UNIVERSITÀ
DEGLI STUDI
FIRENZE

DISIA
DIPARTIMENTO DI STATISTICA,
INFORMATICA, APPLICAZIONI
“GIUSEPPE PARENTI”

‘An Elephant in the Classroom’: Teacher Bias by Student SES or Ability Measurement Bias?

Carlos J. Gil Hernandez,
Mar C. Espadafor

DISIA WORKING PAPER
2024/05

‘An Elephant in the Classroom’:

Teacher Bias by Student SES or Ability Measurement Bias?

Carlos J. Gil-Hernández[1*]
Mar C. Espadafor[2]

[1] University of Florence, Department of Statistics, Informatics, Applications

[2] University of Turku, Invest, Sociology

* Corresponding author: carlos.gil@unifi.it

Abstract: Teachers are academic merit gatekeepers. Yet their potential role in reproducing inequality via assessments was overlooked or not correctly identified, being ‘an elephant in the classroom’. This article teases if teacher grades and track recommendations are biased by student SES or unobserved ability, leading to overestimation in prior research. Using the German NEPS panel across elementary education, we identify student ability with multiple cognitive and noncognitive composite measures and an instrumental variable design. We further assess heterogeneity along the ability distribution to test whether, according to the compensatory hypothesis, teacher bias is largest among low-performers. First, accounting for measurement error, teacher bias declines by 40%, indicating substantial overestimation in previous studies. Second, it concentrates on underperformers, suggesting high-SES parental compensatory strategies to boost teacher assessments. Thus, families and teachers might influence each other in the evaluation process. We discuss the findings’ theoretical and methodological implications for teacher bias as an educational reproduction mechanism.

Keywords: Teacher assessments; teacher bias and discrimination; class inequality; educational transitions; tracking recommendations; standardized testing; grades; longitudinal studies of education.

Funding: This research has been funded through the JRC Centre for Advanced Studies and the project Social Classes in the Digital Age (DIGCLASS).

1. Introduction

Student's skills and socioeconomic status (SES) background are among the strongest predictors of educational attainment (Jackson 2013). Yet students from low-SES families systematically attain less education than equally skilled but better-off peers (Gil-Hernández 2019). An unusual suspect to partially explain these—net-of-ability—SES gaps is the observed divergence between teacher grades and student performance in externally assessed standardized tests—an objective ability proxy (Südkamp, Kaiser, and Möller 2012). This regularity raises concerns about the role of teacher bias in student assessments and its implications for educational inequality, as argued long ago by classical *Cultural Reproduction Theories* (CRT) (Bourdieu and Passeron 1990).

Lately, teachers have received relatively less attention (Jennings and DiPrete 2010) than the well-documented role of schools (Downey and Condron 2016) and families in shaping educational inequalities (Blossfeld et al. 2016). *Rational Action Theories* (RAT) focus on family choice mechanisms (Breen and Goldthorpe 1997), disregarding teachers as an *elephant in the classroom*. Educational systems reward a set of cognitive and noncognitive skills that teachers transform into grades as the evaluators of academic merit they are (Farkas 2003; Bowles and Gintis 2002). In their judgment of student ability, teachers are subject to implicit (Alesina et al. 2018) and explicit biases (Homuth, Thielemann and Wenz 2023), which might lead to self-fulfilling prophecies (Carlana 2019).

Previous observational studies indicate remarkable variation between student rank in blindly assessed test scores and teacher grades (Borghans et al. 2016; Südkamp et al. 2012) or track recommendations (Batrucht et al. 2023; Timmermans et al. 2018). This residual discrepancy, interpreted as teacher bias effects, tends to benefit students with certain ascribed characteristics, such as girls (Lievore and Triventi 2023), native-origin (Lorenz et al. 2024), highbrow cultural capital (Jæger 2022), and high-SES (Gortázar, Martínez de Lafuente and Vega-Bayo 2022; Boone and Van Houtte 2013). Still, the latter findings on student SES-based discrimination are more mixed and less abundant, particularly regarding grading outcomes (Zanga and De Gioannis 2023; Wenz and Hoenig 2020), calling for further evidence we provide here.

Beyond the previous focus on ethnic or gender discrimination (Zanga and De Gioannis 2023), this article is chiefly motivated by three critical methodological limitations of most observational studies: (1) omitted variable bias, (2) measurement error bias, and (3) heterogeneous effects. These make it challenging to tease whether teacher bias is a substantial educational inequality mechanism or a statistical artifact of measurement bias (van Huizen, Jacobs and Oosterveen 2024). Thus, we consider measurement bias an *elephant in the classroom* in teacher bias research that most previous observational studies disregarded (Zanga and De Gioannis 2023).

First, regarding omitted variable bias, most preceding studies overlooked noncognitive skills, while students' classroom behavior is critical for teacher grading practices beyond cognitive competencies (Ferman and Fontes 2023). Noncognitive skills are positively correlated with parental SES (Holtmann, Menze, and Solga 2021), thus leading to teacher bias overestimation. Besides, in low-stakes testing settings, as in most prior investigations measuring performance, students, particularly from low-SES backgrounds, might not exert maximum effort and conceal their true ability (Radl et al. 2024). This, again, implies teacher bias overestimation. Hence, controlling for noncognitive skills is crucial when assessing teacher bias—as the (residual) difference between grades and test scores, especially if stakes are low.

Second, random and systematic measurement error might attenuate the effect of test scores, underestimating true student ability and overestimating parental SES effects, as they positively correlate (van Huizen, Jacobs and Oosterveen 2024). Similarly, noncognitive skills are commonly measured through imprecise self-reports by students or parents (Smithers et al. 2018). There is a large gap between saying and doing when it comes to students exerting effort (Apascaritei, Demel and Radl 2021), while parental reports could misrepresent student classroom behavior. Alternatively, teacher reports might accurately assess student behavior relevant to grading practices. Still, they might also be prejudiced by student SES by, for instance, under-rewarding effort from low-SES students, leading to teacher bias underestimation. Besides, a student's objective ability is usually captured with single test scores or self-reported behavioral measures, particularly subject to measurement error, as the test-retest psychometrics literature on reliability underscores (Lockwood and McCaffrey 2014). Teachers, instead, tend to evaluate student academic trajectories, not just performance snapshots.

Third, even when precisely measuring ability, teacher bias might be heterogeneous across the distribution and hidden behind average effects. According to the *compensatory advantage mechanism* (CAM) (Bernardi 2014), well-off families deploy strategies to prevent their kids' social demotion (Breen and Goldthorpe 1997), particularly if they underperform. Advantaged families might then show off their high expectations (Bernardi and Valdés 2021) while pushing teachers for higher evaluations (Barg 2012). Hence, teacher bias might be heterogeneous across the ability distribution, concentrated at the bottom-medium and absent at the top (Bernardi and Cebolla 2014). Yet, no study has tested whether teachers are actors in the scene of the CAM.

We ask two research questions: (1) Is there a residual effect of parental SES on teacher-assigned grades and track recommendations, net of student test scores? If so, (2) is it explained away by omitted variables and measurement error bias? If not, (3) is teacher bias concentrated among low-performing students?

In answering these questions, we contribute two-fold to the literature on educational inequality. First, to identify student ability, we build composite indexes relying on multiple measures of cognitive competencies (11 externally assessed standardized tests on language and math) and noncognitive skills (20 items on effort and attention skills) reported by teachers and parents. Besides, we exploit multilevel and panel data to implement school fixed effects that minimize unobserved confounding and an instrumental variable (IV) design leveraging random variation in test scores to correct measurement error. We run more than 100 regressions with multiple measurement and model specifications that enhance the consistency of our findings. Second, we explore teacher bias heterogeneity across the ability distribution to test whether teachers *compensate* for a high-SES pupil's low performance by over-grading or recommending attendance to academic school tracks. This way, we analyze the role of teachers in reproducing educational inequality as academic merit gatekeepers in two central assessments for pupil educational careers: grade point average (GPA) and track recommendations.

We draw data from the German *National Educational Panel Study* (NEPS), analyzing a student cohort during elementary school just before transitioning into secondary education. In the German system of early school tracking, the eventual track decision is, in principle, based on student ability and teacher recommendations (Esser 2016). This institutional setting thus represents a stringent test for identifying teacher bias and compensatory mechanisms compared to educational systems without early tracking (Blossfeld et al. 2016), where actors might have more agency.

2. Previous Findings and Theoretical Background

2.1. Previous Findings

Previous scoping reviews on teacher bias generally show that high-SES students tend to get higher grades (Zanga and De Gioannis 2023) and track recommendations (Batruch et al. 2023) when compared to equally-performing low-SES counterparts. Regarding grading bias, only 2 out of 37 reviewed studies by Zanga and Gioannis (2023:4) analyzed SES. Its coverage is larger in the case of track recommendations, with 19 out of 27 studies reviewed by Batruch et al. (2023:4) reporting teacher bias findings by SES. Among these reviewed studies, the vast majority applied an observational research design relying on (low-stakes) standardized test scores as an objective measure of student true ability.

A large meta-analysis of the studies (n=75) examining the association between teacher grades and standardized test scores reported a moderate mean effect size at $r = 0.63$ (Südkamp et al. 2012). This substantial unexplained variance suggests that accounting for test scores alone might not accurately reflect students' true ability and potential to succeed in school. In turn, measurement error in test scores might substantially inflate most previously reported teacher bias estimates. In the case of track recommendations in the Netherlands, van Huizen et al. (2024) used an instrumental variable approach minimizing measurement error bias by exploiting random variation in student-lagged test scores. They document an overestimation of the SES coefficient from 35 to 43%, compared with a model controlling for a single test score (van Huizen et al. 2024:20).

Following this rationale, most previous observational estimates of teacher bias by student SES, ranging from 10 to 20% of an SD unit for grading (Gortázar et al. 2022), might be overestimated by a similar factor. Thus, these figures likely represent an upper-bound benchmark of the unknown true effect we attempt to approximate here.

2.2. Theoretical Background and Mechanisms

This article focuses not on identifying mechanisms causing or mediating the total effect of parental SES on teacher (biased) assessments but on precisely documenting the phenomena. Still, we can get closer to the potential mechanisms at play by doing just that. This section briefly summarises different theoretical perspectives on how teachers, as institutional gatekeepers, might generate—net-of-ability—SES gaps in educational outcomes.

Psychological *Implicit Bias Theories* suggest that individuals automatically associate certain ascribed groups with negative traits (Fazio et al. 2023; Greenwald and Krieger 2006), leading to discriminatory teacher assessments toward low-SES students (Pit–ten Cate and Glock 2019). Similarly, sociological *Status Characteristics Theories* (Melamed et al. 2019) indicate that unconscious competence beliefs internalized in socialization result in differential performance expectations and biased evaluations by status groups (Ridgeway 2014). Accordingly, low-SES students face stricter scrutiny and must outperform high-SES peers to be considered equally competent by teachers (Foschi 2000).

Statistical Discrimination Theories (Arrow 1998) propose instead that, without complete information on students' true ability, teachers rely on group-level characteristics (e.g., average historical, educational outcomes by SES groups) to gauge student potential, leading to biased assessments. These would become fairer as teachers get new input on the individual student (Botelho, Maderia and Rangel 2015; Hanna and Linden 2012). Thus, statistical discrimination might be particularly salient for uncertain long-term outcomes, such as educational expectations or tracking recommendations (Batruch et al. 2023).

In turn, *Cultural Reproduction Theories* (CRT) (Bourdieu and Passeron 1990) argue that teachers favor students socialized in the dominant culture who display high-status cultural signals (Breinholt and Jæger 2019)—typically from high-SES families—by misconceiving these for academic brilliance. Critiques consider that CRT (1) overstate the role of cultural capital due to its endogeneity with (unobserved) cognitive and noncognitive skills rewarded in educational systems that antecedent and largely confound its effect (Jæger 2011; Farkas 2003); (2) they do not precisely identify mechanisms (Jaeger and Breen 2016:1108); and (3) are deterministic, without room for individual choice.

Rational Action Theories (RAT) unravel persistent educational inequalities into *primary* and *secondary effects* (Jackson 2013). *Primary effects* (ability) denote the association between parental SES and children's academic performance (GPA). *Secondary effects* (choice) account for upper-class children's advantage in educational transitions over and above performance due to class-based resource differentials, aversion to social demotion (Breen and Goldthorpe 1997) and success expectations (Barone, Triventi and Assirelli 2018). Most previous research applying the RAT framework measured academic performance through teacher-assigned GPA (Jackson 2013) instead of more *neutral* indicators like externally assessed test scores. Using GPA as the leading ability indicator rules out teacher grading bias as an educational inequality mechanism, as CRT argue. Teachers might also fuel *primary effects* if their assigned GPAs are biased by student SES beyond objective competence (Esser 2016). In turn, *secondary effects* can also generate or reinforce pre-existing teacher biases. Thanks to their cultural resources, high-SES parents effectively navigate the school system (Laureau 2015) by participating in councils (Forster and van de Werfhorst 2020), supporting and monitoring their kids. In teacher meetings, parents might push to inflate competence expectations (Barg 2012). Simultaneously, teachers might contribute to unequal educational choices—termed *tertiary effects* (Esser 2016)—by expressing higher expectations or recommendations for advantaged students than equally performing disadvantaged peers.

Drawing from RAT, the *Compensatory Advantage Mechanism* (CAM) argues that high-SES parents hold *sticky* educational expectations to reproduce their status (Bernardi and Valdés 2021). Social demotion risk peaks among underperforming students, so high-SES families might further reinforce the abovementioned inequality mechanisms. Teachers may perceive low- or average-performing, well-off kids will likely succeed in academic pathways and over-assess them. Contrastingly, low-SES families are more sensitive to performance signals (Holm, Hjorth-Trolle, and Jæger 2019). Thus, their expectations might fall around GPA cut-offs granting access to academic tracks, where information on potential success is particularly unclear (Bernardi and Cebolla 2014). This process might then influence teacher under-assessments of low-SES students. Simultaneously, teachers can impact parental expectations by providing (distorted) ability signals. In sum, the CAM expects educational inequalities to be most prominent among low-average performing students due to parental compensatory strategies (Bernardi 2014). Here, we further argue that teachers might be protagonist actors in addition to families.

Lacking experimental designs, the weight of these theories cannot be disentangled. Still, we argue that multiple actors (parents, students, and teachers) might influence each other in this process. As a bottom line, all theories and mechanisms reviewed expect teacher bias in assessments favoring high-SES students and, according to the CAM, particularly among low performers. All discrimination theories require controlling for true student ability to identify teacher bias. Independently of mechanisms, when teachers assign higher grades and track recommendations based on student SES background rather than competence, they contribute to the reproduction of educational inequality.

3. The German Context

Among most German Federal States, students are tracked from the final year of primary education, usually at age 10 (grade 4). Primary education emphasizes a standardized curriculum focused on mathematics and German, without ability grouping, overseen by a single teacher. Formal grading begins in the third grade with report cards covering academic subjects and classroom behavior. Following primary grade 4, students typically choose from three secondary school paths: lower secondary (*hauptschule*), middle secondary (*realshule*), or upper secondary (*gymnasium*) schools, with the latter offering a rigorous academic trajectory aimed at college enrollment.

In the final primary year (grade 4), core subject teachers recommend suitable secondary school options to families. These are usually given by the end of the first semester (February) or the academic year (June). Teachers base recommendations on a student's learning aptitude, psychological development, academic performance, and work ethic, primarily considering end-of-year grades in mathematics and German. Some states set official GPA thresholds for *gymnasium* recommendations below 2.5 to 2.0 on a (reserved) 6-to-1 scale. According to the intended teacher recommendation, the pupil's end-of-year GPA might be pushed below or above these thresholds. Formal recommendations may be discussed in a meeting. In conflict with parental preferences, the final decision typically lies with the parents or, in some binding states, with the school or supervisory authority. In practice, perceived likelihood of success and potential parental support are crucial factors in teachers' decision-making over and above academic performance (Ashwill 1999).

4. Data, Variables and Methods

4.1. Data

Data comes from the Starting Cohort 2 (SC2) of the *National Educational Panel Study* (NEPS) (NEPS Network 2022; Blossfeld and Roßbach 2019), focusing on the augmentation sample of students entering primary education in 2012/2013, followed up (waves 3-6) during the entire cycle (grades 1-4) until secondary education (wave 7). The analyses use data from wave 3 (97.4% participation rate) to waves 6 (86% participation rate) and 7 (56% participation rate). Using a two-stage approach, the sample was drawn based on a nationwide representative sample of students at elementary schools, including school, teacher (e.g., questionnaire on students), student and parents-level information. We restricted the sample to students with information on parental SES, competence tests and noncognitive skills (reported by teachers and parents) in the last grade (snapshots) and during primary education (composites) and on our two outcomes. Therefore, the sample size varies across outcomes, ranging from 2,152 (310) for GPA to 2,448 (300) students (schools) for track recommendations. Appendix Table A.6. summarizes the shares of missing values for each variable from the initial panel sample at wave 3 (grade 1). Table 2 displays the summary statistics of all variables.

4.2. Variables

Parental SES. SES is measured with the highest parental *International Socioeconomic Index of Occupational Status* (ISEI-08) measured in wave 3 (Ganzeboom and Treiman 1996). Heterogeneity analyses codify it into a dummy by low and high SES (0=q1-q2, and 1=q3-q4). We replicate the main analyses (see Appendix's Table A.4.) relying on the highest parental education (ISCED-97) measured in years and recorded in wave 3, ranging between 9 and 18 (SD=2.25) (see Appendix's Table A.7.).

Table 1. Variables by wave, grade and age

Variables	**Wave 3**	**Wave 4**	**Wave 5**	**Wave 6**
	Grade 1	**Grade 2**	**Grade 3**	**Grade 4**
	2012/2013	2013/2014	2014/2015	2015/2016
	Age 6-7	Age 7-8	Age 8-9	Age 9-10
Socio-demographics				
School	X			
Parental SES	X			
Migration background	X			
Gender	X			
Age	X			
Noncognitive Skills				
Concentration / Persistence (teachers)	**X**	**X**	**X**	**X**
Readiness for Exertion (parents)	X	X	X	X
Test Scores				
Math	X	X		X
German	X	X	X	X
Vocabulary	X		X	
Grammar	X			
Orthography				X X
Reading		X X		X
Outcomes				
Annual GPA (teachers)				**X**[a]
*Annual German grade (*teachers)				**X**[a]
Annual math grade (teachers)				**X**[a]
School track recommendation (teachers)				**X**

Notes: [a] Retrospective self-report of last year annual grade in Wave 7. In bold=assessed by teachers. Readiness for exertion (1-4; 4 items): (i) *Child works carefully with the work materials*, (ii) *Child makes an effort when assignments are difficult*, (iii) *Child gives up easily if something is difficult*, (iv) *Child works diligently in class*. Concentration/persistence (1-5; 1 item): (i) *Persistence and ability to concentrate (e.g. remaining occupied with something for a longer period of time) [compared with other children of the same age]*.

Socio-demographic Controls. All models control for time-constant socio-demographics from wave 3: age in months, gender (1 if female, 0 otherwise), and migration background (0=native origin; 1=first- and second-generation migrant origin).

Test Scores Snapshots. Domain-specific cognitive skills are measured with low-stakes competence test scores on math and language (reading literacy and orthography [two tests]) administered by external evaluators in grade 4 in wave 6.[1] Test scores follow *Item Response Theory*, provided as weighted maximum likelihood estimates (WLEs) or sum scores. To construct a snapshot (grade 4) competence measure averaging math and German literacy, (1) we standardized these domain-specific measures within wave 6 to express students' relative position in the age-specific distribution, and (2) applied factor analysis to estimate the weighted mean z-scores across German (3 domains) and math (1 domain) domains (only one factor retained with Eigenvalue at 2.9 and α=0.87). For the supplementary analysis predicting math and German grades independently, we calculated a weighted standardized average of the former German literacy items (only one factor retained with Eigenvalue at 2.4 and α=0.87). Competence tests in grade 4 were administered between October 2015 and February 2016, before teachers issued track recommendations (February-June 2016) and annual GPA (June 2016).

Noncognitive Skills Snapshots. Noncognitive skills snapshots (wave 6 grade 4) are captured with two measures of student effort reported by teachers (main analysis) and parents (robustness check): (1) teachers' ratings of *concentration and persistence ability* (1-5 scale; 1 item) compared with children of the same age; (2) parental ratings of *readiness for exertion* (1-4; 4 items). See Table 1 for details. Among many other available noncognitive skills indicators, we picked these due to their availability in all survey waves for teachers and parents to build composite indicators (see below) by multiple raters. Teacher questionnaires were administered between October 2015 and January 2016, reporting noncognitive skills before the outcomes. Teacher's reports of pupil *concentration and* *persistence ability* are strongly correlated with conscientiousness (*r* = 0.67; teacher report in wave 5 grade 3), the Big Five personality trait most strongly related to educational performance (Poropat 2009). Besides, it is the noncognitive measure most associated with test score performance in NEPS (*r* = 0.47; wave 6 grade 4). As pointed out above, parental reports could misrepresent students' classroom behavior, as their lower correlation with test score performance (*r* = 0.28; wave 6 grade 4) compared with teacher's reports documents. Parental questionnaires were administered between January and June 2016, partially overlapping with the outcomes. Teachers' reports might accurately assess student classroom behavior relevant to grading practices. Still, they might be biased by student SES. Thus, we replicate the analysis with parental reports (Appendix Table A.5.).

Test Scores and Noncognitive Skills Composites. Applying within-wave standardization and factor analysis, we build composite indexes relying on several measures of domain-specific cognitive competencies (11 test scores comprising the first factor with Eigenvalue at 5.6 and **α**=0.91: 3 tests on math and 8 tests on German literacy [reading, orthography, grammar and vocabulary]) and noncognitive skills reported by teachers (4 items on attention/persistence skills comprising the first factor with Eigenvalue at 2.9 and **α**=0.88) and parents (16 items on student effort comprising the first factor with Eigenvalue at 3 and **α**=0.89). Measures are taken across grades 1-to-4 of elementary education (see Table 1 for details).

Grade Point Average. In wave 7, students are asked for their last year's (2015/2016) annual school report in German and Mathematics in grade 4. In elementary school, GPA is provided on a 1-to-6 scale where 1=very good, 2=good, 3=satisfying, 4=sufficient, 5=inadequate, and 6=unsatisfactory. We averaged German and Math grades (**α** = 0.77) to get GPA at grade 4. We reversed the scale and transformed it into z-scores within the corresponding analytical sample. Analyses using each math and German grade separately (reserved scale transformed into z-scores) are provided in Appendix Table A.2.

Track Recommendation. In wave 6, students attending grade 4 in those federal states tracking students from grade 5 are recommended a school type by teachers (January-June 2016), as reported in the parental questionnaire. Recommendations might be a formal school letter and/or a consultation in a teacher-parent meeting. Track recommendations are recoded into a dummy (1=Gymnasium and 0=other types of schools), and those observations with no recommendation are dropped. Comprehensive schools (around 6%) integrating vocational and academic tracks across grades 5-to-10 are considered lower-rank and included in the denominator with the vocational tracks due to their lower ability and SES composition.

Table 2. Summary Statistics (GPA - Wave 7 Analytical Sample = 2,152)

Variable	Mean	SD	Min	Max
Socio-demographic				
Migrant Origin (W3)	0.20		0	1
Female (W3)	0.51		0	1
Age (W3; shown at W6)	9.75	0.37	8.08	11.17
Parental ISEI (W3)	61.64	18.36	11.56	88.96
Outcomes				
GPA (G4)	2.05	0.73	6	1
Z-GPA (G4)	0.00	1.00	-5.42	1.45
Academic Track Recommendation (G4)[a]	0.65		0	1
Test Scores				
Z-Mean Math/German Test Scores (G4)	0.00	1.00	-3.61	3.24
Z-Mean Math/German Test Scores (G1-G3)	0.00	1.00	-3.24	3.62
Z-Mean Math/German Test Scores (G1-G4)	0.00	1.00	-3.08	3.27
Noncognitive Skills				
Teacher Reports				
Concentration/Persistence (G4)	3.54	1.09	1	5
Z-Concentration/Persistence (G4)	0.00	1.00	-2.33	1.34
Concentration/Persistence (G1-G4)	3.47	0.96	1	5
Z-Concentration/Persistence (G1-G4)	0.00	1.00	-2.57	1.60

Notes: Adjusted by longitudinal weight (W3-W7). [a] Sample size = 2,448, weighted by longitudinal weight (W3-W6). W=Wave; G=Grade.

4.3. Methods

We analyze two outcomes, teacher-assigned GPA in math and German (A) and academic track recommendations (B) represented by A or B following the corresponding model number, with ordinal least squares (OLS), linear probability models (LPM), and two-stage least-squares (2SLS) estimators. Results using separate grades and test scores for German and math are shown in the Appendix (Table A.2.). To account for selective attrition bias, particularly pronounced in wave 7, all models are adjusted by design longitudinal weights provided by NEPS. Standard errors are robust to heteroskedasticity and clustered by schools. All models include socio-demographic controls (age, sex and migration origin)—to estimate the net effect of parental SES independently of other student-ascribed characteristics—and school-fixed effects (FE). School-FE exploits within-school variation to account for all attributes varying between schools and students within. School-FE controls for school-specific ability, SES and ethnic-origin composition, grading distributions and standards (Calsamiglia and Loviglio 2019), and regional heterogeneity. Small sample sizes prevent the implementation of classroom FE, but most school clusters only comprise 1 or 2 classrooms/teachers.

As summarized in Table 3, we implement four main model specifications by outcome attempting to account for measurement and omitted variable bias stepwise: (M1) cognitive snapshots (mean test scores in math and German); (M2) cognitive and noncognitive snapshots; (M3) cognitive and noncognitive composites; and (M4) instrumental variables (IV). In M1, the baseline, we only control for socio-demographic characteristics and a test score snapshot. M2 adds a snapshot of noncognitive ability to account for its critical role in teachers' assessments and its association with (low-stakes) test score performance. For models M1-M2 controlling for ability snapshots, we use measures from grade 4 or the last information available before our outcome of interest. In M3, we use composite measurements of cognitive and noncognitive skills to further approximate true ability, drawing information from all survey waves.

Table 3: Main Model Specifications and Robustness Checks (RC)

Models	Snapshots		Composites	IV
Variables	M1	M2	M3	M4
SOCIO-DEMOGRAPHICS	**Migration** **Gender** **Age** **School-FE**	**Migration** **Gender** **Age** **School-FE**	**Migration** **Gender** **Age** **School-FE**	**Migration** **Gender** **Age** **School-FE**
PARENTAL SES	**ISEI**	**ISEI**	**ISEI**	**ISEI**
RC1: Alternative SES	Years of Education	Years of Education	Years of Education	Years of Education
TEST SCORES	**Math/German (G4)**	**Math/German (G4)**	**Math/German (G1-G4)**	**IV Math/German:** **Instrumented: Math/German (G4)** **Instrument: Math/German (G1-G3)**
RC2: Nonlinearities	Math/German Tertiles (G4)	Math/German Tertiles (G4)	Math/German Tertiles (G1-G4)	IV Math/German: Instrumented: Math/German 3rd Order Polynomials (G4) Instrument: Math/German 3rd Order Polynomials (G1-G3)
RC3A: Math Grade Outcome & Alternative IVs	Math (G4)	Math (G4)	Math (G1, G2, G4)	IV Math: Instrumented: Math (G4) Instrument A: German (G4) *Instrument B: German (G1-G3)* *Instrument C: Math (G1-G3)* *Instrument D: $IV_A + IV_B + IV_C$*
RC3B: German Grade Outcome & Alternative IVs	German (G4)	German (G4)	German (G1-G4)	IV German: Instrumented: German (G4) Instrument A: Math (G4) *Instrument B: Math (G1-G3)* *Instrument C: German (G1-G3)* *Instrument D: $IV_A + IV_B + IV_C$*
NONCOGNITIVE SKILLS		**Concentration/ Persistence (G4: teacher report)**	**Concentration/ Persistence (G1-G4: teacher report)**	**Concentration/ Persistence (G1-G4: teacher report)**
RC2: Nonlinearities		Concentration/ Persistence Tertiles (G4: teacher report)	Concentration/ Persistence Tertiles (G1-G4: teacher report)	Concentration/ Persistence Tertiles (G1-G4: teacher report)
RC4: Parental Reports		Readiness for Exertion (G4: parental report)	Readiness for Exertion (G1-G4: parental report)	Readiness for Exertion (G1-G4: parental report)

Notes: G = Grade; Math/German = Mean performance in math and German test scores (z-scores); IV = Instrumental Variable; FE = Fixed Effects; RC = Robustness Check. **In bold:** main models' specifications. *In italics*: alternative IV approaches output not reported here (available upon request).

M4 adopts an IV design with 2SLS estimators, given that, according to the statistically significant (p-value < 0.000) endogeneity test (Chi-square > 26) for both outcomes, competence scores snapshots (grade 4) are endogenous for correlating with the disturbance. As detailed in Table 3, the IV approach exploits random variation in lagged test scores from the same subjects in the main analysis (mean test scores in math and German) (Zhu 2024) and (contemporaneous or lagged) test scores from different or same domains (supplementary analysis by math and German) (Botelho et al. 2015) to account for measurement error (Van Huizen et al. 2024).

A valid instrument must meet two critical assumptions of *relevance* and *exogeneity.* First, the instrument should strongly predict the explanatory variable. According to the high and statistically significant (p-value < 0.000) *F* statistic (*F* > 580) and *Kleibergen-Paap rk LM* underidentification test (Chi-square > 140) yielded in the first stage implementation for both outcomes (see bottom Table 4), the instruments are relevant, strongly predicting test score performance in grade 4 as the endogenous variable.

Second, the error in the instrument must be independent of the error term (i.e., snapshot test scores error). This assumption is untestable, but it is likely to hold when instrumenting current test score snapshots (mean math and German performance in grade 4) with lagged composite test scores from previous academic years (grades 1-3) (Zhu 2024). Test-specific events affecting performance (e.g., question selection, luck in guessing answers, test-day classroom temperature, student health, mood) should be random, not correlated yearly. Following the same logic, in the supplementary analysis (results upon request), we disaggregate the main GPA analysis by math and German grades using math and German test scores at grade 4 as endogenous ability measures, respectively, instrumented by the lagged composite test scores (grades 1-3) from the same subject (C).

According to the exclusion restriction criterion, the instrument should only indirectly influence the outcome throughout test scores. Teachers' perceptions of students' abilities might be affected by previous performance (even though standardized tests are administered and evaluated by external individuals and scores are not revealed to teachers) as they assess long-term learning progress, potentially influencing later assessments directly. To address this issue, in the supplementary analysis (Table A.2.), we analyze math and German grades independently by correspondingly instrumenting math or German test scores with performance in a different subject in grade 4 (A) (e.g., instrumenting for math scores using simultaneous German scores and vice versa). Yet, a potential limitation of this latter different-subject IV approach is that errors across domains measured contemporaneously might be correlated if, for instance, a student feels ill or is unmotivated during the testing window. Hence, we additionally instrument math and German ability using composite lagged test scores (grades 1-3) from a different subject (B), as well as a joint IV (D) including the three approaches (A: simultaneous snapshot from a different subject; B: lagged composite from a different subject; C: lagged composite from the same subject). This latter approach with multiple IVs, where we can run an overidentification test of all instruments, yields a non-statistically significant Hansen J statistic for both German and math models, supporting the joint null hypothesis on *exogeneity* or valid instruments uncorrelated with the error term.

When these two critical assumptions are met, the IV approach can correct random measurement error in test scores and omitted variable bias by exploiting the exogenous portion of joint variance between the instruments, test score snapshots and the outcomes. Nonetheless, systematic sources of measurement error in standardized test scores—these not directly evaluating school curricula, students not exerting maximum effort in low-stakes settings, or classroom behavior impacting grades beyond test performance—might still impede capturing the whole set of abilities teachers consider in grading and track recommendations. In our IV models, we additionally control for pre-treatment antecedent variables (grade 1 wave 3): student age, sex, migrant origin, and parental SES. We argue that the IV design assumptions would only hold when conditioning on a potential confounder like student noncognitive ability, as it might impact teacher assessments by student SES even after identifying latent ability in test scores. Yet behavioral traits should be stable and independent from previous performance in low-stakes competence tests. Hence, in all IV models, we control for the teacher-reported composite (grades 1-4) on pupils' persistence/concentration ability to account for omitted variable bias.

The general modelling intuition is that the SES gradient or teacher bias on grading and track recommendations should be progressively reduced as we account for more measurement error and better approximate latent ability from M1 to M4. If there remains a residual effect of family SES on teacher's assessments in M3 using ability composites and/or M4 using an IV design, it would be evidence of teacher bias effects. This way, we can assess to what extent teacher bias by student SES was overestimated due to measurement error and omitted variable bias compared to the standard baseline model M1 run in most previous investigations.

Finally, to test the CAM, relying on the control variables as in M3, we include an interaction term between the test scores composite (grades 1-4) and parental ISEI in M5 and the persistence/concentration ability composite (grades 1-4) and parental ISEI in M6. A negative interaction term in M5-M6 would be evidence supporting the CAM. Still, in models M5-M6, test score reliability might vary across the performance distribution and/or SES, potentially inflating teacher bias estimates among low-performing students and over-detecting the CAM. Thus, we additionally run IV specifications as in M4 with heterogeneous models by low (M7) and high (M8) parental ISEI subgroups to assess the CAM while accounting for measurement error bias. A considerably smaller effect size of the cognitive and/or noncognitive ability measure among high- (M8) vs low-SES (M7) students would align with the compensatory hypothesis.

5. Findings

Table 4 displays the main models for each outcome. Before controlling for test score snapshots in M1, the total effect of parental ISEI capturing inequalities in performance stands at β_{ISEI} = 0.012 for Z-GPA and β_{ISEI} = 0.08 for track recommendations. Once test score performance is controlled in M1, inequalities by parental background account for 42.7% and 56.5% (see Table 5, Panel C) of its total association with Z-GPA and track recommendations, respectively.

Next, we use M1 as the baseline to assess to what extent the residual parental ISEI coefficient, net of inequality in test score performance, proxying for teacher bias, might be overestimated. One could argue that teacher evaluations encompass more academic skills than children's performance in a single test score. Besides, students might perform differently in low-stakes test scores according to behavioral traits influenced by parental SES. Indeed, parental ISEI progressively reduces its relative size as we control for noncognitive skills snapshots in M2 ($\beta_{M2A\ ISEI}$ = -12.6%; $\beta_{M2B\ ISEI}$ = -15.7%) and composites in M3 ($\beta_{M3A\ ISEI}$ = -34.8%; $\beta_{M3B\ ISEI}$ = -30.6%), accounting for measurement error and omitted variable biases (see Table 5, Panel B).

Still, composite measures of test scores and noncognitive skills might not fully capture unobserved students' true ability, which might explain away the remaining residual effect of parental ISEI. M4 implements an IV approach to tackle this issue, which exploits random variation between lagged test score performance and the grade 4 snapshot, displaying the second stage. Concerning baseline M1, the residual coefficient of parental ISEI further diminishes by 43.8% for GPA ($\beta_{M4A\ ISEI}$) and 41.2% for track recommendations ($\beta_{M4B\ ISEI}$). Using a similar IV approach with Dutch data, van Huizen et al. (2024:20) reported an overestimation of teacher bias by student SES on track recommendations up to 43%, compared with a model including a test score snapshot. Figure 1 clearly illustrates this stepwise declining pattern, where the residual coefficient of parental ISEI goes from representing over 80% of its size in baseline M1 to below 60% in M4 following the IV identification strategy.

Table 4. Main OLS and 2SLS models on GPA (A) and main LPM and 2SLS models on Track recommendation (B)

	M1A	M2A	M3A	M4A	M1B	M2B	M3B	M4B
	Z-GPA (Grade 4)				Track Recommendation (Grade 4)			
Parental ISEI	0.0053***	0.0047**	0.0035*	0.0030*	0.0043***	0.0036***	0.0030***	0.0025***
	(0.0016)	(0.0015)	(0.0014)	(0.0014)	(0.0008)	(0.0008)	(0.0007)	(0.0007)
Z-Test Scores (G4)	0.594***	0.452***		0.625***	0.233***	0.152***		0.264***
	(0.032)	(0.034)		(0.043)	(0.013)	(0.015)		(0.025)
Z-Non-Cognitive Skills (G4)		0.274***				0.162***		
		(0.027)				(0.012)		
Z-Test Scores (G1-4)			0.460***				0.169***	
			(0.032)				(0.014)	
Z-Non-Cog. Skills (G1-4)			0.277***	0.195***			0.159***	0.110***
			(0.029)	(0.030)			(0.013)	(0.017)
School-FE	✓	✓	✓	✓	✓	✓	✓	✓
IV				✓				✓
Number of observations	2,152	2,152	2,152	2,152	2,448	2,448	2,448	2,448
Number of schools	310	310	310	310	300	300	300	300
R-squared	0.517	0.558	0.568	0.408	0.443	0.507	0.535	0.350
First-stage F Statistic				582.48***				738.68***

Notes: Robust standard errors clustered by schools in parentheses. All models control for migration background, gender and age in months.
G = Grade
Non-Cog. = Noncognitive
*** p<0.001, ** p<0.01, * p<0.05, + p<0.10

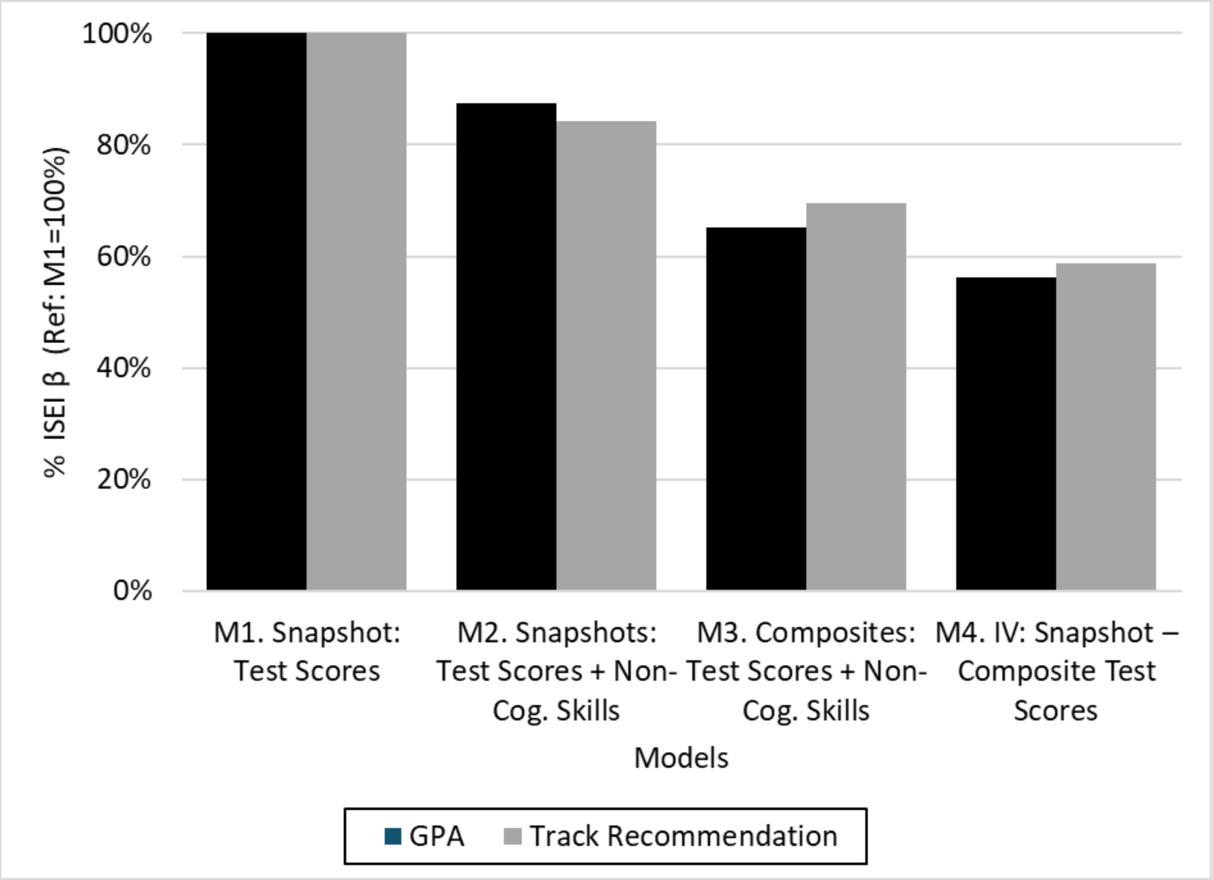
Table 5. ISEI effect size and reduction by model

	Z-GPA	Track Recommendation
A. % ISEI β (SD)[a]		
M1. Snapshot: Test Scores	9.8%	7.9%
M2. Snapshots: Test Scores + Noncognitive Skills	8.6%	6.7%
M3. Composites: Test Scores + Noncognitive Skills	6.4%	5.5%
M4. IV: Snapshot t_0 – Lagged Composite t_{-1} (Test Scores)	5.5%	4.6%
B. % Reduction ISEI β (Ref: M1)[b]		
M2. Snapshots: Test Scores + Noncognitive Skills	-12.6%	-15.7%
M3. Composites: Test Scores + Noncognitive Skills	-34.8%	-30.6%
M4. IV: Snapshot t_0 – Lagged Composite t_{-1} (Test Scores)	-43.8%	-41.2%
C. % *Tertiary Effects*, ISEI β (Ref: M0)[c]		
M1. Snapshot: Test Scores	42.7%	56.5%
M2. Snapshots: Test Scores + Noncognitive Skills	37.3%	47.6%
M3. Composites: Test Scores + Noncognitive Skills	27.9%	39.2%
M4. IV: Snapshot t_0 – Lagged Composite t_{-1} (Test Scores)	24.0%	33.2%

Notes: [a] % ISEI β_{GPA} (SD) = [($M_{1, 2, 3, 4}$ ISEI β x ISEI SD) / Outcome SD[d]] x 100; % ISEI $\beta_{Track\ Recom.}$ (SD) = ($M_{1, 2, 3, 4}$ ISEI β x ISEI SD[e]) x 100; ISEI βs from Table 4; [b] % Reduction ISEI β = [(M1 ISEI β – $M_{2, 3, 4}$ ISEI β) / M1 ISEI β];[c] M0 is a baseline model only including socio-demographic controls to estimate the total ISEI effect, unconditional on ability; [d] ISEI SD = 18.4; GPA SD = 1; [e] ISEI SD = 18.3.

Thus, one would have substantially overestimated teacher bias effects if only controlled for a test score snapshot, as in M1 and most previous observational studies (Batruch et al. 2023; Zanga and De Gioannis 2023). While previous observational teacher bias estimates by parental SES reported effect sizes up to 10-20% an SD, we identified values reducing in size at about 5% an SD for grading and tracking recommendation outcomes once correcting for measurement error and omitted variable bias. Yet, as we aim to approximate true student ability in the IV specifications, the residual effect of SES does not disappear, suggesting that teacher bias in assessments is not just a statistical artifact and might have an identifiable causal basis. However, its role as a foremost mechanism of educational reproduction might have been overstated. Overall, among equally skilled students, an SD increase in parental ISEI (about 20 ISEI points corresponding to, for instance, the difference between a CEO and an office clerk) leads to students getting, on average, 5.5% higher Z-GPA (0.73 points in a 6-to-1 scale) and 4.6% more recommendations to the academic track (average at 65%).

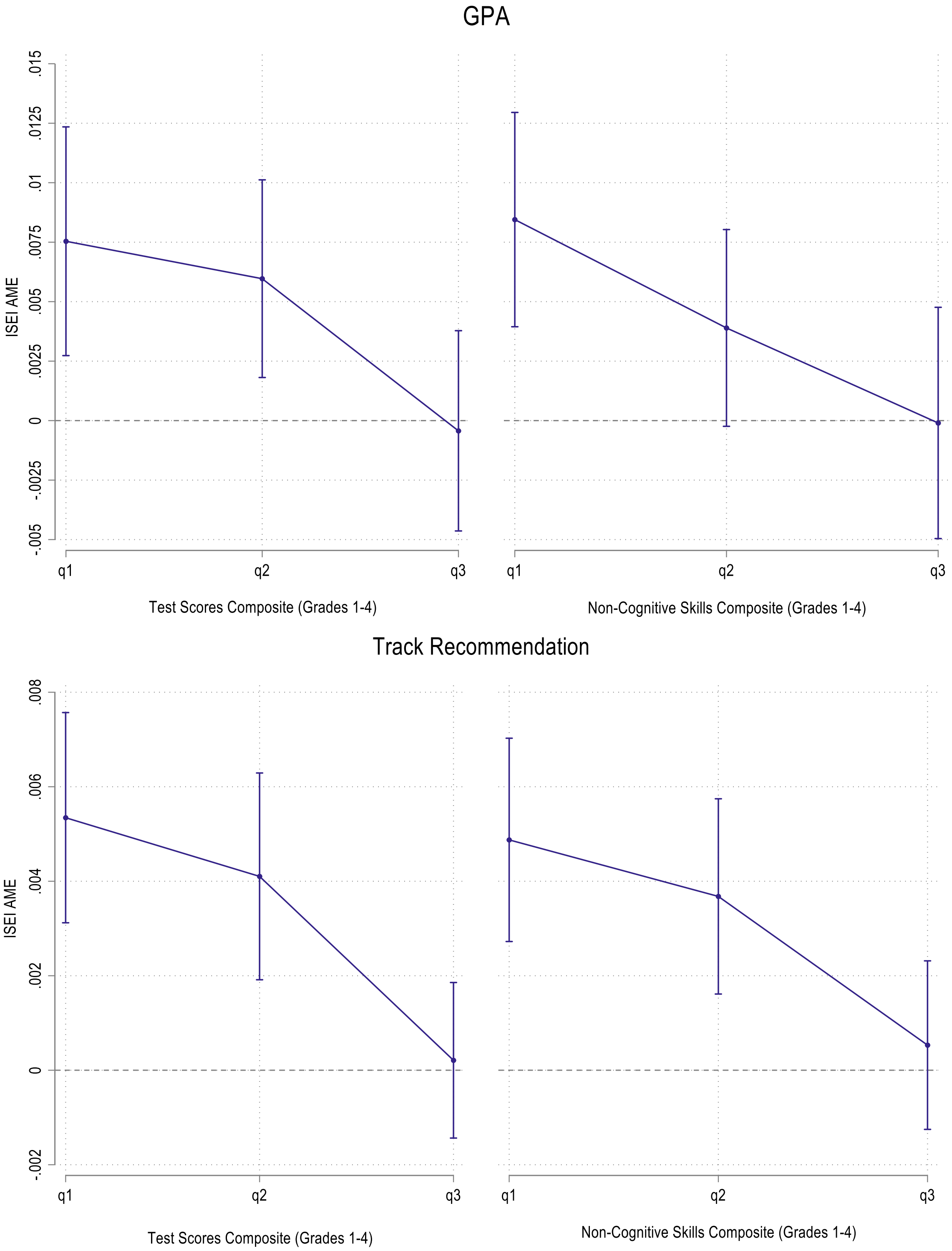
Figure 1. ISEI effect size by model in comparison with M1



Notes: % M1 ISEI β = [(M$_{1, 2, 3, 4}$ ISEI β / M1 ISEI β) x 100]; ISEI βs from Table 4; Non-Cog. = Noncognitive.

If we take IV estimates from M4 as the most reliable benchmark of student true academic ability, *primary effects* of social background—SES-based academic performance inequality—account for up to 67-76 % of the total association between parental ISEI and teacher grades or recommendations. Thus, as shown in Table 5 (Panel C), only the remaining residual share, ranging from 24% to 33% of the total association between parental ISEI and educational outcomes, could be causally attributed to teacher bias—the so-called *tertiary effects* of social background.

Figure 2. Average Marginal Effect (AME) of parental ISEI by test scores (left-panel) and noncognitive skills (right-panel) tertiles on Z-GPA (upper-panel; Table A.3., M5A-M6A; OLS) and track recommendations (bottom panel; Table A.3., M5B-M6B; LPM)



Notes: The figure portrays the AME of parental ISEI from Table A.3. M5A and M6A for GPA (upper-panel) and from M5B and M6B for track recommendations (bottom-panel), including an interaction term between parental ISEI and test scores (left-panels; M5A-B) or noncognitive skills (right panels; M6A-B), respectively. All models include school FE and control for children's gender, migration origin and age. Confidence intervals at the 95% level.

We finally test the CAM by examining whether identified teacher bias effects are heterogeneous across the student ability distribution. M6 and M7 include an interaction term between parental ISEI and student ability (see full output in Appendix Table A.3.), which is of negative sign and statistically significant (p-value < 0.05) as excepted by the CAM. Figure 2 illustrates the Average Marginal Effects (AME) of parental ISEI on each outcome by test scores (M5) and noncognitive skills composites (M6), categorized in tertiles to account for non-linearities. As shown, residual coefficients of parental ISEI, proxying for teacher bias, are most prominent among low- or average-ability students regarding test score performance and noncognitive skills for both outcomes. Instead, among high-performing students, there are null or zero teacher bias or parental ISEI effects on grading or track recommendations.

For clarity, Figure A.1. further displays the marginal effects of student ability tertiles by parental ISEI subgroups (-1SD=40 / +1SD=80) on the original absolute outcomes' scales (non-reversed [lowest] 6-to-1 [highest] scale for GPA). Remarkably, among low-medium-performing students in test scores and noncognitive skills, where the ISEI gap or teacher bias is largest, students from low ISEI families get a GPA just above the mean threshold for getting an academic track recommendation in most German regions (< 2.5 or < 2). Accordingly, as shown in the bottom panel of Figure A.1., inequalities by parental ISEI in academic track recommendations are concentrated among low-medium-performing students.

Still, one could argue that composite ability measures (M3) might conceal students' true ability, as the IV approach in M4 further reduces parental ISEI estimates by 10% compared to M3. To fully address this issue, we run IV specifications (see Appendix Table A.3.) as in M4 but heterogenous by low (q1-q2) (M7) and high (q3-q4) (M8) parental ISEI subgroups. As in M5-M6 above and expected by the CAM, the association between students' ability and educational outcomes is generally less steep for the high ISEI subgroup, except for noncognitive skills on track recommendations. In the concluding section, after summarizing several robustness checks, we discuss these findings' substantial theoretical and methodological implications for the role of teacher bias as an educational inequality mechanism.

5.1. Robustness Checks

We run several robustness checks to assess the credibility of the main findings. First, to account for potential non-linearities (see Appendix Table A.1.), we successfully replicate the main models with non-parametric specifications (tertiles) of test scores (M1-M3) and noncognitive skills (M1-M4) and up to third-order polynomials in the instrumental variables (M4). Second, we disaggregated the main GPA models (M1A-M4A) by subject-specific German and Math grades and test scores, using a simultaneous (grade 4) different-subject IV approach. Results generally mirror the main aggregate models on GPA. Still, results seem more robust for math grades, in line with previous literature identifying larger teacher bias effects for math than language subjects (Alesina et al. 2018). We implemented additional IV approaches, not shown here but available upon request, using lagged (grades 1-3) different-subject scores, lagged (grades 1-3) same-subject scores, and a joint IV including all three different instruments that replicate the simultaneous different-subject IV approach shown in Table A.2. Third, in Appendix Table A.4, we use an alternative measure of parental SES, the highest years of education, to confirm the primary models (M1-M4) findings with parental ISEI. Fifth, given that teachers' reports of students' behavior might contain bias, leading to potential underestimation of SES effects, in Appendix Table A.5., we replicate the primary models (M1-M4) using parental reports of children's noncognitive skills (readiness for exertion), yielding highly equivalent results.

6. Conclusion and Discussion

This article investigates teacher assessment biases by student SES as an educational inequality mechanism. It aims to discern whether student's class-based disparities in teacher grades and recommendations are influenced by actual bias or ability differences. We contribute by implementing methodological strategies to approximate children's true ability that bypass measurement error in test scores and omitted variable bias (e.g., unobserved noncognitive skills), a pervasive problem in most previous observational studies: (1) exploiting a rich longitudinal dataset across the German elementary education system comprising several cognitive and noncognitive ability measures; (2) implementing an instrumental variable approach. Besides, we explore effect heterogeneity across the ability distribution to test the *Compensatory Advantage Mechanism* (CAM).

We report two main findings whose substantial implications we discuss in turn. First, estimating the standard model only controlling for a (low stakes) test score snapshot, teachers evaluate more favourably high-SES students by assigning them around 10% SD higher GPA and recommending about 8% more enrolment in secondary academic schools than low-SES schoolmates. Once we account for measurement error in student ability by implementing composite cognitive and noncognitive measures across elementary education and an IV design, teacher bias estimates are reduced by over 40%, dropping high-SES student's advantage to about 6% SD in GPA and 5% more academic track recommendations. The reduction coefficient for track recommendations (41%) is virtually identical to the one identified van Huizen et al. (2024:20) at 43%, implementing a similar IV approach to predict teacher bias effects on track recommendations in the Netherlands. Thus, most previous teacher bias estimates only controlling for low-stakes test score snapshots might be seriously overestimated, reporting estimates up to 20% an SD (Zanga and De Gioannis 2023; Südkamp et al. 2012). To effectively identify teacher bias in observational data, it is recommended to analyze the residual differences between fully comparable high-stakes blind test scores and teacher-assigned grades that cover the same curricula while accounting for students' noncognitive skills (Ferman and Fontes 2022).

Second, teacher bias effects are concentrated among low-medium performers, just around the grade threshold for academic track recommendations in several German states. This suggests that high-SES parents might deploy compensatory strategies to boost teacher assessments, competence expectations, and/or teacher pervasive stereotypes among low-SES underperforming students. Students from low-

SES backgrounds have less risk aversion to social demotion and lower educational expectations than high-SES peers. Thus, the former might be more sensitive to distorting biases in teacher grades and recommendation signals (Holm et al. 2019), so depressing expectations. This distorting effect may be accentuated among underperforming low-SES students around a pass-or-fail threshold for educational transitions, where information on future success is particularly uncertain (Bernardi and Cebolla 2014). In opposition, high-SES families tend to display 'sticky' high educational expectations to reproduce their status, being inelastic towards low-performance signals (Bernardi and Valdés 2021). Thus, families and teachers might influence each other in the evaluation process.

Taken together, our findings align with previous observational studies documenting sizeable and persistent SES inequalities in educational transitions (Blossfeld et al. 2016) among students at the same level of test scores (Gil-Hernández 2021), GPA (Jackson 2013) or track recommendations. Likewise, the observed biases in track recommendations align with experimental evidence by Wenz and Hoenig (2020) among German elementary teachers, showing expectations of student attendance to the academic track to be unfair by SES. Overall, teacher bias effects are particularly striking for equal educational opportunity in the German educational system of early tracking, which is supposed to apply ability-driven sorting (Esser 2016).

Are teacher bias estimates substantial as an educational inequality mechanism? As illustrated in Table 5 above, we report unconditional (without ability controls) SES gaps in GPA and track recommendations for benchmarking. For instance, our observational teacher bias IV estimates account for 24% and 33% of the total SES gap in GPA and track recommendations, respectively. One can also benchmark our average grading bias effect size (0.06 SD) with school-year learning gains in literacy (0.15-0.2 SD) or educational interventions (0.17-0.47 SD) (Evans and Yuan 2019). These benchmarks indicate that, despite likely being overstated (van Huizen et al. 2024; Jæger 2022), teacher bias effects might be relevant for educational pathways when accumulating (dis)advantages over evaluations and critical transitions (DiPrete and Eirich 2006). Future studies might explore teacher bias by other ascribed characteristics, such as gender and migration background, and across regions to exploit educational legislation variation.

It is not clear-cut to infer discriminatory behavior and its causes from our observational findings, which cannot entirely rule out unobserved heterogeneity. Still, those teachers assessing high-SES students more favorably beyond true competence might act as engines of educational inequality reproduction. The relative weight of teacher's (e.g., implicit bias, status characteristics beliefs, statistical discrimination) and/or student-family (e.g., cultural capital, downward mobility aversion, sticky expectations) mechanisms explaining the black box of residual effects we identified here is an open question that future experimental studies should unpack. In this endeavour, *Cultural Reproduction* and *Rational Action* theories should be combined to better understand why educational inequalities persist.

Research Ethics Statement

This paper uses data from the National Educational Panel Study (NEPS; see Blossfeld & Roßbach, 2019). The NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi, Germany) in cooperation with a nationwide network. All human subjects gave informed consent before participating in the research, and adequate steps were taken to protect participants' confidentiality. The data can be accessed through the Scientific Use File of Starting Cohort Kindergarten upon request at https://doi.org/10.5157/NEPS:SC2:10.0.0. The authors will provide Stata code for recreating the analyzed subsample and replicating the results on request.

Endnotes

1. We assume that noncognitive skills are antecedent to performance in blindly evaluated and low-stakes test scores and that performance in test scores is not causally associated with later changes in baseline noncognitive skills.

References

Alesina, Alberto, Michela Carlana, Eliana La Ferrara, and Paolo Pinotti. 2018. "Revealing Stereotypes: Evidence from Immigrants in Schools." *NBER Working Paper* No. 25333.

Apascaritei, Paula, Simona Demel, and Jonas Radl. 2021. "The Difference Between Saying and Doing: Comparing Subjective and Objective Measures of Effort Among Fifth Graders." *American Behavioral Scientist* 65(11): 1457-1479. https://doi.org/10.1177/0002764221996772

Arrow, Kenneth J. 1998. "What Has Economics to Say about Racial Discrimination?" *Journal of Economic Perspectives* 12(2):91–100.

Ashwill, Mark A. 1999. The educational system in Germany: Case study findings. Washington, DC: National Institute on Student Achievement, Curriculum, and Assessment (ERIC Document Reproduction Service No. ED430906).

Barg, K. 2012. "The influence of students' social background and parental involvement on teachers' school track choices: reasons and consequences." *European Sociological Review 29*(3):565–579.

Barone, Carlo, Moris Triventi, and Giulia Assirelli. 2018. "Explaining Social Inequalities in Access to University: A Test of Rational Choice Mechanisms in Italy." *European Sociological Review*, 34(5):554–569.

Batruch, Anatolia, Sara Geven, Emma Kessenich, and Herman G. van de Werfhorst. 2023. "Are Tracking Recommendations Biased? A Review of Teachers' Role in the Creation of Inequalities in Tracking Decisions." *Teaching and Teacher Education* 123:103985.

Bernardi, Fabrizio. 2014. "Compensatory Advantage as a Mechanism of Educational Inequality: A Regression Discontinuity Based on Month of Birth." *Sociology of Education*, 87(2): 74–88.

Bernardi, Fabrizio, and Héctor-Cebolla Boado. 2014. "Previous School Results and Social Background: Compensation and Imperfect Information in Educational Transitions." *European Sociological Review* 30(2): 207–17.

Bernardi, Fabrizio, and Manuel T. Valdés. 2021. "Sticky Educational Expectations: A Cross-Country Comparison." *Research in Social Stratification and Mobility* 75: 100624. https://www.sciencedirect.com/science/article/pii/S0276562421000445

Blossfeld, Hans-Peter and Hans-Günther Roßbach (Eds.). 2019. *Education as a lifelong process: The German National Educational Panel Study* (NEPS). Edition ZfE (2nd ed.). Springer VS.

Blossfeld, Hans-Peter, Sandra Buchholz, Jan Skopek, and Moris Triventi. 2016. *Models of Secondary Education and Social Inequality: An International Comparison*. Edward Elgar Publishing.

Boone, Simon, and Mieke Van Houtte. 2013. "Why Are Teacher Recommendations at the Transition from Primary to Secondary Education Socially Biased? A Mixed-Methods Research." *British Journal of Sociology of Education* 34(1): 20–38. https://doi.org/10.1080/01425692.2012.704720 (September 20, 2023).

Borghans, Lex, Bart H. H. Golsteyn, James J. Heckman, and John Eric Humphries. 2016. What grades and achievement tests measure. *Proceedings of the National Academy of Sciences 113*(47):13354 LP – 13359.

Botelho, Fernanda, Ricardo A. Madeira, and Marcos A. Rangel. 2015. "Racial Discrimination in Grading: Evidence from Brazil." *American Economic Journal: Applied Economics* 7(4):37–52.

Bourdieu, Pierre, and Jean–Claude Passeron. 1990. *Reproduction in Education, Society, and Culture*. London: Sage.

Bowles, Samuel and Herbert Gintis. 2002. "Schooling in Capitalist America Revisited. *Sociology of Education*." 75:1-18

Breen, Richard, and John H. Goldthorpe. 1997. "Explaining Educational Differentials: Towards a Formal Rational Action Theory." *Rationality and Society* 9(3):275–305.

Breinholt, Asta, and Mads Meier Jæger. 2019. "How Does Cultural Capital Affect Educational Performance: Signals or Skills?" *The British Journal of Sociology* 71(1):28–46.

Calsamiglia, Caterina, and Annalisa Loviglio. 2019. 'Grading on a Curve: When Having Good Peers Is Not Good'. *Economics of Education Review* 73: 101916. http://www.sciencedirect.com/science/article/pii/S0272775718306174 (June 11, 2020).

Carlana, Michela. 2019. "Implicit Stereotypes: Evidence from Teachers' Gender Bias." *Quarterly Journal of Economics* 134(3):1163–1224.

DiPrete, Thomas A., and Gregory M. Eirich. 2006. "Cumulative Advantage as a Mechanism for Inequality: A Review of Theoretical and Empirical Developments." *Annual Review of Sociology* 32:271–297.

Downey, Douglas B., and Dennis J. Condron. 2016. "Fifty Years since the Coleman Report: Rethinking the Relationship between Schools and Inequality". *Sociology of Education* 89(3):207–2020.

Esser, Hartmut. 2016. “The model of ability tracking – Theoretical expectations and empirical findings on how educational systems impact on educational success and inequality.” S. 25-42 in: Hans-Peter Blossfeld, Sandra Buchholz, Otto Friedrich, Jan Skopek, Moris Triventi (Eds.) *Models of Secondary Education and Social Inequality. An International Comparison.* Cheltenham: Edward Elgar.

Evans, David, and Fei Yuan. 2019. “Equivalent Years of Schooling: A Metric to Communicate Learning Gains in Concrete Terms.” World Bank Working Paper No. WPS8752. The World Bank. Retrieved from http://documents.worldbank.org/curated/en/123371–550594320297.

Farkas, George. 2003. “Cognitive Skills and Noncognitive Traits and Behaviors in Stratification Processes.” *Annual Review of Sociology* 29:541–562.

Fazio, Russell H., Javier A. G. Samayoa, Shelby T. Boggs, and Jesse Ladanyi. 2023. “Implicit Bias: What Is It?” In *The Cambridge Handbook of Implicit Bias and Racism*, edited by Jon A. Krosnick, H. Tobias, and A.L. Scott, Cambridge: Cambridge University Press.

Ferman, Bruno, and Luiz F. Fontes. 2022. “Assessing Knowledge or Classroom Behavior? Evidence of Teachers’ Grading Bias.” *Journal of Public Economics*. 216:104773. https://doi.org/10.1016/j.jpubeco.2022.104773

Forster, Andrea G., Herman G van de Werfhorst. 2020. Navigating Institutions: Parents’ Knowledge of the Educational System and Students’ Success in Education, *European Sociological Review* 36(1): 48–64.

Foschi, Martha. 2000. “Double Standards for Competence: Theory and Research.” *Annual Review of Sociology* 26:21–42.

Ganzeboom, Harry B. G., and Donald J. Treiman. 1996. “Internationally Comparable Measures of Occupational Status for the 1988 International Standard Classification of Occupations.” *Social Science Research* 25:201–239.

Gil-Hernández, Carlos J. 2021. “The (Unequal) Interplay Between Cognitive and Noncognitive Skills in Early Educational Attainment.” *American Behavioral Scientist*, 65(11):1577-1598. https://doi.org/10.1177/0002764221996764

Gil-Hernández, Carlos J. 2019. “Do Well-off Families Compensate for Low Cognitive Ability? Evidence on Social Inequality in Early Schooling from a Twin Study.” *Sociology of Education*, 92(2):150-175. https://doi.org/10.1177/0038040719830698

Gortázar, Lucas, David Martínez de Lafuente, and Ainhoa Vega–Bayo. 2022. “Comparing Teacher and External Assessments: Are Boys, Immigrants, and Poorer Students Undergraded?” *Teaching and Teacher Education* 115:103725.

Greenwald, Anthony G., and Linda Hamilton Krieger. 2006. “Implicit Bias: Scientific Foundations.” *California Law Review* 94(4):945–967.

Hanna, Rema N., and Leigh L. Linden. 2012. “Discrimination in Grading.” American *Economic Journal: Economic Policy* 4(4):146–168.

Holm, Anders, Anders Hjorth–Trolle, and Mads Meier Jæger. 2019. “Signals, Educational Decision–Making, and Inequality.” *European Sociological Review* 35(4):447–460.

Holtmann, Anne Christine, Laura Menze, and Heike Solga. 2021. “Intergenerational Transmission of Educational Attainment: How Important Are Children’s Personality Characteristics?” *American Behavioral Scientist 65*(11): 1531-1554. https://doi.org/10.1177/0002764221996779

Homuth, Christoph, Johannes Thielemann, and Sebastian E. Wenz. 2023. “Measuring Elementary School Teachers’ Stereotypes in the NEPS SC2.” NEPS Survey Paper No. 108. Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Jackson, Michelle. 2013. *Determined to succeed? Performance versus choice in educational attainment*. Stanford: Stanford University Press.

Jæger, Mads Meier. 2011. “Does cultural capital really affect academic achievement? New evidence from combined sibling and panel data.” *Sociology of Education*, 84(4), 281–298. doi:10.1177/003804071141701

Jæger, Mads Meier, and Richard Breen. 2016. “A Dynamic Model of Cultural Reproduction.” *American Journal of Sociology* 121(4):1079–1115.

Jæger, Mads Meier. 2022. “Cultural Capital and Educational Inequality: An Assessment of the State of the Art.” Pp. 121–134 in *Handbook of Sociological Science*, edited by K. Gërxhani, N. D. de Graaf, and W. Raub. Edward Elgar Publishing.

Jennings, Jennifer L., and Thomas A. DiPrete. 2010. “Teacher Effects on Social and Behavioral Skills in Early Elementary School.” *Sociology of Education* 83(2):135–159.

Lareau, Annette. 2015. “Cultural Knowledge and Social Inequality.” *American Sociological Review*, 80(1): 1–27.

Lievore, Ilaria, and Moris Triventi. 2023. "Do Teacher and Classroom Characteristics Affect the Way in Which Girls and Boys Are Graded?" *British Journal of Sociology of Education* 44(1):97–122.

Lockwood, J. R., and Daniel F. McCaffrey. 2014. "Correcting for test score measurement error in ANCOVA models for estimating treatment effects." *Journal of Educational and Behavioral Statistics, 39*(1), 22–52. https://doi.org/10.3102/1076998613509405

Lorenz, Georg, Irena Kogan, Sarah Gentrup, and Cornelia Kristen. 2024. "Non–Native Accents Among School Beginners and Teacher Expectations for Future Student Achievements." *Sociology of Education* 97(1):76-96.

Melamed, David, Christopher W. Munn, Leanne Barry, Bradley Montgomery, and Oneya F. Okuwobi. 2019. "Status Characteristics, Implicit Bias, and the Production of Racial Inequality." *American Sociological Review* 84(6):1013–1036.

NEPS Network. 2022. National Educational Panel Study, Scientific Use File of Starting Cohort Kindergarten. Leibniz Institute for Educational Trajectories (LIfBi), Bamberg. https://doi.org/10.5157/NEPS:SC2:10.0.0

Pit–ten Cate, Ineke M., and Sabine Glock. 2019. "Teachers' Implicit Attitudes Toward Students from Different Social Groups: A Meta–Analysis." *Frontiers in Psychology* 10:491099. https://doi.org/10.3389/fpsyg.2019.02832

Poropat, Arthur E. 2009. "A meta-analysis of the five-factor model of personality and academic performance." *Psychological Bulletin*, 135:322–338.

Radl, Jonas et al. 2024. "How socioeconomic status shapes cognitive effort: A laboratory study among fifth graders." UC3M Working Paper. https://hdl.handle.net/10016/43750

Ridgeway, Cecilia L. 2014. "Why Status Matters for Inequality." *American Sociological Review* 79(1):1–16.

Skopek, Jan, and Giampiero Passaretta. 2021. "Socioeconomic Inequality in Children's Achievement from Infancy to Adolescence: The Case of Germany." *Social Forces* 100(1):86–112.

Smithers, Lisa G. et al. 2018. 'A Systematic Review and Meta-Analysis of Effects of Early Life Noncognitive Skills on Academic, Psychosocial, Cognitive and Health Outcomes'. *Nature Human Behaviour* 2(11): 867.

Südkamp, Anna, Johanna Kaiser, and Jens Möller. 2012. “Accuracy of Teachers’ Judgments of Students’ Academic Achievement: A Meta–Analysis.” *Journal of Educational Psychology* 104(3):743–762.

Timmermans, Anneke C., Hester de Boer, Hilda T. A. Amsing, and Marieke P. C. van der Werf. 2018. “Track Recommendation Bias: Gender, Migration Background, and SES Bias Over a 20–Year Period in the Dutch Context.” *British Educational Research Journal* 44(5):847–874.

van Huizen, Thomas, Madelon Jacobs, and Matthijs Oosterveen. 2024. “Teacher Bias or Measurement Error?” arXiv:2401.04200 [econ.EM]. DOI: 10.48550/arXiv.2401.04200.

Wenz, Sebastian E., and Kerstin Hoenig. 2020. “Ethnic and Social Class Discrimination in Education: Experimental Evidence from Germany.” *Research in Social Stratification and Mobility* 65:100461. https://doi.org/10.1016/j.rssm.2019.100461

Zanga, Giulietta, and Elena De Gioannis. 2023. “Discrimination in Grading: A Scoping Review of Studies on Teachers’ Discrimination in School.” *Studies in Educational Evaluation* 78:101284. https://doi.org/10.1016/j.stueduc.2023.101284

Zhu, Maria, 2024. “New Findings on Racial Bias in Teachers’ Evaluations of Student Achievement,” IZA Discussion Papers 16815, Institute of Labor Economics (IZA). https://ideas.repec.org/p/iza/izadps/dp16815.html

APPENDIX

Appendix A. Robustness Checks and Full Output (Table A.3.)

Table A.1. Non-linearities: ability tertiles and higher-order polynomials in IVs

	M1A	M2A	M3A	M4A	M1B	M2B	M3B	M4B
	Z-GPA (Grade 4)				Track Recommendation (Grade 4)			
Parental ISEI	0.0068***	0.0055***	0.0048**	0.0030*	0.0049***	0.0040***	0.0035***	0.0024**
	(0.0015)	(0.0014)	(0.0014)	(0.0015)	(0.0008)	(0.0008)	(0.0007)	(0.0007)
Instrumented Z-Test Scores (G4)				0.666***				0.409***
				(0.042)				(0.051)
Q2 Z-Test Scores (G4)	0.709***	0.570***			0.326***	0.260***		
	(0.065)	(0.065)			(0.034)	(0.034)		
Q3 Z-Test Scores (G4)	1.274***	0.974***			0.536***	0.394***		
	(0.070)	(0.075)			(0.031)	(0.034)		
Q2 Z-Non-Cognitive Skills (G4)		0.442***				0.280***		
		(0.056)				(0.028)		
Q3 Z-Non-Cognitive Skills (G4)		0.669***				0.315***		
		(0.061)				(0.030)		
Q2 Z-Test Scores (G1-4)			0.572***				0.267***	
			(0.058)				(0.028)	
Q3 Z-Test Scores (G1-4)			0.918***				0.418***	
			(0.078)				(0.031)	
Q2 Z-Non-Cognitive Skills (G1-4)			0.518***	0.302***			0.299***	0.223***
			(0.062)	(0.052)			(0.031)	(0.033)
Q3 Z-Non-Cognitive Skills (G1-4)			0.757***	0.385***			0.351***	0.203***
			(0.074)	(0.071)			(0.029)	(0.038)
School-FE	✓	✓	✓	✓	✓	✓	✓	✓
IV				✓				✓
Observations	2,152	2,152	2,152	2,152	2,448	2,448	2,448	2,448
R-squared	0.469	0.516	0.520	0.394	0.432	0.491	0.529	0.339

Notes: Robust standard errors in parentheses, all models control for migration background, gender and age in months.
Reference category for Z-Test Scores and Z-Non-Cognitive Skills = Q1. G = Grade. *** p<0.001, ** p<0.01, * p<0.05, + p<0.10

Table A.2. GPA by German and Math grades

	M1Aa	M2Aa	M3Aa	M4Aa	M1Ab	M2Ab	M3Ab	M4Ab
	Z-German Grade (Grade 4)				Z-Math Grade (Grade 4)			
Parental ISEI	0.0042** (0.0015)	0.0035* (0.0015)	0.0025+ (0.0015)	0.0020 (0.0014)	0.0064*** (0.0016)	0.0053*** (0.0016)	0.0043** (0.0015)	0.0034* (0.0016)
Z-Test Scores German (G4)	0.548*** (0.033)	0.434*** (0.034)		0.622*** (0.059)				
Z-Test Scores Math (G4)					0.507*** (0.025)	0.377*** (0.026)		0.487*** (0.048)
Z-Non-Cognitive Skills (G4)		0.237*** (0.029)				0.299*** (0.027)		
Z-Test Scores German (G1-4)			0.425*** (0.034)					
Z-Test Scores Math (G1, 2, 4)							0.558*** (0.069)	
Z-Non-Cognitive Skills (G1-4)			0.252*** (0.028)	0.152*** (0.035)			0.236*** (0.039)	0.275*** (0.028)
School-FE	✓	✓	✓	✓	✓	✓	✓	✓
IV				✓				✓
Observations	2,152	2,152	2,152	2,152	2,152	2,152	2,152	2,152
R-squared	0.496	0.529	0.528	0.356	0.441	0.496	0.519	0.336

Notes: Robust standard errors in parentheses, all models control for migration background, gender and age in months.
G=Grade.
In M4Aa-M4Ab, Z-Test Scores in German or Math (G4) are instrumented with a different-subject IV (G4), respectively.
*** p<0.001, ** p<0.01, * p<0.05, + p<0.10

Table A.3. Interaction models between parental ISEI and ability and heterogenous IV models by parental ISEI

	M5A	M6A	M7A	M8A	M5B	M6B	M7B	M8B
	Z-GPA (Grade 4)				Track Recommendation (Grade 4)			
			Low ISEI	High ISEI			Low ISEI	High ISEI
Parental ISEI	0.0029*	0.0032*			0.0027***	0.0029***		
	(0.0013)	(0.0014)			(0.0007)	(0.0007)		
Z-Test Scores (G1-4)	0.748***	0.466***	0.700***	0.584***	0.318***	0.169***	0.322***	0.211***
	(0.083)	(0.031)	(0.059)	(0.074)	(0.033)	(0.014)	(0.044)	(0.029)
Z-Non-Cognitive Skills (G1-4)	0.276***	0.527***	0.205***	0.141**	0.158***	0.238***	0.101**	0.119***
	(0.028)	(0.078)	(0.042)	(0.051)	(0.013)	(0.036)	(0.033)	(0.021)
P. ISEI x Z-Test Scores (G1-4)	-0.005***				-0.002***			
	(0.001)				(0.0005)			
P. ISEI x Z-Non-Cog. Skills (G1-4)		-0.004***				-0.001*		
		(0.001)				(0.0006)		
School FE	✓	✓	✓	✓	✓	✓	✓	✓
IV			✓	✓			✓	✓
Observations	2,152	2,152	1,014	1,023	2,448	2,448	1,194	1,167
R-squared	0.574	0.573	0.457	0.280	0.542	0.537	0.311	0.309

Notes: Robust standard errors in parentheses, all models control for migration background, gender and age in months.
G = Grade.
P = Parental; Low ISEI = q1-q2; High ISEI = q3-q4
Non-Cog. = Non-Cognitive
*** p<0.001, ** p<0.01, * p<0.05, + p<0.10

Table A.4. Alternative parental SES measure: highest year of education

	M1A	M2A	M3A	M4A	M1B	M2B	M3B	M4B
	Z-GPA (Grade 4)				Track Recommendation (Grade 4)			
Parental Education (Years)	0.0546***	0.0458***	0.0350*	0.0280*	0.0446***	0.0369***	0.0295***	0.0262***
	(0.0128)	(0.0129)	(0.0141)	(0.0133)	(0.0067)	(0.0061)	(0.0062)	(0.0064)
Z-Test Scores (G4)	0.586***	0.447***		0.616***	0.230***	0.151***		0.258***
	(0.031)	(0.033)		(0.045)	(0.012)	(0.014)		(0.025)
Z-Non-Cog. Skills (G4)		0.272***				0.160***		
		(0.027)				(0.012)		
Z-Test Scores (G1-4)			0.452***				0.166***	
			(0.032)				(0.014)	
Z-Non-Cog. Skills (G1-4)			0.279***	0.199***			0.159***	0.111***
			(0.029)	(0.031)			(0.013)	(0.017)
School-FE	✓	✓	✓	✓	✓	✓	✓	✓
IV				✓				✓
Observations	2,152	2,152	2,152	2,152	2,448	2,448	2,448	2,448
R-squared	0.518	0.558	0.568	0.410	0.447	0.508	0.536	0.354

Notes: Robust standard errors in parentheses, all models control for migration background, gender and age in months
G = Grade
Non-Cog. = Non-Cognitive
*** p<0.001, ** p<0.01, * p<0.05, + p<0.10

Table A.5. Alternative student noncognitive skills measure: parental reports of readiness for exertion

	M1A	M2A	M3A	M4A	M1B	M2B	M3B	M4B
	Z-GPA (Grade 4)				Track Recommendation (Grade 4)			
Parental ISEI	0.0532***	0.0560***	0.0444**	0.0347*	0.0043***	0.0043***	0.0351***	0.0277***
	(0.0016)	(0.0015)	(0.0015)	(0.0015)	(0.0008)	(0.0008)	(0.0007)	(0.0008)
Z-Test Scores (G4)	0.594***	0.520***		0.719***	0.233***	0.191***		0.315***
	(0.032)	(0.032)		(0.034)	(0.013)	(0.014)		(0.018)
Z-Non-Cog. Skills (G4)		0.212***				0.122***		
		(0.024)				(0.012)		
Z-Test Scores (G1-4)			0.554***				0.225***	
			(0.026)				(0.011)	
Z-Non-Cog. Skills (G1-4)			0.212***	0.138***			0.111***	0.078***
			(0.023)	(0.026)			(0.009)	(0.012)
School-FE	✓	✓	✓	✓	✓	✓	✓	✓
IV				✓				✓
Observations	2,152	2,152	2,152	2,152	2,448	2,448	2,448	2,448
R-squared	0.517	0.547	0.560	0.376	0.443	0.489	0.518	0.304

Notes: Robust standard errors in parentheses, all models control for migration background, gender and age in months.
G = Grade
Non-Cog. = Non-Cognitive
*** p<0.001, ** p<0.01, * p<0.05, + p<0.10

Appendix B. Missing Values and Additional Summary Statistics

Table A.6. Missing values by outcomes' subsamples

Variable	Missing (n)	Total (n)	Missing (%)	Missing (n)	Total (n)	Missing (%)
		GPA (W7)			Track Recommendation (W6)	
Longitudinal Weight	440	3,246	13.56	651	5,461	11.92
Migrant Origin (W3)	145	3,246	4.47	504	5,461	9.23
Gender (W3)	1	3,246	0.03	1	5,461	0.02
Age (W3)	0	3,246	0	0	5,461	0
Parental ISEI (W3)	191	3,246	5.88	594	5,461	10.88
Parental Years of Education (W3)	191	3,246	5.88	615	5,461	11.26
GPA (W7)	178	3,246	5.48			
German Grade (W7)	216	3,246	6.65			
Math Grade (W7)	230	3,246	7.09			
Recommendation (W6)[a]				2,430	5,461	44.5
Math Test Scores (W6)	176	3,246	5.42	159	5,461	2.91
Math Test Scores (Last)	5	3,246	0.15	4	5,461	0.07
Math Test Scores (W3-W4)	24	3,246	0.74	37	5,461	0.68
Math Test Scores (W3, W4, W6)	5	3,246	0.15	4	5,461	0.07
German Test Scores (W6)	113	3,246	3.48	31	5,461	0.57
German Test Scores (Last)	4	3,246	0.12	2	5,461	0.04
German Test Scores (W3-W5)	15	3,246	0.46	24	5,461	0.44
German Test Scores (W3-W6)	4	3,246	0.12	2	5,461	0.04
Mean Math/German Test Scores (W6)	108	3,246	3.33	22	5,461	0.4
Mean Math/German Scores (W3-W5)	5	3,246	0.15	11	5,461	0.2
Mean Math/German Scores (W3-W6)	1	3,246	0.03	1	5,461	0.02
Concentration/Persistence (W6)	136	3,246	4.19	244	5,461	4.47
Concentration/Persistence (W3-W6)	136	3,246	4.19	244	5,461	4.47
Effort (W6)	152	3,246	4.68	514	5,461	9.41
Effort (W3-W6)	152	3,246	4.68	514	5,461	9.41
Total Panel Sample	0	6,341	0	0	6,341	0
Total Wave Sample (% Attrition)	3,095	3,246	48.81	880	5,461	13.88
Analytical Sample[b]	1,094	2,152	33.70	3,013	2,448	55.17

Notes: [a] Missing values include Eastern Länders (Berlin, Brandenburg and Mecklenburg-Western Pomerania) without recommendations/tracking until primary grade 6/7 (n=513), and students without a recommendation yet due to grade retention or parents filling in the questionnaire before (n=470); W=Survey Wave

[b] Excluding singleton cases within school clusters.

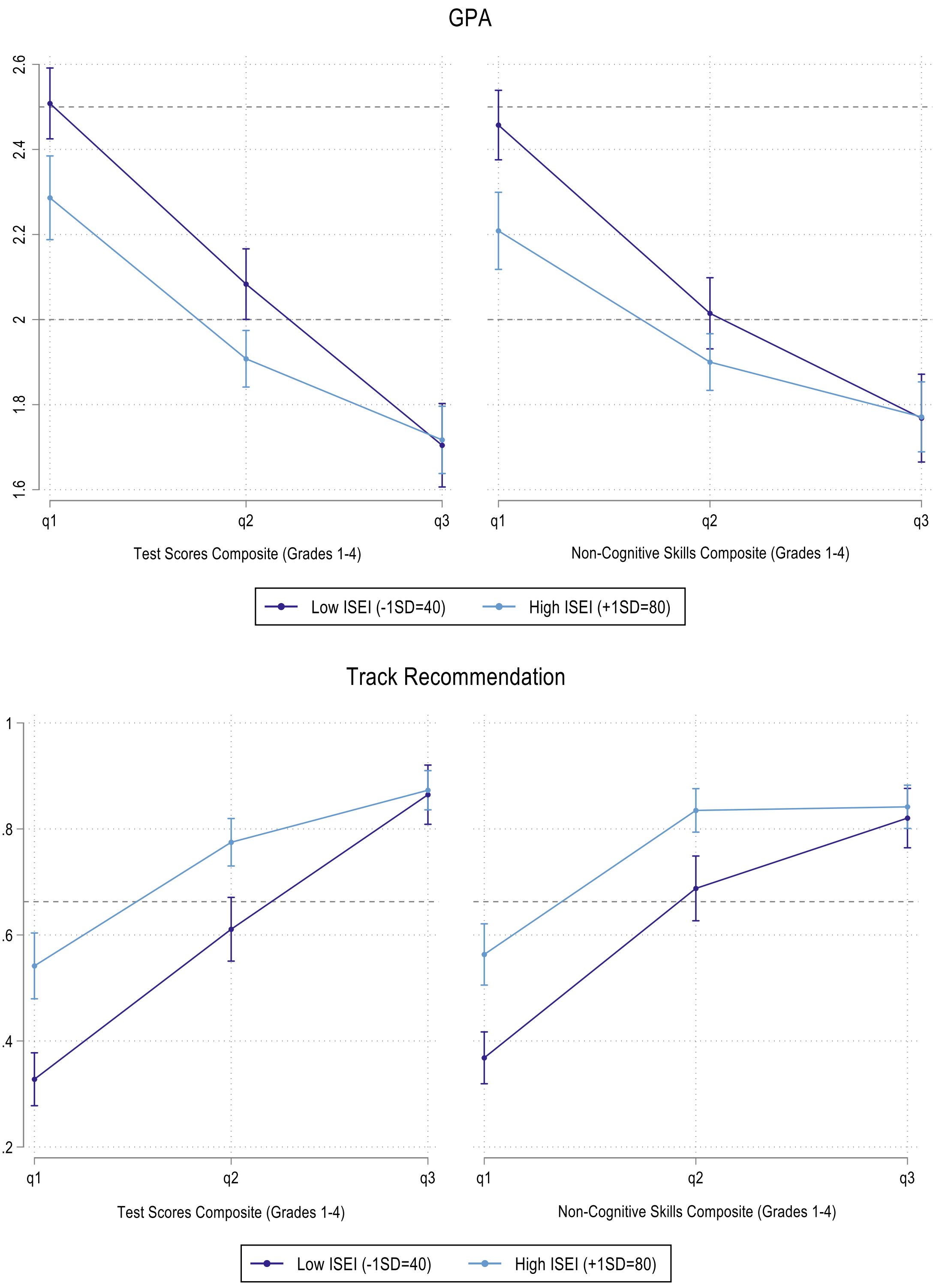
Table A.7. Summary Statistics (GPA - Wave 7 Analytical Sample = 2,152) for variables in robustness checks and additional analyses

Variable	Mean	SD	Min	Max
Socio-demographic				
Parental Years of Education (W3)	14.51	2.05	9	18
Outcomes				
German Grade (G4)	2.05	0.77	6	1
Z-German Grade (G4)	0.00	1.00	-5.14	1.37
Math Grade (G4)	2.06	0.85	6	1
Z-Math Grade (G4)	0.00	1.00	-4.66	1.25
Test Scores				
Z-Math Test Scores (G4)	0.00	1.00	-4.06	4.40
Z-Math Test Scores (G1-G2)	0.00	1.00	-4.17	4.21
Z-Math Test Scores (G1, G2, G4)	0.00	1.00	-4.15	3.52
Z-German Test Scores (G4)	0.00	1.00	-4.24	3.18
Z-German Test Scores (G1-G3)	0.00	1.00	-3.34	4.23
Z-German Test Scores (G1-G4)	0.00	1.00	-2.88	3.90
Noncognitive Skills				
Parental Reports				
Effort (G4)	3.10	0.53	1	4
Z-Effort (G4)	0.00	1.00	-3.95	1.68
Effort (G1-G4)	3.13	0.47	1.44	4
Z-Effort (G1-G4)	0.00	1.00	-3.63	1.86

Notes: All figures are adjusted by longitudinal weight (W3-W7).

Appendix C. Additional Figures

Figure A.1. Marginal effects of student ability by parental ISEI subgroups



Notes: The figure portrays the marginal effects by parental ISEI subgroups from Table A.3. M5A and M6A for GPA (upper-panel) and from M5B and M6B for track recommendations (bottom-panel), including an interaction term between parental ISEI and test scores (left-panels; M5A-B) or noncognitive skills (right panels; M6A-B) tertiles, respectively. GPA is in its original scale, with horizontal lines indicating official thresholds for academic track recommendations (<2; <2.5) in some Federal States. Track recommendation is in a probabiilty scale (0-1), with the horizontal line indicating the sample average academic track reccomendations. All models include school FE and control for children's gender, migration origin and age. Confidence intervals at the 95% level.