Z Erziehungswiss (2023) 26:443–468
https://doi.org/10.1007/s11618-022-01119-7

ZfE

SCHWERPUNKT

Teacher judgements, student social background, and student progress in primary school: a cross-country perspective

Melanie Olczyk · Sarah Jiyoon Kwon · Georg Lorenz · Valentina Perinetti Casoni · Thorsten Schneider · Anna Volodina · Jane Waldfogel · Elizabeth Washbrook

Received: 2 December 2021 / Revised: 20 June 2022 / Accepted: 5 July 2022 / Published online: 16 August 2022

Abstract This study takes a cross-country perspective to examine whether inaccurate teacher judgements of students' math skills correlate with student social origin and whether such bias is associated with math achievement in primary school. We focus on England, Germany, and the US because these countries differ in the teachers' growth mindsets, accountability, the use of standardised tests, and the extent of ability grouping. The data stem from three large-scale surveys, the Millennium Cohort Study for England, the National Educational Panel Study for Germany, and the Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 for the US. At the beginning of primary education, teacher judgements were not entirely consistent with student scores in standardised tests. In England and Germany, teachers underrated students with low-educated parents and overrated those with high-educated parents. In the US, no such differences were found. In all three countries, overrated (or underrated) students performed better (worse) later on. In England and, to a lesser extent, in Germany, we found evidence that biased teacher judgements contribute—over the course of primary school—to widening inequalities in value-added achievement by parental education. Such effects were negligible in the US. Our findings suggest that a cross-country perspective is essential to better understand contextual factors' role in systematic bias in teacher judgements and its relevance for educational achievement. This study can be seen as a starting point for future research to investigate the mechanisms of such contextual effects more thoroughly.

Data availability (data transparency) The data used are available for research at the NEPS Research Data Centre in Bamberg, Germany, the UK Data Archive in Essex, United Kingdom, and at the National Center for Education Statisitics in Washington, United States.

Code availability (software application or custom code) The code for the analysis of the NEPS data (Germany) is available at the Open Science Framework (OSF): https://osf.io/7ebfp/?view_only=b1ac5ecdb43a4320bf91967e27dd5d87.

Extended author information available on the last page of the article.

Keywords Educational inequalities · Socioeconomic status · Self-fulfilling prophecies · Teacher judgement · Primary school

Lehrerurteile, soziale Herkunft der Schülerinnen und Schülern und Lernfortschritte in der Grundschule: Eine länderübergreifende Perspektive

Zusammenfassung In dieser Studie wird aus einer länderübergreifenden Perspektive untersucht, ob Lehrkrafteinschätzungen der Mathematikkenntnisse von Schülerinnen und Schülern über tatsächliche Mathematikfähigkeiten hinaus mit der sozialen Herkunft der Kinder korrelieren und ob diese systematischen Verzerrungen mit der Entwicklung der Mathematikkenntnisse in der Grundschule zusammenhängen. Wir konzentrieren uns auf England, Deutschland und die USA, da sich diese Länder in Bezug auf Überzeugungen zur Beeinflussbarkeit der Leistungsentwicklung (Growth Mindset), Accountability, den Einsatz von standardisierten Testverfahren und die Anwendung von Leistungsgruppierung in der Grundschule unterscheiden. Die Datengrundlage bilden drei Erhebungen: die Millennium Cohort Study für England, das Nationale Bildungspanel für Deutschland und die Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 für die USA. Zu Beginn der Grundschulzeit stimmten die Leistungsurteile nicht vollständig mit den Ergebnissen der Kinder in standardisierten Tests überein. In England und Deutschland zeigte sich, dass Lehrkrafturteile zu Schülerinnen und Schülern aus Familien mit formal niedrigem Bildungsniveau im Durchschnitt negativ, zu denen aus Familien mit formal hohem Bildungsniveau hingegen positiv verzerrt waren. Für die USA ließ sich ein solches Muster nicht beobachten. In allen drei Ländern zeigten positiv (negativ) bewertete Kinder später bessere (schlechtere) Mathematikleistungen. Die Effekte der elterlichen Bildung auf die mathematische Leistungsentwicklung verringerten sich unter Kontrolle der verzerrten Lehrkrafturteile in England und in geringerem Maße auch in Deutschland, aber kaum in den USA. Unsere Ergebnisse verdeutlichen, dass eine länderübergreifende Perspektive wichtig ist, da die Verzerrungen und ihre Bedeutung für die Leistungsentwicklung in den drei Ländern unterschiedlich sind. Das könnte als Ausgangspunkt für künftige Forschungsarbeiten gesehen werden, die zugrundeliegenden Prozesse genauer in den Blick zu nehmen.

Schlüsselwörter Bildungsungleichheiten · Soziale Herkunft · Selbsterfüllende Prophezeiungen · Lehrkrafturteile · Grundschule

1 Introduction

Various dimensions of educational success, such as student achievement, vary by parental socioeconomic status (SES). Discrimination by teachers may account for at least part of the observed socioeconomic inequalities. For instance, research indicates that teacher stereotypes related to family SES can lead to bias in teacher judgement (Jussim et al. 1996; Jussim and Harber 2005; Tenenbaum and Ruck 2007). In addition to affecting grades (e.g., Kiss 2013; Sprietsma 2013), differen-

tial judgements and expectations can result in different verbal (e.g., less warm and supportive, low-quality feedback; Gentrup et al. 2020; Rubie-Davies 2007), non-verbal teacher behaviours (e.g., reduced eye contact; Babad 1990, 1993), and in a self-fulfilling prophecy (Wang et al. 2018). Such processes could exacerbate SES-related achievement gaps and social inequalities in education. While there are many studies examining the extent of such bias, its consequences, and the underlying processes (e.g., Urhahne and Wijnia 2021, for a recent review), only a few studies approach this topic from a cross-country perspective and, hence, consider the wider institutional setting (see, e.g., Geven et al. 2021; Hofer 2015, for exceptions).

This study examines teacher judgements and their association with student achievement in three national contexts: England, Germany, and the US. In all three countries, teachers might shape educational careers by teaching students and grading and sorting them. In some contexts, teachers might have a more significant impact on students' educational careers than in others. In Germany, for example, teachers recommend the type of secondary school a child should attend after primary school. In England and the US, there are comprehensive schooling systems, while ability grouping within schools and/or classes is a common practice from lower grades on. The age at which tracking happens and the type of tracking and ability grouping might impact teacher judgements and their association with later achievement, as might the amount and level of (standardised) testing or further accountability practices used (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). Finally, Geven et al. (2021) point out that cultural beliefs about how effort can overcome original disadvantage (growth mindset) may also shape teacher judgements.

In this study, we first investigate whether SES-related bias in teacher judgements of student skills in early primary school differs between England, Germany, and the US. To this end, we regress the teacher judgements on student test scores. We focus on mathematics, one of the main school subjects in all three countries and one related to skills that strongly affect economic outcomes in later life (e.g., Ritchie and Bates 2013). Second, we examine the effects of these (potentially biased) teacher judgements on student achievement at the end of primary education (age 10–11) using value-added-models (e.g., Gentrup et al. 2020; Hinnant et al. 2009; Madon et al. 1997). We rely on harmonised data from three large-scale 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

Teachers' understanding of the students' achievement, needs, and resources is an essential precondition for efficient teaching and progress in student learning (e.g., Baumert and Kunter 2013; Hattie 2009; Helmke and Schrader 1987; Karing et al. 2011). Research shows that the accuracy of teacher judgements—defined

as teachers' ability to, amongst other things, adequately assess students' characteristics—vary between teachers. Overall, in their meta-analysis, Südkamp et al. (2012) have shown that the shared variance between teacher judgements and students' achievement—including results from standardised achievement tests as well as curriculum-based measures—is around 40%. Moreover, the shared variance is higher when measures of student motivation and cognitive abilities are considered alongside their achievement (e.g., de Boer et al. 2010). The remaining variance might be interpreted as inaccuracy (e.g., Gentrup et al. 2020). Empirically, it has been shown that students from less socioeconomically advantaged families often face lower teacher expectations vis-à-vis their objective achievement measures (for Germany see, e.g., Lorenz et al. 2016; Tobisch and Dresel 2017; for the UK see, e.g., Campbell 2015; Lee and Newton 2021; Plewis 1997; for the US see, e.g., Alvidrez and Weinstein 1999). These statistical effects might then be interpreted as bias. For the US, however, evidence is rather mixed as there are also studies that did not find systematic differences in teacher judgements coinciding with family SES (e.g., Hinnant et al. 2009).

Teacher judgements can be conceptualised as the result of cognitive information processing and diagnostic thinking (see Loibl et al. 2020, for a conceptional framework). Dual-processing theories postulate the presence of two different strategies of information processing: (1) automatic judgements guided by stereotypes that do not include the integration of relevant target information and (2) information-based judgements which involve the deliberate integration of target information into multifaceted judgements (e.g., Fiske et al. 1999). Both strategies could be understood as the end points of a continuum (Fiske et al. 1999).

Which of the two strategies is likely to be implemented depends, among other things, on the teacher's personal characteristics (e.g., attitudes, knowledge, or mindsets) and situational characteristics which include, for example, time pressure, judgement goals, and social cues (Loibl et al. 2020). Social cues are pieces of information that are used to form judgements about objects and other persons. They differ in how easily they are observed and extracted and whether their identification requires inferential information processing (Loibl et al. 2020). Social cues such as names, a dialect or an accent increase the likelihood of categorisation and generalisations (Fiske et al. 1999; Smith and DeCoster 2000). Moreover, teacher judgements might be more likely to be based on categorisation and generalisations when information is ambiguous and when there are only few cognitive resources available to the perceiver (Gawronski et al. 2003).

Above and beyond these personal and situational factors, both information processing and stereotypes are embedded in and affected by the broader institutional context (see, e.g., Geven et al. 2021; Weinstein 2002, for a similar argumentation). This context includes not only conditions and regulations of schools, the education system, and teacher training, but also norms and values, as well as cultural-cognitive beliefs that frame and guide social action (see Geven et al. 2021, for a similar argumentation). For example, conscious and unconscious biases in teacher judgements develop within training programs; within the cultural norms and values prevailing in a country or region as well as the stereotypes and prejudices shared in these contexts (e.g., Geven et al. 2021); and within the setting of schooling and teaching

(e.g., Romijn et al. 2021). Next, we describe aspects of the national context that might affect the extent of bias in teacher judgements.

First, a certain mindset or mindfulness of distortions might shape teachers' information processing. Geven et al. (2021) argue that in "growth mindset" cultures, people assume that initial disadvantages due to family SES can be overcome through effort. The counterpart to a growth mindset is a fixed mindset, i.e., the view that talent and skills are innate. Consequently, Geven and colleagues expect less SES-related bias in teacher expectations in growth mindset cultures than in fixed mindset ones (Geven et al. 2021, p. 6, referring to Dweck and Yeager 2019; Jacovidis et al. 2020).

Second, accountability may frame and affect teachers' diagnostic thinking and information processing (e.g., Krolak-Schwerdt et al. 2018). School accountability policies include, for instance, setting specific goals and sanctioning or rewarding schools based on their performance and governmental goals (Finnigan and Gross 2007). Previous research has shown that accountability policies are often accompanied by higher levels of teacher motivation and effort (e.g., Finnigan and Gross 2007; Kelley et al. 2002; Mintrop 2003) as well as more accurate judgements (e.g., Krolak-Schwerdt et al. 2013; Pit-ten Cate et al. 2020). 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. However, there is also empirical evidence showing that accountability policies can go hand in hand with higher levels of teacher stress and decreased self-efficacy (e.g., Berryhill et al. 2009; Jerrim and Sims 2022), which might offset the expected positive effects on motivation and effort.

Third, (state or nation-wide) regular (standardised) testing—a specific form of accountability—might provide increasing amounts of comprehensive and comparable information to teachers. Such information might help teachers to form more accurate teacher judgements (e.g., Ostermann et al. 2018). Thus, teacher judgements might be more accurate in countries that (regularly) conduct standardised achievement tests or comparable performance assessments.

Fourth, tracking and ability grouping might affect the teachers' information processing. Due to the necessity of assessing which course, stream or track is more suitable for each student, teachers are forced and often trained to thoroughly consider the performance of their students (e.g., Krolak-Schwerdt et al. 2018).[1] Because regular testing might improve teacher accuracy, teacher judgements might be more precise in education systems characterised by ability grouping. This should be particularly the case for teacher judgements preceding the separation of students into tracks and streams.

2.2 Teacher judgements and achievement development

Relying on the concept of the self-fulfilling prophecy (Merton 1948), Rosenthal (1973) proposed four paths through which teacher judgements might affect chil-

[1] Geven et al. (2021, p. 6) discuss additional aspects that might be linked to selection practices such as tracking and that could exacerbate stereotype-based judgements.

dren's learning and achievement: (1) teachers' input, (2) opportunities for output (e.g., calling on students), (3) teacher feedback, and (4) the nature or climate of teacher-student relations. Empirical evidence on these four possible processes and their (relative) relevance for transmitting biased judgements is scarce (see, e.g., Urhahne and Wijnia 2021, for a similar evaluation of the state of the art). One exception is the study conducted by Gentrup et al. (2020) that showed how teacher feedback varied significantly based on the inaccuracy of teachers' expectations. In particular, they reported that—compared to lower-expectancy students with similar achievement—higher-expectancy ones received more performance feedback than behavioural feedback as well as slightly more positive, rather than negative, performance feedback. Although teacher feedback varied with teacher expectations, this did not significantly mediate the significant effect of teacher expectancy on later achievement. In addition, there are studies that, although not explicitly examining teacher behaviour, showed how biased expectations might be related to student achievement by affecting, for example, student's feelings of academic futility (e.g., Agirdag et al. 2013).

From a cross-country perspective, we assume that the potential pathways are relevant in all contexts. However, some of the above-mentioned institutional features might moderate the association between biased teacher judgements and student achievement. In particular, ability grouping might be a powerful way in which teacher bias affects subsequent achievement (see, e.g., Ready and Chu 2015, for a similar argumentation). Students whose ability is underestimated by their teachers will be assigned to less-demanding, lower-quantity courses with slower pace of instruction (e.g., Gamoran 1986; Pallas et al. 1994). Such 'inadequate' placements might de-motivate students, possibly leading to lower achievement than what would have been possible under different conditions. If teacher judgements correlate with students' SES net of abilities and skills, 'inadequate' placements resulting from biased judgements would then contribute to the persistence and even exacerbation of socioeconomic achievement gaps. Another potential moderator could be standardisation (e.g., Klenowski and Wyatt-Smith 2010): the more input factors such as curricular goals, teaching materials, or exercises are predetermined, the less "room" will exist for biased teacher judgements to impact students' skills development.

Research on the effects of biased judgements and expectations (based on student gender or ethnicity) on achievement inequalities often concentrates on later stages of education. Hence, there is little evidence that focuses specifically on primary education as well as on the differential effects of biased teacher judgements for students from varying socioeconomic backgrounds. In the US, Hinnant et al. (2009) found that teacher expectation bias from Grade 1 was related to both third and fifth-grade math achievement, while the effect on achievement in Grade 3 was especially pronounced for students from low-income families. In line with these results, Sorhagen (2013) showed that, amongst other things, teachers' inaccurately low expectations in Grade 1 might foster lower math achievement at age 15, especially for students from low-income families. For Germany, there is evidence that accuracy in teacher judgements measured in Grade 9 was associated with math achievement in Grade 10 when controlling for students' background and prior achievement (Anders et al. 2010). In this study, teacher accuracy was measured with teacher's ability to

rank students in their class in terms of their overall performance in mathematics. This teacher-reported rank was then related to the rank resulting from the actual PISA math test scores from Grade 9 by calculating the rank correlation. This measurement of teacher judgement bias differs from the approach chosen in the referred studies by Hinnant et al. (2009) and Sorhagen (2013), in which teacher perceptions of children's ability were regressed on children's test scores; the resulting residuals from these regressions were interpreted as teacher judgement bias (or discrepancy scores as labelled by the authors; see also Sect. 3.3 in this paper).

2.3 The England, Germany, and US contexts

Mindset According to the data of the World Values Survey (wave 5; Inglehart et al. 2014), people from the US are more likely to strongly agree with the statement that hard work brings success, followed by people from the UK and Germany (own calculations). This suggests that a growth mindset is more prevalent in the US whilst a fixed mindset is more prevalent in Germany, with England somewhere in between. Due to the affinity between the "growth mindset" culture and the ideology of the "American dream", especially when compared to relatively pessimistic European cultural beliefs on educational success and intergenerational mobility (Alesina et al. 2018), Geven et al. (2021, p. 6) expect less SES-related bias in teacher expectations in the US than in European countries.

General notes on the education systems In *England*, compulsory schooling lasts from age 5–16, although most children attend a full-time primary school reception class at age 4. Children in reception classes are in Year 0. One year later, when compulsory school starts, they are in Year 1. Primary schooling is divided into Key Stage (KS) 1 which spans the ages from 5 and 7 (Year 1 and 2) and Key Stage 2 which covers the ages from 7–11 (Year 3–6). Each Key Stage is linked to a national curriculum: ability in a subject is then defined by the attained Key Stage level (Burgess and Greaves 2009, p. 4 f., 2013; Hall and Ozerk 2010, p. 376). In *Germany*, primary education starts on average at age 6 and lasts 4 years (6 years in the states Berlin and Brandenburg). At the end of primary education, around the age of 10, students make the transition into secondary education, which is stratified into one academic track and one or more non-academic tracks. The curricula, which are under the responsibility of the federal states, set specific goals with regard to the performance students have to achieve in each subject (Eckhardt 2019, p. 110). Still, curricula are formulated in such a way that teachers have some room for manoeuvre, although they are supposed to agree on teaching methods and assessment criteria for each specific subject within their school (Eckhardt 2019, p. 110). In the *US*, schooling starts with kindergarten at age 5. Overall, the US education system is characterised by decentralisation (McGuinn and Manna 2013), e.g., the curriculum and funding are determined by school districts, with funding of public schools not being equalised within states (Yanushevsky 2011, p. 40 f.). Hence, there are large differences between schools in different districts with respect to curriculum, school resources etc.

Accountability and testing Overall, *England* has a comprehensive assessment system (Hall and Ozerk 2010) with a high degree of accountability (Bradbury 2014). At the end of Key Stage 2 (age 11), students take national standardised exams. In addition to the standardised tests, teachers give, at the end of the academic year, a judgement for each student in the same subjects where the Key Stage 2 tests take place (Hall and Ozerk 2010, p. 376 f.). This judgement is based, amongst other things, on in-school tests as well as on a set of "probing questions" specifically provided by the central government education authorities to help assess each student's level (Burgess and Greaves 2009, p. 5). The nature and frequency of in-school tests vary greatly across schools as these are at their own discretion. In general, teachers must provide evidence for their judgement. To support teachers in "aligning their judgements systematically with national standards", the Qualifications and Curriculum Authority provides online materials (Burgess and Greaves 2009, p. 5). Key Stage results and teacher judgements—at least during primary education—have no direct impact on students' educational careers (Burgess and Greaves 2009). However, schools have an incentive to award high Key Stage scores as aggregate statistics are published and determine ranking in public school league tables, which then affects school desirability to parents and thus enrolment numbers (Burgess and Greaves 2009; Hall and Ozerk 2010). In *Germany*, during the first years of primary school, students' knowledge and skills are assessed by means of competence-based reports, observation sheets, learning development reports, learning diaries, and portfolios (Eckhardt 2019). From Grade 3 onwards, pupils start taking written tests in subjects such as German or mathematics (Eckhardt 2019). In general, several accountability mechanisms are implemented with the aim of increasing school quality (see chapter 11.2 in Eckhardt 2019, for more information). In primary education, standardised tests in German and mathematics are implemented for the second half-year of Grade 3. These tests assess the level of competency of pupils compared to the binding nation-wide educational standards (the so-called VERA 3; see, e.g., KMK 2015). The central aim of VERA 3 is to support and improve teaching and school development. Feedback based on the test results from VERA 3 for teachers contains information for each test subject at class, task, and student level—each with national comparative scores. In the *US*, there have also been state-wide standardised tests (at least once between Grade 3 and 5) at public schools since the No Child Left Behind Act of 2001 (e.g., Figlio and Loeb 2011; Hanushek and Rivkin 2010). Nevertheless, each state develops its own standards and subject-specific accountability policies (Figlio and Loeb 2011; Hanushek and Rivkin 2010).

Ability grouping In *England* and the *US*, ability grouping within schools is common and rather flexible, with low-threshold opportunities for changes over time (see Boliver and Capsada-Munsech 2021; Hallam and Parsons 2013, for England; see Condron 2008; Loveless 2013, for the US). For the UK, Hallam and Parsons showed that around 16% of children in Year 2 were streamed (Hallam and Parsons 2012, p. 522 ff.), and around 37% were setted (Hallam and Parsons 2013, p. 6). Here, streaming is defined as grouping students from the same year according to their ability into different classes in which most or all lessons are taught. Setting is instead defined as grouping students according to their ability in selected subjects

only (Hallam and Parsons 2012, p. 520), so that students attend courses of different levels for different subjects (e.g., Domina et al. 2017). In the US, ability grouping within classes and subjects is common practice during primary school (e.g., Lleras and Rangel 2009) alongside grade retention (Warren and Saliba 2012). In *Germany*, ability grouping is relatively uncommon during primary education (Ammermueller and Pischke 2009). In most federal states, lessons are taught in grades and only in some federal states, there are age-mixed groups in the first two years of schooling (Eckhardt 2019).

Table 1 summarises key information by country and presents some tentative expectations on cross-country variations in the degree of teacher bias in early primary school. Overall, we expect less pronounced variation according to family SES in the US (due to growth mindset, school accountability policies, and grouping), followed by England (due to school accountability policies and grouping). Conversely, teacher judgements in Germany should be particularly biased according to SES.

With respect to the impact of teacher biases on students' achievement development, we expect stronger effects in England and the US due to the higher prevalence of ability grouping in lower grades (not shown in Table 1). In England, however, this effect might be a bit smaller due to standardised curricula.

Special attention should be paid to the extent of time teachers and children spend together. First, we expect that teacher judgements become more reliable over time as teachers gain additional information on their students (e.g., Paleczek et al. 2017). Second, if students are taught by the same teacher over several years, her or his judgements and behaviour should affect students' achievement more strongly than if the teacher were to change every year (see Raudenbush 1984). Still, teacher bias might nevertheless affect student achievement even with yearly teacher turnover: research has shown that the effects of teacher expectations can persist over years even when teachers change (e.g., Alvidrez and Weinstein 1999; de Boer et al. 2010; Hinnant et al. 2009; Rubie-Davies et al. 2014). Reasons for such long-term, cross-year effects could be that students internalise teachers' positive or negative perceptions of their performance which could then affect, for instance, their motivation and effort.

Table 1 Key country characteristics and expectations on their effect on teacher judgement. (Own compilation)

Cultural and institutional features	Prevalence			Extent of teacher judgement bias		
	England	Germany	US	England	Germany	US
Growth mindset	No	No	Yes	/	/	Lower bias
School accountability	High	Low	(Although state-specific) high	Lower bias	/	Lower bias
Testing	Common	Common	(Although state-specific) common	Lower bias	Lower bias	Lower bias
Grouping/tracking	Streaming and setting relatively common	External tracking after Grade 4 (or 6)	Ability grouping within classes	Lower bias	/	Lower bias

/ indicate that we expect the bias to be higher than in the countries we have specified as having lower bias

It is also possible that students face systematically different learning opportunities (especially where ability grouping exists; e.g., de Boer et al. 2010). The typical situation for England is that, in primary school, the child has a single teacher for all subjects (the "class teacher"; Burgess and Greaves 2009). Teachers then change each year as children transition into a higher grade while students remain in the same class. The situation is similar in the US (e.g., Hill and Jones 2018). In Germany, a class teacher teaches all subjects during primary education and accompanies the children often for more than one year. Mainly from Grade 3 onwards, the likelihood that students are taught by other, subject-specific teachers increases (Eckhardt 2019).

3 Data and operationalisations

3.1 Data

We analysed longitudinal data from England, Germany, and the US covering the period of primary education (see Table 2 for further information).

England The Millennium Cohort Study (MCS) is an ongoing observational, multidisciplinary 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; University of London, Institute of Education, Centre for Longitudinal Studies 2021). The MCS drew a representative sample of 18,552 families from across the UK in the first wave. We restricted the sample to students in state schools in England as only for them we had information on Key Stage test results at the end of primary school from the linked National Pupil Database.

Germany The German National Educational Panel Study (NEPS) is a national multi-cohort study aimed at providing data on the development of a range of skills throughout the lifespan of cohort members (Blossfeld and Roßbach 2019). In our analyses, we used data from the Starting Cohort 2 (NEPS-SC2; NEPS Network 2020).[2] 6733 students from 374 schools were tested in Grade 1, in spring 2013, whereas 5636 parents were interviewed by telephone.

US The Early Childhood Longitudinal Study, Kindergarten Class 2010–11 (ECLS-K: 2011) collected data from a nationally representative sample of about 18,150 students who entered kindergarten in the fall of 2010 in 950 schools across the US (Tourangeau et al. 2015).[3]

[2] This paper uses data from the National Educational Panel Study (NEPS; see Blossfeld and Roßbach 2019). The NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi, Germany) in cooperation with a nationwide network.

[3] All ECLS-K:2011 sample sizes are rounded to the nearest 10 in accordance with National Center for Education Statistics (NCES) regulations.

Table 2 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	2000–2	2005/6	2004/5
T1: beginning of primary school	Y2, age 7	Grade 1, age 6–7	Grade 1, age 6–7
T2: end of primary school	Y6, age 11	Grade 4, age 9–10	Grade 5, age 10–11

MCS Millennium Cohort Study, *NEPS* National Educational Panel Study, *ECLS* Early Childhood Longitudinal Study

3.2 Instruments

3.2.1 Teacher judgements at the beginning of primary education (T1)

In all three studies, teachers were asked to rate the mathematical skills of each student on a 5-point-scale (*much worse, slightly worse, equally as good, slightly better, much better* in Germany; *well/far below average, below average, average, above average*, and *well/far above average* in the UK and US).[4] Teachers in England and Germany were asked to compare the cohort member to children of the same age, whilst teachers in the US compared the child to other children of the same grade level.

3.2.2 Tests on mathematical achievement and cognitive abilities

Mathematical achievement At age 7, the MCS administered an adapted version of the National Foundation for Educational Research (NFER) Progress in Math test. At age 11, Key Stage 2 mathematics test marks from the National Pupil Database were linked to MCS participants. The adapted version of the Progress in Math test assessed mathematical skills and knowledge by asking children 20 questions covering such topics as numbers, spaces, measurement, and data handling. The test was read aloud to children at their homes, and they were asked to complete a series of calculations in a paper and pencil exercise. All children had to complete an initial test and were then routed to an easier, medium, or harder section on the basis of their initial score. Key Stage 2 mathematics test marks are a component of the compulsory standardised assessment based on the national curriculum for all children in state schools in England at the end of Year 6 (age 11).

In Germany, we used results from mathematical tests constructed by the NEPS. The tests covered content-related (i.e., quantity, space and shape, change and relationship, data and change) and process-related components (i.e., applying technical skills, representing, modelling, communicating, problem-solving; Schnittjer et al. 2020). The tests consisted of 22 items in Grade 1 and 24 items in Grade 4. In Grade 1, a picture-based answer format was used, whereas in Grade 4, a paper-pencil format was employed.

For the US, we used results from mathematical tests conducted in Grades 1 and 5. The assessment framework was based on that developed for the National

[4] In England, the original scale goes form *well above* to *well below average,* so we reversed it.

Assessment of Educational Progress and for the Principles and Standards for School Mathematics guidelines of the National Council of Teachers of Mathematics. The assessment was designed to measure skills in conceptual knowledge, procedural knowledge, and problem-solving. The test consisted of questions on number sense, properties, and operations; measurement; geometry and spatial sense; data analysis, statistics, and probability; and patterns, algebra, and functions. At both time points, a set of routing items was administered to all students, and then the students' scores on these items determined which second-stage test (low, middle, or high difficulty) they received.

Cognitive abilities In England, cognitive abilities were measured using the British Ability Score II Pattern Construction test (Elliott et al. 1996; Jones and Schoon 2008), in which children were asked to replicate patterns presented to them using solid plastic cubes. In our analysis, we used the ability score, which accounts for differences in the items answered by children due to differential routing by difficulty.

In Germany, we used results from the NEPS-MAT test administered in Grade 2 to assess nonverbal abilities. The test included horizontally and vertically arranged fields with different geometrical elements. Children were asked to choose 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. The test consisted of 12 items.

In the US, working memory was measured by the Numbers Reversed task (Blackwell 2001). The child was asked to recall an orally presented sequence of numbers and repeat the sequence in reverse order. Although the sequences became progressively longer, they did not exceed eight numbers. 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).

All indicators for student achievement and teacher judgements were z-standardised to allow for cross-country comparisons.

3.2.3 Time aspects

In all three countries, the *windows of information collection* on teacher judgement (T1) spanned several months. We accounted for this variation by creating a dummy variable with the value 0 if a teacher rated the students early and the value 1 if a teacher rated them later in the data collection process. In England, where the school year runs from September to July, we assigned students who had been assessed by teachers between September and January to early, and those assessed between February and July to late. In Germany, early assessments took place between February and March and late assessments between April to June. In the US, assessments between January and March were defined as early 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, when we examined the effects of biased assessment on later achievement, we controlled for the *time span* (in months) between testing at T2, towards the end of primary school, and testing at T1, at the beginning of primary education.

3.2.4 Background variables

We operationalised the SES of the family, relying on the *highest education* of (step-)parents living together with the child at the start 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 where there were two resident parents, the family was categorised based on the more highly educated of the two. In the final variable, we distinguished between high, medium, and low education. In all countries, high education captured a first/bachelor's university degree or higher, requiring 3–4 years of full-time study at the tertiary level. The definition of low education differed between the two countries with comprehensive school systems (England and US) and Germany, an early tracking country. For England and the US, low education was defined as no qualification beyond the expected standard, i.e., the target of the education system for all students in compulsory education. In the US, this was a high school diploma; in England, this was the attainment of at least a grade C qualification at the end of compulsory schooling (age 16). For Germany, low education was defined as no attainment beyond the intermediate/junior secondary track. In the medium education group were all those who did not fall in either the high or low category. Family education could be evident to teachers through, e.g., cultural mannerisms and linguistic patterns (e.g., Ready and Chu 2015, p. 972).

We further considered *immigration status* indicating whether a student and/or at least one parent was foreign-born (0 = *no immigration*; 1 = *foreign-born student*

Table 3 Descriptive statistics by country. (Sources: Own calculations based on MCS, NEPS-SC2, and ECLS-K:2011)

	Time	England (*N* = 4717)		Germany (*N* = 3513)		US[a] (*N* = 3980)	
		M/%	*SD*	*M*/%	*SD*	*M*/%	*SD*
Teacher assessment: math (std.)	T1	0	1	0	1	0	1
Math. achievement (std.)	T2	0	1	0	1	0	1
Math. achievement (std.)	T1	0	1	0	1	0	1
Cognitive abilities (std.)	T1	0	1	0[b]	1	0	1
Late assessment at T1	T1	59.6	–	34.9	–	61.4	–
Age-in-months at T1 testing	T1	86.75	2.91	84.80	4.63	85.65	4.37
Time span testing T2–T1 (in months)	T2–T1	48.46	1.96	32.13	1.49	48.10	1.08
Highest parental education	T1	–	–	–	–	–	–
High	–	32.7	–	37.6	–	43.6	–
Medium	–	27.4	–	52.2	–	27.9	–
Low	–	39.9	–	10.1	–	28.5	–
Student female	T1	50.2	–	51.4	–	49.4	–
Student of immigrant descent	T1	19.3	–	22.1	–	30.8	–

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)

std. z-standardised

[a]Sample sizes rounded to nearest 10, as required by the National Center for Education Statistics

[b]Tested at the beginning of Grade 2, instead of Grade 1

and/or at least one parent). Furthermore, we controlled for the *gender* of the student (0 = *male*; 1 = *female*).

Descriptive statistics of the study variables are displayed in Table 3.

3.3 Analytic approach

We used a stepwise approach as suggested by Madon et al. (1997) and applied in other studies (e.g., Gentrup et al. 2020; Hinnant et al. 2009):

In a first step, we regressed teacher judgement on students' results in a mathematical achievement test (correcting standard errors by clustering students by teacher). In addition, results from tests on cognitive abilities were used as a covariate to reduce the risk of measurement error as well as omitted-variable bias. The concern is that a single mathematical achievement test score may not fully capture a child's "true" ability, either because it is a noisy measure (with random error) or because it is only a partial measure of overall mathematical ability. In this situation, a component of what ends up in residual error term—what we take to represent "bias"—may in fact reflect the teacher's superior knowledge of the child's genuine capacities. To take this into account, we have included a further performance measure. Both test scores—for math and cognitive abilities—should cover to a larger degree students' "true" performance, and variations in teacher ratings beyond these comprehensive indicators of student performance should then map bias. The residuals of these regressions were then standardised to zero-mean unit-variance z-scores and compared across parental education groups in order to identify biased teacher judgements and their SES gradient: a positive residual score represents teacher overestimation and a negative residual score represents teacher underestimation; the prediction of student achievement is more accurate the closer a residual score is to zero (Madon et al. 1997).

In a second step, we estimated linear regression models for maths test scores towards the end of primary school in a value-added model framework. Here, we used information on test results from Grade 1 (Germany and US) or Year 2 (England) as predictors along with the residuals from the previous regression, as well as further controls such as parental education, gender, and immigration status. Standard errors were corrected by clustering students by classes.

All results are based on the complete case analysis.

4 Results

4.1 Teacher judgements and student social background

Results from linear regressions with teacher judgement of students' mathematical skills (z-standardised scores) as the dependent variable are presented in Table 4. Students' math test scores were a much stronger predictor than cognitive abilities in all three countries. Overall, the share of explained variance was highest in England and the US (37% and 40%, respectively) and strikingly lower in Germany (25%).

Table 4 Results of regression models for teacher judgement (z-standardised). (Sources: Own calculations based on MCS, NEPS-SC2, and ECLS-K:2011)

	England	Germany	US[a]
	β (*SE*)	β (*SE*)	β (*SE*)
T1 math. achievement (std.)	0.48 (0.02)*	0.43 (0.02)*	0.54 (0.03)*
T1 cognitive abilities (std.)	0.18 (0.01)*	0.14 (0.02)*	0.12 (0.02)*
Late assessment at T1 (*ref. early*)	0.22 (0.03)*	–0.12 (0.04)*	–0.02 (0.03)
Interaction between late assessment at T1 and T1 math. achievement (std.)	0.00 (0.02)	0.03 (0.03)	0.03 (0.03)
Age-in-months at T1 testing	0.02 (0.00)*	–0.00 (0.00)	–0.00 (0.00)
Constant	–1.66 (0.36)*	0.41 (0.30)	0.32 (0.25)
R^2	0.365	0.249	0.397
N	4717	3513	3980

Results from linear regression models with clustered standard errors
$^{+}p < 0.10$; $^{*}p < 0.05$
std. z-standardised
[a]Sample sizes rounded to nearest 10, as required by the National Center for Education Statistics

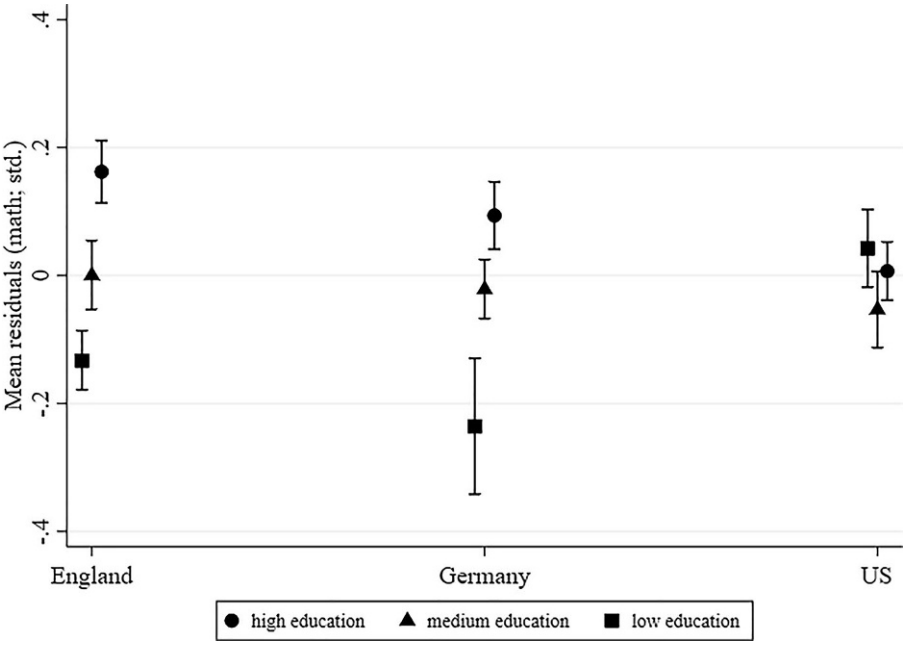

Fig. 1 Teacher judgement bias (mean residuals) by parental education and country (Plotted are the means and 95% confidence intervals for each parental education group (values are in Table S1 in the supplementary material)). (Sources: Own calculations based on MCS, NEPS-SC2, and ECLS-K:2011)

In the next step, we examined whether the standardised residuals from these regressions and, hence, the degree and direction of inaccuracy varied systematically by parental education. Fig. 1 presents the mean residuals for the three education groups with 95% confidence intervals.

In England and Germany, teacher judgements of students with low-educated parents showed on average a negative bias, whilst teacher judgements of students from high-educated families showed a positive bias. In the US, the results revealed a different pattern: family SES was unrelated to teacher judgements.

4.2 Teacher judgements, student social background, and student progress in primary school

Table 5 displays the coefficients of parental education and inaccuracy in teacher judgement (operationalised as the residuals from the regressions presented in Table 4) from the regression models for later student achievement. All models controlled for test results in Grade 1 (Germany and the US) or in Year 2 (England) as well as for socio-demographic characteristics of the student/family and the time elapsed between testing at T1 and T2. In all countries, achievement gaps related to parental education increased during primary school, with students of high-educated parents showing higher gains in mathematical skills, given their initial achievement, compared to students from medium-educated and low-educated parents (see M1

Table 5 Results of regression models for later student mathematical achievement (z-standardised). (Sources: Own calculations based on MCS, NEPS-SC2, and ECLS-K:2011)

	England		Germany		US[a]	
	M1	M2	M1	M2	M1	M2
	β (*SE*)	β (*SE*)	β (*SE*)	β (*SE*)	β (*SE*)	β (*SE*)
Highest parental education (ref. medium)						
High	0.19 (0.03)*	0.13 (0.02)*	0.25 (0.03)*	0.23 (0.03)*	0.19 (0.01)*	0.18 (0.01)*
Low	–0.10 (0.03)*	–0.05 (0.03)*	–0.29 (0.06)*	–0.25 (0.06)*	–0.12 (0.01)*	–0.13 (0.00)*
Teacher judgement residuals (std.)	–	0.34 (0.01)*	–	0.16 (0.01)*	–	0.13 (0.00)*
Controls	X	X	X	X	X	X
Constant	–1.85 (0.27)*	–0.10 (0.25)	–1.77 (0.38)*	–1.82 (0.37)*	–2.28 (0.33)*	–2.24 (0.32)*
R^2	0.458	0.567	0.401	0.427	0.641	0.657
N	4717		3513		3980	

Results from linear regression models with clustered standard errors (complete models in the supplementary material, Table S2). Controls included T1 math. achievement; cognitive abilities; time span between testing; gender; immigration status

$^{+}p < 0.10$; $^{*}p < 0.05$

std. z-standardised

Testing of significant changes between M1 and M2 of the parental education effect revealed:

England: high-educated: $\Delta\beta = -0.06$, $SE = 0.01$, $p < 0.001$; low-educated: $\Delta\beta = 0.05$, $SE = 0.01$, $p < 0.001$

Germany: high-educated: $\Delta\beta = -0.02$, $SE = 0.01$, $p = 0.003$; low-educated: $\Delta\beta = 0.04$, $SE = 0.01$, $p < 0.001$

US: high-educated: $\Delta\beta = -0.01$, $SE = 0.00$, $p = 0.010$; low-educated: $\Delta\beta = -0.00$, $SE = 0.00$, $p = 0.057$

[a]Sample sizes rounded to nearest 10, as required by the National Center for Education Statistics

in Table 5). These parental education-related differences in learning progress were particularly pronounced in Germany.

Our main interest was whether the effects of parental education were at least partly due to biased teacher judgement. Therefore, Model 2 controlled for the standardised residuals from the regressions presented in Table 4. Inaccurate judgement at the beginning of primary education was associated with students' math achievement at later time points (see M2 in Table 5): overrated (or underrated) students performed better (worse) later on. In England and to a lesser extent in Germany, we also observed a clear and significant reduction in the effect of parental education after controlling for biased judgements (England: high-educated: $\Delta\beta = -0.06$, $SE = 0.01$, $p < 0.001$; low-educated: $\Delta\beta = 0.05$, $SE = 0.01$, $p < 0.001$; Germany: high-educated: $\Delta\beta = -0.02$, $SE = 0.01$, $p = 0.003$; low-educated: $\Delta\beta = 0.04$, $SE = 0.01$, $p < 0.001$)[5]. In England, for example, 32% of the reduction of the achievement gap between students from high- and medium-educated families was explained by biased teacher judgements at T1 in the value-added model ($\Delta\beta/\beta_{M1}$); for the reduction of the achievement gap between students from low- and medium-educated families the share was 50%. In the US, in contrast, there was a small, but significant increase in the effect for students with high-educated parents when controlling for teacher judgement residuals (high-educated: $\Delta\beta = -0.01$, $SE = 0.00$, $p = 0.010$; low-educated: $\Delta\beta = -0.00$, $SE = 0.00$, $p = 0.057$). These patterns supported our expectation that biased teacher judgements might contribute to widening SES-related achievement inequalities over time.

4.3 Sensitivity checks and further analysis

Heterogeneous effects of biased teacher judgements In England and the US, we found that the association of biased teacher judgements with math achievement was significantly weaker for students from highly educated families as compared to students from low-educated families (see Table S3 in the supplementary material). In Germany, respective interaction effects between parental education and residuals were nonsignificant.

Teacher change over the course of primary education For Germany, we considered whether effects were stronger for children taught by the same teacher throughout several grades of primary education. Information collected annually on teachers' birth year, month, and gender was used to identify whether the teacher changed over time. We then re-estimated all models for a restricted sample of students who were taught by the same teacher for at least two years (see Table S4). The results were similar to those presented above, both in terms of teacher judgement accuracy and the effects of teacher judgements on student achievement.

Language skills We replicated all analyses using language skills as the outcome variable (see Tables S5–S8 as well as Figure S1). The results regarding biased judgements and their association with achievement development are largely comparable

[5] For the coefficient comparison we used the suest command implemented in Stata (Mize et al. 2019).

to the results presented for mathematics. For Germany, however, we observed a less pronounced association between teacher judgement bias and later language skills than in the mathematical domain. It remains an open research question why there is such a weak association in Germany. One potential explanation might be that the objective measures of language skills at T2 in England and the US are curriculum-oriented, while those in Germany aim to assess general language skills (i.e., receptive vocabulary): Assuming that teacher judgements are manifested in teaching behaviour, then they should have a stronger influence on the acquisition of language skills that are primarily taught at school (e.g., reading skills, spelling, etc.). In contrast, in the acquisition of more general language skills such as vocabulary, parents and peers are also strongly involved. Consequently, effects of teacher judgement bias on later achievement should be more observable when curriculum-oriented tests are used.

5 Conclusion and discussion

In this paper, we asked whether inaccurate teacher judgements of their students' math skills correlate with student social origin and whether such bias is associated with math achievement in primary school. We examined unexplained variance in teacher judgements that remained after controlling for actual student achievement and cognitive abilities, and we interpreted this variance as inaccuracy in teacher judgements. We expected that a growth mindset culture, higher accountability, and more ability grouping lead to lower teacher judgement bias. In consequence, we expected teacher bias to be particularly low in the US, followed by England. In contrast, we expected a more pronounced teacher judgement bias for Germany due to less common growth mindsets, a lower degree of accountability, and non-existing ability grouping during primary education. Empirically, our expectations were confirmed. We showed that the unexplained variance in teacher judgements was systematically linked to family SES, operationalised by the highest parental education, in Germany and to a lesser extent in England but not in the US. This pattern is in line with previous research on systematic variations in teacher judgements based on family SES in those countries (e.g., Campbell 2015; Geven et al. 2021; Hinnant et al. 2009; Tobisch and Dresel 2017).

In a subsequent step, we studied whether teacher judgement bias was associated with later achievement and mediated the effect of parental education. Due to the higher prevalence of ability grouping in lower grades in England and the US, we expected stronger effects on later achievement in these two countries compared to Germany. For England, this effect might be attenuated due to standardised curricula. Empirically, we showed that inaccuracy in teacher judgements predicted students' end of primary school achievement in all three countries, even when considering prior achievement, cognitive abilities, and students' background characteristics. This could be interpreted as a self-fulfilling prophecy. Only in England and Germany did the effect of parental education decrease when controlling for biased judgements. Since no relation was found between teacher inaccuracy and parental education in the US, it is not surprising that, in this country, the parental education effect on math

achievement growth was mediated only partially and to a lesser extent by teacher judgement.

As we observed country differences both in the extent of teacher bias and in the relevance of this bias for the achievement development (see Geven et al. 2021, for similar findings), our findings support the assumption that the institutional and societal settings matter. Hence, a cross-country perspective enriches research on the role of teachers in explaining SES-related inequalities. Our results can be seen as a starting point for future research to investigate cross-country variations and the underlying mechanisms in more detail.

Although our study contributes to the literature on teacher accuracy and bias by providing a cross-country comparison, it has limitations. First, we described the underlying theoretical considerations and linked them to the situation in the three countries under study to derive hypotheses on the extent of family SES-related judgement bias in England, Germany, and the US. However, our expectations referred to general trends for the three countries, but it remains open to what extent within-country variation leads to a weakening or strengthening of observable patterns. Second, we did not consider underlying mechanisms such as ability grouping or actual accountability and monitoring approaches at school and, therefore, we do not know which of the presumed mechanisms have actually led to the observed patterns. For example, we expected ability grouping to be important in mediating the association between biased teacher judgements and later achievement. However, we are not able to explore this fully because ability grouping assignments in our data took place pre-date, and were observable to teachers, when they formed their judgements. Exemplary analyses for England revealed that ability group placement is associated with teacher judgements and predicts how children will progress over time, net of standardised test scores at T1 (see Table S9). Future studies should also examine to what extent there is mutual reinforcement or weakening between, amongst others, growth mindset, accountability, and ability grouping. Furthermore, further institutional or societal factors that might (simultaneously) affect teacher judgement bias should be considered. For example, another social mechanism possibly responsible for the observed country differences might be variation in the awareness of expectation effects among the teachers. In the US, research on self-fulfilling prophecies in schools has a much longer tradition (initiated by the experiment Pygmalion in the Classrooms, see Rosenthal and Jacobsen 1968) compared to Europe. Consequently, teachers might be informed about this phenomenon in the US, for example, during teacher education, but to a lesser extent in Europe. Third, concerns might be raised about linking results from standardised assessments with global teacher judgements (e.g., Arens et al. 2017; Hübner et al. 2022; Jussim and Harber 2005). Teacher judgements might be more accurate than test results as teachers might have "valid" information above and beyond what a (single) test captures. This additional information would also account for the fact that children with higher teacher judgements perform better in later achievement tests. Fourth, previous research showed that indirect teacher judgements of general mathematical performance, like the ones we drew on, correlate less strongly with actual student skills and abilities than direct judgements such as those referring to the expected number of correctly solved tasks of a math test (Hoge and Coladarci 1989; see also Südkamp et al. 2012). This

result indicates that direct teacher judgements are more accurate. Consequently, the patterns we reported in this study might have been less pronounced, if one were to use more specific measures of teacher judgement measures. Fifth, possible criticism might pertain to the two-step approach used: standard errors in the second regression will tend to be underestimated as residuals are treated as observed variables, ignoring the imprecision that comes from estimating them (Murphy and Topel 2002). Sixth, we did not consider the role of race or ethnicity. For the US, in particular, there is mixed evidence that the discrepancy between teacher judgement or expectations and student test scores systematically varies with students' race or ethnicity (see, e.g., Geven et al. 2018; Wang et al. 2018, for an overview). While some studies documented variation in teacher ratings by race and/or ethnicity (e.g., McKown and Weinstein 2008; Ready and Wright 2011), others suggest much of this variation to reflect actual performance differences between various race and/or ethnic groups (e.g., Jussim et al. 1996; Madon et al. 1998). Understanding how the country differences in SES-related teacher bias are confounded with race- or ethnicity-related bias is one important topic for future research. Finally, as in all cross-country studies with post-hoc harmonised data, survey designs and instruments differ in measurement points, test material, or wording of questions (Law et al. 2021). We tried to make the data as comparable as possible; however, some issues remain. For example, although all mathematics achievement tests seemed to measure similar facets of math skills (e.g., number knowledge, knowledge of geometry, and spatial sense), we did not have access to single test items.

Supplementary Information The online version of this article (https://doi.org/10.1007/s11618-022-01119-7) contains supplementary material, which is available to authorised users.

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.

Funding Open Access funding enabled and organized by Projekt DEAL.

Conflict of interest M. Olczyk, S.J. Kwon, G. Lorenz, V. Perinetti Casoni, T. Schneider, A. Volodina, J. Waldfogel and E. Washbrook declare that they have no competing interests.

References

Agirdag, O., van Avermaet, P., & van Houtte, M. (2013). School segregation and math achievement: a mixed-method study on the role of self-fulfilling prophecies. *Teachers College Record*, *115*(3), 30305.

Alesina, A., Stantcheva, S., & Teso, E. (2018). Intergenerational mobility and preferences for redistribution. *American Economic Review*, *108*(2), 521–554. https://doi.org/10.1257/aer.20162015.

Alvidrez, J., & Weinstein, R. S. (1999). Early teacher perceptions and later student academic achievement. *Journal of Educational Psychology*, *91*(4), 731–746. https://doi.org/10.1037/0022-0663.91.4.731.

Ammermueller, A., & Pischke, J. S. (2009). Peer effects in European primary schools: evidence from the Progress in International Reading Literacy Study. *Journal of Labor Economics*, *27*(3), 315–348. https://doi.org/10.1086/603650.

Anders, Y., Kunter, M., Brunner, M., Krauss, S., & Baumert, J. (2010). Diagnostische Fähigkeiten von Mathematiklehrkräften und ihre Auswirkungen auf die Leistungen ihrer Schülerinnen und Schüler. *Psychologie in Erziehung und Unterricht*, *57*(3), 175–193. https://doi.org/10.2378/peu2010.art13d.

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. *Journal of Educational Psychology*, *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. *Journal of Educational Psychology*, *82*(4), 683–690. https://doi.org/10.1037/0022-0663.82.4.683.

Babad, E. (1993). Teachers' differential behavior. *Educational Psychology Review*, *5*(4), 347–376. https://doi.org/10.1007/BF01320223.

Baumert, J., & Kunter, M. (2013). The COACTIV model of teachers' professional competence. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss & M. Neubrand (Eds.), *Cognitive activation in the mathematics classroom and professional competence of teachers* (pp. 25–48). New York: Springer Science + Business Media.

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), 1–14. 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. *Rehabilitation Counseling Bulletin*, *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)* (2nd edn.). Wiesbaden: Springer VS. https://doi.org/10.1007/978-3-658-23162-0.

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

Boliver, V., & Capsada-Munsech, Q. (2021). Does ability grouping affect UK primary school pupils' enjoyment of Maths and English? *Research in Social Stratification and Mobility*. https://doi.org/10.1016/j.rssm.2021.100629.

Bradbury, A. (2014). 'Slimmed down' assessment or increased accountability? Teachers, elections and UK government assessment policy. *Oxford Review of Education*, *40*(5), 610–627. https://doi.org/10.1080/03054985.2014.963038.

Burgess, S., & Greaves, E. (2009). *Test scores, subjective assessment and stereotyping of ethnic minorities*. Working Paper 09/221. Bristol: The Centre for Market and Public Organisation, University of Bristol.

Burgess, S., & Greaves, E. (2013). Test scores, subjective assessment, and stereotyping of ethnic minorities. *Journal of Labor Economics*, *31*(3), 535–576. https://doi.org/10.1086/669340.

Campbell, T. (2015). Stereotyped at seven? Biases in teacher judgement of pupils' ability and attainment. *Journal of Social Policy*, *44*(3), 517–547. https://doi.org/10.1017/S0047279415000227.

Condron, D. J. (2008). An early start: skill grouping and unequal reading gains in the elementary years. *The Sociological Quarterly*, *49*(2), 363–394. https://doi.org/10.1111/j.1533-8525.2008.00119.x.

Domina, T., Penner, A., & Penner, E. (2017). Categorical inequality: schools as sorting machines. *Annual Review of Sociology*, *43*, 311–330. https://doi.org/10.1146/annurev-soc-060116-053354.

Dweck, C. S., & Yeager, D. S. (2019). Mindsets: a view from two eras. *Perspectives on Psychological Science*, *14*(3), 481–496. https://doi.org/10.1177/1745691618804166.

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*. Bonn: KMK.

Elliott, C. D., Smith, P., & McCulloch, K. (1996). *British ability scales (2nd ed.)*. Windsor: NFER-NELSON.

Figlio, D., & Loeb, S. (2011). School accountability. In *Handbook of the economics of education* (Vol. 3, pp. 383–421). Amsterdam: Elsevier. 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. *American Educational Research Journal*, *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 S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 231–254). New York: Guilford.

Gamoran, A. (1986). Instructional and institutional effects of ability grouping. *Sociology of Education*, *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. *European Journal of Social Psychology*, *33*(5), 573–589. https://doi.org/10.1002/ejsp.166.

Gentrup, S., Lorenz, G., Kristen, C., & Kogan, I. (2020). Self-fulfilling prophecies in the classroom: teacher expectations, teacher feedback and student achievement. *Learning and Instruction*, *66*, 101296. https://doi.org/10.1016/j.learninstruc.2019.101296.

Geven, S., Batruch, A., & van de Werfhorst, H. (2018). *Inequality in teacher judgements, expectations and track recommendations: a review study*. Amsterdam: University of Amsterdam. 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. *Social Science Research*. https://doi.org/10.1016/j.ssresearch.2021.102599.

Hall, K., & Ozerk, K. (2010). Primary curriculum and assessment: England and other countries. In R. J. Alexander (Ed.), *The Cambridge Primary Review Research Surveys* (pp. 375–414). London: Routledge.

Hallam, S., & Parsons, S. (2012). Prevalence of streaming in UK primary schools: evidence from the Millennium Cohort Study. *British Educational Research Journal*. https://doi.org/10.1080/01411926.2012.659721.

Hallam, S., & Parsons, S. (2013). The incidence and make up of ability grouped sets in the UK primary school. *Research Papers in Education*, *28*(4), 393–420. https://doi.org/10.1080/02671522.2012.729079.

Hanushek, E. A., & Rivkin, S. G. (2010). The quality and distribution of teachers under the No Child Left Behind Act. *Journal of Economic Perspectives*, *24*(3), 133–150. https://doi.org/10.1257/jep.24.3.133.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London: Routledge. https://doi.org/10.4324/9780203887332.

Helmke, A., & Schrader, F. W. (1987). Interactional effects of instructional quality and teacher judgement accuracy on achievement. *Teaching and Teacher Education*, *3*(2), 91–98. https://doi.org/10.1016/0742-051X(87)90010-2.

Hill, A. J., & Jones, D. B. (2018). A teacher who knows me: the academic benefits of repeat student-teacher matches. *Economics of Education Review*, *64*, 1–12. https://doi.org/10.1016/j.econedurev.2018.03.004.

Hinnant, J. B., O'Brien, M., & Ghazarian, S. R. (2009). The longitudinal relations of teacher expectations to achievement in the early school years. *Journal of Educational Psychology*, *101*(3), 662–670. https://doi.org/10.1037/a0014306.

Hofer, S. I. (2015). Studying gender bias in physics grading: the role of teaching experience and country. *International Journal of Science Education*, *37*(17), 2879–2905. https://doi.org/10.1080/09500693.2015.1114190.

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: a review of literature. *Review of Educational Research*, *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. *Journal of Educational Psychology*, *114*(2), 326–345. https://doi.org/10.1037/edu0000571.

Inglehart, R., Haerpfer, C., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., Lagos, M., Norris, P., Ponarin, E., & Puranen, B. (2014). *World Values Survey: round five—country-pooled datafile version: www.worldvaluessurvey.org/WVSDocumentationWV5.jsp*. Madrid: JD Systems Institute.

Jacovidis, J. N., Anderson, R. C., Beach, P. T., & Chadwick, K. L. (2020). *Growth mindset thinking and beliefs in teaching and learning*. London: inflexion.

Jerrim, J., & Sims, S. (2022). School accountability and teacher stress: international evidence from the OECD TALIS study. *Educational Assessment, Evaluation and Accountability*, *34*(1), 5–32. https://doi.org/10.1007/s11092-021-09360-0.

Jones, E. M., & Schoon, I. (2008). Child cognition and behaviour. In K. Hansen & H. Joshi (Eds.), *Millennium cohort study: third survey: a user's guide to initial findings* (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., & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review*, *9*(2), 131–155. https://doi.org/10.1207/s15327957pspr0902_3.

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 M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 28, pp. 281–388). Cambridge: Academic Press. https://doi.org/10.1016/S0065-2601(08)60240-3.

Karing, C., Pfost, M., & Artelt, C. (2011). Hängt die diagnostische Kompetenz von Sekundarstufenlehrkräften mit der Entwicklung der Lesekompetenz und der mathematischen Kompetenz ihrer Schülerinnen und Schüler zusammen. *Journal for Educational Research Online*, *3*(2), 119–147.

Kelley, C., Heneman, H., & Milanowski, A. (2002). Teacher motivation and school-based performance Awards. *Educational Administration Quarterly*, *38*(3), 372–401. https://doi.org/10.1177/0013161X02383004.

Kelly, S., & Carbonaro, W. (2012). Curriculum tracking and teacher expectations: evidence from discrepant course taking models. *Social Psychology of Education*, *15*(3), 271–294. https://doi.org/10.1007/s11218-012-9182-6.

Kiss, D. (2013). Are immigrants and girls graded worse? Results of a matching approach. *Education Economics*, *21*(5), 447–463. https://doi.org/10.1080/09645292.2011.585019.

Klenowski, V., & Wyatt-Smith, C. M. (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—The Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany (2015). Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring. https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2015/2015_06_11-Gesamtstrategie-Bildungsmonitoring.pdf. Accessed: 1 June 2022.

Krolak-Schwerdt, S., Böhmer, M., & Gräsel, C. (2013). The impact of accountability on teachers' assessments of student performance: a social cognitive analysis. *Social Psychology of Education*, *16*(2), 215–239. https://doi.org/10.1007/s11218-013-9215-9.

Krolak-Schwerdt, S., Pit-ten, C. 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 O. Zlatkin-Troitschanskaia, M. Toepper, H. A. Pant, C. Lautenbach & C. Kuhn (Eds.), *Assessment of learning outcomes in higher education.* Methodology of educational measurement and assessment (Vol. 27, pp. 73–101). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-74338-7_5.

Law, J., Wareham, H., Volodina, A., & Rush, R. (2021). The pros and cons of combining birth cohort data. https://dynamicsofinequality.org/publication/the-pros-and-cons-of-combining-birth-cohort-data/. Accessed: 1 June 2022.

Lee, M. W., & Newton, P. (2021). *Systematic divergence between teacher and test-based assessment: literature review*. Coventry: Ofqual.

Lee, J., Liu, X., Amo, L. C., & Wang, W. L. (2014). Multilevel linkages between state standards, teacher standards, and student achievement. *Educational Policy*, *28*(6), 780–811. https://doi.org/10.1177/0895904813475708.

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, *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. *American Journal of Education*, *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 modeling (DiacoM). *Teaching and Teacher Education*, *91*, 103059. https://doi.org/10.1016/j.tate.2020.103059.

Lorenz, G., Gentrup, S., Kristen, C., Stanat, P., & Kogan, I. (2016). Stereotype bei Lehrkräften? Eine Untersuchung systematisch verzerrter Lehrererwartungen. *Kölner Zeitschrift Für Soziologie Und Sozialpsychologie*, *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?* Washington, DC: The Brookings Institution.

Madon, S., Jussim, L., & Eccles, J. (1997). In search of the powerful self-fulfilling prophecy. *Journal of Personality and Social Psychology*, *72*(4), 791–809. https://doi.org/10.1037/0022-3514.72.4.791.

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

McGuinn, P., & Manna, P. (2013). Education governance in America: Who leads when everyone is in charge? In P. Manna & P. McGuinn (Eds.), *Education governance for the twenty-first century: overcoming the structural barriers to school reform* (pp. 1–18). Washington, DC: Brookings Institution Press.

McKown, C., & Weinstein, R. S. (2008). Teacher expectations, classroom context, and the achievement gap. *Journal of School Psychology*, *46*(3), 235–261. https://doi.org/10.1016/j.jsp.2007.05.001.

Merton, R. K. (1948). The self-fulfilling prophecy. *The Antioch Review*, *8*, 193–210.

Mintrop, H. (2003). The limits of sanctions in low-performing schools. *Education Policy Analysis Archives*, *11*(3), 1–30. https://doi.org/10.14507/epaa.v11n3.2003.

Mize, T. D., Doan, L., & Long, S. (2019). A general framework for comparing predictions and marginal effects across models. *Sociological Methodology*, *49*(1), 152–189. https://doi.org/10.1177/0081175019852763.

Murphy, K. M., & Topel, R. H. (2002). Estimation and inference in two-step econometric models. *Journal of Business and Economic Statistics*, *20*(1), 88–97. https://doi.org/10.1198/073500102753410417.

Network, N. E. P. S. (2020). *National educational panel study, scientific use file of starting cohort Kindergarten*. Bamberg: Leibniz Institute for Educational Trajectories (LIfBi). https://doi.org/10.5157/NEPS:SC2:9.0.0.

Ostermann, A., Leuders, T., & Nückles, M. (2018). Improving the judgment of task difficulties: prospective teachers' diagnostic competence in the area of functions and graphs. *Journal of Mathematics Teacher Education*, *21*(6), 579–605. https://doi.org/10.1007/s10857-017-9369-z.

Paleczek, L., Seifert, S., & Gasteiger-Klicpera, B. (2017). Influence on teachers' judgment accuracy of reading abilities on second and third grade students: A multilevel analysis. *Psychology in the Schools*, *54*(3), 228–245. https://doi.org/10.1002/pits.21993.

Pallas, A. M., Entwisle, D. R., Alexander, K. L., & Stluka, M. F. (1994). Ability-group effects: Instructional, social, or institutional? *Sociology of Education*, *67*(1), 27–46. https://doi.org/10.2307/2112748.

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. *European Journal of Psychology of Education*, *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. *British Journal of Educational Psychology*, *67*(2), 235–247. https://doi.org/10.1111/j.2044-8279.1997.tb01240.x.

Raudenbush, S. W. (1984). Magnitude of teacher expectancy effects on pupil IQ as a function of the credibility of expectancy induction: a synthesis of findings from 18 experiments. *Journal of Educational Psychology*, *76*(1), 85–97. https://doi.org/10.1037/0022-0663.76.1.85.

Ready, D. D., & Chu, E. M. (2015). Sociodemographic inequality in early literacy development: the role of teacher perceptual accuracy. *Early Education and Development*, *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. *American Educational Research Journal*, *48*(2), 335–360. https://doi.org/10.3102/0002831210374874.

Ritchie, S.J., & Bates, T.C. (2013). Enduring links from childhood mathematics and reading achievement to adult socioeconomic status. *Psychological Science*, *24*(7), 1301–1308. https://doi.org/10.1177/0956797612466268.

Romijn, B.R., Slot, P.L., & Leseman, P.P. (2021). Increasing teachers' intercultural competences in teacher preparation programs and through professional development: a review. *Teaching and Teacher Education*, *98*, 103236. https://doi.org/10.1016/j.tate.2020.103236.

Rosenthal, R. (1973). The mediation of Pygmalion effects: a four factor "theory". *Papua New Guinea Journal of Education*, *9*(1), 1–12.

Rosenthal, R., & Jacobsen, L. (1968). *Pygmalion in the classroom: teacher expectation and pupils' intellectual development*. New York: Holt, Rinehart and Winston.

Rubie-Davies, C.M. (2007). Classroom interactions: exploring the practices of high- and low-expectation teachers. *The British Journal of Educational Psychology*, *77*(2), 289–306. https://doi.org/10.1348/000709906X101601.

Rubie-Davies, C.M., Weinstein, R.S., Huang, F.L., Gregory, A., Cowan, P.A., & Cowan, C.P. (2014). Successive teacher expectation effects across the early school years. *Journal of Applied Developmental Psychology*, *35*(3), 181–191. https://doi.org/10.1016/j.appdev.2014.03.006.

Schnittjer, I., Gerken, A.-L., & Petersen, L.-A. (2020). *NEPS technical report for mathematics: scaling results of Starting Cohort 2 in grade 4* (NEPS Survey Paper No. 69). Bamberg: Leibniz Institute for Educational Trajectories.

Smith, E.R., & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review*, *4*(2), 108–131. https://doi.org/10.1207/S15327957PSPR0402_01.

Sorhagen, N.S. (2013). Early teacher expectations disproportionately affect poor children's high school performance. *Journal of Educational Psychology*, *105*(2), 465–477. https://doi.org/10.1037/a0031754.

Sprietsma, M. (2013). Discrimination in grading: experimental evidence from primary school teachers. *Empirical Economics*, *45*(1), 523–538. https://doi.org/10.1007/s00181-012-0609-x.

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: a meta-analysis. *Journal of Educational Psychology*, *104*(3), 743–762. 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. *Journal of Educational Psychology*, *99*(2), 253–273. https://doi.org/10.1037/0022-0663.99.2.253.

Tobisch, A., & Dresel, M. (2017). Negatively or positively biased? Dependencies of teachers' judgments and expectations based on students' ethnic and social backgrounds. *Social Psychology of Education*, *20*(4), 731–752. https://doi.org/10.1007/s11218-017-9392-z.

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 (NCES 2015-078)*. Washington, DC: National Center for Education Statistics.

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), England: secure access*. UK Data Services. 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, 2008*. https://doi.org/10.5255/UKDA-SN-6411-8.

Urhahne, D., & Wijnia, L. (2021). A review on the accuracy of teacher judgments. *Educational Research Review*, *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. *Educational Research and Evaluation*, *24*(3–5), 124–179. https://doi.org/10.1080/13803611.2018.1548798.

Warren, J.R., & Saliba, J. (2012). First through eighth grade retention rates for all 50 states: a new method and initial results. *Educational Researcher*, *41*(8), 320–329. https://doi.org/10.3102/0013189X12457813.

Weinstein, R.S. (2002). *Reaching higher: the power of expectations in schooling* (1. Paperback Ed.). Cambridge: Harvard University Press.

Yanushevsky, R. (2011). *Improving education in the US: a political paradox*. New York: Algora Publishing.

Affiliations

Dr. Melanie Olczyk (✉)
Institute of Sociology, Martin-Luther-Universität Halle-Wittenberg, Paracelsusstraße 22, 06114 Halle (Saale), Germany
E-Mail: melanie.olczyk@soziologie.uni-halle.de

Sarah Jiyoon Kwon · Prof. Dr. Jane Waldfogel
Columbia University, 1255 Amsterdam Avenue, New York, NY 10027, USA
Sarah Jiyoon Kwon
E-Mail: sarah.jiyoon.kwon@columbia.edu
Prof. Dr. Jane Waldfogel
E-Mail: j.waldfogel@columbia.edu

Dr. Georg Lorenz
Institute for Educational Quality Improvement, Humboldt-Universität zu Berlin, Hannoversche Str. 19, 10115 Berlin, Germany
E-Mail: georg.lorenz@iqb.hu-berlin.de

Institute of Sociology, Leipzig University, Beethovenstraße 15, 04107 Leipzig, Germany

Valentina Perinetti Casoni · Dr. Elizabeth Washbrook
University of Bristol, 35 Berkeley Square, Bristol, UK
Valentina Perinetti Casoni
E-Mail: valentina.perinetticasoni@bristol.ac.uk
Dr. Elizabeth Washbrook
E-Mail: Liz.Washbrook@bristol.ac.uk

Prof. Dr. Thorsten Schneider
Institute of Sociology, Leipzig University, Beethovenstraße 15, 04107 Leipzig, Germany
E-Mail: thorsten.schneider@uni-leipzig.de

Dr. Anna Volodina
University of Bamberg, Augustenstraße 6, 96047 Bamberg, Germany
E-Mail: anna.volodina@uni-bamberg.de