mathematics, the reversed pattern was observable, indicating on average a positive bias for boys and a negative bias for girls.

However, the extent of gender-specific bias varied between the three countries (see Table S2b in the supplementary material): For language skills, the extent of gender-specific bias was largest in England ($\beta = 0.30$, $SE = 0.03$, $p < .001$), followed by Germany ($\beta = 0.15$, $SE = 0.03$, $p < .001$), and the US ($\beta = 0.07$, $SE = 0.02$, $p = .002$). For mathematics, the bias was largest for Germany ($\beta = -0.27$, $SE = 0.03$, $p < .001$), followed by England ($\beta = -0.10$, $SE = 0.03$, $p < .001$), and the US ($\beta = -0.06$, $SE = 0.02$, $p = .005$).

### 4.2. *Teacher judgement bias, students' gender, and student progress in primary school*

Our second research question pertained to the association between teacher judgement bias and later student achievement (see Table 4). M0 shows the difference between boys' and girls' achievement at T2 without considering any controls. All further models (M1-M2) controlled for test results measured at T1, at the beginning of primary education, as well as for socio-demographic characteristics of the student/family and the time elapsed between testing at T1 and T2. In all three countries, at T2, there was a gender achievement gap in advantage of girls in the language domain and in advantage of boys in mathematics (see M0 in Table 4). In the language domain, girls' advantage increased throughout primary school in England and Germany, whereas the gender gap narrowed in the US. In mathematics, the gender achievement gap in advantage of boys increased in all three countries throughout primary education (see M1 in Table 4).

Next, we added the standardised residuals (see M2 in Table 4). Inaccurate teacher judgements at the beginning of primary education were associated with students' achievement at later time points, such that overrated students performed better later on. With respect to the gender effect, for language skills in England and Germany, we observed a significant reduction in the effect of students' gender after controlling for biased judgements (England: $\Delta\beta = 0.14 - 0.22 = -0.08$, $SE = 0.01$, $p < .001$; Germany: $\Delta\beta = 0.08 - 0.11 = -0.03$, $SE = 0.01$, $p < .001$). Hence, 36% (England) resp. 27% (Germany) of the gender achievement gap between T1 and T2 was explained here by biased teacher judgements at T1 ($\Delta\beta/\beta_{M1}$). In the US, in contrast, there was a significant increase of the gender effect indicating a more pronounced narrowing of the gender gap throughout primary education when teacher judgement bias was

**Table 4**
Results of regression models for later student achievement (z-standardised) by country.

| | England | | | Germany | | | US | | |
|---|---|---|---|---|---|---|---|---|---|
| | M0[a] | M1 | M2 | M0 | M1 | M2 | M0 | M1 | M2 |
| | β (*SE*) | β (*SE*) | β (*SE*) | β (*SE*) | β (*SE*) | β (*SE*) | β (*SE*) | β (*SE*) | β (*SE*) |
| **Language skills** | | | | | | | | | |
| Student female (*ref. male*) | .28* | .22* | .14* | .17* | .11* | .08* | .07* | −.06* | −.07* |
| | (.03) | (.02) | (.02) | (.03) | (.03) | (.03) | (.02) | (.02) | (.01) |
| Teacher judgement residuals (std.) | | | .25* | | | .19* | | | .11* |
| | | | (.01) | | | (.02) | | | (.01) |
| Controls | | ✓ | ✓ | | ✓ | ✓ | | ✓ | ✓ |
| Constant | | −1.83* | −0.43 | −0.09* | −0.86* | −0.88* | −0.04 | −1.38* | −1.18* |
| | | (.28) | (.28) | (.03) | (.31) | (.30) | (.02) | (.33) | (.33) |
| $R^2$ | .020 | .418 | .474 | .007 | .350 | .383 | .001 | .568 | .581 |
| *N* | | 4,721 | | | 3,870 | | | 8,420[b] | |
| **Mathematics** | | | | | | | | | |
| Student female (*ref. male*) | −.11* | −.10* | −.07* | −.15* | −.06* | −.01 | −.11* | −.06* | −.05* |
| | (.03) | (.02) | (.02) | (.03) | (.03) | (.03) | (.02) | (.01) | (.01) |
| Teacher judgement residuals (std.) | | | .34* | | | .17* | | | .11* |
| | | | (.01) | | | (.01) | | | (.01) |
| Controls | | ✓ | ✓ | | ✓ | ✓ | | ✓ | ✓ |
| Constant | | −1.85* | −0.10 | 0.08* | −1.71* | −1.75* | .06* | −1.69* | −1.69* |
| | | (.27) | (.25) | (.03) | (.31) | (.30) | (.02) | (.33) | (.32) |
| $R^2$ | .003 | .458 | .567 | .006 | .406 | .433 | .003 | .651 | .663 |
| *N* | | 4,717 | | | 3,738 | | | 8,410[b] | |

*Notes*. *$p < .05$. std.: z-standardised.
Controls included T1 achievement, cognitive abilities, time span between testing, highest parental education, immigration status. Complete models in the supplementary material, Table S3.
Testing of significant changes between M1 and M2 of the gender effect revealed.
England: language skills: $\Delta\beta = -0.08$, $SE = 0.01$, $p < .001$; mathematics: $\Delta\beta = 0.03$, $SE = 0.01$, $p = .003$.
Germany: language skills: $\Delta\beta = -0.03$, $SE = 0.01$, $p < .001$; mathematics: $\Delta\beta = 0.05$, $SE = 0.01$, $p < .001$.
US: language skills $\Delta\beta = -0.01$, $SE = 0.00$, $p = .005$; mathematics: $\Delta\beta = 0.01$, $SE = 0.00$, $p = .004$.
[a] The estimated constant for England is suppressed under the statistical disclosure rules of the UK Data Service.
[b] Sample sizes are rounded to nearest 10, as required by the National Center for Education Statistics.
Sources: Own calculations based on MCS, NEPS-SC2, and ECLS-K:2011.

controlled for ($\Delta\beta = -.07 - (-0.06) = -0.01$, $SE = 0.00$, $p = .005$).[6] For mathematics, the gender effect significantly decreased in all three countries after controlling for biased judgement (England: $\Delta\beta = 0.03$, $SE = 0.01$, $p = .003$; Germany: $\Delta\beta = 0.05$, $SE = 0.01$, $p < .001$; the US: $\Delta\beta = 0.01$, $SE = 0.00$, $p = .004$). Overall, this pattern supported our expectation that biased teacher judgements would contribute to widening gender achievement gaps over time.

### 4.3. Sensitivity checks and further analysis

*Students' motivation and work habits.* As earlier research has shown, teacher beliefs about students' motivation and work habits vary substantially by student gender and are related to teacher judgements of student abilities. Thus, girls are seen by their teachers as academically more motivated and more eager to learn than boys (e.g., Gentrup et al., 2018; Jussim et al., 1996; Trautwein and Baeriswyl, 2007). These differences also relate to gender bias in ability judgements: Whereas in the language domain teacher judgements of motivation and work habits have been shown to fully account for the advantage of girls, in mathematics, the bias in favour of boys increased when teacher judgement of student motivation and work habits were controlled for (Duckworth and Seligman, 2006; Gentrup et al., 2018; Robinson-Cimpian et al., 2014). Therefore, to check the robustness of the results presented above, we included teacher perceptions of students' motivation and work habits and re-calculated the models for Germany and the US (see Table S4 and Table S5 in the supplementary material; for England, no information on teacher perceptions of students' motivation and work habits at T1 are available in the data). With respect to judgement bias, results for both countries supported previous findings. In particular, first, teacher beliefs about students' motivation and work habits were positively related to teacher judgement bias of student achievement (see Table S4). Second, the gender effect in the language domain reduced, while it increased in mathematics (see Table S4). Interestingly and inconsistent to the findings of earlier studies, a weak, but statistically significant male advantage became visible in the language domain after controlling for the teachers' perception of students' motivation and work habits. With respect to later achievement, the results changed slightly compared to the baseline findings presented above. In the language domain, in Germany, controlling for teacher perceptions of students' motivation and work habits reduced the gender effect on student later

[6] We used the suest command implemented in Stata to compare the coefficients (Mize et al., 2019).

achievement and adding teacher judgement bias to the model did not further reduce the gender effect (see Table S5). In the US, the narrowing of the gender gap in the language domain was somewhat more pronounced when teacher perceptions of student motivation and work habits were considered. But, contradicting our results presented above, this effect did not increase but decreased when teacher judgements of student motivation and work habits were controlled for (see Table S5). However, the change in the coefficient is very small in both cases (without teacher perceptions of student motivation and work habits: $\Delta\beta = -.01$; with teacher perceptions of student motivation and work habits $\Delta\beta = 0.01$). In mathematics, in both countries, the gender achievement gap in the disadvantage of girls was somewhat more pronounced when teacher perceptions of students' motivation and work habits were controlled for (see Table S5). Nevertheless, teacher judgement bias of student achievement accounted for an additional amount of the gender gap. In Germany, this additional explanatory effect was of the same magnitude as in our baseline results presented in Table 4 (both without and with teacher perception of student motivation and work habits: $\Delta\beta = .05$), indicating robust findings. In the US, an even larger amount of the gender gap was explained by teacher judgement bias when teacher perception of student motivation and work habits were held constant ($\Delta\beta = .03$ compared to $\Delta\beta = 0.01$).

*Language skills.* In the German survey, the language tests in Grade 1 focus on more specific skills (i.e., vocabulary and grammar skills) than the tests in England and the US. As for Germany data regarding students' early reading skills are also available (measured in Grade 2 with the modified ELFE; Lenhard and Schneider, 2006), we additionally considered this more direct measure of early reading skills in sensitivity checks. Test results for vocabulary, grammar, early reading skills and cognitive abilities should then cover to a larger degree students' "true" language skills that are observed and assessed by the teachers. Variations in teacher ratings beyond these comprehensive indicators of student performance should then map bias. The results regarding biased judgements and their association with achievement development are largely comparable to the results presented in the main text (see Tables S6–S8), although the effect of teacher judgement bias on later achievement was smaller in these models where additionally early reading skills were considered. However, as the ELFE test took place in Grade 2, the performance in this test may already have been influenced by teacher judgement bias.

*Heterogeneous effects of biased teacher judgements.* As some student groups have been found to be more sensitive to teacher judgements than others (Gentrup and Rjosk, 2018; Jussim et al., 1996; McKown and Weinstein, 2002), we tested whether the effects of teacher judgements on later student achievement differed for boys and girls. Only in the US, we found that the association of biased teacher judgements with math achievement was significantly stronger for female students as compared to male students (see Table S9). In England and Germany, the effects of teacher judgement bias on language and mathematical skills did not significantly vary with students' gender.

## 5. Conclusion and discussion

In the present study, we examined teacher judgements of student achievement and their role in gender achievement gaps. We analysed, first, whether inaccurate teacher judgements of students' language and mathematical skills correlate with students' gender and, second, whether such bias is associated with achievement development in primary school.

With respect to the first research question, for the language domain, we found that teacher judgements of female students showed on average a positive bias, while male students on average were underestimated by their teachers. In mathematics, the reversed pattern was observed with positive bias for boys and negative bias for girls. This finding is in line with previous research on systematic variations in teacher judgements based on students' gender in the UK, Germany, and in the US (e.g., Campbell, 2015; Gentrup and Rjosk, 2018; Geven et al., 2021; Hinnant et al., 2009).

With respect to the second research question, the association between teacher judgement bias and later achievement, in the language domain, we observed that the girls' advantage in achievement widened throughout primary school in England and Germany (for similar findings, see, e.g., Cavaglia et al., 2020; Machin and McNally, 2005, for England; see, e.g., Ehrtmann and Wolter, 2018, for Germany), whereas the gender gap slightly narrowed in the US (see, e.g., Robinson and Lubienski, 2011, for similar evidence; for contrasting evidence reporting an increasing gender effect, see, e.g., Petersen, 2018; Reilly et al., 2019). In mathematics, the gender achievement gap in advantage of boys increased in all three countries throughout primary education (for similar findings, see, e.g., Cavaglia et al., 2020; Machin and McNally, 2005, for England; see, e.g., Cimpian et al., 2016; Robinson and Lubienski, 2011, for the US; for deviating results, see, e.g., Winkelmann et al., 2008, for Germany). Furthermore, the results showed that teacher judgement bias predicted students' end of primary school achievement in all three countries, even when considering prior achievement, cognitive abilities, and students' background characteristics. This finding could be understood as the product of a self-fulfilling prophecy. Regarding the contribution of teacher judgement bias to gender achievement gaps, in all three countries the effect of gender on later student achievement was partly mediated by inaccurate teacher judgements.

Besides these robust patterns, there were also cross-country variations (see also Table 5). For judgement bias, we found that the

**Table 5**
Summary of results on teacher judgement bias in comparison between countries.

| Domain | Degree of bias | | | Bias and later achievement | | | Contribution of bias to gender gap in achievement | | |
|---|---|---|---|---|---|---|---|---|---|
| | England | Germany | US | England | Germany | US | England | Germany | US |
| Language | high | middle | low | high | middle | low | high | middle | low |
| Mathematics | middle | high | low | high | middle | low | middle | high | low |

Source. Own compilation based on findings presented in Sections 4.1 and 4.2.

extent of biased teacher judgements differed by country and domain: While in the language domain the gender bias was strongest in England, followed by Germany, and was lowest in the US, in mathematics, the gender bias was strongest in Germany, followed by England and the US. Overall, gender-related teacher judgement bias was less pronounced in the US compared to England and Germany.

With respect to the impact of teacher judgement bias on subsequent achievement, we found that judgement bias in general contributed to achievement development. However, in both domains, these effects were strongest in England, followed by Germany, and least pronounced in the US. Furthermore, we observed, for England and Germany, that judgement bias partly mediated the effect of gender on later achievement. In the US, only a negligible mediation was observable, which is not surprising since there was a very weak relation between inaccurate teacher judgements and gender as well as between inaccurate teacher judgements and later student achievement.

Unfortunately, we could not directly account for the underlying mechanisms such as the degree of gender (in)equality in society or ability grouping and actual accountability approaches used in schools and/or classes, and, hence, we do not know which of the supposed mechanisms actually lead to the patterns observed and if and how the different processes interplay. Also, the patterns found for the cross-country variations do not allow direct conclusions on the presumed underlying processes as they partly deviate from the theoretical considerations: For example, England and the US are characterised by ability grouping and school accountability, which should support a lower teacher judgement bias according to the theoretical considerations. While the findings for the US support these assumptions, they do not for England.

The situation is further complicated by the fact that we examined post-hoc harmonised data which differed in survey designs, measurement points, test materials, and wording of questions (Law et al., 2021). Even though we have tried to make the data as comparable as possible, some issues remain. For example, despite the fact that mathematics achievement tests should measure similar facets of mathematical skills (e.g., number knowledge, knowledge of geometry, and spatial sense), single test items were not accessible. Hence, we could not eliminate variations that are grounded in different instruments (e.g., Winkelmann et al., 2008). Consequently, country differences in the inaccuracy of teacher judgements could be due to, for example, country differences in the measures used (but also the time of measurement) to capture teacher judgements and student ability.

Nevertheless, we were able to show that there are generally robust associations and that teacher bias – if present – contributes to gender differences in achievement, so more attention to this topic is merited. Further limitations of our study are:

First, we linked results from standardised assessments with global teacher judgements which might raise concerns (e.g., Arens et al., 2017; Hübner et al., 2022; Jussim and Harber, 2005). The reason is that teacher judgements could be more accurate than test results as teachers might have "valid" information above and beyond what tests measure. The additional information teachers potentially possess would also account for the fact that children with higher teacher judgements perform better in later achievement tests.

Second, analyses of gender bias in teacher judgements should consider teachers' assessments of students' motivation and work habits as our sensitivity checks for Germany and the US showed. Unfortunately, we were not able to consider teacher assessments of students' motivation and work habits for England. In consequence, a systematic comparison of the results with and without considering the teachers' views of the students' motivation between the three countries under consideration was not possible. Future studies should deal with this desideratum and examine whether the findings occur in the same way in the UK/England as well as further countries.

Third, previous research revealed that direct teacher judgements such as those referring to the expected number of correctly solved tasks of a math or language test correlate more strongly with actual student skills and abilities compared to the indirect measures we used (Hoge and Coladarci, 1989; see also Südkamp et al., 2012). Hence, direct teacher judgements tend to be more accurate than indirect ones. According to this finding, it might be the case that (cross-country variation in) teacher judgement bias would be less pronounced if more specific measures of teacher judgement measures are used.

Fourth, we did not consider teachers' gender. Primary education tends to be dominated by female teachers (OECD.Stat, 2022), as reflected in our data with, for example, a share of 94% resp. 93% at T1 in Germany resp. England. However, as empirical findings from earlier studies have not tended to find that the match between teacher and student gender matters substantively for students' achievement development (see, e.g., Antecol et al., 2015; Coenen and van Klaveren, 2016; Neugebauer et al., 2011; Watson et al., 2019), this should be less relevant for our findings.

Despite these limitations our findings stress the existence of gender bias in teacher judgements and its relevance for boys' and girls' achievements in reading and mathematics. They highlight the need for future research, in particular, to consider the implications of the institutional and societal setting for bias in teacher judgements and its effects on student achievement. This refers not only to cross-country studies, but also to within-country studies, especially when the educational setting varies within a country between specific entities such as states or even schools and classes. Besides studying the underlying processes, future studies should also examine to what extent there is mutual reinforcement or weakening between the societal and institutional features under study such as, for example, accountability and ability grouping. Related to this, different mechanisms may operate in different settings, but could lead to similar gender achievement gaps between these entities. Hence, it is worthwhile to examine specific mechanisms for various entities such as countries – even if there is no variation between them in the extent of gender achievement gaps.

## Funding

The Development of Inequalities in Child Educational Achievement: A Six Country Study (DICE) is an Open Research Area (ORA)-funded project. We gratefully acknowledge funding support from the Economic and Social Research Council (ESRC Grant ES/S015191/1, United Kingdom) and the Deutsche Forschungsgemeinschaft (DFG, Germany, SCHN 1116/1-1; WE 1478/12–1). Jane Waldfogel also gratefully acknowledges support from the Columbia Population Research Center which is funded by NICHD;

2P2CHD058486.

## Code availability

The code for the analysis of the NEPS data (Germany) is available at the Open Science Framework (OSF): https://osf.io/vn56m/?view_only=b7273c5e2f5d4ed8be85d50649f10b1b.

## Declarations of competing interest

Authors declare that they have no competing interests.

## Acknowledgements

We are grateful for the reviewer's detailed and valuable comments.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ssresearch.2023.102938.

## References

Abernathy, S.F., 2007. No Child Left behind and the Public Schools. University of Michigan Press, Ann Arbor, MI. https://doi.org/10.3998/mpub.184344.
Ammermueller, A., Pischke, J.-S., 2009. Peer effects in European primary schools: evidence from the progress in international reading literacy study. J. Labor Econ. 27 (3), 315–348. https://doi.org/10.1086/603650.
Antecol, H., Eren, O., Ozbeklik, S., 2015. The effect of teacher gender on student achievement in primary school. J. Labor Econ. 33 (1), 63–89. https://doi.org/10.1086/677391.
Arens, A.K., Marsh, H.W., Pekrun, R., Lichtenfeld, S., Murayama, K., vom Hofe, R., 2017. Math self-concept, grades, and achievement test scores: long-term reciprocal effects across five waves and three achievement tracks. J. Educ. Psychol. 109 (5), 621–634. https://doi.org/10.1037/edu0000163.
Babad, E., 1990. Measuring and changing teachers' differential behavior as perceived by students and teachers. J. Educ. Psychol. 82 (4), 683–690. https://doi.org/10.1037/0022-0663.82.4.683.
Babad, E., 1993. Teachers' differential behavior. Educ. Psychol. Rev. 5 (4), 347–376. https://doi.org/10.1007/BF01320223.
Barone, C., 2011. Some things never change: gender segregation in higher education across eight nations and three decades. Sociol. Educ. 84 (2), 157–176. https://doi.org/10.1177/0038040711402099.
Berendes, K., Weinert, S., Zimmermann, S., Artelt, C., 2013. Assessing language indicators across the lifespan within the German national educational panel study (NEPS). Journal for Educational Research Online 5 (2), 15–49.
Berryhill, J., Linney, J.A., Fromewick, J., 2009. The effects of educational accountability on teachers: are policies too stress provoking for their own good? International Journal of Education Policy and Leadership 4 (5). https://doi.org/10.22230/ijepl.2009v4n5a99.
Blackwell, T.L., 2001. Test review: woodcock, R. W., McGrew, K. S., & mather, N. (2001). Woodcock-Johnson® III test. Riverside publishing company. Itasca, IL. Rehabil. Counsel. Bull. 44 (4), 232–235. https://doi.org/10.1177/003435520104400407.
Blossfeld, H.-P., Roßbach, H.-G., 2019. Education as a Lifelong Process: the German National Educational Panel Study (NEPS), second ed. Springer VS, Wiesbaden, Germany. https://doi.org/10.1007/978-3-658-23162-0. *Edition ZfE*.
Bodenhausen, G.V., Macrae, C.N., Sherman, J.W., 1999. On the dialectics of discrimination: dual processes in social stereotyping. In: Chaiken, S., Trope, Y. (Eds.), Dual-process Theories in Social Psychology. Guilford Press, New York, NY, pp. 271–290.
Boliver, V., Capsada-Munsech, Q., 2021. Does ability grouping affect UK primary school pupils' enjoyment of maths and English? Res. Soc. Stratif. Mobil. 76, 100629 https://doi.org/10.1016/j.rssm.2021.100629.
Breda, T., Jouini, E., Napp, C., Thebault, G., 2020. Gender stereotypes can explain the gender-equality paradox. Proc. Natl. Acad. Sci. U.S.A. 117 (49), 31063–31069. https://doi.org/10.1073/pnas.2008704117.
Burgess, S., Greaves, E., 2009. Test Scores, Subjective Assessment and Stereotyping of Ethnic Minorities. Working Paper 09/221). Bristol, England.
Burgess, S., Greaves, E., 2013. Test scores, subjective assessment, and stereotyping of ethnic minorities. J. Labor Econ. 31 (3), 535–576. https://doi.org/10.1086/669340.
Bussey, K., Bandura, A., 1999. Social cognitive theory of gender development and differentiation. Psychol. Rev. 106 (4), 676–713. https://doi.org/10.1037/0033-295X.106.4.676.
Campbell, T., 2015. Stereotyped at seven? Biases in teacher judgement of pupils' ability and attainment. J. Soc. Pol. 44 (3), 517–547. https://doi.org/10.1017/S0047279415000227.
Carlana, M., 2019. Implicit stereotypes: evidence from teachers' gender bias. Q. J. Econ. 134 (3), 1163–1224. https://doi.org/10.1093/qje/qjz008.
Cavaglia, C., Machin, S., McNally, S., Ruiz-Valenzuela, J., 2020. Gender, achievement, and subject choice in English education. Oxf. Rev. Econ. Pol. 36 (4), 816–835. https://doi.org/10.1093/oxrep/graa050.
Cimpian, J.R., Lubienski, S.T., Timmer, J.D., Makowski, M.B., Miller, E.K., 2016. Have gender gaps in math closed? Achievement, teacher perceptions, and learning behaviors across two ECLS-K cohorts. AERA Open 2 (4), 233285841667361. https://doi.org/10.1177/2332858416673617.
Coenen, J., van Klaveren, C., 2016. Better test scores with a same-gender teacher? Eur. Socio Rev. 32 (3), 452–464. https://doi.org/10.1093/esr/jcw012.
Condron, D.J., 2008. An early start: skill grouping and unequal reading gains in the elementary years. Socio. Q. 49 (2), 363–394. https://doi.org/10.1111/j.1533-8525.2008.00119.x.
Dersch, A.-S., Heyder, A., Eitel, A., 2022. Exploring the nature of teachers' math-gender stereotypes: the Math-Gender Misconception Questionnaire. Front. Psychol. 13, 820254 https://doi.org/10.3389/fpsyg.2022.820254.
Doherty, J., Conolly, M., 1985. How accurately can primary school teachers predict the scores of their pupils in standardised tests of attainment? A study of some non-cognitive factors that influence specific judgements. Educ. Stud. 11 (1), 41–60. https://doi.org/10.1080/0305569850110105.
Duckworth, A.L., Seligman, M.E.P., 2006. Self-discipline gives girls the edge: gender in self-discipline, grades, and achievement test scores. J. Educ. Psychol. 98 (1), 198–208. https://doi.org/10.1037/0022-0663.98.1.198.
Eagly, A.H., 1987. Sex Differences in Social Behavior: A Social-Role Interpretation. John M. MacEachran Memorial Lecture Series, vol. 1985. Taylor & Francis Group, New York, NY.

Eckhardt, T., 2019. The Education System in the Federal Republic of Germany 2016/2017: A Description of the Responsibilities, Structures and Developments in Education Policy for the Exchange of Information in Europe. KMK, Bonn, Germany.
Ehrtmann, L., Wolter, I., 2018. The impact of students' gender-role orientation on competence development in mathematics and reading in secondary school. Learn. Indiv Differ 61, 256–264. https://doi.org/10.1016/j.lindif.2018.01.004.
Elliott, C.D., Smith, P., McCulloch, K., 1996. British Ability Scales II. NFER-NELSON Publishing, Windsor, Berkshire.
Espeland, W.N., Sauder, M., 2007. Rankings and reactivity: how public measures recreate social worlds. Am. J. Sociol. 113 (1), 1–40. https://doi.org/10.1086/517897.
Figlio, D., Loeb, S., 2011. School accountability. In: Handbook of the Economics of Education, vol. 3. Elsevier, Amsterdam, The Netherlands, pp. 383–421. https://doi.org/10.1016/B978-0-444-53429-3.00008-9.
Finnigan, K.S., Gross, B., 2007. Do accountability policy sanctions influence teacher motivation? Lessons from Chicago's low-performing schools. Am. Educ. Res. J. 44 (3), 594–630. https://doi.org/10.3102/0002831207306767.
Fiske, S.T., Lin, M., Neuberg, S.L., 1999. The continuum model: ten years later. In: Chaiken, S., Trope, Y. (Eds.), Dual-process Theories in Social Psychology. Guilford Press, New York, NY, pp. 231–254.
Fors Connolly, F., Goossen, M., Hjerm, M., 2020. Does gender equality cause gender differences in values? Reassessing the gender-equality-personality paradox. Sex. Roles 83, 101–113. https://doi.org/10.1007/s11199-019-01097-x.
Gamoran, A., 1986. Instructional and institutional effects of ability grouping. Sociol. Educ. 59 (4), 185–198. https://doi.org/10.2307/2112346.
Gawronski, B., Geschke, D., Banse, R., 2003. Implicit bias in impression formation: associations influence the construal of individuating information. Eur. J. Soc. Psychol. 33 (5), 573–589. https://doi.org/10.1002/ejsp.166.
Gehrer, K., Zimmermann, S., Artelt, C., Weinert, S., 2013. NEPS framework for assessing reading competence and results from an adult pilot study. Journal for Educational Research Online 5 (2), 50–79. https://doi.org/10.25656/01:8424.
Gentrup, S., Lorenz, G., Kristen, C., Kogan, I., 2020. Self-fulfilling prophecies in the classroom: teacher expectations, teacher feedback and student achievement. Learn. InStruct. 66, 101296 https://doi.org/10.1016/j.learninstruc.2019.101296.
Gentrup, S., Rjosk, C., 2018. Pygmalion and the gender gap: do teacher expectations contribute to differences in achievement between boys and girls at the beginning of schooling? Educ. Res. Eval. 24 (3–5), 295–323. https://doi.org/10.1080/13803611.2018.1550840.
Gentrup, S., Rjosk, C., Stanat, P., Lorenz, G., 2018. Einschätzungen der schulischen Motivation und des Arbeitsverhaltens durch Grundschullehrkräfte und deren Bedeutung für Verzerrungen in Leistungserwartungen. Z. für Erziehungswiss. (ZfE) 21 (4), 867–891. https://doi.org/10.1007/s11618-018-0806-2.
Geven, S., Batruch, A., van de Werfhorst, H., 2018. Inequality in Teacher Judgements, Expectations and Track Recommendations: A Review Study. University of Amsterdam, Amsterdam, The Netherlands. Retrieved from. https://zoek.officielebekendmakingen.nl/blg-864911.
Geven, S., Wiborg, Ø.N., Fish, R.E., van de Werfhorst, H.G., 2021. How teachers form educational expectations for students: a comparative factorial survey experiment in three institutional contexts. Soc. Sci. Res., 102599 https://doi.org/10.1016/j.ssresearch.2021.102599.
Hall, K., Ozerk, K., 2010. Primary curriculum and assessment: England and other countries. In: Alexander, R.J. (Ed.), The Cambridge Primary Review Research Surveys. Routledge, London, England, pp. 375–414.
Hallam, S., Parsons, S., 2013. The incidence and make up of ability grouped sets in the UK primary school. Res. Pap. Educ. 28 (4), 393–420. https://doi.org/10.1080/02671522.2012.729079.
Halpern, D.F., 2011. Sex Differences in Cognitive Abilities, fourth ed. Taylor & Francis Group, London, England.
Hannover, B., Wolter, I., 2019. Geschlechtsstereotype: wie sie entstehen und sich auswirken. In: Kortendiek, B., Riegraf, B., Sabisch, K. (Eds.), Handbuch Interdisziplinäre Geschlechterforschung. Springer VS, Wiesbaden, Germany, pp. 201–210.
Hanushek, E.A., Rivkin, S.G., 2010. The quality and distribution of teachers under the No Child Left behind Act. J. Econ. Perspect. 24 (3), 133–150. https://doi.org/10.1257/jep.24.3.133.
Harris, M.J., Rosenthal, R., 1985. Mediation of interpersonal expectancy effects: 31 meta-analyses. Psychol. Bull. 97 (3), 363–386. https://doi.org/10.1037/0033-2909.97.3.363.
Hartong, S., 2014. Neue Bildungsregulierung im Zeitalter der »governance by numbers«. Das Beispiel standardisierter Bildungsreformen in Deutschland und den USA. Leviathan 42 (4), 606–634. https://doi.org/10.5771/0340-0425-2014-4-606.
Hilton, J.L., von Hippel, W., 1996. Stereotypes. Annu. Rev. Psychol. 47, 237–271. https://doi.org/10.1146/annurev.psych.47.1.237.
Hinnant, J.B., O'Brien, M., Ghazarian, S.R., 2009. The longitudinal relations of teacher expectations to achievement in the early school years. J. Educ. Psychol. 101 (3), 662–670. https://doi.org/10.1037/a0014306.
Hoge, R.D., Butcher, R., 1984. Analysis of teacher judgments of pupil achievement levels. J. Educ. Psychol. 76 (5), 777–781. https://doi.org/10.1037//0022-0663.76.5.777.
Hoge, R.D., Coladarci, T., 1989. Teacher-based judgments of academic achievement: a review of literature. Rev. Educ. Res. 59 (3), 297–313. https://doi.org/10.3102/00346543059003297.
Hübner, N., Spengler, M., Nagengast, B., Borghans, L., Schils, T., Trautwein, U., 2022. When academic achievement (also) reflects personality: using the personality-achievement saturation hypothesis (PASH) to explain differential associations between achievement measures and personality traits. J. Educ. Psychol. 114 (2), 326–345. https://doi.org/10.1037/edu0000571.
Hyde, J.S., 2014. Gender similarities and differences. Annu. Rev. Psychol. 65, 373–398. https://doi.org/10.1146/annurev-psych-010213-115057.
Jerrim, J., Sims, S., 2022. School accountability and teacher stress: international evidence from the OECD TALIS study. Educ. Assess. Eval. Account. 34 (1), 5–32. https://doi.org/10.1007/s11092-021-09360-0.
Johnson, J., Rosenberg, R., Platt, L., Parsons, S., 2011. Millennium Cohort Study Fourth Survey. A Guide to the Teacher Survey Dataset. Centre for Longitudinal Studies, London, England.
Jones, E.M., Schoon, I., 2008. Child cognition and behaviour. In: Hansen, K., Joshi, H. (Eds.), Millennium Cohort Study: Third Survey: A User's Guide to Initial Findings. Institute for Longitudinal Studies, London, England, pp. 118–126.
Joshi, H., Fitzsimons, E., 2016. The Millennium Cohort Study: the making of a multi-purpose resource for social science and policy. Longitudinal and Life Course Studies 7 (4), 409–430. https://doi.org/10.14301/llcs.v7i4.410.
Jussim, L., Eccles, J.S., 1992. Teacher expectations: II. Construction and reflection of student achievement. J. Pers. Soc. Psychol. 63 (6), 947–961. https://doi.org/10.1037/0022-3514.63.6.947.
Jussim, L.J., Eccles, J., 1995. Are teacher expectations biased by students' gender, social class, or ethnicity? In: Lee, Y.-T., Jussim, L.J., McCauley, C.R. (Eds.), Stereotype Accuracy: toward Appreciating Group Differences. American Psychological Association, Washington, DC, pp. 245–271. https://doi.org/10.1037/10495-010.
Jussim, L., Eccles, J., Madon, S., 1996. Social perception, social stereotypes, and teacher expectations: accuracy and the quest for the powerful self-fulfilling prophecy. In: Zanna, M.P. (Ed.), Advances in Experimental Social Psychology, vol. 28. Academic Press, San Diego, CA, pp. 281–388. https://doi.org/10.1016/S0065-2601(08)60240-3.
Jussim, L., Harber, K.D., 2005. Teacher expectations and self-fulfilling prophecies: knowns and unknowns, resolved and unresolved controversies. Pers. Soc. Psychol. Rev. 9 (2), 131–155. https://doi.org/10.1207/s15327957pspr0902_3.
Karing, C., Matthäi, J., Artelt, C., 2011. Genauigkeit von Lehrerurteilen über die Lesekompetenz ihrer Schülerinnen und Schüler in der Sekundarstufe I – eine Frage der Spezifität? Z. für Padagogische Psychol. 25 (3), 159–172. https://doi.org/10.1024/1010-0652/a000041.
Kelly, S., Carbonaro, W., 2012. Curriculum tracking and teacher expectations: evidence from discrepant course taking models. Soc. Psychol. Educ. 15 (3), 271–294. https://doi.org/10.1007/s11218-012-9182-6.
Klenowski, V., Wyatt-Smith, C., 2010. Standards, teacher judgement and moderation in contexts of national curriculum and assessment reform. Assessment Matters 2, 107–131. https://doi.org/10.18296/am.0078.

KMK, 2015. Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring. Retrieved from. https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2015/2015_06_11-Gesamtstrategie-Bildungsmonitoring.pdf.
Krolak-Schwerdt, S., Böhmer, M., Gräsel, C., 2013. The impact of accountability on teachers' assessments of student performance: a social cognitive analysis. Soc. Psychol. Educ. 16 (2), 215–239. https://doi.org/10.1007/s11218-013-9215-9.
Krolak-Schwerdt, S., Pit-ten Cate, I.M., Hörstermann, T., 2018. Teachers' judgments and decision-making: studies concerning the transition from primary to secondary education and their implications for teacher education. In: Zlatkin-Troitschanskaia, O., Toepper, M., Pant, H.A., Lautenbach, C., Kuhn, C. (Eds.), Methodology of Educational Measurement and Assessment: Assessment of Learning Outcomes in Higher Education, vol. 27. Springer, Cham, Switzerland, pp. 73–101. https://doi.org/10.1007/978-3-319-74338-7_5.
Lang, F.R., Kamin, S., Rohr, M., Stünkel, C., Williger, B., 2014. Erfassung der fluiden kognitiven Leistungsfähigkeit über die Lebensspanne im Rahmen des Nationalen Bildungspanels: Abschlussbericht zu einer NEPS-Ergänzungsstudie. (NEPS Working Paper No. 43), Bamberg, Germany.
Law, J., Wareham, H., Volodina, A., Rush, R., 2021. The Pros and Cons of Combining Birth Cohort Data. Publications Archive – DIAL (dynamicsofinequality.org). Retrieved from. https://dynamicsofinequality.org/publication/the-pros-and-cons-of-combining-birth-cohort-data/.
Lee, J., Liu, X., Amo, L.C., Wang, W.L., 2014. Multilevel linkages between state standards, teacher standards, and student achievement. Educ. Pol. 28 (6), 780–811. https://doi.org/10.1177/0895904813475708.
Lee, M.W., Newton, P., 2021. Systematic Divergence between Teacher and Test-Based Assessment: Literature Review. Ofqual, Coventry, England.
Lenhard, W., Schneider, W., 2006. ELFE 1-6: Ein Leseverständnistest für Erst- bis Sechstklässler. Hogrefe, Göttingen, Germany.
Lerner, J.S., Tetlock, P.E., 1999. Accounting for the effects of accountability. Psychol. Bull. 125 (2), 255–275. https://doi.org/10.1037/0033-2909.125.2.255.
Lleras, C., Rangel, C., 2009. Ability grouping practices in elementary school and African American/Hispanic achievement. Am. J. Educ. 115 (2), 279–304. https://doi.org/10.1086/595667.
Loibl, K., Leuders, T., Dörfler, T., 2020. A framework for explaining teachers' diagnostic judgements by cognitive modelling (DiaCoM). Teach. Teach. Educ. 91, 103059 https://doi.org/10.1016/j.tate.2020.103059.
Lorenz, C., Berendes, K., Weinert, S., 2017. Measuring Receptive Grammar in Kindergarten and Elementary School Children in the German National Educational Panel Study. NEPS Working Papers/Survey Papers No. 24, Bamberg, Germany.
Lorenz, G., 2018. Selbsterfüllende Prophezeiungen in der Schule: Leistungserwartungen von Lehrkräften und Kompetenzen von Kindern mit Zuwanderungshintergrund. Springer VS, Wiesbaden, Germany. https://doi.org/10.1007/978-3-658-19881-7.
Lorenz, G., Gentrup, S., Kristen, C., Stanat, P., Kogan, I., 2016. Stereotype bei Lehrkräften? Eine Untersuchung systematisch verzerrter Lehrererwartungen. Kölner Z. Soziol. Sozialpsychol. 68 (1), 89–111. https://doi.org/10.1007/s11577-015-0352-3.
Loveless, T., 2013. The 2013 Brown Center Report on American Education: How Well Are American Students Learning? Brown Center on Education Policy at Brookings, Washington, DC.
Machin, S., McNally, S., 2005. Gender and student achievement in English schools. Oxf. Rev. Econ. Pol. 21 (3), 357–372. https://doi.org/10.1093/oxrep/gri021.
Macqueen, S.E., 2013. Grouping for inequity. Int. J. Incl. Educ. 17 (3), 295–309. https://doi.org/10.1080/13603116.2012.676088.
Madon, S., Jussim, L., Eccles, J., 1997. In search of the powerful self-fulfilling prophecy. J. Pers. Soc. Psychol. 72 (4), 791–809. https://doi.org/10.1037/0022-3514.72.4.791.
Madon, S., Jussim, L., Keiper, S., Eccles, J., Smith, A., Palumbo, P., 1998. The accuracy and power of sex, social class, and ethnic stereotypes: a naturalistic study in person perception. Pers. Soc. Psychol. Bull. 24 (12), 1304–1318. https://doi.org/10.1177/01461672982412005.
Mast, M.S., Krings, F., 2008. Stereotype und Informationsverarbeitung. In: Petersen, L.-E., Six, B. (Eds.), Stereotype, Vorurteile und soziale Diskriminierung: Theorien, Befunde und Interventionen, first ed. Beltz, Weinheim, Germany, pp. 33–44.
McElvany, N., Kessels, U., Schwabe, F., Kasper, D., 2017. Geschlecht und Lesekompetenz. In: Hußmann, A., Wendt, H., Bos, W., Bremerich-Vos, A., Kasper, D., Lankes, E.-M., McElvany, N., Stubbe, T.C., Valtin, R. (Eds.), IGLU 2016. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich. Waxmann, Münster, Germany, pp. 177–194.
McGuinn, P., Manna, P., 2013. Education governance in America: who leads when everyone is in charge? In: Manna, P., McGuinn, P. (Eds.), Education Governance for the Twenty-First Century: Overcoming the Structural Barriers to School Reform. Brookings Institution Press, Washington, DC, pp. 1–18.
McKown, C., Weinstein, R.S., 2002. Modeling the role of child ethnicity and gender in children's differential response to teacher expectations. J. Appl. Soc. Psychol. 32 (1), 159–184. https://doi.org/10.1111/j.1559-1816.2002.tb01425.x.
McKown, C., Weinstein, R.S., 2008. Teacher expectations, classroom context, and the achievement gap. J. Sch. Psychol. 46 (3), 235–261. https://doi.org/10.1016/j.jsp.2007.05.001.
Merton, R.K., 1948. The self-fulfilling prophecy. Antioch Rev. 8 (2), 193–210.
Mize, T.D., Doan, L., Long, J.S., 2019. A general framework for comparing predictions and marginal effects across models. Socio. Methodol. 49 (1), 152–189. https://doi.org/10.1177/0081175019852763.
Muench, R., Wieczorek, O., 2022. In search of quality and equity: the United Kingdom and Germany in the struggle for PISA scores. International Journal of Educational Research Open 3, 100165. https://doi.org/10.1016/j.ijedro.2022.100165.
Mullis, I.V.S., Martin, M.O., Goh, S., Cotter, K. (Eds.), 2016. TIMSS 2015 Encyclopedia: Education Policy and Curriculum in Mathematics and Science. Retrieved from Boston College. TIMSS & PIRLS International Study Center website. http://timssandpirls.bc.edu/timss2015/encyclopedia/.
Mullis, I.V., Martin, M.O., Foy, P., Hooper, M., 2017a. PIRLS 2016: International Results in Reading. TIMSS & PIRLS International Study Center, Lynch School of Education.
Mullis, I.V.S., Martin, M.O., Goh, S., Prendergast, C. (Eds.), 2017b. PIRLS 2016 Encyclopedia: Education Policy and Curriculum in Reading. Retrieved from Boston College. TIMSS & PIRLS International Study Center website. http://timssandpirls.bc.edu/pirls2016/encyclopedia/.
Muntoni, F., Retelsdorf, J., 2018. Gender-specific teacher expectations in reading—the role of teachers' gender stereotypes. Contemp. Educ. Psychol. 54, 212–220. https://doi.org/10.1016/j.cedpsych.2018.06.012.
Najarian, M., Tourangeau, K., Nord, C., Wallner-Allen, K., 2018. Early Childhood Longitudinal Study, Kindergarten Class of 2010–11: First- and Second-Grade Psychometric Report (NCES 2018-183). National Center for Education Statistics, Washington, DC.
Najarian, M., Tourangeau, K., Nord, C., Wallner-Allen, K., Vaden-Kiernan, N., 2020. *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11: Third-grade, fourth-grade, and fifth-grade psychometric report (NCES 2020-123)*. National Center for Education Statistics, Washington, DC.
Napp, C., Breda, T., 2022. The stereotype that girls lack talent: a worldwide investigation. Sci. Adv. 8 (10), eabm3689 https://doi.org/10.1126/sciadv.abm3689.
NEPS Network, 2020. NEPS-startkohorte 2. https://doi.org/10.5157/NEPS:SC2:9.0.0. *Kindergarten (SC2 9.0.0)*.
Neugebauer, M., Helbig, M., Landmann, A., 2011. Unmasking the myth of the same-sex teacher advantage. Eur. Socio Rev. 27 (5), 669–689. https://doi.org/10.1093/esr/jcq038.
OECD, 2016. PISA 2015 Results, vol. I. OECD, Paris, France. https://doi.org/10.1787/9789264266490-en.
OECD, 2019. PISA 2018 Results (Volume II): where All Students Can Succeed. OECD, Paris, France. https://doi.org/10.1787/888934038723.
OECD, 2023. Gender Wage Gap (Indicator). https://doi.org/10.1787/7cee77aa-en. (Accessed 23 May 2023).
OECD.Stat, 2022. Distribution of Teachers by Age and Gender. Retrieved from. https://stats.oecd.org/Index.aspx?DataSetCode=EAG_PERS_SHARE_AGE#. (Accessed 23 May 2023).
Olczyk, M., Kwon, S.J., Lorenz, G., Perinetti Casoni, V., Schneider, T., Volodina, A., Waldfogel, J., Washbrook, E., 2022. Teacher judgements, student social background, and student progress in primary school: a cross-country perspective. Zeitschrift für Erziehungswissenschaft. Advance online publication. https://doi.org/10.1007/s11618-022-01119-7.
Pallas, A.M., Entwisle, D.R., Alexander, K.L., Stluka, M.F., 1994. Ability-group effects: instructional, social, or institutional? Sociol. Educ. 67 (1), 27–46. https://doi.org/10.2307/2112748.
Perryman, J., Calvert, G., 2020. What motivates people to teach, and why do they leave? Accountability, performativity and teacher retension. Br. J. Educ. Stud. 68 (1), 3–23. https://doi.org/10.1080/00071005.2019.1589417.

Petersen, J., 2018. Gender difference in verbal performance: a meta-analysis of United States state performance assessments. Educ. Psychol. Rev. 30 (4), 1269–1281. https://doi.org/10.1007/s10648-018-9450-x.
Pit-ten Cate, I.M., Hörstermann, T., Krolak-Schwerdt, S., Gräsel, C., Böhmer, I., Glock, S., 2020. Teachers' information processing and judgement accuracy: effects of information consistency and accountability. Eur. J. Psychol. Educ. 35 (3), 675–702. https://doi.org/10.1007/s10212-019-00436-6.
Plewis, I., 1997. Inferences about teacher expectations from national assessment at key stage one. Br. J. Educ. Psychol. 67 (2), 235–247. https://doi.org/10.1111/j.2044-8279.1997.tb01240.x.
Ready, D.D., Chu, E.M., 2015. Sociodemographic inequality in early literacy development: the role of teacher perceptual accuracy. Early Educ. Dev. 26 (7), 970–987. https://doi.org/10.1080/10409289.2015.1004516.
Ready, D.D., Wright, D.L., 2011. Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities: the role of child background and classroom context. Am. Educ. Res. J. 48 (2), 335–360. https://doi.org/10.3102/0002831210374874.
Reilly, D., Neumann, D.L., Andrews, G., 2019. Gender differences in reading and writing achievement: evidence from the national assessment of educational progress (NAEP). Am. Psychol. 74 (4), 445–458. https://doi.org/10.1037/amp0000356.
Riegle-Crumb, C., Humphries, M., 2012. Exploring bias in math teachers' perceptions of students' ability by gender and race/ethnicity. Gend. Soc. 26 (2), 290–322. https://doi.org/10.1177/0891243211434614.
Rihal, S., Gomes, D., 2021. Millennium Cohort Study: A Guide to the Linked Education Administrative Datasets, second ed. UCL Centre for Longitudinal Studies, London.
Robinson, J.P., Lubienski, S.T., 2011. The development of gender achievement gaps in mathematics and reading during elementary and middle school. Am. Educ. Res. J. 48 (2), 268–302. https://doi.org/10.3102/0002831210372249.
Robinson-Cimpian, J.P., Lubienski, S.T., Ganley, C.M., Copur-Gencturk, Y., 2014. Teachers' perceptions of students' mathematics proficiency may exacerbate early gender gaps in achievement. Dev. Psychol. 50 (4), 1262–1281. https://doi.org/10.1037/a0035073.
Rosenthal, R., 1973. The mediation of Pygmalion effects: a four factor "theory". Papua New Guinea Journal of Education 9 (1), 1–12.
Rubie-Davies, C.M., 2007. Classroom interactions: exploring the practices of high- and low-expectation teachers. Br. J. Educ. Psychol. 77 (Pt 2), 289–306. https://doi.org/10.1348/000709906X101601.
Schmitt, D.P., Long, A.E., McPhearson, A., O'Brien, K., Remmert, B., Shah, S.H., 2017. Personality and gender differences in global perspective. Int. J. Psychol. 52 (Suppl. 1), 45–56. https://doi.org/10.1002/ijop.12265.
Schnittjer, I., Gerken, A.-L., Petersen, L.A., 2020. NEPS Technical Report for Mathematics – Scaling Results of Starting Cohort 2 in Fourth Grade. NEPS Survey Paper No. 69). Bamberg, Germany.
Schrader, F.-W., Helmke, A., 1990. Lassen sich Lehrer bei der Leistungsbeurteilung von sachfremden Gesichtspunkten leiten? Eine Untersuchung zu Determinanten diagnostischer Lehrerurteile. Z. für Entwicklungspsychol. Pädagogische Psychol. 22 (4), 312–324.
Stahl, N., 2007. Schülerwahrnehmung und -beurteilung durch Lehrkräfte. In: Ditton, H. (Ed.), Kompetenzaufbau und Laufbahnen im Schulsystem. Waxmann, Münster, Germany, pp. 171–198.
Stoet, G., Bailey, D.H., Moore, A.M., Geary, D.C., 2016. Countries with higher levels of gender equality show larger national sex differences in mathematics anxiety and relatively lower parental mathematics valuation for girls. PLoS One 11 (4), e0153857. https://doi.org/10.1371/journal.pone.0153857.
Stoet, G., Geary, D.C., 2018. The gender-equality paradox in science, technology, engineering, and mathematics education. Psychol. Sci. 29 (4), 581–593. https://doi.org/10.1177/0956797617741719.
Stoet, G., Geary, D.C., 2022. Sex differences in adolescents' occupational aspirations: variations across time and place. PLoS One 17 (1). https://doi.org/10.1371/journal.pone.0261438.
Südkamp, A., Kaiser, J., Möller, J., 2012. Accuracy of teachers' judgments of students' academic achievement: a meta-analysis. J. Educ. Psychol. 104 (3), 743–762. https://doi.org/10.1037/a0027627.
Tenenbaum, H.R., Ruck, M.D., 2007. Are teachers' expectations different for racial minority than for European American students? A meta-analysis. J. Educ. Psychol. 99 (2), 253–273. https://doi.org/10.1037/0022-0663.99.2.253.
Tiedemann, J., 2000. Gender-related beliefs of teachers in elementary school mathematics. Educ. Stud. Math. 41 (2), 191–207.
Tiedemann, J., 2002. Teachers' gender stereotypes as determinants of teacher perceptions in elementary school mathematics. Educ. Stud. Math. 50 (1), 49–62. https://doi.org/10.1023/A:1020518104346.
Tourangeau, K., Nord, C., Lê, T., Wallner-Allen, K., Hagedorn, M.C., Leggitt, J., 2015. *User's Manual for the ECLS-K:2011 Kindergarten–First Grade: Data File and Electronic Codebook, Public Version* (2015-078). National Center for Education Statistics, Washington, DC.
Tourangeau, K., Nord, C., Lê, T., Wallner-Allen, K., Vaden-Kiernan, N., Blaker, L., Najarian, M., 2019. Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011) User's Manual for the ECLS-K:2011 Kindergarten–Fifth Grade Data File and Electronic Codebook, Public Version (NCES 2019-051). National Center for Education Statistics, Washington, DC.
Trautwein, U., Baeriswyl, F., 2007. Wenn leistungsstarke Klassenkameraden ein Nachteil sind. Z. für Padagogische Psychol. 21 (2), 119–133. https://doi.org/10.1024/1010-0652.21.2.119.
University College London, UCL Institute of Education, Centre for Longitudinal Studies, & Department for Education, 2021. Millennium Cohort Study: Linked Education Administrative Datasets (National Pupil Database). Secure Access, England. https://doi.org/10.5255/UKDA-SN-8481-2.
University of London, Institute of Education, Centre for Longitudinal Studies, 2021. Millennium Cohort Study: Fourth Survey, p. 2008. https://doi.org/10.5255/UKDA-SN-6411-8.
Urhahne, D., Wijnia, L., 2021. A review on the accuracy of teacher judgments. Educ. Res. Rev. 32, 100374 https://doi.org/10.1016/j.edurev.2020.100374.
Wang, S., Rubie-Davies, C.M., Meissel, K., 2018. A systematic review of the teacher expectation literature over the past 30 years. Educ. Res. Eval. 24 (3–5), 124–179. https://doi.org/10.1080/13803611.2018.1548798.
Watson, P.W. St J., Rubie-Davies, C.M., Meissel, K., Peterson, E.R., Flint, A., Garrett, L., McDonald, L., 2019. Teacher gender, and expectation of reading achievement in New Zealand elementary school students: essentially a barrier? Gend. Educ. 31 (8), 1000–1019. https://doi.org/10.1080/09540253.2017.1410108.
Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., Carstensen, C.H., 2011. 5 Development of competencies across the life span. Z. für Erziehungswiss. (ZfE) 14 (S2), 67–86. https://doi.org/10.1007/s11618-011-0182-7.
Weinstein, R.S., 2002. *Reaching Higher: The Power of Expectations in Schooling* (1, Paperback ed. Harvard University Press, Cambridge, MA.
Wendt, H., Steinmayr, R., Kasper, D., 2016. Geschlechterunterschiede in mathematischen und naturwissenschaftlichen Kompetenzen. In: Wendt, H., Bos, W., Selter, C., Köller, O., Schwippert, K., Kasper, D. (Eds.), TIMSS 2015. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich. Waxmann, Münster, Germany, pp. 257–298.
Winkelmann, H., Heuvel-Panhuizen, M., Robitzsch, A., 2008. Gender differences in the mathematics achievements of German primary school students: results from a German large-scale study. ZDM 40 (4), 601–616. https://doi.org/10.1007/s11858-008-0124-x.
WorldBank, 2023. Share of Graduates by Field, Female (%). https://genderdata.worldbank.org/indicators/se-ter-grad-fe-zs/?fieldOfStudy=Science%2C%20Technology%2C%20Engineering%20and%20Mathematics%20%28STEM%29&geos=DEU_USA_GBR&view=bar. (Accessed 23 May 2023).
World Economic Forum, 2019. The Global Gender Gap Report 2020. World Economic Forum, Geneva, Switzerland.
Yanushevsky, R., 2011. Improving Education in the US: A Political Paradox. Algora Publishing, New York, NY.
Zinn, S., Würbach, A., Steinhauer, H.W., Hammon, A., 2020. Attrition and selectivity of the NEPS Starting Cohorts: an overview of the past 8 years. AStA Wirtschafts- und Sozialstatistisches Archiv 14 (2), 163–206.
Zuckerman, M., Li, C., Hall, J.A., 2016. When men and women differ in self-esteem and when they don't: a meta-analysis. J. Res. Pers. 64, 34–51. https://doi.org/10.1016/j.jrp.2016.07.007.