

# Teacher bias or measurement error?

Thomas van Huizen\*    Madelon Jacobs<sup>†</sup>    Matthijs Oosterveen<sup>‡</sup>

September 16, 2025

## Abstract

Subjective teacher evaluations play a key role in shaping students' educational trajectories. Previous studies have shown that students of low socioeconomic status (SES) receive worse subjective evaluations than their high SES peers, even when they score similarly on objective standardized tests. This is often interpreted as evidence of teacher bias. Measurement error in test scores challenges this interpretation. We discuss how both classical and non-classical measurement error in test scores generate a biased coefficient of the conditional SES gap, and consider three empirical strategies to address this bias. Using administrative data from the Netherlands, where secondary school track recommendations are pivotal teacher judgments, we find that measurement error explains 35 to 43% of the conditional SES gap in track recommendations.

**JEL:** I24, C36

**Keywords:** Teacher evaluation, Track recommendation, Measurement error, Instrumental variables, Errors-in-variables

---

\*Department of Applied Economics, Utrecht University School of Economics, Utrecht University, Utrecht, The Netherlands. [t.m.vanhuizen@uu.nl](mailto:t.m.vanhuizen@uu.nl)

<sup>†</sup>Research Centre for Education and the Labour Market (ROA), School of Business and Economics, Maastricht University, Maastricht, The Netherlands. [mce.jacobs@maastrichtuniversity.nl](mailto:mce.jacobs@maastrichtuniversity.nl)

<sup>‡</sup>Department of Economics, Lisbon School of Economics and Management, and Advance/ISEG Research, University of Lisbon, Lisbon, Portugal. [oosterveen@iseg.ulisboa.pt](mailto:oosterveen@iseg.ulisboa.pt)

<sup>§</sup>We thank Sönke Matthewes, Joppe de Ree, Astrid Sandsør, and Dinand Webbink for valuable comments. The paper has also benefited from the comments of participants at the EQOP meeting at the University of Oslo, the workshop on exam performance and test-taking behavior at University Carlos III of Madrid, KVS new paper sessions, LMG Lisbon, the EALE, LEER, LESE, and PPE conference, and the brown-bag seminars at the University of the Balearic Islands, the University of Bergen, and the Utrecht University School of Economics. The data used in this study are housed within the secure virtual environment at Statistics Netherlands and cannot be publicly shared due to privacy restrictions. However, all the underlying data supporting the presented results are securely accessible through Statistics Netherlands. National and international researchers can request access to the administrative data through Statistics Netherlands, and to the test score data through the Netherlands Cohort Study on Education (NCO) by contacting [info@nationaalcohortonderzoek.nl](mailto:info@nationaalcohortonderzoek.nl). Matthijs Oosterveen gratefully acknowledges financial support from FCT, Fundação para a Ciência e a Tecnologia (Portugal), national funding through research grant UIDB/04521/2020 & UID06522. The authors have no relevant or material financial interests that relate to the research described in this paper. All omissions and errors are our own. Author names are ordered alphabetically.

# 1 Introduction

Teacher bias may be central to inequalities in educational attainment and later life outcomes. To estimate the extent of such bias, researchers typically compare student assessments in which teachers exercise discretion, such as teacher-assigned grades or track recommendations, to more objective indicators of student ability, such as standardized test scores (Alesina et al., 2024; Falk et al., 2025). However, the presence of systematic gaps in subjective teacher evaluations between groups, conditional on objective ability measures, is not necessarily evidence of teacher bias. This approach faces two main concerns: omitted variables and measurement error in the ability measure. Studies often address the first concern by controlling for additional factors (Burgess and Greaves, 2013). In contrast, the second concern is typically overlooked or addressed under strict assumptions that may not hold in practice, a shortcoming shared by the broader literature on discrimination.

This paper focuses on the second concern: measurement error in the objective measure of ability. In classical testing theory, test scores are typically based on the number of correct answers and provide an unbiased but noisy estimate of ability. The resulting “classical” measurement error in the test score is considered purely random and hence uncorrelated with ability. In modern item response theory (IRT), statistical models produce test scores that may be biased estimates of ability but are unbiased as predictors. The resulting “non-classical” measurement error may correlate with ability. For instance, in case the test score is estimated with a posterior mean score, students with a low and high signal are drawn towards the population mean, such that the error correlates negatively with ability, yet is uncorrelated with the test score.

To examine the implications, consider a regression in which a teacher’s subjective evaluation of a student is the dependent variable, and the independent variables include a group characteristic, such as family socioeconomic status (SES), and a test score as a measure of student ability. In case of classical measurement error, the test score coefficient is biased towards zero (the well-known attenuation bias). Moreover, if student ability is associated with SES and matters for teacher evaluations, this bias also contaminates the coefficient of the conditional SES gap. In particular, when ability is positively associated with both SES and teacher evaluations, the conditional SES gap is overestimated. In case of non-classical

measurement error, the attenuation bias in the test score coefficient may be partially or fully eliminated, though this leads to a biased coefficient of the association between ability and SES. This bias then also contaminates the estimate for the conditional SES gap.

We first show how both classical and certain forms of non-classical measurement error, particularly relevant to modern test scores, generate identical contamination bias in the coefficient measuring the conditional gap between groups. We then discuss three empirical strategies to address measurement error. First, we consider an instrumental variable (IV) approach, a well-established method in econometrics to correct for measurement error ([Hausman, 2001](#)). Second, we discuss a method in the spirit of errors-in-variables (EIV) models, an approach that is common in psychometrics ([Fuller, 2009](#)). The additional parameter required for the EIV strategy, compared to OLS, is the reliability ratio of the test that is used to measure ability. We argue that, under weaker assumptions than the IV strategy, the IV first stage can be used to identify this ratio. Third, we combine the EIV strategy with a novel approach to identify the reliability ratio under yet weaker assumptions.

We apply these three strategies to estimate conditional gaps in teacher track recommendations in the Netherlands, focusing on differences by family SES. The Dutch tracking regime provides a compelling setting for this analysis. While nearly all OECD countries have some form of educational tracking, the Netherlands assigns students to secondary-school tracks at a relatively young age of twelve. Assignment is based on binding track recommendations from the primary school teacher ([WVO, 2014](#)). These recommendations may be highly consequential, as roughly 70% of students does not experience track mobility four years into secondary education ([de Ree et al., 2023](#)). Track placement determines the level of the curriculum and peer group abilities, and may affect student achievement ([Matthewes, 2021](#)). As only the highest tracks provide access to university, track placement may also have long-term consequences for attainment and income ([Dustmann et al., 2017](#); [Borghans et al., 2019](#)). Moreover, negatively biased recommendations signal low teacher expectations, which can in turn harm student outcomes ([Carlana, 2019](#); [Papageorge et al., 2020](#)).

Given the consequential nature of these recommendations, potential teacher bias has been a major policy concern in the Netherlands. This concern gained momentum in 2016, when a report by the Dutch [Inspectorate of Education \(2016\)](#) demonstrated that students from low SES families receive significantly lower track recommendations, conditional on

scores from a standardized end-of-primary school test. This finding fueled an ongoing public debate on teachers as gatekeepers of opportunities and contributed to a national policy reform, effective from 2023-24. The reform reduces the role of teachers in track placement, and increases the role of the end-of-primary school test. Similar findings in academic studies reinforce concerns about unequal opportunities in education, both in the Netherlands (Zumbuehl et al., 2025) and elsewhere (Carlana et al., 2022; Falk et al., 2025).

Using new administrative data, we first replicate previous findings from the Netherlands, and find that children with lower family SES receive systematically lower track recommendations, conditional on standardized test scores. We discuss that our test scores are derived from an IRT model, with ability parameters estimated using weighted maximum likelihood (Eggen and Sanders, 1993). This method is known to produce largely unbiased estimates of student ability (Warm, 1989), and the measurement error is considered classical. We address concerns about measurement error using three strategies. Our novel strategy combines the EIV method with an unbiased estimate for the test’s reliability ratio under weak assumptions, based on students’ standardized test scores across all primary school grades. Across the three approaches, we find that measurement error can explain between 35 and 43% of the conditional SES gap.

Our paper contributes to the broad literature on measurement error (Schennach, 2016). In particular, we focus on contamination bias due to measurement error. As Modalsli and Vosters (2024) point out, “*[a]lthough the notion of bias in one coefficient arising from error in another regressor is a well-known econometric result, it is seldom addressed in practice with empirical studies.*” Modalsli and Vosters (2024) examine the implications of classical measurement error in multigenerational income regressions and use an IV approach to mitigate the resulting bias. Similarly, Gillen et al. (2019) use IV to show that the gender gap in competitiveness is well explained by risk attitudes and overconfidence once classical measurement error is addressed. Our study extends this literature by applying three strategies to address contamination bias and by incorporating both classical and non-classical measurement error. The latter is pervasive in education, where test scores are often derived from IRT models (Jacob and Rothstein, 2016). It is also common in the labor market, as many individuals misreport hours worked, leading to biased estimates of wage discrimination and inequality (Borjas and Hamermesh, 2024). More generally, it arises in survey data when respondents

base their answers on incomplete information (Hyslop and Imbens, 2001)

We also contribute to the literature on teacher bias by providing new evidence on the conditional SES gap in track recommendations. Teacher bias in track recommendations has been studied in Germany (Falk et al., 2025), Italy (Carlana et al., 2022), and the Netherlands (Zumbuehl et al., 2025). Previous studies have also examined teacher bias in grading by gender (Cornwell et al., 2013; Lavy and Sand, 2018; Terrier, 2020), ethnicity (Burgess and Greaves, 2013), and migration background (Alesina et al., 2024). Consistent with the observation above, contamination bias due to measurement error has been largely ignored in this literature. Three notable exceptions are Botelho et al. (2015), Ferman and Fontes (2022), and Zhu (2024), who study teacher bias in grading in Brazil and the US. All three studies find that the conditional gaps are substantially reduced or even reversed when addressing measurement error in test scores. However, as these studies assume classical measurement error and exclusively apply an IV strategy, it remains unclear whether the results hold under non-classical error and are robust to alternative strategies relying on weaker assumptions.

## 2 Educational tracking in the Netherlands

In the Netherlands, the transition to tracked secondary education takes place around the age of twelve when students leave the sixth and final grade of primary education. The secondary school system consists of three main tracks: (i) a university track (*vwo*), which has a duration of six years and gives students access to university education, (ii) a college track (*havo*), which has a duration of five years and prepares students for college education, and (iii) a vocational track (*vmbo*), which has a duration of four years and serves as a preparation for vocational education. The vocational track consists of four sub-tracks, which can be ordered from more practice-oriented to more theory-oriented.

The allocation of students to tracks is based on the recommendation provided by the sixth-grade primary school teacher. The recommendation process consists of two main steps. First, by March of the sixth grade the teacher provides an initial track recommendation. Then, in April or May, each student takes a standardized end-of-primary education test, which the teacher may use to consider an upward revision (WPO, 2014). Hence, the final track recommendation may be higher (but not lower) than the initial one. The teacher

track recommendation is binding in the sense that, as a rule, secondary schools cannot place students in a higher track than indicated by the recommendation (WVO, 2014).

We use the initial, instead of the final, track recommendation as our measure for the teacher’s subjective evaluation. This choice reflects that the initial recommendation is often the most influential, as many secondary schools begin allocating students to first-year classes as early as March, relying on these initial recommendations. Secondary schools may face challenges when changing their allocation by the time the final track recommendation is provided in May (Inspectorate of Education, 2019). Moreover, in practice, teachers upgrade only 1 in 10 students, meaning the initial and final track recommendations are identical for 9 out of 10 students (de Ree et al., 2023).

Teachers are expected to make holistic track recommendations by considering both objective standardized test results and their subjective judgments on factors such as motivation, attitude towards school, and classroom behavior. While there are no strict national rules describing how to combine these types of information, general guidelines from the Ministry of Education (2023) suggest that standardized test results provide a “good starting point”. A recent survey among 400 primary school teachers shows that 85% of the respondents attach “considerable” to “a lot of” value to standardized test results when formulating their recommendations (DUO, 2023).

## 2.1 Standardized test scores

All Dutch primary school students, from grade one to six, are expected to take two standardized tests per year. The first test is administered in the middle of the school year around February and the second at the end of the school year in June. We will refer to these tests as the midterm and end term test, respectively. We use the fifth-grade end term test as the main measure of student ability to estimate the conditional SES gap in track recommendations. This test can be considered the last score that teachers observe before making the initial recommendation. Alternatively, one could use the end-of-primary school test score (the sixth-grade end term test) as the ability measure and the final track recommendation as the teacher’s subjective evaluation. This alternative approach is problematic since the final track recommendation can only be adjusted upwards based on performance on the end-of-

primary school test. Hence, this test is low (high) stakes for students who are (un)satisfied with the initial recommendation.

Primary schools have to select one of the standardized test systems that are approved by the Ministry of Education. Most schools opt for the test system provided by *Cito*, which has a market share of roughly 80% in the period we study. The biannual *Cito* tests focus on two domains: mathematics and reading. The tests consist of approximately 70 to 100 items, include both multiple-choice and open-ended questions, and are scored by the teacher using standardized answer keys. The item scores are entered into the *Cito* software and used to generate students' test scores through an item response theory (IRT) model.

An IRT model specifies the probability that a student answers a test item correctly as a function of latent student ability and item parameters. The IRT model used by *Cito* is the Birnbaum model (Cito, 2017), which includes two item parameters: difficulty and discrimination. The item difficulty parameter reflects the student ability level at which there is a 50% probability of answering the item correctly. The discrimination parameter reflects how sharply the item distinguishes between students of different ability levels. These parameters may be estimated for a selection of items to form a so-called calibrated item bank, which can be used to create multiple tests of varying difficulty that aim to measure student ability on a common scale. This enables, for instance, observing the development of student ability from grade 1 to 6, one of the main goals of standardized testing in the Netherlands.

*Cito* constructs its calibrated item bank through national norming studies. First, sample tests are administered to representative student populations. Second, the item parameters are estimated: the discrimination parameters are held fixed while the difficulty parameters are estimated using conditional maximum likelihood. Both parameters are then iteratively adjusted to improve model fit. The resulting calibrated item bank enables the construction of the biannual tests for various grade levels. Although more difficult items are used in tests for higher grades, student ability can supposedly be measured on a common scale.

Latent student ability is estimated from these biannual tests as a “fixed effect” using weighted maximum likelihood (Eggen and Sanders, 1993). In this approach, the two item parameters are held fixed at their calibrated values, and the likelihood function is weighted by the test information function. This function indicates how much information the test provides about student ability, with items of difficulty close to the student's ability offering

more information. [Warm \(1989\)](#) shows that this procedure yields (nearly) unbiased estimates of student ability. The estimated ability scores obtained through this procedure are the test scores we use in our analysis. Unbiased, yet noisy, estimates of ability implies that the measurement error in the biannual *Cito* test scores is assumed to be classical.

### 3 Data

We use proprietary administrative data from Statistics Netherlands. The data on the biannual standardized tests is gathered by the Netherlands Cohort Study on Education (NCO) and subsequently made available to Statistics Netherlands. The NCO collected the test score data only from schools using the tests developed by *Cito*. Since 2018–19, data were obtained by individually contacting primary schools, each of which had to grant permission. As a result, the test score data covers about 50% of all primary schools in the Netherlands. See [Haelermans et al. \(2020\)](#) for details on the NCO data collection. This test score data can be linked to standard administrative data using an anonymized personal ID, allowing us to also observe teacher track recommendations, parental income and education, and several other student background characteristics.

The test score history, from grade one to six, is available for the students who were enrolled in sixth grade from 2018-19 onward. Our sample includes four cohorts: students who are in sixth grade in the school years 2018-19, 2019-20, 2020-21, and 2021-22. The standard administrative data covers 686,309 students across these four cohorts, with almost no missing values. The biannual test score data includes 318,630 students, although not all students took both standardized tests in each grade. After merging these two datasets, we observe 276,841 students in 5,802 schools. After dropping students with one or both grade 5 test scores missing, as well as those with extreme household parental income values (discussed below), our main estimation sample consists of 148,019 students in 3,295 schools.

#### 3.1 Descriptive statistics

The first four columns in Table 1 present the descriptive statistics for our estimation sample. Panel A focuses on the measures used for the subjective teacher evaluation. Our primary measure is a dummy variable that equals one if the initial track recommendation is at least



Table 1: Descriptive statistics

|                                 | Estimation sample |        |          |          | Samples w/ missings |            |         |
|---------------------------------|-------------------|--------|----------|----------|---------------------|------------|---------|
|                                 |                   |        |          |          | Admin               | Test score | Merged  |
|                                 | (1)               | (2)    | (3)      | (4)      | (5)                 | (6)        | (7)     |
|                                 | Mean              | SD     | Min      | Max      |                     | Mean       |         |
| Panel A: Teacher recommendation |                   |        |          |          |                     |            |         |
| 1(recom. ≥ upper voc. tr.)      | 0.770             | 0.421  | 0.000    | 1.000    | 0.754               |            | 0.758   |
| 1(recom. ≥ college tr.)         | 0.497             | 0.500  | 0.000    | 1.000    | 0.478               |            | 0.489   |
| 1(recom. ≥ university tr.)      | 0.211             | 0.408  | 0.000    | 1.000    | 0.197               |            | 0.209   |
| Panel B: SES measures           |                   |        |          |          |                     |            |         |
| SES income                      | 6.439             | 3.772  | <-0.410  | >55.500  | 6.394               |            | 6.472   |
| 1(SES inc. ≥ median)            | 0.517             | 0.500  | 0.000    | 1.000    | 0.500               |            | 0.506   |
| SES years of schooling          | 15.975            | 2.334  | 8.000    | 18.000   | 16.004              |            | 15.945  |
| 1(SES edu. ≥ college)           | 0.559             | 0.497  | 0.000    | 1.000    | 0.618               |            | 0.553   |
| Panel C: Test scores            |                   |        |          |          |                     |            |         |
| Reading midterm grade 5         | 191.618           | 26.766 | <81.000  | >340.000 |                     | 191.614    | 192.536 |
| Reading end term grade 5        | 197.279           | 26.983 | <84.000  | >354.000 |                     | 195.374    | 196.367 |
| Math midterm grade 5            | 254.142           | 27.142 | <118.000 | >379.000 |                     | 253.982    | 254.564 |
| Math end term grade 5           | 261.959           | 26.085 | <118.000 | >394.000 |                     | 262.583    | 263.179 |
| Avg. midterm grade 5            | 0.005             | 0.900  | -5.177   | 5.077    |                     |            |         |
| Avg. end term grade 5           | 0.004             | 0.905  | -5.438   | 5.130    |                     |            |         |
| Panel D: Control variables      |                   |        |          |          |                     |            |         |
| 1(female)                       | 0.499             | 0.500  | 0.000    | 1.000    | 0.503               |            | 0.503   |
| 1(Western imm.)                 | 0.076             | 0.265  | 0.000    | 1.000    | 0.077               |            | 0.077   |
| 1(non-Western imm.)             | 0.200             | 0.400  | 0.000    | 1.000    | 0.178               |            | 0.198   |
| Age                             | 11.590            | 0.647  | <9.250   | >14.160  | 11.704              |            | 11.660  |
| Observations                    | 148019            |        |          |          | 686309              | 318630     | 276841  |

Notes: this table shows the descriptive statistics across four samples. Income is measured in €10,000. We translated the schooling levels in the Netherlands to years of schooling as follows: primary school = 8, mbo 1 = 13, havo or vwo = 14, mbo 2 or 3 = 15, mbo 4 = 16, college or university bachelor = 17, and university master = 18 years. Statistics Netherlands makes a distinction between individuals with a Western and non-Western migration background. Non-Western immigrants have a background from all countries in Africa, Latin America or Asia (including Turkey, excluding Indonesia and Japan), and Western immigrants from the remaining countries. Statistics Netherlands prevents reporting minima and maxima that are based upon less than 10 observations. Therefore, the minimum (maximum) may contain the smallest (largest) number, such that at least ten observations have values lower (higher) than that number.

equal to the college track. College or university track enrollment is highly policy relevant, since it implies the student is tracked towards higher education. Nearly half (49.7%) of the students receive a track recommendation equal to, or higher than, the college track.

Panel B presents the statistics for the different measures of SES. Our baseline SES measure is gross annual household parental income, which averages €64,390. We dropped observations with a yearly parental income above (below) the 99.9th (0.1st) percentile. In our analysis below we standardize parental income per cohort. For robustness checks, we use three alternative SES measures: parental years of schooling of the highest educated parent, a binary income indicator denoting whether parental income exceeds the student’s cohort median, and a binary education indicator reflecting whether the highest educated parent holds at least a college degree.

Panel C shows the statistics for the biannual *Cito* tests in fifth grade, per domain and the average across domains. The domain-specific scores are the raw scores estimated on a common scale by the IRT model. The average score is calculated by first standardizing the domain-specific scores per cohort, and subsequently taking the mean across domains. Table 3 in Appendix D provides the descriptive statistics for the raw domain-specific scores on the biannual tests in grades two through four. We exclude grade 1 data since a relatively high fraction of the test scores for that grade are missing. Panel D presents the statistics for the variables used as controls in the robustness checks.

We also compare the statistics of the students in our estimation sample with those from three larger samples representing the different steps of our selection process. Columns (5) to (7) correspond to the samples from the standard administrative data covering the full population, the test score data, and the merged dataset, respectively. The statistics in these columns are based on different numbers of observations due to missing values in some variables, particularly in the grade 5 test scores of columns (6) and (7). The comparison indicates that our estimation sample closely resembles the student population in the Netherlands.

## 4 Methodology

Let  $y_t$  be a variable that measures a teacher’s subjective evaluation of a student in time period  $t$ ,  $SES$  a variable that measures the socioeconomic status of the student, and  $s_t$  a vari-

able that reflects student ability in time period  $t$ . The long regression equation equals,

$$y_t = \beta SES + \gamma s_t + e_t. \quad (1)$$

Throughout the paper we suppress the subscript for student  $i$  and the intercept in regression equations. Scholars and policy makers are generally interested in the parameter  $\beta$ . That is, they are interested in whether subjective teacher evaluations differ systematically by SES, or another group characteristic such as gender or ethnicity, conditional on student ability. If individuals with similar student ability experience similar causal effects of the subjective evaluation on outcomes of interest, then  $\beta$  can also be interpreted as a measure of principal fairness, as recently introduced by [Imai and Jiang \(2023\)](#).

#### 4.1 Classical versus non-classical measurement error

We cannot estimate (1) as we do not observe ability. Instead, students produce an unbiased but noisy signal of ability by taking a test, denoted by  $\tilde{s}_t$ . The signal equals ability plus measurement error,  $\tilde{s}_t = s_t + m_t$ , where  $m_t$  is the signal's measurement error. An unbiased signal, meaning one that equals ability in expectation, implies that the error is classical.

**Assumption 1.** *Classical measurement error in the signal*

$$(Zero\ mean\ and\ cov)\ \mathbb{E}[m_t] = 0\ and\ Cov[m_t, x] = 0\ \forall x \neq \tilde{s}_t, t.$$

This classical error reflects, for instance, that students may guess answers, are more or less familiar with the specific selection of test items, or feel (un)well on the day of the test. According to classical testing theory, measures such as the fraction of correctly answered items follow Assumption 1 and are unbiased but noisy estimates of ability.

In modern assessment systems, test scores are rarely that simple. Instead, they are often the product of statistical IRT models designed to extract more information from the pattern of responses. For instance, [Jacob and Rothstein \(2016\)](#) discuss that IRT test scores in many assessment systems in the US, such as the ECLS, the NELS:88, the ELS, and the HSLs are constructed from posterior means. Let  $s_t^m$  denote the test score, which is then computed by

$$s_t^m = \mathbb{E}[s_t | \tilde{s}_t] = \int s_t f(s_t | \tilde{s}_t) ds_t \quad (2)$$

$$\simeq (1 - \tilde{\lambda})\mathbb{E}[\tilde{s}_t] + \tilde{\lambda}\tilde{s}_t, \quad \forall t. \quad (3)$$

Students are treated as a random sample from a population of ability values  $s_t$ , characterized by the posterior density function  $f(s_t | \tilde{s}_t)$  from the IRT model. The posterior mean score is the mean of that population. Hence, instead of using the signal as the test score directly, the signal is used as a predictor for ability to generate a test score. This prediction is not perfect, so we can similarly write that the test score equals ability plus the score's measurement error,  $s_t^m = s_t + m_t^m$ .

Posterior mean scores can be interpreted as Empirical Bayes estimates that, roughly, shrink the student's (unbiased) maximum likelihood score toward the population mean in proportion to the noisiness of the maximum likelihood score (Jacob and Rothstein, 2016). Building on this interpretation, we approximate the posterior mean in (2) by the Kelley (1947) estimator in (3), where  $\tilde{\lambda}$  is the shrinkage factor. With  $\tilde{\lambda} \in [0, 1]$ , students with low (high) ability signal are pushed up (down) towards the population mean. As a result, the measurement error of the test score  $m_t^m$  correlates negatively with ability. This feature of the non-classical error in (3) is similar to the error generated by any posterior mean in (2).

We will analyze contamination bias in  $\beta$  when replacing ability in (1) by test scores measured through (3). Our results hold (approximately) for two types of test scores. The first are test scores estimated classically or through an IRT model using (weighted) maximum likelihood under a “fixed effect” approach. Both provide an unbiased estimate of ability. This corresponds to (3) with  $\tilde{\lambda} = 1$ : the test score is equal to the signal,  $s_t^m = \tilde{s}_t$ , and measurement error is classical,  $m_t^m = m_t$ . The second are test scores estimated from an IRT model under a “random-effects” approach, where the scores are posterior means. Equation (3) holds exactly only for the Kelley estimator, which corresponds to normality for both the prior and likelihood distributions. Generally, it is an approximation, and depending on the choice of these two distributions, the posterior mean in (2) may even lack a closed form. Because the Kelley estimator corresponds to an OLS regression of  $s_t$  on  $\tilde{s}_t$ , by the properties of OLS it provides the best linear approximation to the (possibly nonlinear) posterior mean.

Our results speak less to settings where test scores are reported as posterior medians or modes, where multiple plausible values are drawn from a student's posterior distribution, or where the prior distribution on ability incorporates student characteristics. The latter approach is used in the major US and international assessments NAEP and PISA. An alternative strategy is to jointly estimate the research model in (1) alongside the IRT model, treating the

latter as a direct model of measurement (Junker et al., 2012). This requires access to item-level response data, which is typically unavailable. IRT models often produce noisier test scores at the tails of the distribution, reflecting the limited information available to these parametric methods when student ability does not align well with item difficulty. Our approach allows for the potential difference between so-called local and global reliability.

Non-classical measurement generated by (3) with  $\tilde{\lambda} \in [0, 1)$  is also considered by Hyslop and Imbens (2001) in survey data. They consider a setting in which respondents are asked questions like “what is the value of  $s_t$ ?” given an information set that consists of the unbiased but noisy signal  $\tilde{s}_t$ . Respondents may passively report the “flawed” but unbiased value ( $\tilde{s}_t$ ), or may actively seek to provide an “optimal” response ( $s_t^m$ ) given their information set. An argument for the latter is that respondents are likely aware of the lack of precision in the signal. Our results on the contamination bias below may thus also be relevant for the analysis of survey data more generally.

## 4.2 Contamination bias

The medium regression equation replaces  $s_t$  with  $s_t^m$ ,

$$y_t = \beta^m SES + \gamma^m s_t^m + e_t^m. \quad (4)$$

To analyze what  $\beta^m$  and  $\gamma^m$  identify, we also introduce the balancing regression equation,

$$\begin{pmatrix} s_t \\ s_t^m \end{pmatrix} = \begin{pmatrix} s_t \\ \tilde{\lambda}(s_t + m_t) \end{pmatrix} = \begin{pmatrix} \delta \\ \delta^m \end{pmatrix} SES + \begin{pmatrix} u_t \\ u_t^m \end{pmatrix}. \quad (5)$$

Note that  $(1 - \tilde{\lambda})\mathbb{E}[\tilde{s}_t]$  is fixed and gets soaked up by the intercept of the regression. The following proposition formalizes what OLS identifies.

**Proposition 1.** *Under Assumption 1 and Equation (3), it holds that:*

$$\delta^m = \delta \tilde{\lambda}, \quad (6)$$

$$\gamma^m = \gamma \left( \frac{\lambda}{\tilde{\lambda}} \right), \quad (7)$$

$$\beta^m = \beta + \gamma \delta (1 - \lambda). \quad (8)$$

With  $\lambda = \left( \frac{\sigma_{u_t}^2}{\sigma_{u_t}^2 + \sigma_{m_t}^2} \right)$ , where  $\sigma_{\bullet}^2$  denotes the variance of the variable in the subscript.

All proofs are provided in Appendix A to C. First consider (6) and (7) under classical measurement error, where  $\tilde{\lambda} = 1$ . Classical error on the left-hand side does not introduce bias in the estimate of  $\delta^m$ . However, as shown in (7), the estimate of  $\gamma^m$  is subject to the well-known attenuation bias. The multivariate reliability (or signal-to-total variance) ratio of the test, denoted by  $\lambda$ , decreases from one toward zero as the measurement error variance increases, leading to attenuation bias in  $\gamma^m$ . Now consider the case of non-classical measurement error, where  $\tilde{\lambda} \in [0, 1)$ . Equation (7) shows that the attenuation bias in  $\gamma^m$  may be smaller and disappears completely when the shrinkage factor equals the reliability ratio,  $\tilde{\lambda} = \lambda$ . However, (6) shows that non-classical error on the left-hand side results in attenuation of  $\delta^m$ .

Equation (8) clarifies that both types of measurement error contaminate the estimate of  $\beta^m$ . Under classical measurement error, attenuation of  $\gamma^m$  spills over to  $\beta^m$ . Under non-classical measurement error, the attenuation bias in  $\gamma^m$  may be corrected, but at the expense of attenuation in  $\delta^m$ , which again spills over to  $\beta^m$ . Importantly, the contamination is identical for both types of measurement error, as the bias term involves the product of  $\gamma^m$  and  $\delta^m$ , where the former is divided and the latter is multiplied by  $\tilde{\lambda}$ . Although this may seem intuitive, to our knowledge this result has not been reported in the literature. In particular, [Hyslop and Imbens \(2001\)](#) analyze (6) and (7) under both classical and non-classical measurement error, while [Pei et al. \(2019\)](#); [Gillen et al. \(2019\)](#); [Ferman and Fontes \(2022\)](#) examine (8) in the case of classical measurement error only. Note that the contamination bias grows larger as  $\gamma$  and  $\delta$  increase and  $\lambda$  decreases.

### 4.3 The IV strategy

To address measurement error, the instrumental variables (IV) strategy aims to use a second objective measure of ability in period  $t$  as an instrument for the first measure. In practice, a lagged test score from period  $t - 1$  is often used instead ([Botelho et al., 2015](#); [Ferman and Fontes, 2022](#); [Zhu, 2024](#)). Similar to above, students produce an unbiased but noisy signal of ability by taking a test in  $t - 1$ , where  $\tilde{s}_{t-1} = s_{t-1} + m_{t-1}$ . The corresponding test score satisfies (3), where  $s_{t-1}^m = \mathbb{E}[s_{t-1} | \tilde{s}_{t-1}]$  and  $s_{t-1}^m = s_{t-1} + m_{t-1}^m$ . We do not index the shrinking parameter  $\tilde{\lambda}$  by time, thereby assuming it remains constant across tests. Our main results hold without this simplification, though at the expense of additional notational clutter and

reduced intuition. Using  $s_{t-1}^m$  as an instrument for  $s_t^m$  generates the following first and second stage regression equations respectively,

$$s_t^m = \kappa SES + \pi s_{t-1}^m + e_t^{fs}, \quad (9)$$

$$y_t = \beta^{iv} SES + \gamma^{iv} \hat{s}_t^m + e_t^{iv}. \quad (10)$$

To show how this IV strategy can address measurement error, it is useful to also introduce the balancing regression for  $t - 1$ ,

$$\begin{pmatrix} s_{t-1} \\ s_{t-1}^m \end{pmatrix} = \begin{pmatrix} s_{t-1} \\ \tilde{\lambda}(s_{t-1} + m_{t-1}) \end{pmatrix} = \begin{pmatrix} \delta_{t-1} \\ \delta_{t-1}^m \end{pmatrix} SES + \begin{pmatrix} u_{t-1} \\ u_{t-1}^m \end{pmatrix}, \quad (11)$$

and the variable  $\Delta_{s_t} = s_t - s_{t-1}$ , which captures the change in unobserved ability between time period  $t$  and  $t - 1$ . We describe two additional assumptions that place increasing restrictions on  $\Delta_{s_t}$ .

**Assumption 2.** *The covariance between  $\Delta_{s_t}$ , SES and  $s_t$  is zero*

a.  $(SES) \text{Cov}[\Delta s_t, SES] = 0.$

b.  $(Ability) \text{Cov}[\Delta s_t, s_t] = 0.$

Assumption 2(a) requires the change in ability to be the same, on average, across low and high SES students. Assumption 2(b) requires this for low and high ability students in time  $t$ .

**Assumption 3.** *Constant  $\Delta_{s_t}$*

*(Constant)  $\Delta s_t$  is the same for each student.*

Assumption 3 requires that the change in ability is the same for each student, which is stronger than Assumption 2. The following proposition summarizes what the IV strategy identifies while combining Assumption 1 with either 2 or 3.

**Proposition 2.** *(i) Under Assumption 1 and 2, and Equation (3), it holds that:*

$$\pi' = \lambda, \quad (12)$$

$$\gamma^{iv} = \gamma \left( \frac{1}{\tilde{\lambda}} \right) B_\gamma, \quad (13)$$

$$\beta^{iv} = \beta + \gamma \delta (1 - B_\gamma). \quad (14)$$

With  $\pi' = \pi \left( \frac{\sigma_{u_t}^2 + \sigma_{\Delta s_t}^2 + \sigma_{m_{t-1}}^2}{\sigma_{u_t}^2 + \sigma_{m_t}^2} \right)$  and  $B_\gamma = 1 - \left( \frac{\text{Cov}[e_t, \Delta s_t]}{\text{Cov}[y_t, u_t]} \right)$ .

(ii) Under Assumption 1 and 3, and Equation (3), it holds that:

$$\pi' = \lambda, \quad (15)$$

$$\gamma^{iv} = \gamma \left( \frac{1}{\lambda} \right), \quad (16)$$

$$\beta^{iv} = \beta. \quad (17)$$

With  $\pi' = \pi \left( \frac{\sigma_{u_t}^2 + \sigma_{m_{t-1}}^2}{\sigma_{u_t}^2 + \sigma_{m_t}^2} \right)$ .

Under classical measurement error with  $\tilde{\lambda} = 1$ , Proposition 2 formalizes that the IV strategy yields consistent estimates under Assumption 3. This result is not new. For instance, [Gillen et al. \(2019\)](#) use IV to address contamination bias from classical error in (17) and review similar applications in the literature. Intuitively, a constant  $\Delta s_t$  implies that the test score in period  $t - 1$  is truly a second objective measure of ability in period  $t$ . The first stage regression of  $s_t^m$  upon  $s_{t-1}^m$  measures to which extent this relationship is diluted due to measurement error, so that it reveals the reliability ratio  $\lambda$ . Furthermore, the reduced form equation is similar to the medium regression equation (4) while replacing  $s_t^m$  with  $s_{t-1}^m$ . With a constant  $\Delta s_t$  the reduced form coefficient reveals the estimate  $\gamma^m = \gamma\lambda$ , which was attenuated towards zero by the reliability ratio. Dividing the reduced form by the first stage corrects for this attenuation. Since  $\gamma^{iv} = \gamma$  and  $\delta^m = \delta$ , we also have that  $\beta^{iv} = \beta$ .

In the presence of non-classical error with  $\tilde{\lambda} \in [0, 1)$ , the IV strategy also addresses contamination bias under Assumption 3, though for a different reason. To illustrate, suppose the shrinkage factor equals the reliability ratio, such that  $\tilde{\lambda} = \lambda$ . In this case, the OLS coefficient on test scores  $\gamma^m$  does not suffer from attenuation bias. However, the IV strategy still divides by the first stage, which means that the IV coefficient  $\gamma^{iv}$  is overestimated by a factor equal to the inverse of the reliability ratio. As noted by [Hyslop and Imbens \(2001\)](#), IV has a bias away from zero in the presence of non-classical measurement error. However, this upward bias in  $\gamma^{iv}$  is exactly what is needed to correct for the contamination bias to  $\beta^{iv}$ , which arises from attenuation in the estimate of  $\delta^m$  from the balancing regression. Since the contamination bias term involves the product of the test score and balancing coefficients, the upward bias in the former offsets the downward bias in the latter. As a result, the IV



strategy also effectively addresses the contamination bias in the presence of non-classical measurement error.

#### 4.4 The EIV strategy

Based on the results of Proposition 1, the errors-in-variables (EIV) strategy directly formulates expressions for the parameters of interest,

$$\gamma^{eiv} = \frac{\gamma^m}{\lambda}, \quad (18)$$

$$\beta^{eiv} = \beta^m - \gamma^m \delta^m \left( \frac{1 - \lambda}{\lambda} \right). \quad (19)$$

All the elements on the right-hand side are known under Assumption 1, except for the reliability ratio. The EIV strategy then plugs a reliable estimate for  $\lambda$  into (18) and (19) to estimate the parameters of interest.

##### 4.4.1 The first stage

Proposition 2 shows that the first stage can be used as an estimate for  $\lambda$  under Assumption 2. The intuition for this result follows more easily in a univariate setting without  $SES$  as an independent variable. The univariate reliability ratio is defined by replacing  $\sigma_{u_t}^2$  in  $\lambda$  with  $\sigma_{s_t}^2$ . Assumption 2(b) implies that  $\text{Cov}[s_t, s_{t-1}] = \sigma_{s_t}^2$ , so the coefficient from the univariate first stage regression of  $s_t^m$  on  $s_{t-1}^m$  is proportional to the univariate reliability ratio. This result extends to the multivariate setting since together Assumption 2(a)-(b) imply that  $\text{Cov}[u_t, u_{t-1}] = \sigma_{u_t}^2$ .

Our EIV first stage (EIV FS) strategy uses the first stage as an estimate for  $\lambda$ . The following proposition formalizes what EIV FS identifies under Assumption 1 and 2.

**Proposition 3.** *Under Assumption 1 and 2, and Equation (3), it holds that:*

$$\gamma^{eiv} = \gamma \left( \frac{1}{\hat{\lambda}} \right), \quad (20)$$

$$\beta^{eiv} = \beta. \quad (21)$$

Where  $\gamma^{eiv}$  and  $\beta^{eiv}$  are defined by replacing  $\lambda$  with  $\pi'$  in (18) and (19).

Similar to the IV strategy, EIV FS corrects attenuation bias in the test score coefficient under classical measurement error, overestimates it under non-classical error, and addresses contamination bias in the SES coefficient in both cases. Unlike the IV strategy, EIV FS achieves these results under the weaker Assumption 2. It relies directly on the medium regression estimate, whereas the IV strategy uses the reduced form estimate. The two are equivalent only if  $\text{Cov}[e_t, \Delta_{s_t}] = 0$ , which holds when  $\Delta_{s_t}$  is constant. Whereas Assumption 3 only seems likely when the two tests are administered arbitrarily shortly after one another, Assumption 2 appears generally more plausible. In particular, under Assumption 2 the teachers may pay attention to  $\Delta_{s_t}$  in their subjective evaluations, so that it is part of the error term  $e_t$ , but  $\Delta_{s_t}$  cannot correlate with the included variable  $SES$  and unobserved ability  $s_t$ .

#### 4.4.2 Test score history

We propose a novel EIV test score history (EIV TH) strategy that recovers the reliability ratio under Assumption 1 only, while making use of data on students' test score histories. The intuition of this strategy is to approximate the ideal conditions of the test-retest method, which estimates the reliability ratio using a test administered at time  $t - \epsilon$ , arbitrarily short before time  $t$  (Cito, 2017). With this ideal test from period  $t - \epsilon$ , the first stage equation is,

$$s_t^m = \kappa_{t-\epsilon} SES + \pi_{t-\epsilon} s_{t-\epsilon}^m + e_t^{f s_\epsilon}, \quad (22)$$

where we include a time subscript on the coefficients. For a test administered arbitrarily close to time  $t$ , Assumption 3 holds trivially, and Proposition 2 shows that  $\pi'_{t-\epsilon} = \lambda$ . While a test score that close to time  $t$  may not be available, we may observe multiple test scores further away. It is plausible that such scores satisfy Assumption 1, even if Assumption 2 and 3 may not hold. Nevertheless, the EIV TH strategy leverages these test scores observed at earlier points in time to identify the reliability ratio.

We begin by repeating the first stage equation, sequentially replacing the test score from period  $t - 1$  on the right-hand side with scores from earlier periods ,

$$s_t^m = \kappa_{t-\tau} SES + \pi_{t-\tau} s_{t-\tau}^m + e_t^{f s_\tau}, \quad \forall \tau \geq 1. \quad (23)$$

It follows from the proof of Proposition 2 that under Assumption 1 only,  $\pi'_{t-\tau}$  is a biased

estimate for the reliability ratio,

$$\pi'_{t-\tau} = \lambda B_{\pi_{t-\tau}}, \quad \forall \tau \geq 1, \quad (24)$$

with  $B_{\pi_{t-\tau}} = \left( \frac{\text{Cov}[u_t, u_{t-\tau}]}{\sigma_{u_t}^2} \right)$ . Therefore, we use  $\pi'_{t-\tau}$  as a dependent variable in a forecasting regression equation of the form,

$$\pi'_{t-\tau} = f(t - \tau) + \varepsilon_{t-\tau}. \quad (25)$$

Subsequently, we make a one-period out-of-sample forecast towards  $\tau = \epsilon$ , denoted by  $\hat{f}(t - \epsilon)$ . Without forecasting error this resembles the ideal conditions of the test-retest method, such that  $\hat{f}(t - \epsilon) = f(t - \epsilon) = \pi'_{t-\epsilon} = \lambda$ . In the absence of forecasting error, the chosen polynomial in (25) accurately captures the changes in  $\pi'_{t-\tau}$  over time. While the EIV FS strategy estimates  $\gamma^{eiv}$  and  $\beta^{eiv}$  by replacing  $\lambda$  with  $\pi'$  in (18) and (19), EIV TH estimates them by replacing  $\lambda$  with  $f(t - \epsilon)$ .

## 5 Results

The baseline OLS estimates are presented in columns (1) and (2) of Table 2. Column (1) reports estimates from a regression where the dependent variable is a dummy equal to one if the teacher's track recommendation is at or above the college track, and the independent variables are parental income (SES) and the fifth-grade end term test score. Consistent with previous evidence from the Netherlands, the findings show that students with higher income parents are more likely to receive at least a college track recommendation, conditional on the fifth-grade end term test. The estimate for the conditional SES gap implies that a one standard deviation increase in parental income is associated with a 2.8 percentage point increase in receiving at least a college track recommendation.

Column (1) further shows that test scores are positively associated with teacher track recommendations. As our test scores are subject to classical measurement error, this estimate is attenuated. Whether and how this contaminates the OLS estimate for the conditional SES gap depends upon the relationship between test scores and SES. This relationship is presented in column (2). As expected, SES is positively associated with the fifth-grade end term test score. We conclude that OLS overestimates the conditional SES gap, where the magnitude of this positive contamination bias also depends on the reliability ratio of the test.

Table 2: OLS and IV estimates

|  | OLS strategy  |                     |                     | IV strategy   |   |
|--|---|---------------------|---------------------|---|---|
|  | (1)   | (2)                 | (3)                 | (4)   | (5)   |
|  | $\mathbb{1}(\text{recom.} \geq \text{college tr.})$ | End term grade 5    | End term grade 5    | $\mathbb{1}(\text{recom.} \geq \text{college tr.})$ | $\mathbb{1}(\text{recom.} \geq \text{college tr.})$ |
| SES income   | 0.028***<br>(0.001)                                 | 0.229***<br>(0.004) | 0.031***<br>(0.001) | 0.030***<br>(0.001)                                 | 0.016***<br>(0.001)                                 |
| End term grade 5   | 0.381***<br>(0.001)                                 |                     |                     |   | 0.432***<br>(0.002)                                 |
| Midterm grade 5  |   |                     | 0.887***<br>(0.002) | 0.383***<br>(0.001)                                 |   |
| Observations   | 148019  | 148019              | 148019              | 148019  | 148019  |
| $R^2$  | 0.498   | 0.064               | 0.794               | 0.500   | 0.490   |
| $\left( \frac{\sigma_{u_{t-1}}^2 + \sigma_{m_{t-1}}^2}{\sigma_{u_t}^2 + \sigma_{m_t}^2} \right)$ |   |                     | 0.992               |   |   |

Notes: \*\*\*, \*\*, \* refers to statistical significance at the 1, 5, and 10% level. Coefficients in columns (1) through (4) are estimated with OLS, and in column (5) with 2SLS. Standard errors (in parentheses) are clustered at the school level.

Table 4 in Appendix D shows OLS estimates of the conditional SES gap while including additional control variables: gender, migration dummies, age, and cohort fixed effects. The estimated gap only changes marginally, from 0.028 to 0.029. It may seem unsurprising that the estimate changes little, given that most included controls correlate weakly with parental income, with migration background as a clear exception. In general, these results suggest that our main estimates are robust to controlling for additional variables typically observed in administrative data.

## 5.1 The IV strategy

Columns (3) through (5) of Table 2 show the results of the IV strategy, which uses the fifth-grade midterm test as an instrument for the fifth-grade end term test. The first stage in

column (3) shows that the estimated coefficient on the midterm test is equal to 0.887. This estimate measures the reliability ratio after multiplication with the fraction containing the variances of the error terms from the balancing regression. Hence, the first stage estimate for the reliability ratio is  $0.887 \times 0.992 = 0.880$ .

The IV strategy addresses measurement error by using the first stage to correct the reduced form estimate on the midterm term test shown in column (4). The second stage estimate on the end term test in column (5) can be calculated by dividing the reduced form by the first stage,  $\frac{0.383}{0.887} = 0.432$ . The corresponding IV estimate for the conditional SES gap is equal to 0.016, which implies that a one standard deviation increase in parental income is associated with a 1.6 percentage point increase in receiving at least a college track recommendation.

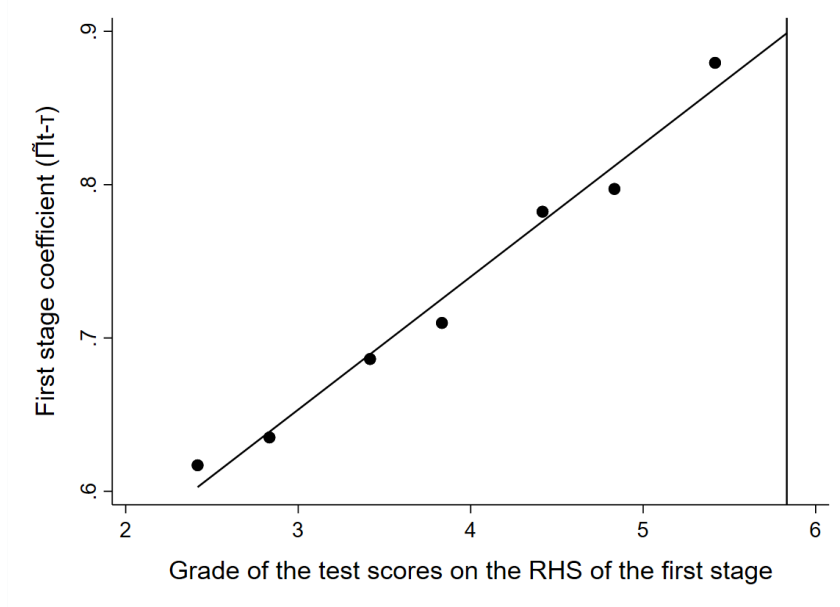
## 5.2 The EIV strategy

The EIV strategy requires a value for the reliability ratio of the end term test. From the IV strategy we know that the first stage estimate for the reliability ratio is 0.880. Reliability ratios near 0.90 are considered high (Cito, 2017). Depending on the topic, the GMAT is advertised to have a reliability ratio between 0.89 and 0.90 (GMAT, 2023), the GRE between 0.87 and 0.95 (ETS, 2023), and the SAT between 0.89 and 0.93 (College Board, 2013).

The EIV FS strategy uses the first stage as an estimate for the reliability ratio to directly address the bias in the OLS estimates. In particular, with the reliability ratio equal to 0.880, the OLS estimate on the fifth-grade end term test is attenuated by a factor of 0.880, and the conditional SES gap is overestimated by a value of  $\gamma^m \delta^m \left( \frac{1-\lambda}{\lambda} \right) = 0.381 \times 0.229 \times \left( \frac{1-0.880}{0.880} \right) = 0.012$ . Hence, the EIV FS estimates are equal to  $\left( \frac{0.381}{0.880} \right) = 0.433$  and  $0.028 - 0.012 = 0.016$ , respectively.

The novel EIV TH strategy uses an alternative estimate for the reliability ratio, which is obtained by using the biannual test score data throughout grade two to five. Figure 1 presents the estimates from the first stage regressions that separately use the midterm and end term tests from grade two to five as independent variable. Instead of using the right-most estimate on the fifth-grade midterm test as the reliability ratio, the EIV TH approach relies on the first stage estimates of all previous test scores to predict the first stage estimate

Figure 1: Estimate for the reliability ratio using students' test score histories



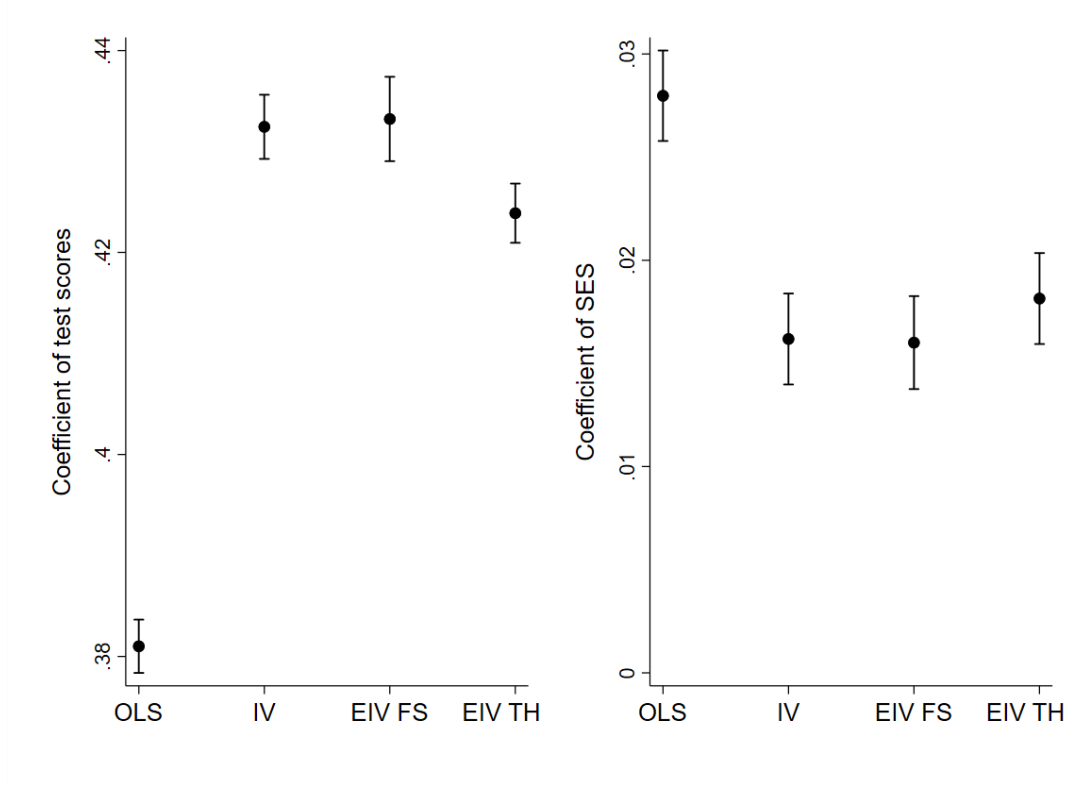
Notes: this figure plots the estimates for  $\pi'_{t-\tau}$  and the forecasting regression used to estimate  $\lambda$ . We use OLS to estimate  $\pi_{t-\tau}$  from (23) and the variance of the error term from (5), to obtain  $\pi'_{t-\tau}$  for each  $\tau$ . The forecasting regression equation (25) is specified as a polynomial of degree one and estimated with OLS.

of a test score very close to the fifth-grade end term test. This approximates the estimate obtained under the ideal conditions of the test-retest method. We use a linear polynomial for the forecasting regression equation. With an  $R^2$  of 0.981 this provides a good fit for the development of the first stage estimates. The estimate for the reliability ratio is the one-period out-of-sample forecast and is equal to 0.899. The EIV TH strategy produces an estimate for the fifth-grade end term test of 0.424 and an estimate for the conditional SES gap of 0.018.

### 5.3 Comparing results

Figure 2 shows the estimates with 95% confidence intervals for the OLS strategy together with all three strategies that aim to address measurement error. The left figure presents the estimates for the end term test and visualizes the attenuation bias of OLS with classical measurement error. The IV and EIV FS strategy inflate the test score estimate by a comparable amount, since the reduced form estimate on the midterm test scores ( $0.383 \times 0.992 = 0.380$ ) is similar to the medium regression estimate on the end term test scores (0.381). The right

Figure 2: OLS, IV, and EIV estimates



Notes: this figure shows the estimated coefficients for the OLS, IV, and EIV strategy. The lines are 95% confidence intervals. The estimates and standard errors for the OLS and IV strategy can be found in Table 2. The estimates and standard errors for EIV FS are obtained via a stacking approach. We triplicate the data and estimate (4), (5), and (9) with OLS via a single regression. We cluster standard errors at the school id that is similarly triplicated. For EIV TH we use the same stacking approach to estimate (4) and (5), while  $\lambda$  is estimated from the forecasting regression equation (25) (see Figure 1 for details) and considered fixed.

figure presents the estimates for the conditional SES gap. Similar attenuation bias translates into a similar contamination bias, and so the IV and EIV FS strategy also produce essentially the same estimate for the conditional SES gap. The attenuation and contamination bias are somewhat smaller for the EIV TH strategy, since the estimate for the reliability ratio is somewhat higher when using the one-period out-of-sample forecast.

The formula  $\left( \frac{\beta^m - \beta^*}{\beta^m} \right) \times 100$  quantifies the share of the conditional SES gap in track recommendations attributable to measurement error. Here,  $\beta^m$  is the OLS estimate and  $\beta^*$  refers to the IV and EIV estimates. The IV strategy attributes 42.14% of the SES gap to measurement error, while the EIV strategies yield estimates ranging from 35.13% to 42.78%, depending on

how the reliability ratio is estimated. Overall, the results highlight the importance of addressing contamination bias and that, in our setting, the three strategies produce broadly consistent findings.

## 5.4 Testing assumptions

Assumption 2(a) and 3 have direct testable implications. Define  $\Delta_{s_t^m} = s_t^m - s_{t-1}^m$  as the change in grade 5 test scores, and let  $\alpha_x = \left( \frac{\text{Cov}[\Delta_{s_t^m}, x]}{\sigma_x^2} \right)$  be the OLS estimator of a model that regresses  $\Delta_{s_t^m}$  upon a single student characteristic  $x$ . The classical measurement error in our test scores, when they appear on the left-hand side, does not affect the OLS estimate. Hence, Assumption 2(a) requires that  $\alpha_{SES}$  is zero and Assumption 3 requires that  $\alpha_x$  is zero for every student characteristic  $x$ , including  $SES$ .

Table 5 in Appendix D shows OLS estimates of the change in grade 5 test scores on  $SES$  and all additional control variables. Column (1) shows that the estimate of  $SES$  is statistically significant at the 1% level. Although we can reject Assumption 2(a), the  $SES$  estimate of 0.005 seems economically small. In particular, the estimate is reduced by 97.82% compared to the  $SES$  estimate from the regression of the fifth-grade end term test score on  $SES$  in Table 2. Column (2) to (5) further show that the gender dummy, the immigrant dummies, and age also correlate with the change in test scores at the 1% significance level. The coefficients for the immigrant dummies seem of non-negligible size. These results cast doubt on Assumption 3. In general, Assumption 3 seems unlikely in settings where the difference between time period  $t$  and  $t - 1$  is large, especially if the variable is highly dynamic.

Though it is difficult to empirically test Assumption 2(b), one may discuss its implications for the change in ability over time. In particular, if  $\Delta_{s_t}$  is uncorrelated with  $s_t$ , the following two equalities directly follow:

$$-\sigma_{\Delta_{s_t}}^2 = \text{Cov}[\Delta_{s_t}, s_{t-1}], \quad (26)$$

$$\sigma_{\Delta_{s_t}}^2 = \sigma_{s_{t-1}}^2 - \sigma_{s_t}^2. \quad (27)$$

Low ability students in  $t - 1$  must have had larger positive changes in ability than high ability students in  $t - 1$  (from (26)) and the variance of ability must be larger in  $t - 1$  than in  $t$  (from (27)). Whether these equalities are likely depends on the setting at hand. In an education context, (26) and (27) are consistent with diminishing marginal returns to ability.



The EIV TH strategy also relies, in addition to Assumption 1, on an accurate prediction of changes in the first stage estimate over time. Figure 1 supports this, showing that a linear polynomial explains 98.1% of the variation in the first stage estimates. This also suggests that the factors driving the first stage away from the reliability ratio, which are  $\text{Cov}[\Delta s_t, SES] \neq 0$  and  $\text{Cov}[\Delta s_t, s_t] \neq 0$ , diminish over time in a predictable way. This is further supported by results presented in Figure 3 in Appendix D, which plots OLS estimates of  $\Delta s_t^m$  on  $SES$  across time and shows that  $\text{Cov}[\Delta s_t^m, SES]$  converges to zero in an approximately linear fashion.

## 5.5 Robustness checks

We test the robustness of our findings in four ways. First, Figure 4 in Appendix D shows the results when using two alternative measures for the teacher track recommendation: a dummy that equals one if the teacher recommendation is equal to or higher than the theory-oriented vocational track (Figure 4(a)) or the university track (Figure 4(b)). Second, Figure 5 in Appendix D shows the results when using three alternative SES measures: a dummy that equals one if parental income is above the median of a student’s cohort (Figure 5(a)), years of schooling of the highest educated parent (Figure 5(b)), and a dummy that equals one if the highest educated parent completed at least college education (Figure 5(c)). Third, Figure 6 in Appendix D shows the results when separately estimating the models per cohort (Figure 6(a)-(d)). All results are similar to our baseline findings.

Fourth, as SES is strongly correlated with migration background, a potential concern is that the conditional SES gap at least partly reflects a gap by migration background. To test whether this is the case, we estimated our baseline models for immigrants and natives separately. Figure 7 in Appendix D shows that the SES estimates are similar in both subgroups.

## 6 Conclusion

Teacher bias is typically assessed by examining gaps in teacher’s subjective evaluations, conditional on objective measures of student abilities. A key challenge in identifying such bias is the presence of measurement error in test scores, which are used as proxies for student ability. We show that both classical and certain forms of non-classical measurement error introduce equivalent bias in the estimated conditional gap in subjective evaluations. We

discuss three different strategies to address this contamination bias. While an IV approach is more commonly used in applied econometric studies, we demonstrate that an EIV approach can address measurement error under weaker assumptions.

The empirical analysis focuses on conditional SES gaps in teacher track recommendations in the Netherlands. This is a setting where potential teacher bias is highly consequential and findings on conditional gaps with respect to parental SES are widely accepted. Using new administrative data, we find that measurement error in test scores can explain a substantial portion of the conditional SES gap, between 35 and 43%.

It may seem striking that measurement error explains such a large share of the conditional SES gap, especially since our test scores are reliable by conventional norms. Yet when student ability is strongly correlated with SES and matters for teacher track recommendations, contamination bias can still be substantial. While our findings show that teacher bias may be smaller than previously reported, the strong correlation between ability and SES points to significant inequalities of opportunity before and during primary education.

## References

- Alesina, A., M. Carlana, E. La Ferrara, and P. Pinotti (2024). Revealing stereotypes: Evidence from immigrants in schools. *American Economic Review* 114(7), 1916–1948.
- Borghans, L., R. Diris, W. Smits, and J. de Vries (2019). The long-run effects of secondary school track assignment. *PLOS ONE* 14(10), 1–29.
- Borjas, G. J. and D. S. Hamermesh (2024). The mismeasurement of work time: Implications for wage discrimination and inequality. Working Paper 32025, NBER.
- Botelho, F., R. A. Madeira, and M. A. Rangel (2015). Racial discrimination in grading: Evidence from Brazil. *American Economic Journal: Applied Economics* 7(4), 37–52.
- Burgess, S. and E. Greaves (2013). Test scores, subjective assessment, and stereotyping of ethnic minorities. *Journal of Labor Economics* 31(3), 535–576.
- Carlana, M. (2019). Implicit stereotypes: Evidence from teachers’ gender bias. *The Quarterly Journal of Economics* 134(3), 1163–1224.

- Carlana, M., E. La Ferrara, and P. Pinotti (2022). Implicit stereotypes in teachers' track recommendations. *AEA Papers and Proceedings* 112, 409–14.
- Cito (2017). Rekenen-wiskunde 3.0 voor groep 7. Centraal Instituut voor Toetsontwikkeling. Wetenschappelijke verantwoording.
- College Board (2013). Test characteristics of the SAT: Reliability, difficulty levels, completion rates. <https://secure-media.collegeboard.org/digitalServices/pdf/research/Test-Characteristics-of-SAT-2013.pdf>, Accessed August 2023.
- Cornwell, C., D. B. Mustard, and J. Van Parys (2013). Noncognitive skills and the gender disparities in test scores and teacher assessments: Evidence from primary school. *Journal of Human resources* 48(1), 236–264.
- de Ree, J., M. Oosterveen, and D. Webbink (2023). The quality of school track assignment decisions by teachers. Working paper, arXiv preprint arXiv:2304.10636.
- DUO (2023). Onderzoek LVS en eindtoets. Dienst Uitvoering Onderwijs. Rapportage.
- Dustmann, C., P. A. Puhani, and U. Schönberg (2017). The long-term effects of early track choice. *The Economic Journal* 127(603), 1348–1380.
- Eggen, T. and P. Sanders (1993). *Psychometrie in de praktijk*. Cito, Arnhem.
- ETS (2023). Guide to the use of scores. Educational Testing Service. <https://www.ets.org/pdfs/gre/gre-guide-to-the-use-of-scores.pdf>, Accessed August 2023.
- Falk, A., E. Kosse, and P. Pinger (2025). Mentoring and schooling decisions: Causal evidence. *Journal of Political Economy*, forthcoming.
- Ferman, B. and L. F. Fontes (2022). Assessing knowledge or classroom behavior? Evidence of teachers' grading bias. *Journal of Public Economics* 216, 104773.
- Fuller, W. A. (2009). *Measurement error models*. John Wiley & Sons.
- Gillen, B., E. Snowberg, and L. Yariv (2019). Experimenting with measurement error: Techniques with applications to the Caltech Cohort Study. *Journal of Political Economy* 127(4), 1826–1863.

- GMAT (2023). Validity, reliability & fairness. Graduate Management Admission Council. <https://www.gmac.com/gmat-other-assessments/about-the-gmat-exam/validity-reliability-fairness>, Accessed August 2023.
- Haelermans, C., T. Huijgen, M. Jacobs, M. Levels, R. van der Velden, L. van Vugt, and S. van Wetten (2020). Using data to advance educational research, policy, and practice: Design, content, and research potential of the Netherlands Cohort Study on Education. *European Sociological Review* 36(4), 643–662.
- Hausman, J. (2001). Mismeasured variables in econometric analysis: Problems from the right and problems from the left. *Journal of Economic Perspectives* 15(4), 57–67.
- Hyslop, D. R. and G. W. Imbens (2001). Bias from classical and other forms of measurement error. *Journal of Business & Economic Statistics* 19(4), 475–481.
- Imai, K. and Z. Jiang (2023). Principal fairness for human and algorithmic decision-making. *Statistical Science* 38(2), 317–328.
- Inspectorate of Education (2016). De staat van het onderwijs: Onderwijsverslag 2014/2015. Technical report, Inspectie van het Onderwijs.
- Inspectorate of Education (2019). Plaatsing en toelating in het VO. Technical report, Inspectie van het Onderwijs.
- Jacob, B. and J. Rothstein (2016). The measurement of student ability in modern assessment systems. *Journal of Economic Perspectives* 30(3), 85–108.
- Junker, B., L. S. Schofield, and L. J. Taylor (2012). The use of cognitive ability measures as explanatory variables in regression analysis. *IZA Journal of Labor Economics* 1(1), 4.
- Kelley, T. L. (1947). *Fundamentals of statistics*. Harvard University Press.
- Lavy, V. and E. Sand (2018). On the origins of gender gaps in human capital: Short-and long-term consequences of teachers' biases. *Journal of Public Economics* 167, 263–279.
- Matthewes, S. H. (2021). Better together? Heterogeneous effects of tracking on student achievement. *The Economic Journal* 131(635), 1269–1307.

- Ministry of Education (2023). De overstap van PO naar VO: Handreiking schooladvisering. Ministerie van Onderwijs, Cultuur en Wetenschap.
- Modalsli, J. and K. Vosters (2024). Spillover bias in multigenerational income regressions. *Journal of Human Resources* 59(3), 743–776.
- Papageorge, N. W., S. Gershenson, and K. M. Kang (2020). Teacher expectations matter. *Review of Economics and Statistics* 102(2), 234–251.
- Pei, Z., J.-S. Pischke, and H. Schwandt (2019). Poorly measured confounders are more useful on the left than on the right. *Journal of Business & Economic Statistics* 37(2), 205–216.
- Schennach, S. M. (2016). Recent advances in the measurement error literature. *Annual review of economics* 8(1), 341–377.
- Terrier, C. (2020). Boys lag behind: How teachers’ gender biases affect student achievement. *Economics of Education Review* 77, 101981.
- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika* 54(3), 427–450.
- WPO (2014). Artikel 42, lid 2, Wet op het primair onderwijs. Datum inwerkingtreding: 01/08/2014. [wetten.overheid.nl](http://wetten.overheid.nl).
- WVO (2014). Artikel 3, lid 2, Inrichtingsbesluit Wet op het voortgezet onderwijs. Datum inwerkingtreding: 01/08/2014. [wetten.overheid.nl](http://wetten.overheid.nl).
- Zhu, M. (2024). New findings on racial bias in teachers’ evaluations of student achievement. Working Paper No. 16815, IZA.
- Zumbuehl, M., N. Chehber, and R. Dillingh (2025). Can skills differences explain the gap in track recommendation by socio-economic status? *Education Economics* 33(2), 161–179.

## A Proof Proposition 1

We start with  $\delta^m$  from the balancing regression equation (5). Using Assumption 1 we derive

$$\begin{aligned}\delta^m &= \left( \frac{\text{Cov}[\tilde{\lambda}(s_t + m_t), SES]}{\text{Var}[SES]} \right) = \left( \frac{\text{Cov}[s_t + m_t, SES]}{\text{Var}[SES]} \right) \tilde{\lambda} = \left( \frac{\text{Cov}[s_t, SES]}{\text{Var}[SES]} \right) \tilde{\lambda} \\ &= \delta \tilde{\lambda}.\end{aligned}\quad (28)$$

Hence, we can write the balancing regression equation as

$$\begin{pmatrix} s_t \\ s_t^m \end{pmatrix} = \begin{pmatrix} s_t \\ \tilde{\lambda}(s_t + m_t) \end{pmatrix} = \begin{pmatrix} \delta \\ \tilde{\lambda}\delta \end{pmatrix} SES + \begin{pmatrix} u_t \\ \tilde{\lambda}(u_t + m_t) \end{pmatrix}.\quad (29)$$

Applying the Frisch-Waugh logic and Assumption 1, we can now derive the expression for  $\gamma^m$  as follows,

$$\begin{aligned}\gamma^m &= \left( \frac{\text{Cov}[y_t, u_t^m]}{\text{Var}[u_t^m]} \right) = \left( \frac{\text{Cov}[y_t, \tilde{\lambda}(u_t + m_t)]}{\text{Var}[\tilde{\lambda}(u_t + m_t)]} \right) = \left( \frac{\text{Cov}[y_t, (u_t + m_t)]}{\text{Var}[u_t + m_t]} \right) \left( \frac{\tilde{\lambda}}{\tilde{\lambda}^2} \right) \\ &= \left( \frac{\text{Cov}[y_t, u_t]}{\sigma_{u_t}^2 + \sigma_{m_t}^2} \right) \left( \frac{1}{\tilde{\lambda}} \right) = \left( \frac{\text{Cov}[y_t, u_t]}{\sigma_{u_t}^2} \right) \left( \frac{\sigma_{u_t}^2}{\sigma_{u_t}^2 + \sigma_{m_t}^2} \right) \left( \frac{1}{\tilde{\lambda}} \right) \\ &= \gamma \left( \frac{\lambda}{\tilde{\lambda}} \right),\end{aligned}\quad (30)$$

where  $\lambda = \left( \frac{\sigma_{u_t}^2}{\sigma_{u_t}^2 + \sigma_{m_t}^2} \right)$ . To derive the expression for  $\beta^m$ , we first introduce the short regression equation:

$$y_t = \beta^s SES + e_t^s.\quad (31)$$

Next, we use the omitted variable bias formula for  $\beta^s$  to write

$$\beta^m + \gamma^m \delta^m = \beta + \gamma \delta.\quad (32)$$

Rewriting for  $\beta^m$ , substituting for  $\delta^m = \delta \tilde{\lambda}$  and  $\gamma^m = \gamma \left( \frac{\lambda}{\tilde{\lambda}} \right)$ , and simplifying yields

$$\beta^m = \beta + \gamma \delta (1 - \lambda).\quad (33)$$

## B Proof Proposition 2

We first derive results using only Assumption 1. Subsequently, we simplify these results by additionally invoking Assumption 2 and 3 in Appendix Sections B.1 and B.2, respectively.

We start with  $\delta_{t-1}^m$  from balancing regression equation (11). Using Assumption 1 we derive

$$\begin{aligned}\delta_{t-1}^m &= \left( \frac{\text{Cov}[\tilde{\lambda}(s_{t-1} + m_{t-1}), SES]}{\text{Var}[SES]} \right) \\ &= \delta_{t-1} \tilde{\lambda}.\end{aligned}\tag{34}$$

Hence, we can write the balancing regression equation as

$$\begin{pmatrix} s_{t-1} \\ s_{t-1}^m \end{pmatrix} = \begin{pmatrix} s_{t-1} \\ \tilde{\lambda}(s_{t-1} + m_{t-1}) \end{pmatrix} = \begin{pmatrix} \delta_{t-1} \\ \tilde{\lambda}\delta_{t-1} \end{pmatrix} SES + \begin{pmatrix} u_{t-1} \\ \tilde{\lambda}(u_{t-1} + m_{t-1}) \end{pmatrix}.\tag{35}$$

Applying the Frisch-Waugh logic and Assumption 1, we can now derive the expression for the first stage coefficient  $\pi$  as follows,

$$\begin{aligned}\pi &= \left( \frac{\text{Cov}[s_t^m, u_{t-1}^m]}{\text{Var}[u_{t-1}^m]} \right) = \left( \frac{\text{Cov}[\tilde{\lambda}(s_t + m_t), \tilde{\lambda}(u_{t-1} + m_{t-1})]}{\text{Var}[\tilde{\lambda}(u_{t-1} + m_{t-1})]} \right) = \left( \frac{\text{Cov}[s_t, u_{t-1}]}{\sigma_{u_{t-1}}^2 + \sigma_{m_{t-1}}^2} \right) \left( \frac{\tilde{\lambda}\tilde{\lambda}}{\tilde{\lambda}^2} \right) \\ &= \left( \frac{\text{Cov}[s_t, u_{t-1}]}{\sigma_{u_t}^2 + \sigma_{m_t}^2} \right) \left( \frac{\sigma_{u_t}^2 + \sigma_{m_t}^2}{\sigma_{u_{t-1}}^2 + \sigma_{m_{t-1}}^2} \right) \\ &= \left( \frac{\sigma_{u_t}^2}{\sigma_{u_t}^2 + \sigma_{m_t}^2} \right) \left( \frac{\text{Cov}[s_t, u_{t-1}]}{\sigma_{u_t}^2} \right) \left( \frac{\sigma_{u_t}^2 + \sigma_{m_t}^2}{\sigma_{u_{t-1}}^2 + \sigma_{m_{t-1}}^2} \right) \\ &= \left( \frac{\sigma_{u_t}^2}{\sigma_{u_t}^2 + \sigma_{m_t}^2} \right) \left( \frac{\text{Cov}[\delta SES + u_t, u_{t-1}]}{\sigma_{u_t}^2} \right) \left( \frac{\sigma_{u_t}^2 + \sigma_{m_t}^2}{\sigma_{u_{t-1}}^2 + \sigma_{m_{t-1}}^2} \right) \\ &= \lambda \left( \frac{\text{Cov}[u_t, u_{t-1}]}{\sigma_{u_t}^2} \right) \left( \frac{\sigma_{u_t}^2 + \sigma_{m_t}^2}{\sigma_{u_{t-1}}^2 + \sigma_{m_{t-1}}^2} \right).\end{aligned}\tag{36}$$

Next rewrite the second stage equation (10) to obtain the reduced form regression equation,

$$y_t = \theta_0 SES + \theta_1 s_{t-1}^m + e_t^{iv},\tag{37}$$

where  $\frac{\theta_1}{\pi} = \frac{\gamma^{iv}\pi}{\pi} = \gamma^{iv}$ . Applying the Frisch-Waugh logic and Assumption 1, we can similarly derive an expression for the reduced form coefficient  $\theta_1$ ,

$$\begin{aligned}\theta_1 &= \left( \frac{\text{Cov}[y_t, u_{t-1}^m]}{\text{Var}[u_{t-1}^m]} \right) = \left( \frac{\text{Cov}[y_t, \tilde{\lambda}(u_{t-1} + m_{t-1})]}{\text{Var}[\tilde{\lambda}(u_{t-1} + m_{t-1})]} \right) = \left( \frac{\text{Cov}[y_t, u_{t-1}]}{\sigma_{u_{t-1}}^2 + \sigma_{m_{t-1}}^2} \right) \left( \frac{\tilde{\lambda}}{\tilde{\lambda}^2} \right) \\ &= \left( \frac{\text{Cov}[y_t, u_{t-1}]}{\sigma_{u_t}^2} \right) \left( \frac{\sigma_{u_t}^2}{\sigma_{u_{t-1}}^2 + \sigma_{m_{t-1}}^2} \right) \left( \frac{1}{\tilde{\lambda}} \right) \\ &= \left( \frac{\text{Cov}[y_t, u_t]}{\sigma_{u_t}^2} \right) \left( \frac{\text{Cov}[y_t, u_{t-1}]}{\text{Cov}[y_t, u_t]} \right) \left( \frac{\sigma_{u_t}^2}{\sigma_{u_t}^2 + \sigma_{m_t}^2} \right) \left( \frac{\sigma_{u_t}^2 + \sigma_{m_t}^2}{\sigma_{u_{t-1}}^2 + \sigma_{m_{t-1}}^2} \right) \left( \frac{1}{\tilde{\lambda}} \right) \\ &= \gamma \left( \frac{\lambda}{\tilde{\lambda}} \right) \left( \frac{\text{Cov}[y_t, u_{t-1}]}{\text{Cov}[y_t, u_t]} \right) \left( \frac{\sigma_{u_t}^2 + \sigma_{m_t}^2}{\sigma_{u_{t-1}}^2 + \sigma_{m_{t-1}}^2} \right).\end{aligned}\tag{38}$$

We can now combine the results in (36) and (38) to derive an expression for the second stage estimate  $\gamma^{iv}$ ,

$$\gamma^{iv} = \frac{\theta_1}{\pi} = \gamma \left( \frac{1}{\tilde{\lambda}} \right) \left( \frac{\text{Cov}[y_t, u_{t-1}]}{\text{Cov}[y_t, u_t]} \right) \left( \frac{\sigma_{u_t}^2}{\text{Cov}[u_t, u_{t-1}]} \right). \quad (39)$$

To derive an expression for the second stage estimate  $\beta^{iv}$ , we first introduce the balancing regression equation for  $\hat{s}_t^m$ ,

$$\begin{pmatrix} s_t^m \\ \hat{s}_t^m \end{pmatrix} = \begin{pmatrix} s_t^m \\ s_t^m - e_t^{fs} \end{pmatrix} = \begin{pmatrix} \delta^m \\ \delta^{iv} \end{pmatrix} SES + \begin{pmatrix} u_t^m \\ u_t^{iv} \end{pmatrix}. \quad (40)$$

We have that  $\delta^{iv} = \delta^m$ ,

$$\begin{aligned} \delta^{iv} &= \left( \frac{\text{Cov}[\hat{s}_t^m, SES]}{\text{Var}[SES]} \right) = \left( \frac{\text{Cov}[s_t^m - e_t^{fs}, SES]}{\sigma_{SES}^2} \right) = \left( \frac{\text{Cov}[s_t^m, SES]}{\sigma_{SES}^2} \right) \\ &= \delta^m, \end{aligned} \quad (41)$$

and from Proposition 1 we know that under Assumption 1,  $\delta^{iv} = \delta^m = \delta \tilde{\lambda}$ . We use the omitted variable bias formula for  $\beta^s$  from (31) to write

$$\beta^{iv} + \gamma^{iv} \delta^{iv} = \beta + \gamma \delta. \quad (42)$$

Rewriting for  $\beta^{iv}$  and substituting for  $\delta^{iv} = \delta \tilde{\lambda}$  and  $\gamma^{iv}$  from (39), it follows that,

$$\beta^{iv} = \beta + \gamma \delta \left( 1 - \left( \frac{\text{Cov}[y_t, u_{t-1}]}{\text{Cov}[y_t, u_t]} \right) \left( \frac{\sigma_{u_t}^2}{\text{Cov}[u_t, u_{t-1}]} \right) \right). \quad (43)$$

## B.1 Case (i)

Here we use the results above derived under Assumption 1 to simplify the expressions while additionally invoking Assumption 2. First, we use Assumption 2 to find a link between  $u_{t-1}$  and  $u_t$ . We can write,

$$\begin{aligned} \delta &= \left( \frac{\text{Cov}[s_t, SES]}{\text{Var}[SES]} \right) = \left( \frac{\text{Cov}[s_{t-1} + \Delta_{s_t}, SES]}{\sigma_{SES}^2} \right) = \left( \frac{\text{Cov}[s_{t-1}, SES]}{\sigma_{SES}^2} \right) + \left( \frac{\text{Cov}[\Delta_{s_t}, SES]}{\sigma_{SES}^2} \right) \\ &= \delta_{t-1} + \alpha, \end{aligned} \quad (44)$$

with  $\alpha = \left( \frac{\text{Cov}[\Delta_{s_t}, SES]}{\sigma_{SES}^2} \right)$ . Assumption 2(a) guarantees that  $\delta = \delta_{t-1}$ . This allows us to write (5) and (11) as

$$\begin{pmatrix} s_t \\ s_{t-1} \end{pmatrix} = \begin{pmatrix} s_t \\ s_t - \Delta_{s_t} \end{pmatrix} = \delta SES + \begin{pmatrix} u_t \\ u_t - (\Delta_{s_t} - \bar{\Delta}_{s_t}) \end{pmatrix}. \quad (45)$$



The mean of  $\Delta_{s_t}$ , denoted by  $\bar{\Delta}_{s_t}$ , will be absorbed by the intercept of (11), so that  $u_{t-1} = u_t - (\Delta_{s_t} - \bar{\Delta}_{s_t})$ . We use this to write  $\text{Cov}[u_t, u_{t-1}] = \sigma_{u_t}^2 - \text{Cov}[u_t, \Delta_{s_t}]$ . Subsequently, Assumption 2(a)-(b) together guarantee that  $\text{Cov}[u_t, \Delta_{s_t}] = \text{Cov}[s_t - \delta SES, \Delta_{s_t}] = 0$ . Thus Assumption 2 allows us to write  $\text{Cov}[u_t, u_{t-1}] = \sigma_{u_t}^2$ . Assumption 2 further allows to derive an expression for  $\sigma_{u_{t-1}}^2$ . In particular, we can write  $\sigma_{u_{t-1}}^2 = \text{Var}[u_t - (\Delta_{s_t} - \bar{\Delta}_{s_t})] = \text{Var}[u_t - \Delta_{s_t}] = \sigma_{u_t}^2 + \sigma_{\Delta_{s_t}}^2 - 2 \times \text{cov}[u_t, \Delta_{s_t}] = \sigma_{u_t}^2 + \sigma_{\Delta_{s_t}}^2$ . In short,  $\sigma_{u_{t-1}}^2 = \sigma_{u_t}^2 + \sigma_{\Delta_{s_t}}^2$ .

We can use these two results to write the first stage coefficient  $\pi$  from (36) as follows,

$$\pi = \lambda \left( \frac{\sigma_{u_t}^2 + \sigma_{m_t}^2}{\sigma_{u_t}^2 + \sigma_{\Delta_{s_t}}^2 + \sigma_{m_{t-1}}^2} \right). \quad (46)$$

For the second stage estimate  $\gamma^{iv}$ , we first use that  $u_{t-1} = u_t - (\Delta_{s_t} - \bar{\Delta}_{s_t})$  to additionally simplify the following term,

$$\begin{aligned} \left( \frac{\text{Cov}[y_t, u_{t-1}]}{\text{Cov}[y_t, u_t]} \right) &= \left( \frac{\text{Cov}[y_t, u_t]}{\text{Cov}[y_t, u_t]} \right) - \left( \frac{\text{Cov}[y_t, \Delta_{s_t}]}{\text{Cov}[y_t, u_t]} \right) \\ &= 1 - \left( \frac{\text{Cov}[\beta SES + \gamma s_t + e_t, \Delta_{s_t}]}{\text{Cov}[y_t, u_t]} \right) \\ &= 1 - \left( \frac{\text{Cov}[e_t, \Delta_{s_t}]}{\text{Cov}[y_t, u_t]} \right). \end{aligned} \quad (47)$$

We can now write  $\gamma^{iv}$  from (39) as

$$\gamma^{iv} = \gamma \left( \frac{1}{\bar{\lambda}} \right) \left( 1 - \left( \frac{\text{Cov}[e_t, \Delta_{s_t}]}{\text{Cov}[y_t, u_t]} \right) \right), \quad (48)$$

and  $\beta^{iv}$  from (43) as,

$$\beta^{iv} = \beta + \gamma \delta \left( \frac{\text{Cov}[e_t, \Delta_{s_t}]}{\text{Cov}[y_t, u_t]} \right). \quad (49)$$

## B.2 Case (ii)

Here we use the results above derived under Assumption 1 to simplify the expressions while additionally invoking Assumption 3. First, we use Assumption 3 to find a link between  $u_{t-1}$  and  $u_t$ . Similar to (44), we write

$$\delta = \delta_{t-1} + \alpha, \quad (50)$$

with  $\alpha = \left( \frac{\text{Cov}[\Delta_{s_t}, SES]}{\sigma_{SES}^2} \right)$ . Assumption 3 guarantees again that  $\delta = \delta_{t-1}$ . Moreover, we can write (5) and (11) as

$$\begin{pmatrix} s_t \\ s_{t-1} \end{pmatrix} = \begin{pmatrix} s_t \\ s_t - \Delta_{s_t} \end{pmatrix} = \delta SES + \begin{pmatrix} u_t \\ u_t \end{pmatrix}, \quad (51)$$

where the constant  $\Delta s_t$  is completely absorbed by the intercept of (11), so that  $u_{t-1} = u_t$ . We use this to write  $\text{Cov}[u_t, u_{t-1}] = \sigma_{u_t}^2$  and  $\sigma_{u_{t-1}}^2 = \sigma_{u_t}^2$ .

We can use these two results to write the first stage coefficient  $\pi$  from (36) as follows,

$$\pi = \lambda \left( \frac{\sigma_{u_t}^2 + \sigma_{m_t}^2}{\sigma_{u_t}^2 + \sigma_{m_{t-1}}^2} \right). \quad (52)$$

For the second stage estimate  $\gamma^{iv}$ , we also use that  $u_{t-1} = u_t$  additionally implies that  $\left( \frac{\text{Cov}[y_t, u_{t-1}]}{\text{Cov}[y_t, u_t]} \right) =$

1. We can now write  $\gamma^{iv}$  from (39) as

$$\gamma^{iv} = \gamma \left( \frac{1}{\tilde{\lambda}} \right), \quad (53)$$

and  $\beta^{iv}$  from (43) as,

$$\beta^{iv} = \beta. \quad (54)$$

## C Proof Proposition 3

Proposition 2 shows that under Assumption 1 and 2 we have that  $\pi' = \lambda$ . It immediately follows that,

$$\begin{aligned} \gamma^{ev} &= \frac{\gamma^m}{\pi'} = \gamma^m \left( \frac{1}{\lambda} \right) = \gamma \left( \frac{\lambda}{\tilde{\lambda}} \right) \left( \frac{1}{\lambda} \right) \\ &= \gamma \left( \frac{1}{\tilde{\lambda}} \right), \end{aligned} \quad (55)$$

and that

$$\begin{aligned} \beta^{ev} &= \beta^m - \gamma^m \delta^m \left( \frac{1 - \pi'}{\pi'} \right) = \beta + \gamma \delta (1 - \lambda) - \gamma^m \delta^m \left( \frac{1 - \lambda}{\lambda} \right) \\ &= \beta + \gamma \delta (1 - \lambda) - \gamma \left( \frac{\lambda}{\tilde{\lambda}} \right) \delta \tilde{\lambda} \left( \frac{1 - \lambda}{\lambda} \right) = \beta + \gamma \delta (1 - \lambda) - \gamma \delta (1 - \lambda) \\ &= \beta. \end{aligned} \quad (56)$$

## D Additional empirical results

Table 3: Descriptive statistics on the test scores across all grades

|                        | (1)    | (2)     | (3)    | (4)      | (5)      |
|------------------------|--------|---------|--------|----------|----------|
|                        | Obs    | Mean    | SD     | Min      | Max      |
| Panel A: Reading tests |        |         |        |          |          |
| Midterm grade 2        | 114743 | 139.354 | 27.933 | <42.000  | >264.000 |
| End term grade 2       | 126688 | 146.532 | 26.752 | <55.000  | >282.000 |
| Midterm grade 3        | 137499 | 157.415 | 26.149 | <66.000  | >279.000 |
| End term grade 3       | 121392 | 161.807 | 27.028 | <71.000  | >280.000 |
| Midterm grade 4        | 142069 | 173.447 | 25.339 | <73.000  | >309.000 |
| End term grade 4       | 111460 | 182.413 | 27.147 | <79.000  | >337.000 |
| Midterm grade 5        | 148019 | 191.618 | 26.766 | <81.000  | >340.000 |
| End term grade 5       | 148019 | 197.279 | 26.983 | <84.000  | >354.000 |
| Panel B: Math tests    |        |         |        |          |          |
| Midterm grade 2        | 132855 | 167.815 | 30.385 | <30.000  | >278.000 |
| End term grade 2       | 132976 | 189.235 | 29.063 | <53.000  | >289.000 |
| Midterm grade 3        | 137502 | 206.446 | 28.042 | <50.000  | >313.000 |
| End term grade 3       | 137643 | 218.366 | 27.299 | <64.000  | >334.000 |
| Midterm grade 4        | 141986 | 230.064 | 26.684 | <89.000  | >356.000 |
| End term grade 4       | 123702 | 241.663 | 26.635 | <100.000 | >361.000 |
| Midterm grade 5        | 148019 | 254.142 | 27.142 | <118.000 | >379.000 |
| End term grade 5       | 148019 | 261.959 | 26.085 | <118.000 | >394.000 |

Notes: this table shows the descriptive statistics for the test scores across all grades. Statistics Netherlands prevents reporting minima and maxima that are based upon less than 10 observations. Therefore, the minimum (maximum) may contain the smallest (largest) number, such that at least ten observations have values lower (higher) than that number.

Table 4: OLS estimates with control variables

|  | (1)   | (2)                 | (3)                  |
|--|---|---------------------|----------------------|
|  | $\mathbb{1}(\text{recom.} \geq \text{college tr.})$ |                     |                      |
| SES income                                 | 0.115***<br>(0.002)                                 | 0.028***<br>(0.001) | 0.029***<br>(0.001)  |
| End term grade 5                           |   | 0.381***<br>(0.001) | 0.375***<br>(0.001)  |
| $\mathbb{1}(\text{female})$                |   |                     | 0.011***<br>(0.002)  |
| $\mathbb{1}(\text{Western imm.})$          |   |                     | 0.036***<br>(0.004)  |
| $\mathbb{1}(\text{non-Western imm.})$      |   |                     | 0.036***<br>(0.003)  |
| Age  |   |                     | -0.072***<br>(0.002) |
| $\mathbb{1}(\text{Cohort}=2019\text{-}20)$ |   |                     | -0.042***<br>(0.005) |
| $\mathbb{1}(\text{Cohort}=2020\text{-}21)$ |   |                     | 0.030***<br>(0.005)  |
| $\mathbb{1}(\text{Cohort}=2021\text{-}22)$ |   |                     | -0.072***<br>(0.005) |
| Observations                               | 148019  | 148019              | 148019               |
| $R^2$                                      | 0.053   | 0.498               | 0.506                |

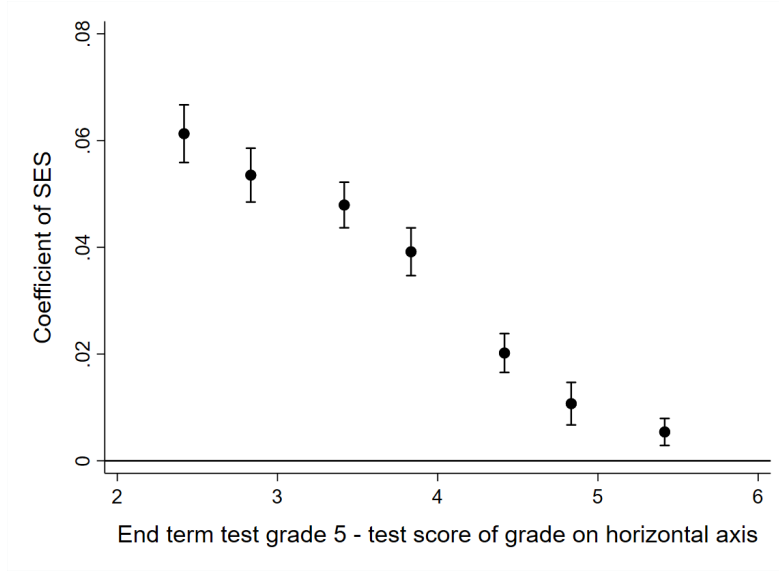
Notes: \*\*\*, \*\*, \* refers to statistical significance at the 1, 5, and 10% level. Coefficients are estimated with OLS. Standard errors (in parentheses) are clustered at the school level.

Table 5: Cov[ $\Delta s_t^m, x$ ] for observable characteristics  $x$

|                                       | (1)                                | (2)                 | (3)                 | (4)                 | (5)                 |
|---------------------------------------|------------------------------------|---------------------|---------------------|---------------------|---------------------|
|                                       | $\Delta s_t^m = s_t^m - s_{t-1}^m$ |                     |                     |                     |                     |
| SES income                            | 0.005***<br>(0.001)                |                     |                     |                     |                     |
| $\mathbb{1}(\text{female})$           |                                    | 0.007***<br>(0.002) |                     |                     |                     |
| $\mathbb{1}(\text{Western imm.})$     |                                    |                     | 0.018***<br>(0.004) |                     |                     |
| $\mathbb{1}(\text{non-Western imm.})$ |                                    |                     |                     | 0.012***<br>(0.004) |                     |
| Age                                   |                                    |                     |                     |                     | -0.005**<br>(0.002) |
| Observations                          | 148019                             | 148019              | 148019              | 148019              | 148019              |
| $R^2$                                 | 0.000                              | 0.000               | 0.000               | 0.000               | 0.000               |

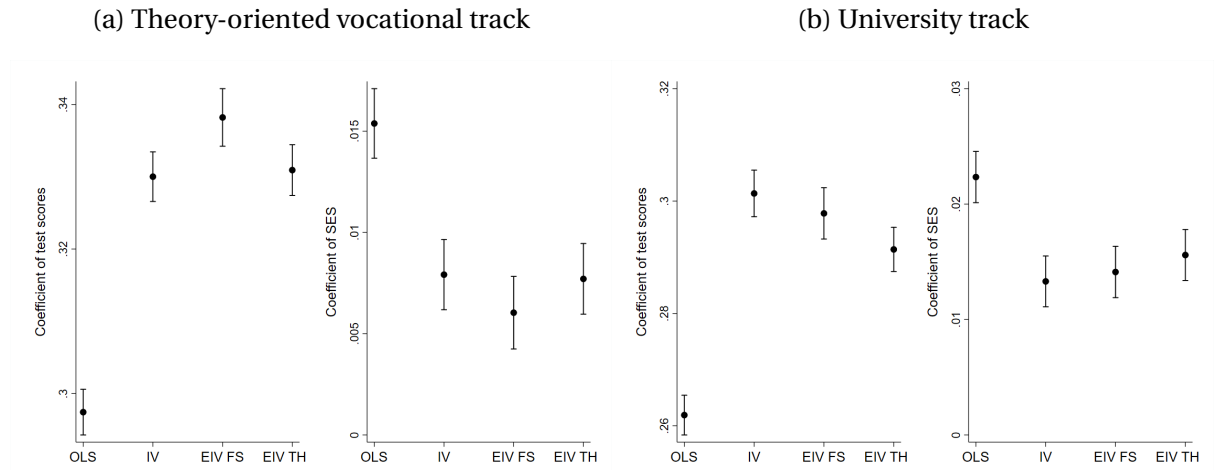
Notes: \*\*\*, \*\*, \* refers to statistical significance at the 1, 5, and 10% level. Coefficients are estimated with OLS. Standard errors (in parentheses) are clustered at the school level.

Figure 3:  $\text{Cov}[\Delta s_t^m, \text{SES}]$  over time



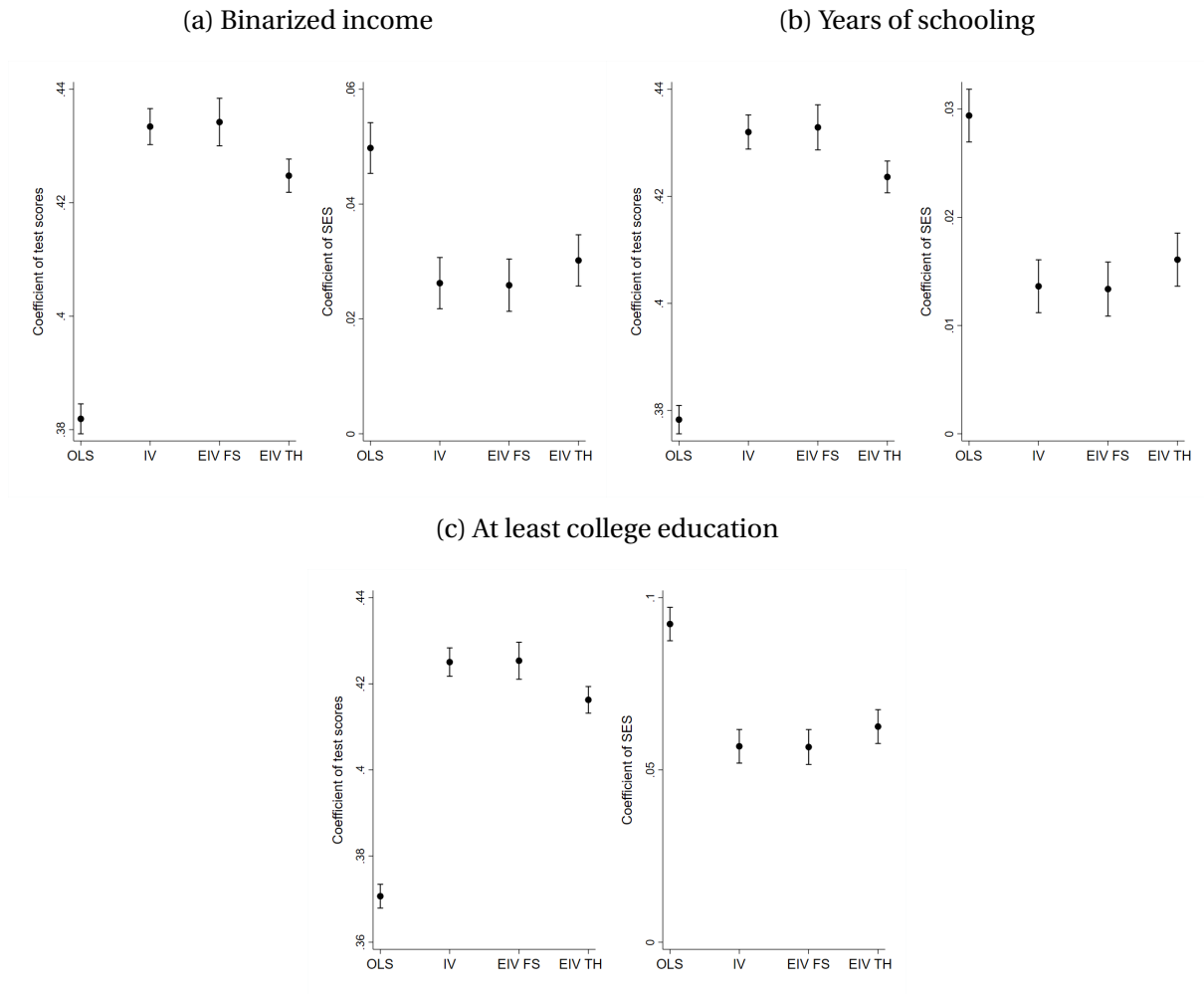
Notes: this figure plots the estimates for the coefficients of SES from regressions of the change in test scores ( $s_t^m - s_{t-\tau}^m$ ) upon SES over time  $\tau$ . Coefficients are estimated with OLS. Standard errors are clustered on the school id.

Figure 4: OLS, IV, and EIV estimates with two alternative measures of teacher recommendation



Notes: this figure shows the estimated coefficients for the OLS, IV, and EIV strategy with two alternative measures of teacher recommendation. The lines are 95% confidence intervals. The two alternative measures are a dummy variable that is equal to one if the recommendation is equal to or higher than the theory-oriented vocational track (Figure 4(a)) or the university track (Figure 4(b)). See Figure 2 for details on the estimation.

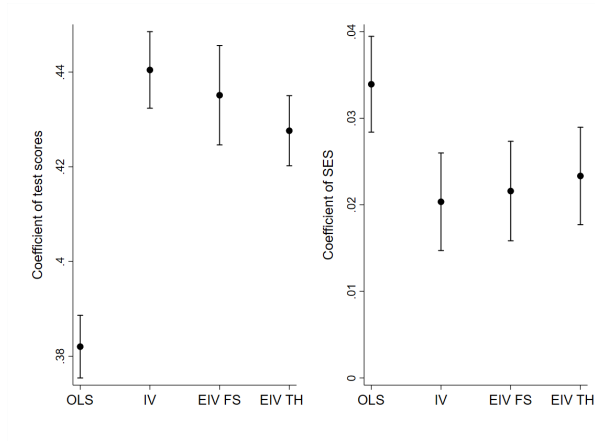
Figure 5: OLS, IV, and EIV estimates with three alternative SES measures



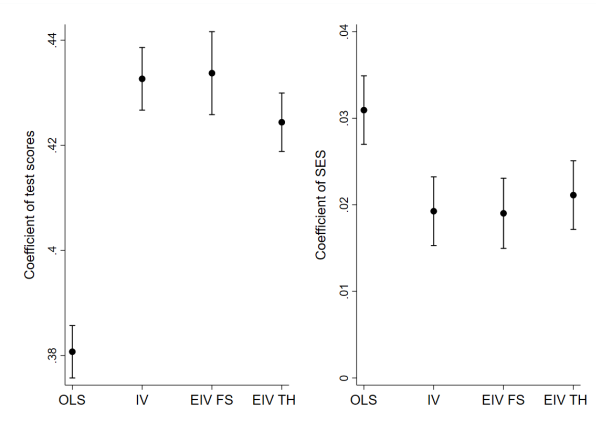
Notes: this figure shows the estimated coefficients for the OLS, IV, and EIV strategy with three alternative SES measures. The lines are 95% confidence intervals. The three alternative measures are a dummy that equals one if parental income is above the median (Figure 5(a)), years of schooling of the highest educated parent (Figure 5(b)), and a dummy that equals one if the highest educated parent completed at least college education (Figure 5(c)). See Figure 2 for details on the estimation.

Figure 6: OLS, IV, and EIV estimates per cohort

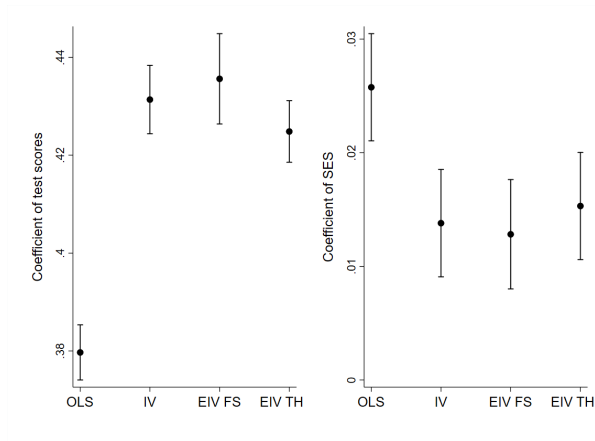
(a) Cohort 2018-19



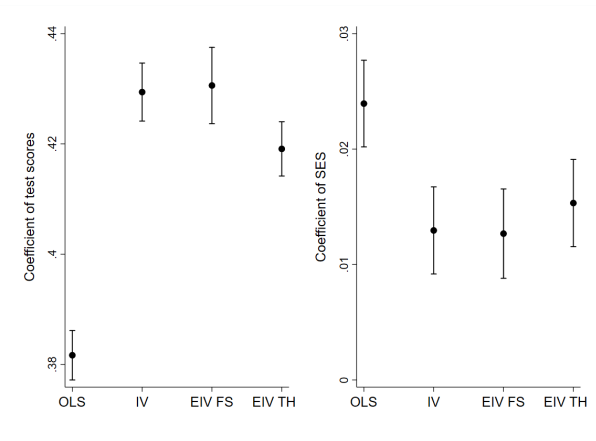
(b) Cohort 2019-20



(c) Cohort 2020-21



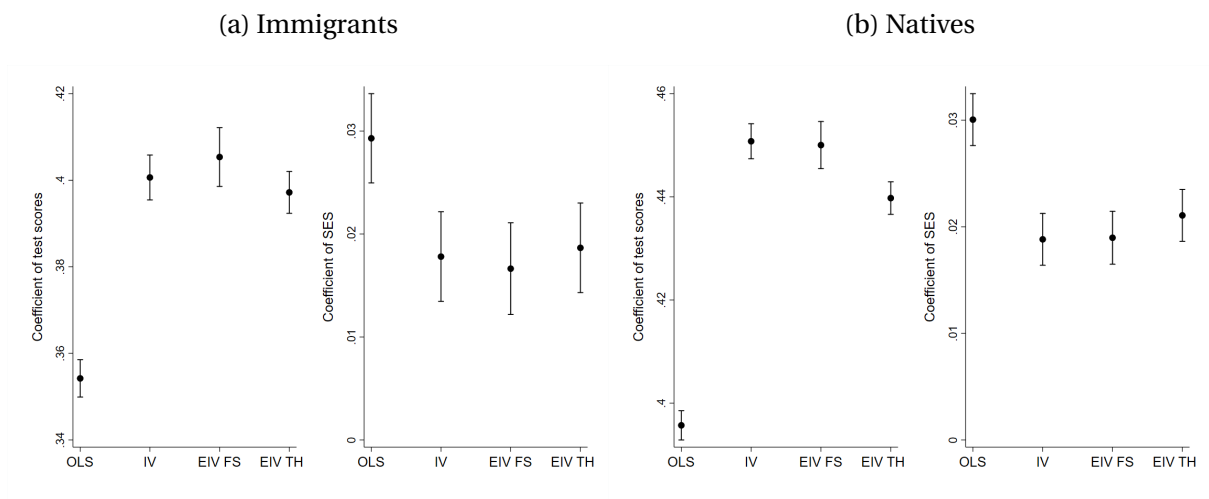
(d) Cohort 2021-22



Notes: this figure shows the estimated coefficients for the OLS, IV, and EIV strategy, separately for each cohort (Figure 6(a)-(d)). The lines are 95% confidence intervals. See Figure 2 for details on the estimation.



Figure 7: OLS, IV, and EIV estimates separately for immigrants and natives



Notes: this figure shows the estimated coefficients for the OLS, IV, and EIV strategy, separately for immigrants (Figure 7(a)) and natives (Figure 7(b)). The lines are 95% confidence intervals. See Figure 2 for details on the estimation.