University of BRISTOL

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

A cross-country perspective

Melanie Olczyz - *MLU Halle-Wittenberg, Germany*

Sarah Kwon - Columbia University, USA
Georg Lorenz - *IQB, Humboldt-Universität zu Berlin, Germany*
Valentina Perinetti Casoni - *University of Bristol, UK*
Thorsten Schneider - *University Leipzig, Germany*
Anna Volodina - *University of Bamberg, Germany*
Jane Waldfogel - Columbia University, USA
Elizabeth Washbrook - *University of Bristol, UK*

SLLS Conference – Cleveland, October 2022

CONTENT

- Background
- Theoretical considerations
- Country contexts: England, Germany, and the US
- Data
- Analytical approach
- RESULTS: Step 1
- RESULTS: Step 2
- Discussion
- Sensitivity checks & future research

University of BRISTOL

BACKGROUND

Various dimensions of educational success, such as student achievement, vary by parental socioeconomic status (SES)

Stereotypes held by teachers can bias teacher judgement of pupils' ability *(Jussim et al., 1996; Jussim & Harber, 2005; Tenenbaum & Ruck, 2007)*

Differential teacher judgements & expectations can:

- affect given grades *(Kiss, 2013; Sprietsma, 2013)*
- lead to less-warm and supportive feedback from teachers *(Gentrup et al., 2020; Rubie-Davies, 2007)*
- result in different non-verbal teacher behaviours (e.g., reduced eye contact) *(Babad, 1990, 1993)*

EXACERBATE or (partially) ACCOUNT FOR SES-related achievement gaps and social inequalities in education

Few studies:

- take a cross-country perspective and consider the wider institutional setting *(see, e.g., Geven et al., 2021; Hofer, 2015)*
- focus specifically on primary education *(see, e.g., Hinnant et. al, 2009; Sorhagen, 2013; Anders et al. 2010).*

University of BRISTOL

THEORETICAL CONSIDERATION - I

Teacher judgements & judgement bias

ACCURACY of teacher judgement varies between teachers

From meta-analysis by Sudkamp et al, 2012

shared variance between → teacher judgement; students' achievement } is around 40%

The remaining variance is **INACCURACY**, (positively or negatively) biased teacher judgement

In all three countries of this study, it has been empirically shown that students from more socioeconomically disadvantaged families often face lower teacher expectations vis-à-vis their objective achievement measures

→ **JUDGED MORE INACCURATELY**

(e.g., Lorenz et al., 2016; Tobisch & Dresel, 2017; Campbell, 2015; Lee & Newton, 2021; Plewis, 1997; Alvidrez & Weinstein, 1999)

University of BRISTOL

THEORETICAL CONSIDERATION - II

Teacher judgements & judgement bias

Automatic judgements guided by stereotypes that don't include the integration of relevant target information

TEACHER JUDGEMENT

Information-based judgements that involves the deliberate integration of target information into multifaceted judgements

PERSONAL CHARACTERISTICS:
attitudes, knowledge, mindset, etc.

SITUATIONAL CHARACTERISTICS:
time pressure, judgement goals, social cues, etc.

INSTITUTIONAL CONTEXT:
conditions and regulations on schools, school system, teacher training, norms & values, cultural-cognitive beliefs

MINDSET = in 'growth mindset' cultures, it is believed that initial disadvantages due to family SES can be overcome through effort (vs 'fixed mindset' cultures where talent and skills are viewed as innate (Geven et al., 2021)

ACCOUNTABILITY = teachers could be expected to have more incentive to judge student achievement accurately in systems in which they are held accountable for their work (Krolak-Schwerdt et al., 2018)

STANDARDISED TESTING = might provide teachers with increasing amounts of comprehensive and comparable information + specific form of accountability

TRACKING & ABILITY GROUPING = teachers might be better trained at judging students due to the necessity of assessing which course, stream, or track is more suitable

University of BRISTOL

THEORETICAL CONSIDERATION - III

Teacher judgement & achievement development

How can teacher judgements affect children's learning and achievement?

Rosenthal (1973) relying on Merton's (1948) concept of the SELF-FULFILLING PROPHECY proposed 4 paths

1. Teacher's input
2. Opportunities for output (e.g., calling on students)
3. Teacher feedback
4. Nature or climate of teacher-student relations

PLUS some of the institutional context feature that might affect teacher judgement might also moderate the association between teacher judgement and achievement development

ABILITY GROUPING: (see Ready and Chu, 2015)
Students whose abilities are underestimated will be assigned to less-demanding, lower-quantity, more slowly-paced course which might then demotivate students, possibly leading to lower achievement

STANDARDISATION: (see Klenowski and Wyatt-Smith, 2010)
the more input factors such as curricular goals, teaching materials, or exercise are predetermined, the less room will exist for biased teachers judgement

Might contribute to the persistence or even exacerbation of SES achievement gaps

University of BRISTOL

COUNTRY CONTEXTS

Key country characteristics and expectations on their effect on teacher judgement

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	(state-specific) high	lower bias	/	lower bias
Testing	common	common	(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

Note. Own compilation. / indicates that we expect the bias to be higher than in the countries we have specified as having lower bias.

EXPECTATIONS:

1) *Extent of teacher bias* (systematic variation according to SES): less in the US, followed by England, and then Germany.

2) *Effect of teacher bias*: stronger effects in England and the US

University of BRISTOL

DATA

	ENGLAND	GERMANY	UNITED STATES
SURVEY	Millennium Cohort Study* MCS	National Educational Panel Study – Starting Cohort 2 NEPS-SC2	Early Childhood Longitudinal Study: Kindergarten Class of 2010-2011** ECLS-K:2011
BIRTH COHORT	2000 – 2002	2005 – 2006	2004 – 2005
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
SAMPLING: PSU	Electoral wards	Schools	schools

* *Sample restricted to students in state schools in England*

** *Sample sizes are rounded to nearest 10, as required by the National Center for Education Statistics.*

University of BRISTOL

INSTRUMENTS

	ENGLAND	GERMANY	UNITED STATES
T1 Teacher assessment: math. (std.)	Teachers rating on pupil's mathematical skills on a 5-point scale		
T2 Math. achievement (std.)	KS2 Total Math marks	NEPS Grade 4 Math test	ECLS:K Grade 5 Maths test
T1 Math. achievement (std.)	NFER PiM	NEPS Grade 1 Math test	ECLS:K Grade 1 Maths test
T1 Cognitive abilities (std.)	BAS II Pattern Construction	NEPS-MAT Grade 2	Working Memory

	SES
T1	HIGHEST PARENTAL EDUCATION [High, Medium, Low]
	TIME CONTROLS
T1	Late assessment at T1
T1	Age-in-months at T1 testing
T2-T1	Time span testing T2-T1 (in months)
	OTHER CONTROLS
T1	Immigration status
T1	Female student

University of BRISTOL

METHODOLOGY

Stepwise approach:

(1) Is teacher assessment at T1 (positively or negatively) biased?

RESIDUAL APPROACH: regress T1 teacher assessment on T1 achievement (and T1 cognitive abilities + controls)

-> POSITIVE residuals = teacher overestimation of pupil's ability

-> NEGATIVE residuals = teacher underestimation of pupil's ability

Is there a SES gradient in (biased) teacher assessment?

(2) Does T1 (biased) teacher assessment predict achievement at T2?

Regress T2 achievement on (std) T1 residuals (and SES + controls)

See Madon et al., (1997); Gentrup et al., (2020); and Hinnant et al., (2009)

University of BRISTOL

STEP 1: Is teacher assessment biased?

Results of regression models for teacher judgement (z-standardised)

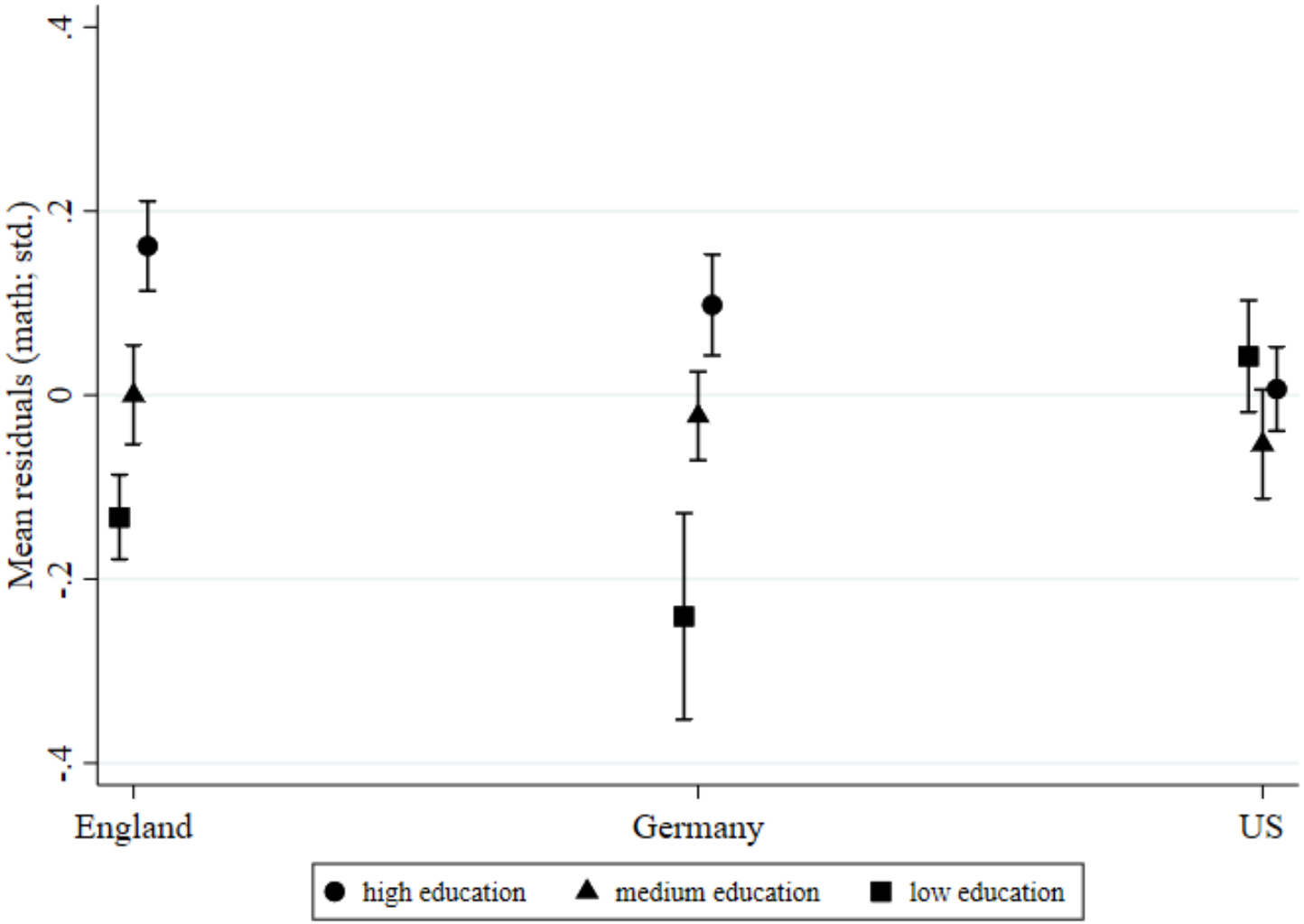
	England β (*SE*)	Germany β (*SE*)	US[1] β (*SE*)
T1 math. achievement (std.)	.48 *	.44 *	.54 *
	(.02)	(.02)	(.03)
T1 cognitive abilities (std.)	.18 *	.13 *	.12 *
	(.01)	(.02)	(.02)
Late assessment at T1 (*ref. early*)	.22 *	−.11 *	−.02
	(.03)	(.04)	(.03)
Interaction between late assessment at T1 and T1 math. achievement (std.)	.00	.01	.03
	(.02)	(.03)	(.03)
Age-in-months at T1 testing	.02 *	−.01	−.00
	(.00)	(.00)	(.00)
Constant	−1.66*	0.60*	0.32
	(.36)	(.30)	(.25)
R^2	.365	.255	.397
N	4,717	3,213	3,980

Notes. Results from linear regression models with clustered standard errors. $^{+}p < .10$; $^{*}p < .05$. Abbr. std.: z-standardised.
[1]Sample sizes 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.

University of BRISTOL

SES gradient in (biased) teacher assessment

Teacher judgement bias (mean residuals), by SES



	ENGLAND	GERMANY	US
	m	m	m
SES			
High	.16	.10	.01
Medium	.00	−.02	−.05
Low	−.13	−.24	.04

University of BRISTOL

STEP 2: Does T1 teacher assessment predict T2 achievement?

Results of regression models for T1 student mathematical achievement (z-standardised)

	England		Germany		US[1]	
	M1 β (*SE*)	M2 β (*SE*)	M1 β (*SE*)	M2 β (*SE*)	M1 β (*SE*)	M2 β (*SE*)
Highest parental education (*ref. medium*)						
High	.19 *	.13 *	.24 *	.22 *	.19 *	.18 *
	(.03)	(.02)	(.03)	(.03)	(.01)	(.01)
Low	−.10 *	−.05 *	−.28 *	−.24 *	−.12 *	−.13 *
	(.03)	(.03)	(.06)	(.06)	(.01)	(.00)
Teacher judgement residuals (std.)		.34 *		.17 *		.13 *
		(.01)		(.01)		(.00)
Controls	X	X	X	X	X	X
Constant	−1.85*	−.10	−1.63*	−1.70*	−2.28*	−2.24*
	(.27)	(.25)	(.34)	(.34)	(.33)	(.32)
R^2	.458	.567	.402	.428	.641	.657
N	4,717		3,213		3,980	

Notes. Results from linear regression models with clustered standard errors. $^{+}p < .10$; $^{*}p < .05$.
Abbr. std.: z-standardised.
Controls included T1 achievement; T1 cognitive abilities; time span between T1 & T2 testing; gender; immigration status.
[1]Sample sizes rounded to nearest 10, as required by the National Center for Education Statistics.

University of BRISTOL

DISCUSSION

1. We suspected that an existing growth mindset, as well as accountability, and ability grouping, lead to a lower teacher judgement bias.
2. We expected the bias to be particularly low in the US, followed by England. For Germany, in contrast, we expected a more pronounced teacher judgement bias due to a lower observable growth mindset, a lower degree of accountability, and missing ability grouping during primary education.
3. We expected stronger effects on later achievement in England and the US due to ability grouping, although standardised curricula might attenuate this effect in England.

CONFIRMED! Unexplained variance in teacher judgement was systematically linked to family SES in Germany and England (not in the US)

CONFIRMED! In all three countries, the inaccuracy in teacher judgment predicted student's later achievement (even considering prior achievement, cognitive abilities, socio-demographic controls)

ONLY IN ENGLAND & GERMANY (biased) teacher judgement (partially) mediated the effect of SES (i.e., the effect of SES decreased when controlling for biased judgements)

INSTITUTIONAL AND SOCIETAL SETTINGS, alongside SCHOOL POLICIES, MATTER!

University of BRISTOL

SENSITIVITY CHECKS

- *Heterogenous effects of biased teacher judgement*: (England, US) the association of biased teacher judgement with achievement was significantly weaker for high-SES students as compared to low-SES students.
- *Teacher change over the course of primary education* (Germany): results were very similar
- *Language skills:* largely comparable results. Although for Germany – less pronounced association between teacher judgment and later language skills

FURTHER RESEARCH

Mechanisms thought which (biased) teacher judgement affects later students' achievement

University of BRISTOL

University of BRISTOL

Thank you for you attention

valentina.perinetticasoni@bristol.ac.uk

University of BRISTOL

APPENDIX

A1:Unweighted descriptive statistics

		ENGLAND		GERMANY		UNITED STATES[1]	
		(N = 4,717)		(N = 3,213)		(N = 3,980)	
	time	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[2]	1	0	1
Late assessment at T1	T1	59.6		38.2		61.4	
Age-in-months at T1 testing	T1	86.75	2.91	84.92	4.68	85.65	4.37
Time span testing T2-T1 (in months)	T2-T1	48.46	1.96	32.03	1.50	48.10	1.08
HIGHEST PARENTAL EDUCATION	T1						
High		32.7		37.7		43.6	
Medium		27.4		51.9		27.9	
Low		39.9		10.5		28.5	
Female student	T1	50.2		51.5		49.4	
Immigration status	T1	19.3		23.1		30.8	

University of BRISTOL

A2: Complete Step2 regression model

Results of regression models for later student mathematical achievement (z-standardised)

	England		Germany		US[1]	
	M1 β (*SE*)	M2 β (*SE*)	M1 β (*SE*)	M2 β (*SE*)	M1 β (*SE*)	M2 β (*SE*)
Highest parental education (*ref. medium*)						
High	.19 *	.13 *	.24 *	.22 *	.19 *	.18 *
	(.03)	(.02)	(.03)	(.03)	(.01)	(.01)
Low	−.10 *	−.05 *	−.28 *	−.24 *	−.12 *	−.13 *
	(.03)	(.03)	(.06)	(.06)	(.01)	(.00)
Teacher judgement residuals (std.)		.34 *		.17 *		.13 *
		(.01)		(.01)		(.00)
T1 achievement (std.)	.49 *	.49 *	.47 *	.48 *	.69 *	.69 *
	(.01)	(.01)	(.02)	(.02)	(.00)	(.00)
Cognitive abilities (std.)	.23 *	.24 *	.20 *	.20 *	.10 *	.10 *
	(.02)	(.01)	(.02)	(.01)	(.00)	(.00)
Time span testing T2-T1 (in months)	.04 *	.00	.05 *	.05 *	.05 *	.05 *
	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)
Student female	−.10 *	−.07 *	−.07 *	−.02	−.07 *	−.07 *
	(.02)	(.02)	(.03)	(.03)	(.01)	(.01)
Immigration status	.21 *	.20 *	−.00	−.02	.12 *	.09 *
	(.03)	(.03)	(.04)	(.03)	(.01)	(.01)
Constant	−1.85*	−.10	−1.63*	−1.70*	−2.28*	−2.24*
	(.27)	(.25)	(.34)	(.34)	(.33)	(.32)
R^2	.458	.567	.402	.428	.641	.657
N	4,717		3,213		3,980	

Notes. Results from linear regression models with clustered standard errors. $^{+}p < .10$; $^{*}p < .05$. Abbr. std.: z-standardised.
[1]Sample sizes 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.

University of BRISTOL

A3: Statistical significance of the difference between M1 and M2

Is the SES gradient in T2 achievement at least partially due to (biased) teacher assessment?

Testing of significant changes between M1 and M2 in the effect of SES

	England		Germany		US[1]	
	Δb (SE)		Δb (SE)		Δb (SE)	
Highest parental education (*ref. medium*)						
High	-.06	***	-.02	**	-.01	*
	(.01)		(.01)		(.00)	
Low	.05	***	.04	***	-.00	
	(.01)		(.01)		(.00)	
N	4,717		3,213		3,980	

*$p < .05$; **$p < .01$; *** p<.001

Sources: Own calculations based on MCS, NEPS-SC2, and ECLS-K:2011.

University of BRISTOL

A3: Heterogenous effects of teacher judgement

Results of regression models for later student mathematical achievement (z-standardised) when considering heterogenous effects of biased teacher judgements

	England β (*SE*)	Germany β (*SE*)	US[1] β (*SE*)
Highest parental education (*ref. medium*)			
High	.14 *	.21 *	.18 *
	(.02)	(.03)	(.01)
Low	−.05	−.24 *	−.13 *
	(.03)	(.06)	(.00)
Teacher judgement residuals (std.)	.36 *	.17 *	.15 *
	(.02)	(.02)	(.02)
Interaction between parental education and teacher judgement residuals (std.)			
Residuals## high-educated	−.07 *	−.00	−.07 *
	(.02)	(.03)	(.02)
Residuals## Low-educated	.01	−.01	.02
	(.02)	(.05)	(.02)
T1 achievement (std.)	.49 *	.48 *	.69 *
	(.01)	(.02)	(.00)
Cognitive abilities (std.)	.24 *	.20 *	.10 *
	(.01)	(.01)	(.00)
Time span testing T2-T1 (in months)	.00	.05 *	.05 *
	(.01)	(.01)	(.01)
Student female	−.07 *	−.02	−.07 *
	(.02)	(.03)	(.01)
Immigration status	.20 *	−.02	.09 *
	(.03)	(.04)	(.01)
Constant	−.10	−1.70*	−2.24*
	(.25)	(.33)	(.31)
R^2	.569	.428	.658
N	4,717	3,213	3,980

University of BRISTOL

A4:Unweighted descriptive statistics (language skills)

		ENGLAND		GERMANY		UNITED STATES[1]	
		(N = 4,717)		(N = 3,213)		(N = 3,980)	
	time	M/%	SD	M/%	SD	M/%	SD
Teacher assessment: language skills (std.)	T1	0	1	0	1	0	1
Language skills achievement (std.)	T2	0	1	0	1	0	1
Language skills achievement (std.)	T1	0	1	0	1	0	1
Language skills achievement, grammar	T1	n/a	n/a	0	1	n/a	n/a
Cognitive abilities (std.)	T1	0	1	0[2]	1	0	1
Late assessment at T1	T1	59.7		39.2		61.3	
Age-in-months at T1 testing	T1	86.75	2.90	85.0	4.66	85.62	4.40
Time span testing T2-T1 (in months)	T2-T1	48.47	1.96	20.11	1.48	48.09	1.08
HIGHEST PARENTAL EDUCATION	T1						
High		32.9		38.8		43.0	
Medium		27.4		51.8		29.5	
Low		39.6		9.4		27.5	
Female student	T1	50.5		51.4		49.4	
Immigration status	T1	19.3		22.3		30.8	

University of BRISTOL

A5: Is teacher assessment biased? (language skills)

Results of regression models for teacher judgement (z-standardised; language skills)

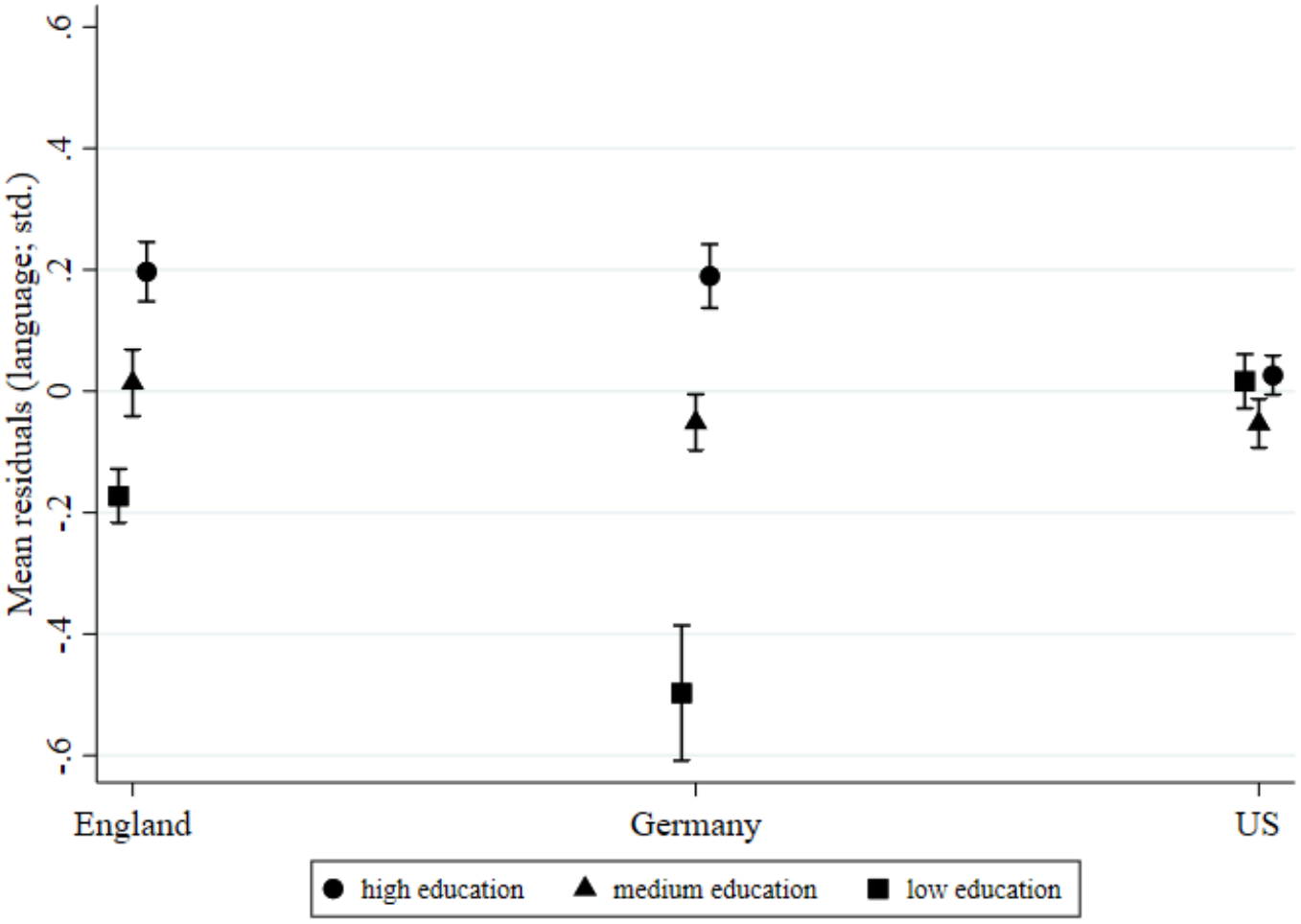
	England β (*SE*)	Germany β (*SE*)	US[1] β (*SE*)
T1 lang. achievement (std.)	.64 * (0.02)	.18 * (.03)	.68 * (.03)
T1 cognitive abilities (std.)	.16 * (0.01)	.14 * (.02)	.06 * (.01)
Late assessment at T1 (*ref. early*)	.15 * (0.02)	.00 (.04)	−.04 + (.02)
Interaction between late assessment at T1 and T1 lang. achievement (std.)	.03 (0.02)	.01 (.04)	.01 (.04)
Language ach., grammar (std.)	n.a.	.28 * (.03)	n.a.
Interaction between late assessment at T1 and T1 lang. achievement, grammar (std.)		.02 (.04)	
Age-in-months at T1 testing	.01 * (0.00)	−.01 * (.00)	−.00 (.00)
Constant	−.94* (.31)	1.10* (.29)	0.05 (.17)
R^2	.539	.230	.507
N	4,721	3,361	7,990

Notes. Results from linear regression models with clustered standard errors. $^{+}p < .10$; $^{*}p < .05$.
Abbr. std.: z-standardised. n.a.: not applicable.
[1]Sample sizes 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.

University of BRISTOL

A6: SES gradient in (biased) teacher assessment (language skills)

Teacher judgement bias (mean residuals), by SES



	ENGLAND	GERMANY	US
	m	m	m
SES			
High	.20	.19	.03
Medium	.01	−.05	−.05
Low	−.17	−.50	.02

University of BRISTOL

A7: Does T1 teacher assessment predict T2 achievement? (language skills)

Results of regression models for later student mathematical achievement (z-standardised)

	England		Germany		US[1]	
	M1 β (*SE*)	M2 β (*SE*)	M1 β (*SE*)	M2 β (*SE*)	M1 β (*SE*)	M2 β (*SE*)
Highest parental education (*ref. medium*)						
High	.25 *	.19 *	.13 *	.12 *	.23 *	.22 *
	(.03)	(.03)	(.03)	(.03)	(.01)	(.01)
Low	−.14 *	−.10 *	−.12 *	−.11 *	−.12 *	−.12 *
	(.03)	(.03)	(.04)	(.04)	(.01)	(.01)
Teacher judgement residuals (std.)		.25 *		.03 *		.12 *
		(.01)		(.01)		(.00)
Controls	X	X	X	X	X	X
Constant	−1.83*	−.43	−.37*	−.37*	−1.51*	−1.30*
	(.28)	(.28)	(.17)	(.17)	(.34)	(.37)
R^2	.418	.474	.593	.594	.565	.579
N	4,721		3,361		7,990	

Testing of significant changes of the parental education effect between M1 and M2 revealed:
England: high-educated: $\Delta b = -.05$, $SE = .01$, $p < .001$; low-educated: $\Delta b = .04$, $SE = .01$, $p < .001$;
Germany: high-educated: $\Delta b = -.01$, $SE = .00$, $p = .010$; low-educated: $\Delta b = .02$, $SE = .01$, $p = .007$; US: high-educated: $\Delta b = -.01$, $SE = .00$, $p < .001$; low-educated: $\Delta b = .00$, $SE = .00$, $p = .516$.

University of BRISTOL

A8: Ability grouping in England

Results of regression models for later student achievement (z-standardised) considering within-class ability grouping at T1 (England only)

	Mathematics			Language skills		
	M1a β (*SE*)	M2a β (*SE*)	M3 β (*SE*)	M1a β (*SE*)	M2a β (*SE*)	M3 β (*SE*)
Highest parental education (*ref. medium*)						
High	.18 * (.02)	.14 * (.02)	.14 * (.02)	.24 * (.03)	.20 * (.03)	.20 * (.03)
Low	−.07 * (.03)	−.05 * (.03)	−.05 * (.05)	−.14 * (.03)	−.10 * (.03)	−.10 * (.03)
Teacher judgement residuals (std.)		.25 * (.01)	.25 * (.04)		.22 * (.01)	.22 * (.05)
Ability grouping in numeracy (*ref.: no group*)[1]						
Bottom	−.72 * (.06)	−.43 * (.06)	−.42 * (.06)			
Middle	−.11 * (.05)	−.06 (.05)	−.05 (.05)			
Top	.34 * (.05)	.16 * (.05)	.18 * (.05)			
Missing	.03 (.08)	.02 (.07)	.03 (.07)			

University of BRISTOL

A8: Ability grouping in England

Interaction between residuals and ability grouping			
Residuals#	.01		
#Bottom	(.05)		
Residuals#	.02		
#Middle	(.05)		
Residuals#	−.04		
#Top	(.05)		
Residuals#	.07		
#Missing	(.07)		
Ability grouping in literacy (*ref.: no group*)[1]			
Bottom		−.32 *	−.12	−.11
		(.06)	(.06)	(.07)
Middle		.01	.02	.02
		(.06)	(.05)	(.05)
Top		.31 *	.12 *	.13 *
		(.06)	(.05)	(.05)
Missing		.11	.09	.09
		(.09)	(.09)	(.09)
Interaction between residuals and ability grouping				
Residuals#				.00
#Bottom				(.06)

University of BRISTOL

A8: Ability grouping in England

	Mathematics			Language skills		
	M1a β (*SE*)	M2a β (*SE*)	M3 β (*SE*)	M1a β (*SE*)	M2a β (*SE*)	M3 β (*SE*)
Residuals# #Middle						.03 (.05)
Residuals# #Top						−.02 (.05)
Residuals# #Missing						−.01 (.09)
T1 achievement	.33 * (.02)	.41 * (.01)	.41 * (.01)	.38 * (.02)	.46 * (.02)	.46 * (.02)
Cognitive abilities	.18 * (.01)	.20 * (.01)	.21 * (.01)	.11 * (.01)	.14 * (.01)	.14 * (.01)
Time span testing T2-T1 (in months)	.01 (.01)	−.00 (.01)	−.00 (.01)	.01 * (.01)	.00 (.01)	.00 (.01)
Student female	−.11 * (.02)	−.09 * (.02)	−.09 * (.02)	.17 * (.02)	.13 * (.02)	.13 * (.02)
Immigration status	.16 * (.03)	.18 * (.03)	.18 * (.03)	−.06 (.03)	−.01 (.03)	−.01 * (.03)
Constant	−.37 (.26)	.22 (.25)	.24 (.25)	−.87* (.29)	−.27 (.28)	−.24 (.28)
R^2	.55	.59	.59	0.45	0.48	0.48
N		4,717			4,721	

Notes. Results from linear regression models with clustered standard errors. $^{+}p < .10$; $^{*}p < .05$.

University of BRISTOL

REFERENCES

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

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

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

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

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., 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*, 102599. https://doi.org/10.1016/j.ssresearch.2021.102599

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

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). Academic Press. 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. *Personality and Social Psychology Review*, *9*(2), 131–155. https://doi.org/10.1207/s15327957pspr0902_3

University of BRISTOL

REFERENCES

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.

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

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

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

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

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

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

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

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

University of BRISTOL

REFERENCES

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. *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

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

University of BRISTOL