Teacher bias or measurement error?

Teacher bias or measurement error?

Thomas van Huizen[*] Madelon Jacobs[†] Matthijs Oosterveen[‡]

February 19, 2024

Abstract

In many countries, teachers' track recommendations are used to allocate students to secondary school tracks. Previous studies have shown that students from families with low socioeconomic status (SES) receive lower track recommendations than their peers from high SES families, conditional on standardized test scores. It is often argued that this indicates teacher bias. However, this claim is invalid in the presence of measurement error in test scores. We discuss how measurement error in test scores generates a biased coefficient of the conditional SES gap, and consider three empirical strategies to address this bias. Using administrative data from the Netherlands, 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

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

[‡]Department of Economics, Lisbon School of Economics and Management, and Advance/CSG, University of Lisbon, Lisbon 1200-781, Portugal. oosterveen@iseg.ulisboa.pt

[§]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, KVS new paper sessions, the LEER conference, the LESE conference, LMG Lisbon, the PPE conference, and the Utrecht University School of Economics seminar. 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 accessible through Statistics Netherlands. National and international researchers can request access via the Netherlands Cohort Study on Education (NCO) and Statistics Netherlands. Requests for data access should be directed to info@nationaalcohortonderzoek.nl. A comprehensive description of the data and the access procedure can be found in Haelermans et al. (2020). 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 constitute an important source of inequality in educational attainment and later life outcomes. To study teacher bias, studies generally rely on comparisons between subjective teacher evaluations, such as track recommendations, and objective measures of student abilities, such as performance on standardized tests (Botelho et al., 2015; Alesina et al., 2018; Timmermans et al., 2018). However, the presence of systematic gaps in teacher evaluations between groups that condition on objective measures is not necessarily evidence of teacher bias. The identification of teacher bias with this approach is complicated by two main concerns: omitted variable bias and measurement error in the objective measure of student abilities. As omitted variables could (at least partly) explain the potentially remaining conditional gap, the literature tries to address this concern by controlling for additional factors such as non-cognitive skills (Burgess and Greaves, 2013; Cornwell et al., 2013; Triventi, 2020).[1]

This paper focuses on the second methodological concern: measurement error in the objective measure of ability. Ability is often measured by standardized tests and classical measurement error in test scores leads to attenuation bias in the test score coefficient. Moreover, to the extent that student background characteristics are associated with ability, this bias contaminates the main parameter of interest that measures the conditional gap. For instance, if socioeconomic status (SES) of the family is positively related to student ability, analyses that control for standardized test scores without addressing measurement error will overestimate teacher bias.[2] Hence, systematic differences in teacher evaluations conditional on test scores could be merely a statistical artefact and can therefore not automatically be interpreted as evidence of teacher bias.

Following Pei et al. (2019), we first discuss how classical measurement error in test scores generates a bias that contaminates the coefficient measuring the conditional gap between groups. We then discuss three 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; Gillen et al., 2019; Ward, 2023). Second, we discuss a method in the spirit of errors-in-variables (EIV) models, an approach that is common in psychometrics (Fuller, 2009; Culpepper, 2012). The additional parameter required for the

[1]For instance, Cornwell et al. (2013) show that girls are graded more favorably by their teachers than boys, even if they perform equally well on several standardized tests. However, this differential treatment disappears after controlling for differences in non-cognitive skills.

[2]This methodological problem has been coined Kelley's Paradox (Wainer and Brown, 2006), after statistician Truman Kelley who described this phenomenon in the first half of the twentieth century (Kelley, 1927).

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 the reliability ratio. Third, we combine the EIV strategy with a novel approach to estimate the reliability ratio under yet weaker assumptions, while making use of students' complete test score histories.

We apply these three strategies to estimate conditional gaps in teacher track recommendations in the Netherlands, focusing on differences by family SES. The Dutch early tracking regime provides an interesting case for several reasons. First, existing studies claim that there is evidence of teacher bias in track recommendations against low SES children, both in the Netherlands (Timmermans et al., 2018) and other countries (Falk et al., 2020; Batruch et al., 2023). Second, teacher bias in track recommendations may be highly consequential for the educational paths students take and therefore may reduce equal opportunities. Third, we can make use of detailed administrative data, covering roughly half of the Dutch primary schools, to apply the three different approaches mentioned above.

While almost all OECD countries have some form of educational tracking, the Netherlands is considered an early tracking country. Students are assigned to different educational programs in secondary school at the age of twelve. In contrast to some other early tracking regimes, track assignment is based upon binding track recommendations from the primary school teacher (WVO, 2014). These recommendations may be highly consequential: although students are able to move between tracks at least in the first grades of secondary school, roughly 70% of the students do not experience track mobility four years into secondary education (de Ree et al., 2023). Track placement determines among other things the level of the curriculum and peer group abilities, and as such may affect student achievement (Matthewes, 2021). As only the highest tracks give access to college and university education, track placement may also have long-term consequences in terms of attainment and income (Dustmann et al., 2017; Borghans et al., 2019, 2020). Moreover, negatively biased track recommendations signal low teacher expectations, which may by itself negatively affect student outcomes (Rosenthal and Jacobson, 1968; Carlana, 2019; Papageorge et al., 2020; Hill and Jones, 2021).

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 with lower family SES receive significantly lower track recommendations, conditional on scores from a standardized end-of-primary school test. This finding fueled an ongoing public debate on the role of teachers as gatekeepers of opportunities and contributed to the implementation of a national policy reform, effective in school year 2023-24. The reform reduces the role of teachers in track placement, and increases the role of the standardized end-of-primary school test (Ministry of Education, 2023b).[3] Concerns about unequal opportunities have also been expressed in other countries based on similar findings (Falk et al., 2020; Carlana et al., 2022).

Using new administrative data, we first replicate previous findings from the Netherlands (Inspectorate of Education, 2016; Timmermans et al., 2018; Zumbuehl et al., 2022) and find that children with lower family SES receive systematically lower track recommendations, conditional on standardized test scores. We then address concerns about measurement error using our three strategies. The IV strategy uses the score from a standardized test administered in the middle of the school year as an instrumental variable for the standardized test administered at the end of the year. The two EIV strategies use different methods to identify the reliability ratio of the end-of-year test: the first stage from the IV strategy and a more novel approach that uses a student's standardized test scores across all primary school grades. Our results indicate that measurement error can explain between 35 and 43% of the conditional SES gap.

Our paper contributes to two strands of literature. First, our study speaks to the broader literature on measurement error (Pei et al., 2019; Ward, 2023; Borjas and Hamermesh, 2024). In particular, we focus on contamination bias due to measurement error. As Modalsli and Vosters (2022) 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 (2022) examine the implications of measurement error in multigenerational income regressions and use an IV approach to mitigate the resulting bias. They demonstrate that measurement error in parents' income leads to an upwards bias in the grandparent coefficient. Gillen et al. (2019) replicate three classic experiments and show that the findings change substantially when measurement error is addressed with an IV approach. Gillen et al. (2019) argue that "*these results show that failing to properly account for measurement error may cause a field-wide bias leading scholars to identify "new" phenomena.*" A contribution of our study is to discuss how different strategies to address measurement error are related and how they produce valid estimates under different assumptions. This paper thereby also contributes to a larger literature studying bias and discrimination, where similar methodological concerns apply.

Second, we contribute to the literature on teacher bias by providing new evidence on

[3] Moreover, the reform includes a revision of national guidelines for the formulation of track recommendations (Ministry of Education, 2023a). These guidelines pay explicit attention to (implicit) teacher bias in track recommendations, especially with respect to SES.

the conditional SES gap in track recommendations. Teacher bias in track recommendations have been recently studied in Germany (Falk et al., 2020), Italy (Carlana et al., 2022), and the Netherlands (Timmermans et al., 2018; Zumbuehl et al., 2022). Previous studies have also examined teacher bias in grading by gender (Cornwell et al., 2013; Lavy and Sand, 2018; Terrier, 2020), race and ethnicity (Burgess and Greaves, 2013; Botelho et al., 2015), and migration background (Alesina et al., 2018; Triventi, 2020). Consistent with the observation of Modalsli and Vosters (2022), the issue of contamination bias due to measurement error has been largely ignored in this literature. A notable exception is the study by Botelho et al. (2015) on racial discrimination in grading in Brazil, who use an IV approach to address contamination bias. Consistent with our findings, they find that the conditional gap is reduced by around 50%. However, Botelho et al. (2015) do not discuss or test the assumptions under which the IV approach is valid, and do not examine alternative strategies to address measurement error.

The remainder of the paper is organized as follows. Section 2 discusses contamination bias in the presence of measurement error and explains the three strategies to address this bias. Section 3 describes the Dutch education system and administrative data. Section 4 shows our empirical results. Section 5 concludes.

2 Methodology

2.1 Contamination bias

Let y_t be a variable that measures a teacher's subjective evaluation of a student in time period t (*e.g.*, a track recommendation), SES a variable that measures the socioeconomic status of the student, and s_t a variable that reflects student ability in time period t. The long regression equation equals,

$$y_t = \beta SES + \gamma s_t + e_t. \tag{1}$$

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 β. 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.

We cannot estimate (1) as we do not observe ability. Instead we observe a test score s_t^m as an objective measure for student ability. This test score is equal to ability plus measurement error m_t,

$$s_t^m = s_t + m_t. \tag{2}$$

The measurement error reflects, for instance, that students may guess several answers, are more or less familiar with the specific selection of test items, or feel (un)well on the day of the test. These processes could be random, reflected by the assumption of classical measurement error.

Assumption 1. *Classical measurement error*

a. (Zero mean) $\mathbb{E}[m_t] = 0 \; \forall t$.

b. (Zero covariance) $\text{Cov}[m_t, x] = 0, \; \forall t, x$.

The medium regression equation replaces s_t with s_t^m,

$$y_t = \beta^m SES + \gamma^m s_t^m + e_t^m. \tag{3}$$

To analyze what β^m and γ^m identify, we also introduce the balancing regression equation,

$$\begin{pmatrix} s_t \\ s_t^m \end{pmatrix} = \begin{pmatrix} s_t \\ s_t + m_t \end{pmatrix} = \begin{pmatrix} \delta \\ \delta^m \end{pmatrix} SES + \begin{pmatrix} u_t \\ u_t^m \end{pmatrix}. \tag{4}$$

Classical measurement error on the left-hand side does not affect OLS estimates, and thus Assumption 1 guarantees that $\delta^m = \delta$ and $u_t^m = u_t + m_t$.

The following proposition formalizes what OLS identifies under Assumption 1.

Proposition 1. *Under Assumption 1, it holds that:*

$$\gamma^m = \gamma\lambda, \tag{5}$$

$$\beta^m = \beta + \gamma\delta(1 - \lambda). \tag{6}$$

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

All proofs are provided in Appendix A to C. Equation (5) reflects the attenuation bias in a multivariate setting. λ is the multivariate reliability (or signal-to-total variance) ratio. If $\sigma^2_{m_t} = 0$, measurement error is absent and the test score delivers a perfect signal on ability, so that $\lambda = 1$. If $\sigma^2_{m_t} > 0$, measurement error is present and the test score delivers an imperfect signal, so that $\lambda < 1$. As $\sigma^2_{m_t}$ grows larger, λ approaches zero. Hence, the presence of measurement error implies the estimate γ^m is attenuated.

Equation (6) clarifies how the attenuation bias in γ^m contaminates β^m. If measurement error is absent, $\lambda = 1$ and contamination is also absent, so that $\beta^m = \beta$. If measurement error is present, $\lambda < 1$ and $\beta^m = \beta + \gamma\delta(1 - \lambda)$. As λ approaches zero the test does not deliver any signal on ability and we have that $\beta^m = \beta + \gamma\delta$, which is the OLS estimate when not controlling for test scores. Hence, given the presence of measurement error, β^m is larger than β when γ and δ are positive. This bias grows as γ and δ become larger, and λ becomes smaller. See also Pei et al. (2019) for a discussion of the result in Proposition 1.

2.2 The IV strategy

The instrumental variables (IV) strategy uses a second objective measure of ability in period t as an instrumental variable for the first objective measure. Intuitively, under Assumption 1 the measurement errors of the two objective measures are uncorrelated, and the second stage IV estimates are consistent since they only use the variation generated by unobserved ability. In practice, one may use the scores from a test administered in period $t-1$ as the second objective measure. This test score, s_{t-1}^m, also contains classical measurement error according to Assumption 1,

$$s_{t-1}^m = s_{t-1} + m_{t-1}. \tag{7}$$

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

$$s_t^m = \kappa SES + \pi s_{t-1}^m + e_t^{fs}, \tag{8}$$

$$y_t = \beta^{iv} SES + \gamma^{iv} \widehat{s}_t^m + e_t^{iv}. \tag{9}$$

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

$$\begin{pmatrix} s_{t-1} \\ s_{t-1}^m \end{pmatrix} = \begin{pmatrix} s_{t-1} \\ s_{t-1} + m_{t-1} \end{pmatrix} = \begin{pmatrix} \delta_{t-1} \\ \delta_{t-1}^m \end{pmatrix} SES + \begin{pmatrix} u_{t-1} \\ u_{t-1}^m \end{pmatrix}, \tag{10}$$

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 Δ_{s_t}.

Assumption 2. *The covariance between Δ_{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 Δ_{s_t}*

(Constant) Δs_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 under only Assumption 1 and while combining Assumption 1 with either 2 or 3.

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

$$\tilde{\pi} = \lambda\left(\frac{1}{B_{\pi}^{1}}\right), \tag{11}$$

$$\gamma^{iv} = \gamma B_{\gamma}^{1} B_{\pi}^{1}, \tag{12}$$

$$\beta^{iv} = \beta + \gamma\delta\left(1 - B_{\gamma}^{1} B_{\pi}^{1}\right). \tag{13}$$

With $\tilde{\pi} = \pi\left(\frac{\sigma^2_{u_{t-1}} + \sigma^2_{m_{t-1}}}{\sigma^2_{u_t} + \sigma^2_{m_t}}\right)$, $B_{\pi}^{1} = \left(\frac{\sigma^2_{u_t}}{\text{Cov}[u_t, u_{t-1}]}\right)$, *and* $B_{\gamma}^{1} = \left(\frac{\text{Cov}[y_t, u_{t-1}]}{\text{Cov}[y_t, u_t]}\right)$.

(ii) Under Assumption 1 and 2, it holds that:

$$\tilde{\pi} = \lambda, \tag{14}$$

$$\gamma^{iv} = \gamma B_{\gamma}^{2}, \tag{15}$$

$$\beta^{iv} = \beta + \gamma\delta\left(1 - B_{\gamma}^{2}\right). \tag{16}$$

With $\tilde{\pi} = \pi\left(\frac{\sigma^2_{u_t} + \sigma^2_{\Delta_{s_t}} + \sigma^2_{m_{t-1}}}{\sigma^2_{u_t} + \sigma^2_{m_t}}\right)$ *and* $B_{\gamma}^{2} = 1 - \left(\frac{\text{Cov}[e_t, \Delta_{s_t}]}{\text{Cov}[y_t, u_t]}\right)$.

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

$$\tilde{\pi} = \lambda, \tag{17}$$

$$\gamma^{iv} = \gamma, \tag{18}$$

$$\beta^{iv} = \beta. \tag{19}$$

With $\tilde{\pi} = \pi\left(\frac{\sigma^2_{u_t} + \sigma^2_{m_{t-1}}}{\sigma^2_{u_t} + \sigma^2_{m_t}}\right)$.

Proposition 2 formalizes that the IV strategy only addresses measurement error under Assumption 3. This assumption, and the corresponding result, is not new. Gillen et al. (2019) and Hausman (2001) provide a review of the literature that uses the IV strategy to address mismeasurement, which dates back to Reiersøl (1941).

Intuitively, a constant Δs_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 λ. This resembles the test-retest method to estimate test reliability, where the reliability ratio is identified by administering the same test to the same students within a relatively short period of time (Guttman, 1945). Furthermore, note that the reduced form equation is similar to the medium regression equation (3) while replacing s_t^m with s_{t-1}^m,

$$y_t = \theta_0 SES + \theta_1 s_{t-1}^m + e_t^{iv}. \tag{20}$$

With a constant Δs_t the reduced form coefficient θ_1 therefore reveals the estimate $\gamma^m = \gamma\lambda$ of the medium regression, 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 $\widehat{s_t}^m$ still spans the relevant aspects of s_t^m in the data, we also have that $\beta^{iv} = \beta$.

2.3 The EIV strategy

The errors-in-variables (EIV) strategy uses the results in Proposition 1 to express the estimates of interest γ and β as follows,

$$\gamma = \frac{\gamma^m}{\lambda}, \tag{21}$$

$$\beta = \beta^m - \gamma^m \delta\left(\frac{1-\lambda}{\lambda}\right). \tag{22}$$

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 value for λ into (21) and (22) to identify the estimates of interest.

2.3.1 The first stage

Case (ii) of Proposition 2 shows that the first stage can be used as an estimate for λ 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 λ^u is defined by replacing $\sigma^2_{u_t}$ in λ with $\sigma^2_{s_t}$. Assumption 2(b) implies that $\text{Cov}[s_t, s_{t-1}] = \sigma^2_{s_t}$, so that the univariate first stage coefficient π^u is equal to $\lambda^u\left(\frac{\sigma^2_{s_t}+\sigma^2_{m_t}}{\sigma^2_{s_{t-1}}+\sigma^2_{m_{t-1}}}\right)$. If we define $\tilde{\pi}^u$ as the multiplication of π^u with the inverse of the second term, which contains the variances of the test scores instead of the variances of the error terms from the balancing regressions as in Proposition 2, we have that $\tilde{\pi}^u = \lambda^u$. This result extends to the multivariate setting since together Assumption 2(a)-(b) imply that $\text{Cov}[u_t, u_{t-1}] = \sigma^2_{u_t}$ (instead of only $\text{Cov}[s_t, s_{t-1}] = \sigma^2_{s_t}$).

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

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

$$\gamma^{fs} = \gamma B^1_\pi, \tag{23}$$

$$\beta^{fs} = \beta + \gamma\delta\left(1 - B^1_\pi\right). \tag{24}$$

(ii) Under Assumption 1 and 2, it holds that:

$$\gamma^{fs} = \gamma, \tag{25}$$

$$\beta^{fs} = \beta. \tag{26}$$

Where γ^{fs} *and* β^{fs} *are defined by replacing* λ *with* $\tilde{\pi}$ *in* (21) *and* (22).

The EIV FS strategy only requires the weaker Assumption 2 since it directly uses the medium regression estimate. In contrast, the IV strategy uses the estimate from the reduced form, which is equal to the medium regression estimate only if $\text{Cov}[e_t, \Delta_{s_t}] = 0$. The latter holds when Δ_{s_t} is constant. Whereas Assumption 3 only seems likely when the two tests are administered shortly after each other, Assumption 2 seems generally more plausible. In particular, under Assumption 2 the teachers may pay attention to Δ_{s_t} in their subjective evaluations, so that it is part of the error term e_t, but Δ_{s_t} cannot correlate with the included variable SES and unobserved ability s_t.

2.3.2 Test score history

The more novel EIV test score history (EIV TH) strategy 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 mimic the ideal of the test-retest method, which uses a test in time $t-\epsilon$ close to time t to estimate the reliability ratio. The first stage equation with a test in $t-\epsilon$ is,

$$s_t^m = \kappa_{t-\epsilon} SES + \pi_{t-\epsilon} s_{t-\epsilon}^m + e_t^{fs_\epsilon}, \tag{27}$$

where we include a time subscript on the coefficients. Proposition 2 shows that $\tilde{\pi}_{t-\epsilon} = \lambda$ under Assumption 3. While such a test score close to time t (*i.e.*, $s_{t-\epsilon}^m$) may not be available, we may observe test scores further away from time t (*i.e.*, $s_{t-\tau}^m$ for $\tau \geq 1$). It seems plausible that all observed test scores contain classical measurement error according to Assumption 1. However, the stronger Assumption 2 or 3 may be less likely to hold.

The EIV TH strategy consists of four steps. First, we repeat the first stage equation with $s_{t-\tau}^m$ for all $\tau \geq 1$,

$$s_t^m = \kappa_{t-\tau} SES + \pi_{t-\tau} s_{t-\tau}^m + e_t^{fs_\tau}, \quad \forall \tau \geq 1. \tag{28}$$

It follows from Proposition 2 that under Assumption 1 only, $\tilde{\pi}_{t-\tau}$ can be expressed as,

$$\tilde{\pi}_{t-\tau} = \lambda \left(\frac{1}{B^1_{\pi_{t-\tau}}} \right), \quad \forall \tau \geq 1, \tag{29}$$

with $B^1_{\pi_{t-\tau}} = \left(\frac{\sigma^2_{u_t}}{\text{Cov}[u_t, u_{t-\tau}]}\right)$. Second, we use $\widetilde{\pi}_{t-\tau}$ as a dependent variable in a forecasting regression equation of the form,

$$\widetilde{\pi}_{t-\tau} = f(t-\tau) + \varepsilon_{t-\tau}. \tag{30}$$

Third, we make a one-period out-of-sample forecast towards $\tau = \epsilon$. Without forecasting error this equals,

$$\widehat{f}(t-\epsilon) = f(t-\epsilon) = \widetilde{\pi}_{t-\epsilon} = \lambda. \tag{31}$$

Fourth, we define γ^{th} and β^{th} as the parameters that result after replacing λ with $f(t-\epsilon)$ in (21) and (22). Proposition 3 shows that $\gamma^{th} = \gamma$ and $\beta^{th} = \beta$.

Although not explicitly written down as an assumption, the absence of forecasting error implies that the chosen polynomial for $f(t-\tau)$ accurately captures the changes in $\widetilde{\pi}_{t-\tau}$ over time. This, in turn, implies that the underlying factors that drive $B^1_{\pi_{t-\tau}}$ away from one, which are $\text{Cov}[\Delta s_t, SES] \neq 0$ and $\text{Cov}[\Delta s_t, s_t] \neq 0$, should close over time in a predictable way.

3 Institutional setting and data

3.1 The Dutch education system

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 also be ordered from more practice-oriented to more theory-oriented tracks.

The allocation of students to secondary school tracks is based on the track 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. Each student then takes a standardized end-of-primary education test in April or May. This end-of-primary education test also maps into a track recommendation. If this test-based recommendation is higher than the teacher recommendation, the teacher should consider an upwards 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 secondary schools cannot place students in a lower track than indicated by the track recommendation (WVO, 2014).

An important characteristic of the Dutch education system is that primary schools are required to implement a student monitoring system. This includes the requirement that all primary school students take two standardized tests per school year, from grade one to six. The first test is administered in the middle of the 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. Note that the end term test in sixth grade is the end-of-primary school test. The biannual tests generally focus on two domains: mathematics and reading. The mathematics domain contains both abstract problems and contextual problems that describe a concrete task. The reading domain assesses the ability to understand written texts, including both factual and literary content.

There are no strict national rules or procedures on which information to use when formulating the track recommendation. However, the guidelines from the Ministry of Education (2023a) state that standardized test results provide a "good starting point". Consistent with this guideline, a recent survey among 400 primary school teachers shows that 85% of the respondents attach "considerable" to "a lot of" value to the standardized test results when formulating track recommendations (DUO, 2023). Teachers are also expected to make a holistic assessment and may therefore take into account their subjective judgements on factors such as motivation, attitude towards school, and classroom behavior. This discretionary power of primary school teachers in the Dutch track assignment process has been criticized for creating inconsistency and bias (OECD, 2016).

In this paper we use the initial, instead of the final, track recommendation as the teacher's subjective evaluation. The reason is that the initial track recommendation can be considered the most important. Many secondary schools already start allocating students to their first-year classes in March using the initial track 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, the initial and final track recommendation are identical for 9 out of 10 students.[4]

We use the fifth-grade end term test as a measure of ability because it can be considered the final test score that the teacher observes before formulating the initial track recommendation. Alternatively, one could use the final track recommendation as the teacher's subjective evaluation and the end-of-primary school test result as the ability measure. This ap-

[4]For instance, in 2018-19, 41.50% of all students obtained a test-based recommendation above the initial track recommendation, and 22.96% of them received an upwards revision (Inspectorate of Education, 2022). This implies that the initial and final recommendation were different for 9.53% of the students in this year.

proach is problematic since the final track recommendation can only be adjusted upwards, introducing additional sources of measurement error for the end-of-primary school test: the test is low (high) stakes for students who are (un)satisfied with the initial recommendation.

3.2 Data

We use proprietary administrative data from Statistics Netherlands on primary school students in the 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 test score data can be linked to other administrative data using an anonymized personal ID, allowing us to also observe the initial teacher track recommendation, parental income and education, and several other student background characteristics.

The biannual standardized test scores are available for roughly 50% of the primary schools in the Netherlands. The NCO collected this data from 2018-19 onwards by individually contacting primary schools. Each school had to provide permission for the data collection, where parents were given the possibility to raise objections. The NCO only collected the data of schools that made use of standardized tests from the test developer *Cito*. With roughly 80% of the primary schools using *Cito* across years, this is the largest test provider in the Netherlands. See Haelermans et al. (2020) for more details on the NCO data collection.

The complete standardized test score history, from grade one to six, is available for the students who were enrolled in sixth grade from 2018-19 onwards. Our sample therefore covers four cohorts: students who are in sixth grade in the school years 2018-19, 2019-20, 2020-21, and 2021-22. For these four cohorts, the biannual test score data contains information on 318,630 students in 5,887 schools. After merging the biannual test score data with the data on track recommendations and student background characteristics, we observe 276,841 students in 5,802 schools. After dropping the students with missing values for any variable used in our analyses, and with extreme values for household income (discussed below), our estimation sample includes 148,019 students in 3,295 schools.

3.3 Descriptive statistics

Table 1 presents the descriptive statistics for our estimation sample. Panel A focuses on several measures for the initial track recommendation. Our baseline outcome variable y_t is a dummy variable that equals one if the track recommendation is at least equal to the college track. Enrollment in the college or university track is most relevant from a policy perspective, since it implies the student is tracked towards higher education. Panel A shows

Table 1: Descriptive statistics

	(1) Mean	(2) SD	(3) Min	(4) Max
Panel A: Teacher recommendation				
$\mathbb{1}$(recom. ≥ upper voc. tr.)	0.770	0.421	0.000	1.000
$\mathbb{1}$(recom. ≥ college tr.)	0.497	0.500	0.000	1.000
$\mathbb{1}$(recom. ≥ university tr.)	0.211	0.408	0.000	1.000
Panel B: SES measures				
SES income	6.439	3.772	<-0.410	>55.500
$\mathbb{1}$(SES inc. ≥ median)	0.517	0.500	0.000	1.000
SES years of schooling	15.975	2.334	8.000	18.000
$\mathbb{1}$(SES edu. ≥ college)	0.559	0.497	0.000	1.000
Panel C: Test scores				
Reading midterm grade 5	191.618	26.766	<81.000	>340.000
Reading end term grade 5	197.279	26.983	<84.000	>354.000
Math midterm grade 5	254.142	27.142	<118.000	>379.000
Math end term grade 5	261.959	26.085	<118.000	>394.000
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				
$\mathbb{1}$(female)	0.499	0.500	0.000	1.000
$\mathbb{1}$(Western imm.)	0.076	0.265	0.000	1.000
$\mathbb{1}$(non-Western imm.)	0.200	0.400	0.000	1.000
Age	11.590	0.647	<9.250	>14.160
Observations	148019			

Notes: this table shows the descriptive statistics for our estimation sample. 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. Due to confidentiality reasons we cannot show 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.

that 49.7% of the students receive a track recommendation equal to, or higher than, the college track.

Panel B shows the statistics for four different measures of SES. Our baseline measure of SES will be gross annual household income, which is €64,390 on average. Household income is right-skewed with a maximum (minimum) of €550,000 (€-4,100). We dropped observations with a yearly household income above (below) the 99.9th (0.1st) percentile. To facilitate interpretation, for our analysis below we standardize household income per cohort. In our robustness checks we use three alternative SES measures. First, we use parental years of schooling of the highest educated parent, which is 16 years on average. Moreover, both the income and education measure are binarized. The binary income measure equals one if household income is above the median of a student's cohort, while the binary education measure equals one if the highest educated parent holds at least a college degree.

Panel C shows the statistics for the midterm and end term test in fifth grade, per domain and the average across domains. The latter is calculated by first standardizing the domain-specific test scores per cohort and subsequently taking the average. We use this average score on the fifth-grade end term test as our main measure of ability s_t. We address measurement error, among others, by using the average score on the fifth-grade midterm test as our measure for s_{t-1}. Table 5 in Appendix D shows the descriptive statistics on the midterm and end term tests from grades two to four, which form our measures for $s_{t-\tau}$ with $\tau > 1$. We do not use the data from grade one since the biannual reading test is not observed in that grade.

While our baseline analysis does not include any control variables, in robustness tests we control for gender, migration background, age, and cohort dummies. Panel D shows that 49.9% of the students are female, 7.6% (20%) of the students have a Western (non-Western) immigration background, and students are on average 11.6 years old in March of the sixth grade.

4 Results

4.1 The OLS strategy

The baseline OLS estimates are presented in Table 2. Column (1) shows the OLS estimates of the medium regression equation (3). Consistent with previous evidence from the Netherlands (Inspectorate of Education, 2016; Timmermans et al., 2018; Zumbuehl et al., 2022), 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

Table 2: OLS estimates

	(1)	(2)
	$\mathbb{1}$(recom. ≥ college tr.)	End term grade 5
SES income	0.028***	0.229***
	(0.001)	(0.004)
End term grade 5	0.381***	
	(0.001)	
Observations	148019	148019
R^2	0.498	0.064

Notes: ***, **, * refers to statistical significance at the 1, 5, and 10% level. The numbers represent the estimated coefficients and the numbers in parentheses the corresponding standard errors. Coefficients are estimated with OLS. Standard errors are clustered on the school id.

for β^m indicates 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.

As expected, column (1) further shows that test scores are positively associated with teacher track recommendations. This estimate is attenuated in the presence of measurement error. Whether and how this contaminates the estimate for the conditional SES gap depends upon the results of the balancing regression equation (4), which is presented in column (2). The OLS estimate for δ is positive, meaning that parental income is positively associated with the fifth-grade end term test score. We can therefore conclude that our estimate for $\gamma\delta$ is positive and that, in the presence of measurement error, OLS overestimates the conditional SES gap.

Our baseline models do not include any controls other than the fifth-grade end term test score. Table 6 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 results show that the additional controls hardly affect the conditional SES gap: the estimate changes from 0.028 to 0.029. This may be expected as many of the controls are not associated with SES. However, this is not the case for migration background, which is strongly correlated with SES.[5] In general, these results suggest that our main estimates

[5]Note that the results in Table 6 indicate that students with a migration background are more likely to receive at least a college track recommendation, conditional on SES and test scores. This is not inconsistent with the literature, which reports mixed evidence on teacher bias by migration status (Batruch et al., 2023).

Table 3: IV estimates

	(1) End term grade 5	(2) $\mathbb{1}$(recom. ≥ college tr.)	(3) $\mathbb{1}$(recom. ≥ college tr.)
Midterm grade 5	0.887***	0.383***	
	(0.002)	(0.001)	
SES income	0.031***	0.030***	0.016***
	(0.001)	(0.001)	(0.001)
End term grade 5			0.432***
			(0.002)
Observations	148019	148019	148019
R^2	0.794	0.500	0.490
$(\sigma^2_{u_{t-1}} + \sigma^2_{m_{t-1}})/(\sigma^2_{u_t} + \sigma^2_{m_t})$	0.992		

Notes: ***, **, * refers to statistical significance at the 1, 5, and 10% level. The numbers represent the estimated coefficients and the numbers in parentheses the corresponding standard errors. Coefficients in column (1) and (2) are estimated with OLS and in column (3) with 2SLS. Standard errors are clustered on the school id.

are rather robust to controlling for additional variables typically observed in administrative data.

4.2 The IV strategy

Table 3 shows 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 (1) shows that the coefficient estimate of 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.[6] Hence, the first stage estimate for the reliability ratio is $0.887 \times 0.992 = 0.880$. The presence of measurement error implies that the OLS estimate of 0.028 overestimates the conditional SES gap.

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 (2). In particular, the sec-

[6]As shown in Proposition 2, under Assumption 2 this fraction can be written as $(\sigma^2_{u_t} + \sigma^2_{\Delta_{s_t}} + \sigma^2_{m_{t-1}})/(\sigma^2_{u_t} + \sigma^2_{m_t})$. Hence, whether this fraction is bigger or smaller than one depends on the variance of the change in ability and the change in the variance of measurement error over time.

ond stage estimate on the end term test in column (3) can be calculated by dividing the reduced form by the first stage, $\frac{0.383}{0.887} = 0.432$. If the second stage estimate γ^{iv} is equal to γ, there cannot be any contamination bias, and the second stage estimate β^{iv} identifies the conditional SES gap β. The 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.

4.3 The EIV strategy

The EIV strategy requires a value for the reliability ratio of the end term test to address measurement error. 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 (Feld and Zölitz, 2017). Depending on the topic, the Graduate Management Admission Test (GMAT) is advertised to have a reliability ratio between 0.89 and 0.90 (Graduate Management Admission Council, 2023), the Graduate Record Examination (GRE) between 0.87 and 0.95 (Educational Testing Service, 2023), and the Scholastic Assessment Test (SAT) between 0.89 and 0.93 (College Board, 2013).

The EIV FS strategy uses the first stage estimate for the reliability ratio to directly estimate γ and β. In particular, with the reliability ratio equal to 0.880 the OLS estimate on the fifth-grade end term test γ^m is attenuated by a factor of 0.880. Hence, the EIV FS estimate γ^{fs} is equal to $\frac{0.381}{0.880} = 0.433$. This attenuation bias contaminated the OLS estimate for β^m from Table 2. Accordingly, the conditional SES gap is overestimated by a value of $\gamma\delta(1-\lambda) = 0.433 \times 0.229 \times (1-0.880) = 0.012$. The EIV strategy subtracts this value from β^m, and so the estimate for β^{fs} is equal to $0.028 - 0.012 = 0.016$.

An alternative estimate for the reliability ratio can be 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 rightmost 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 of a test score very close to the fifth-grade end term test. We use a linear polynomial for the forecasting regression equation (30). With an R^2 of 0.981 this fits the development of the first stage estimates relatively well. 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.

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

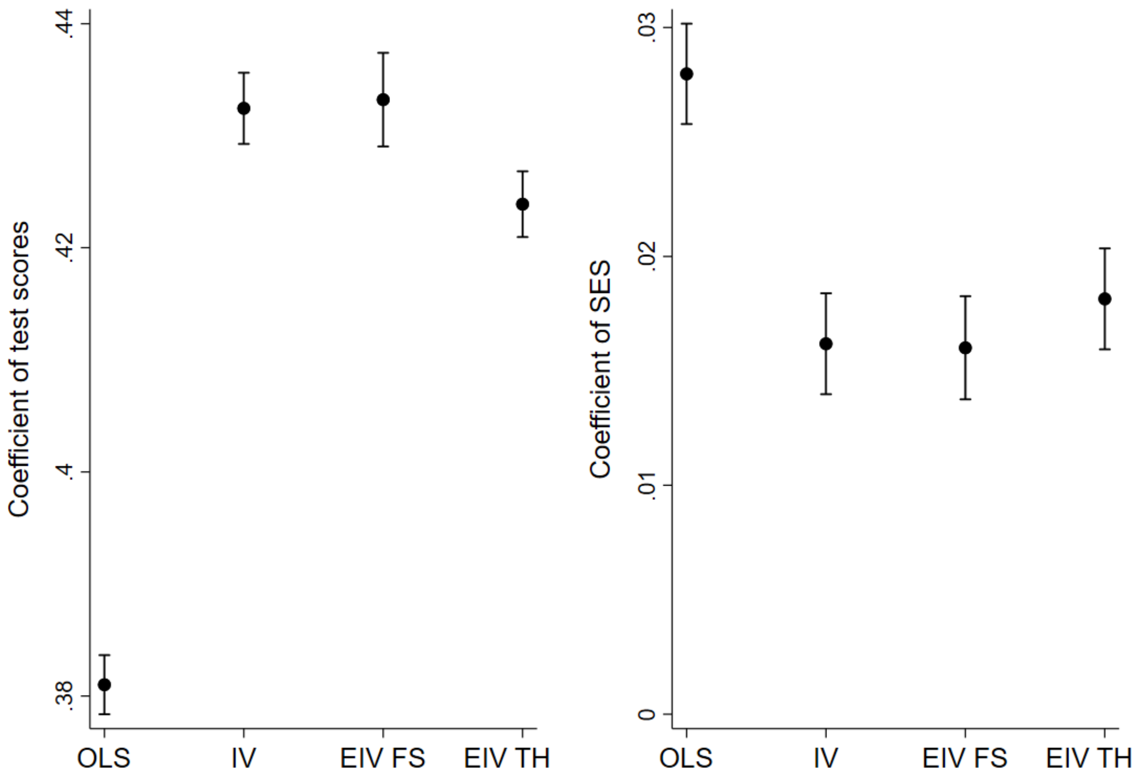
Notes: this figure plots the estimates for $\widetilde{\pi}_{t-\tau}$ and the forecasting regression equation to estimate $\lambda = f(t-\epsilon)$. We use OLS to estimate $\pi_{t-\tau}$ from (28), and the variance of the error term from (4), to obtain $\widetilde{\pi}_{t-\tau}$ for each τ. The forecasting regression equation (30) is specified as a polynomial of degree one, estimated with OLS, and then used to obtain the one-period out-of-sample forecast $f(t-\epsilon)$.

4.4 Comparing the 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. 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 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. Since the estimate for the reliability ratio is somewhat higher when using the one-period out-of-sample forecast instead of the first stage (0.899 versus 0.880), the attenuation and contamination bias are somewhat smaller for the EIV TH strategy.

Table 4 summarizes how much of the conditional SES gap in track recommendations can be explained by measurement error in test scores across the three strategies. Column

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 and 3 respectively. The estimates and standard errors for the EIV FS are obtained via a stacking approach. We triplicate the data and estimate (3), (4), and (8) with OLS via a single regression equation. We cluster on the school id that is similarly triplicated. The EIV TH strategy uses the same stacking approach to estimate (3) and (4). See Figure 1 for details on the estimation of the one-period out-of-sample forecast $f(t-\epsilon)$. To obtain standard errors, we assume that the estimate for $f(t-\epsilon)$ is fixed.

(1) shows that the IV strategy estimates that 42.14% of the SES gap is due to measurement error. Column (2) and (3) show that the EIV strategies estimate this percentage between 35.13 and 42.78, depending on the method used to estimate the reliability ratio. In general, the results show that mitigating contamination bias matters substantially and that the three strategies addressing measurement yield similar results.

Table 4: Percentage change in the conditional SES gap

	(1) $\beta^{\bullet} = \beta^{iv}$	(2) $\beta^{\bullet} = \beta^{fs}$	(3) $\beta^{\bullet} = \beta^{th}$
$\left(\frac{\beta^m - \beta^{\bullet}}{\beta^m}\right) \times 100$	42.14%	42.78%	35.13%

Notes: the numbers represent the estimated values for$\left(\frac{\beta^m-\beta^{\bullet}}{\beta^m}\right) \times 100$. The estimate for β^m is shown in Table 2, for β^{iv} in Table 3, and for β^{fs} and β^{th} in Figure 2.

4.5 Testing the assumptions

Since measurement error and ability are unobserved, we are unaware of testable implications for Assumption 1 and 2(b). However, Assumption 2(a) and 3 do have immediate testable implications. Define $\Delta_{m_t} = m_t - m_{t-1}$, so that $s_t^m - s_{t-1}^m = \Delta_{s_t} + \Delta_{m_t} = \Delta_{s_t^m}$. 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. Classical measurement error on the left-hand side does not affect the OLS estimate, and so Assumption 2(a) requires that α_{SES} is zero. In turn, Assumption 3 requires that α_x is zero for every student characteristic x, including SES.[7]

Table 7 in Appendix D shows OLS estimates of the change in 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% level. The immigrant dummies seem of non-negligible size. These results cast strong 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 mismeasured variable is highly dynamic.

[7] Using test scores s_t^m as student characteristic x, instead of unobserved ability s_t, may produce an overestimate for α_{s_t} and thus a misleading test for Assumption 2(b). To see this, we specify the following equations,

$$\begin{pmatrix} \Delta_{s_t^m} \\ \Delta_{s_t^m} \end{pmatrix} = \begin{pmatrix} \alpha_{s_t} \\ \alpha_{s_t}^m \end{pmatrix} \begin{pmatrix} s_t \\ s_t^m \end{pmatrix} + \begin{pmatrix} v_t \\ v_t^m \end{pmatrix}.$$

It can be shown that the OLS estimate for $\alpha_{s_t}^m = 1 + \left(\alpha_{s_t} - 1\right)\lambda^u$, where $\lambda^u = \left(\frac{\sigma_{s_t}^2}{\sigma_{s_t}^2 + \sigma_{m_t}^2}\right)$ is the univariate reliability ratio. Given $\lambda^u < 1$, we have that $\alpha_{s_t}^m > 0$ even if $\alpha_{s_t} = 0$. The intuition for this positive bias is that the same measurement error m_t enters both the left-hand and right-hand side of the equation.

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 Δ_{s_t} is uncorrelated with s_t, the following two equalities on the process $s_t = s_{t-1} + \Delta_{s_t}$ directly follow:

$$-\sigma^2_{\Delta s_t} = \text{Cov}[\Delta_{s_t}, s_{t-1}], \tag{32}$$

$$\sigma^2_{\Delta s_t} = \sigma^2_{s_{t-1}} - \sigma^2_{s_t}. \tag{33}$$

Low ability students in $t-1$ must have had larger positive changes in ability than high ability students in $t-1$ (from (32)) and the variance of ability must be larger in $t-1$ than in t (from (33)). Whether these equalities are likely depends on the setting at hand. In an education context, (32) and (33) would be consistent with diminishing marginal returns to ability.

Although not explicitly written down as an assumption, using students' test score histories to estimate the reliability ratio requires that we can accurately capture the changes in the first stage estimate over time. Figure 1 shows this is the case, as the linear polynomial captures 98.1% of the variance in the first stage estimate over time. This also suggests that the underlying factors that drive the first stage away from the reliability ratio, which are $\text{Cov}[\Delta s_t, SES] \neq 0$ and $\text{Cov}[\Delta s_t, s_t] \neq 0$, close over time in a predictable way. Figure 3 in Appendix D plots OLS estimates of Δs^m_t on SES over time, and shows that $\text{Cov}[\Delta s^m_t, SES]$ indeed approaches zero in a near linear fashion.

4.6 Explaining the difference in results

The results above suggest that both Assumption 2 and 3 are violated. Can such violations explain the relatively small difference in results across the various strategies? To analyze this question, recall that the EIV TH strategy identifies the true estimates under Assumption 1 only. Our empirical results can be explained by violations of Assumption 2 and 3 with $B^1_\gamma \approx 1$ and $B^1_\pi > 1$, as this generates:

$$\gamma^m < \gamma^{th} = \gamma < \gamma^{iv} \approx \gamma^{fs}, \tag{34}$$

$$\beta^{fs} \approx \beta^{iv} < \beta^{th} = \beta < \beta^m. \tag{35}$$

First, Proposition 1 shows that OLS underestimates γ, and in turn overestimates β, given the presence of classical measurement error. Second, case (i) of Proposition 2 shows that the IV strategy overestimates γ, and in turn underestimates β, if $B^1_\gamma \approx 1$ and $B^1_\pi > 1$. The reason is that the reduced form will be equal to the medium regression estimate $\gamma^m = \gamma\lambda$ if $B^1_\gamma \approx 1$, but the first stage will be an underestimate for the reliability ratio λ if $B^1_\pi > 1$. Note that $B^1_\gamma \approx 1$ if $\text{Cov}[y_t, u_t] \approx \text{Cov}[y_t, u_{t-1}]$ and $B^1_\pi > 1$ if $\sigma^2_{u_t} > \text{Cov}[u_t, u_{t-1}]$. The latter may happen if for instance Assumption 2(b) is violated, so that $\sigma^2_{u_t} = \text{Cov}[u_t, u_{t-1}] + \text{Cov}[\Delta_{s_t}, s_t]$, and $\text{Cov}[\Delta_{s_t}, s_t] > 0$. This implies that high ability students in t must have had larger positive changes in ability than low ability students in t.

Third, case (i) of Proposition 3 shows that the EIV FS strategy also overestimates γ, and in turn underestimates β, if $B^1_\pi > 1$. A comparison between case (i) of Proposition 2 and 3 shows that the bias to the EIV FS and IV strategy are similar if $B^1_\gamma \approx 1$. The reason is that in this case the medium regression estimate, which is used by the EIV FS strategy, is similar to the reduced form estimate, which is used by the IV strategy.

Our empirical results can be interpreted in this light. Both the IV and EIV FS strategy use the first stage estimate of 0.880 for the reliability ratio. The IV strategy uses this to correct the reduced form estimate of 0.380, and produces $\widehat{\gamma}^{iv} = \frac{0.380}{0.880} = 0.432$. In contrast, the EIV FS strategy uses this to correct the medium regression estimate of 0.381, and produces $\widehat{\gamma}^{fs} = \frac{0.381}{0.880} = 0.433$. As the medium and reduced form regression estimates are similar, the results for the IV and EIV FS are similar. However, the correction for the attenuation bias by both strategies is slightly too large, since with data on students' test score histories we estimate a somewhat larger reliability ratio of 0.892. Using this to correct the medium regression estimate produces $\widehat{\gamma}^{th} = \frac{0.381}{0.899} = 0.424$.

4.7 Robustness checks

We test the robustness of our results in four ways. First, Figure 4 in Appendix D shows the results when using two alternative measures for teacher 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 results: the pattern of the estimates is identical in each figure and measurement error explains a sizeable share of the conditional SES gap.

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. This is consistent with our OLS estimates in Table 6 of Appendix D that show the conditional SES gap is hardly affected by controlling for migration background.

5 Conclusion

This paper examines the role of measurement error in explaining conditional gaps in teacher evaluations. Teacher bias is generally measured by estimating gaps in subjective teacher evaluations, while conditioning on objective measures of student abilities. A major challenge in the identification of teacher bias is the presence of measurement error in test scores that are supposed to reflect student abilities. We discuss three different strategies to address measurement error and explain under which assumptions these strategies are valid. Whereas an instrumental variable approach is more commonly used in applied econometric studies, we show that an errors-in-variables approach requires weaker assumptions to address measurement error and contamination bias.

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 well established in the literature. Using recent administrative data covering about half of the Dutch primary schools, 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 such a high share of the conditional SES gap can be explained by measurement error, especially since the test scores we use in our analyses are highly reliable according to conventional norms. However, when student abilities are strongly correlated with SES and matter for teacher track recommendations, the magnitude of contamination bias may be nontrivial - even when tests are highly reliable. While our findings show that teacher bias may be smaller than previous studies report, the strong correlation between SES and ability does suggest large inequalities of opportunity before and during primary education.

Our findings also have important implications for the larger literature studying bias and discrimination. Studies often suggest that systematic differences in outcomes across groups (e.g., student grades, job promotion, wages) conditional on control variables (e.g., test scores, education, experience) indicate bias. However, this claim is invalid in the presence of measurement error in the controls. It is therefore key that studies testing the presence of systematic differences in outcomes properly address measurement error in control variables.

References

Alesina, A., M. Carlana, E. La Ferrara, and P. Pinotti (2018). Revealing stereotypes: Evidence from immigrants in schools. Working Paper 25333, National Bureau of Economic

Research.

Batruch, A., S. Geven, E. Kessenich, and H. G. van de Werfhorst (2023). Are tracking recommendations biased? A review of teachers' role in the creation of inequalities in tracking decisions. *Teaching and Teacher Education 123*, 103985.

Borghans, L., R. Diris, W. Smits, and J. de Vries (2019). The long-run effects of secondary school track assignment. *PloS one 14*(10), e0215493.

Borghans, L., R. Diris, W. Smits, and J. de Vries (2020). Should we sort it out later? The effect of tracking age on long-run outcomes. *Economics of Education Review 75*, 101973.

Borjas, G. J. and D. S. Hamermesh (2024). The mismeasurement of work time: Implications for wage discrimination and inequality. Working Paper 32025, National Bureau of Economic Research.

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.

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.

Culpepper, S. A. (2012). Evaluating EIV, OLS, and SEM estimators of group slope differences in the presence of measurement error: The single-indicator case. *Applied Psychological Measurement 36*(5), 349–374.

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. 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.

Educational Testing Service (2023). Guide to the use of scores. https://www.ets.org/pdfs/gre/gre-guide-to-the-use-of-scores.pdf, Accessed August 2023.

Falk, A., F. Kosse, and P. Pinger (2020). Mentoring and schooling decisions: Causal evidence. Working Paper No. 13387, IZA.

Feld, J. and U. Zölitz (2017). Understanding peer effects: On the nature, estimation, and channels of peer effects. *Journal of Labor Economics 35*(2), 387–428.

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.

Graduate Management Admission Council (2023). Validity, reliability & fairness. https://www.gmac.com/gmat-other-assessments/about-the-gmat-exam/validity-reliability-fairness, Accessed August 2023.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika 10*(4), 255–282.

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.

Hill, A. J. and D. B. Jones (2021). Self-fulfilling prophecies in the classroom. *Journal of Human Capital 15*(3), 400–431.

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.

Inspectorate of Education (2022). Heroverwegen en bijstellen van schooladviezen na de eindtoets. Technical report, Inspectie van het Onderwijs.

Kelley, T. L. (1927). *Interpretation of educational measurements.* World Book Company.

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 (2023a). De overstap van PO naar VO: Handreiking schooladvisering. Ministerie van Onderwijs, Cultuur en Wetenschap.

Ministry of Education (2023b). Kamerbrief over schooladvisering en doorstroomtoetsen PO 2023, 3 april 2023. Ministerie van Onderwijs, Cultuur en Wetenschap.

Modalsli, J. and K. Vosters (2022). Spillover bias in multigenerational income regressions. *Journal of Human Resources*, 0520–10919R2.

OECD (2016). *Netherlands 2016: Foundations for the Future, Reviews of National Policies for Education.* OECD publishing.

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.

Reiersøl, O. (1941). Confluence analysis by means of lag moments and other methods of confluence analysis. *Econometrica 9*(1), 1–24.

Rosenthal, R. and L. Jacobson (1968). Pygmalion in the classroom. *The urban review 3*(1), 16–20.

Terrier, C. (2020). Boys lag behind: How teachers' gender biases affect student achievement. *Economics of Education Review 77*, 101981.

Timmermans, A. C., H. de Boer, H. T. Amsing, and M. Van Der Werf (2018). Track recommendation bias: Gender, migration background and SES bias over a 20-year period in the Dutch context. *British Educational Research Journal 44*(5), 847–874.

Triventi, M. (2020). Are children of immigrants graded less generously by their teachers than natives, and why? Evidence from student population data in Italy. *International Migration Review 54*(3), 765–795.

Wainer, H. and L. M. Brown (2006). 28 Three statistical paradoxes in the interpretation of group differences: Illustrated with medical school admission and licensing data. *Handbook of statistics 26*, 893–918.

Ward, Z. (2023). Intergenerational mobility in American history: Accounting for race and measurement error. *American Economic Review 113*(12), 3213–48.

WPO (2014). Artikel 42, lid 2, Wet op het primair onderwijs. Datum inwerkingtreding: 01/08/2014. wetten.overheid.nl.

WVO (2014). Artikel 3, lid 2, Inrichtingsbesluit Wet op het voortgezet onderwijs. Datum inwerkingtreding: 01/08/2014. wetten.overheid.nl.

Zumbuehl, M., N. Chehber, and R. Dillingh (2022). Can skills differences explain the gap in track recommendation by socio-economic status? Technical report, CPB Netherlands Bureau for Economic Policy Analysis.

A Proof Proposition 1

Using Assumption 1, we can write the balancing regression equation (4) as

$$\begin{pmatrix} s_t \\ s_t^m \end{pmatrix} = \begin{pmatrix} s_t \\ s_t + m_t \end{pmatrix} = \delta SES + \begin{pmatrix} u_t \\ u_t + m_t \end{pmatrix}. \tag{36}$$

Applying the Frisch-Waugh logic, we can now derive the expression for γ^m as follows,

$$\begin{aligned} \gamma^m =& \left(\frac{\text{Cov}[y_t, u_t^m]}{\text{Var}[u_t^m]}\right) = \left(\frac{\text{Cov}[y_t, u_t + m_t]}{\text{Var}[u_t + m_t]}\right) = \left(\frac{\text{Cov}[y_t, u_t]}{\sigma^2_{u_t} + \sigma^2_{m_t}}\right) \\ =& \left(\frac{\text{Cov}[y_t, u_t]}{\sigma^2_{u_t}}\right)\left(\frac{\sigma^2_{u_t}}{\sigma^2_{u_t} + \sigma^2_{m_t}}\right) \\ =& \gamma\lambda, \end{aligned} \tag{37}$$

where $\lambda = \left(\frac{\sigma^2_{u_t}}{\sigma^2_{u_t} + \sigma^2_{m_t}}\right)$.

To derive the expression for β^m, we first introduce the short regression equation:

$$y_t = \beta^s SES + e_t^s. \tag{38}$$

Next, we use the omitted variable bias formula for β^s to write

$$\beta^m + \gamma^m \delta^m = \beta + \gamma\delta. \tag{39}$$

Using that $\delta^m = \delta$ and substituting for $\gamma^m = \gamma\lambda$, results in

$$\beta^m = \beta + \gamma\delta\left(1 - \lambda\right). \tag{40}$$

B Proof Proposition 2

B.1 Case (i)

We start by rewriting the second stage equation (9) to obtain the reduced form regression equation,

$$y_t = \theta_0 SES + \theta_1 s_{t-1}^m + e_t^{iv}, \tag{41}$$

where $\frac{\theta_1}{\pi} = \frac{\gamma^{iv}\pi}{\pi} = \gamma^{iv}$.

Using Assumption 1, we can write the balancing regression equation (10) as

$$\begin{pmatrix} s_{t-1} \\ s_{t-1}^m \end{pmatrix} = \begin{pmatrix} s_{t-1} \\ s_{t-1} + m_{t-1} \end{pmatrix} = \delta_{t-1} SES + \begin{pmatrix} u_{t-1} \\ u_{t-1} + m_{t-1} \end{pmatrix}, \tag{42}$$

Next we derive the expression for the first stage coefficient π. Applying the Frisch-Waugh logic, results in

$$\begin{aligned}\pi =&\left(\frac{\text{Cov}[s_t^m, u_{t-1}^m]}{\text{Var}[u_{t-1}^m]}\right) = \left(\frac{\text{Cov}[s_t + m_t, u_{t-1} + m_{t-1}]}{\text{Var}[u_{t-1} + m_{t-1}]}\right) = \left(\frac{\text{Cov}[s_t, u_{t-1}]}{\sigma^2_{u_{t-1}} + \sigma^2_{m_{t-1}}}\right) \\ =&\left(\frac{\text{Cov}[s_t, u_{t-1}]}{\sigma^2_{u_t} + \sigma^2_{m_t}}\right)\left(\frac{\sigma^2_{u_t} + \sigma^2_{m_t}}{\sigma^2_{u_{t-1}} + \sigma^2_{m_{t-1}}}\right) \\ =&\left(\frac{\sigma^2_{u_t}}{\sigma^2_{u_t} + \sigma^2_{m_t}}\right)\left(\frac{\text{Cov}[s_t, u_{t-1}]}{\sigma^2_{u_t}}\right)\left(\frac{\sigma^2_{u_t} + \sigma^2_{m_t}}{\sigma^2_{u_{t-1}} + \sigma^2_{m_{t-1}}}\right) \\ =&\left(\frac{\sigma^2_{u_t}}{\sigma^2_{u_t} + \sigma^2_{m_t}}\right)\left(\frac{\text{Cov}[\delta SES + u_t, u_{t-1}]}{\sigma^2_{u_t}}\right)\left(\frac{\sigma^2_{u_t} + \sigma^2_{m_t}}{\sigma^2_{u_{t-1}} + \sigma^2_{m_{t-1}}}\right) \\ =&\lambda\left(\frac{\text{Cov}[u_t, u_{t-1}]}{\sigma^2_{u_t}}\right)\left(\frac{\sigma^2_{u_t} + \sigma^2_{m_t}}{\sigma^2_{u_{t-1}} + \sigma^2_{m_{t-1}}}\right).\end{aligned} \tag{43}$$

We continue with the reduced form coefficient θ_1. Applying the Frisch-Waugh logic, results in

$$\begin{aligned}\theta_1 =&\left(\frac{\text{Cov}[y_t, u_{t-1}^m]}{\text{Var}[u_{t-1}^m]}\right) = \left(\frac{\text{Cov}[y_t, u_{t-1} + m_{t-1}]}{\text{Var}[u_{t-1} + m_{t-1}]}\right) = \left(\frac{\text{Cov}[y_t, u_{t-1}]}{\sigma^2_{u_{t-1}} + \sigma^2_{m_{t-1}}}\right) \\ =&\left(\frac{\text{Cov}[y_t, u_{t-1}]}{\sigma^2_{u_t}}\right)\left(\frac{\sigma^2_{u_t}}{\sigma^2_{u_{t-1}} + \sigma^2_{m_{t-1}}}\right) \\ =&\left(\frac{\text{Cov}[y_t, u_t]}{\sigma^2_{u_t}}\right)\left(\frac{\text{Cov}[y_t, u_{t-1}]}{\text{Cov}[y_t, u_t]}\right)\left(\frac{\sigma^2_{u_t}}{\sigma^2_{u_t} + \sigma^2_{m_t}}\right)\left(\frac{\sigma^2_{u_t} + \sigma^2_{m_t}}{\sigma^2_{u_{t-1}} + \sigma^2_{m_{t-1}}}\right) \\ =&\gamma\lambda\left(\frac{\text{Cov}[y_t, u_{t-1}]}{\text{Cov}[y_t, u_t]}\right)\left(\frac{\sigma^2_{u_t} + \sigma^2_{m_t}}{\sigma^2_{u_{t-1}} + \sigma^2_{m_{t-1}}}\right).\end{aligned} \tag{44}$$

We finish with the second stage estimates γ^{iv} and β^{iv}. Combining (43) and (44), results in

$$\gamma^{iv} = \frac{\theta_1}{\pi} = \gamma\left(\frac{\text{Cov}[y_t, u_{t-1}]}{\text{Cov}[y_t, u_t]}\right)\left(\frac{\sigma^2_{u_t}}{\text{Cov}[u_t, u_{t-1}]}\right). \tag{45}$$

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

$$\begin{pmatrix} s_t^m \\ \hat{s}_t^m \end{pmatrix} = \begin{pmatrix} s_t^m \\ s_t^m - e_t^{fs} \end{pmatrix} = \begin{pmatrix} \delta^m \\ \delta^{iv} \end{pmatrix} SES + \begin{pmatrix} u_t^m \\ u_t^{iv} \end{pmatrix}. \tag{46}$$

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

$$\begin{aligned}\delta^{iv} =&\left(\frac{\text{Cov}[\hat{s}_t^m, SES]}{\text{Var}[SES]}\right) = \left(\frac{\text{Cov}[s_t^m - e_t^{fs}, SES]}{\sigma^2_{SES}}\right) = \left(\frac{\text{Cov}[s_t^m, SES]}{\sigma^2_{SES}}\right) \\ =&\delta^m,\end{aligned} \tag{47}$$

and under Assumption 1 we have that $\delta^{iv} = \delta^m = \delta$. We use the omitted variable bias formula for β^s from (38) to write

$$\beta^{iv} + \gamma^{iv}\delta^{iv} = \beta + \gamma\delta. \tag{48}$$

Using $\delta^{iv} = \delta$ and (45) it follows that,

$$\beta^{iv} = \beta + \gamma\delta\left(1 - \left(\frac{\text{Cov}[y_t, u_{t-1}]}{\text{Cov}[y_t, u_t]}\right)\left(\frac{\sigma^2_{u_t}}{\text{Cov}[u_t, u_{t-1}]}\right)\right). \tag{49}$$

B.2 Case (ii)

First we use Assumption 2 to find a link between u_{t-1} and u_t. We can write,

$$\begin{aligned}\delta &= \left(\frac{\text{Cov}[s_t, SES]}{\text{Var}[SES]}\right) = \left(\frac{\text{Cov}[s_{t-1} + \Delta_{s_t}, SES]}{\sigma^2_{SES}}\right) \\ &= \left(\frac{\text{Cov}[s_{t-1}, SES]}{\sigma^2_{SES}}\right) + \left(\frac{\text{Cov}[\Delta_{s_t}, SES]}{\sigma^2_{SES}}\right), \\ &= \delta_{t-1} + \alpha,\end{aligned} \tag{50}$$

with $\alpha = \left(\frac{\text{Cov}[\Delta_{s_t}, SES]}{\sigma^2_{SES}}\right)$. Assumption 2(a) guarantees that $\delta = \delta_{t-1}$. This allows us to write (4) and (10) as

$$\begin{pmatrix} s_t \\ s_{t-1} \end{pmatrix} = \begin{pmatrix} s_t \\ s_t - \Delta s_t \end{pmatrix} = \delta SES + \begin{pmatrix} u_t \\ u_t - \left(\Delta_{s_t} - \overline{\Delta}_{s_t}\right) \end{pmatrix}. \tag{51}$$

The mean of Δ_{s_t}, denoted by $\overline{\Delta}_{s_t}$, will be absorbed by the intercept of (10), so that $u_{t-1} = u_t - \left(\Delta_{s_t} - \overline{\Delta}_{s_t}\right)$. We use this to write $\text{Cov}[u_t, u_{t-1}] = \sigma^2_{u_t} - \text{Cov}[u_t, \Delta_{s_t}]$. 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 $\text{Cov}[u_t, u_{t-1}] = \sigma^2_{u_t}$.

Case (i) assumes Assumption 1 and case (ii) assumes both Assumption 1 and 2. Hence, we use the results above derived under Assumption 2 to simplify the expressions in case (i).

We start with the first stage coefficient π derived in (43). We note that the term $\left(\frac{\text{Cov}[u_t, u_{t-1}]}{\sigma^2_{u_t}}\right)$ equals 1 since $\text{Cov}[u_t, u_{t-1}] = \sigma^2_{u_t}$. Assumption 2 further allows to substitute for $\sigma^2_{u_{t-1}} = \text{Var}[u_t - \left(\Delta_{s_t} - \overline{\Delta}_{s_t}\right)] = \text{Var}[u_t - \Delta_{s_t}] = \sigma^2_{u_t} + \sigma^2_{\Delta_{s_t}} - 2 \times \text{cov}[u_t, \Delta_{s_t}] = \sigma^2_{u_t} + \sigma^2_{\Delta_{s_t}} - 2 \times \text{cov}[s_t - \delta SES, \Delta_{s_t}] = \sigma^2_{u_t} + \sigma^2_{\Delta_{s_t}}$ in the term $\left(\frac{\sigma^2_{u_t} + \sigma^2_{m_t}}{\sigma^2_{u_{t-1}} + \sigma^2_{m_{t-1}}}\right)$. This allow us to write (43) as follows,

$$\pi = \lambda\left(\frac{\sigma^2_{u_t} + \sigma^2_{m_t}}{\sigma^2_{u_t} + \sigma^2_{\Delta_{s_t}} + \sigma^2_{m_{t-1}}}\right). \tag{52}$$

We continue with the reduced form coefficient θ_1 derived in (44). We substitute for $u_{t-1} = u_t - \left(\Delta_{s_t} - \overline{\Delta}_{s_t}\right)$ in the term $\left(\frac{\text{Cov}[y_t, u_{t-1}]}{\text{Cov}[y_t, u_t]}\right)$. It follows that,

$$\begin{aligned}\left(\frac{\text{Cov}[y_t, u_{t-1}]}{\text{Cov}[y_t, u_t]}\right) =& \left(\frac{\text{Cov}[y_t, u_t]}{\text{Cov}[y_t, u_t]}\right) - \left(\frac{\text{Cov}[y_t, \Delta_{s_t}]}{\text{Cov}[y_t, u_t]}\right) \\ =& 1 - \left(\frac{\text{Cov}[\beta SES + \gamma s_t + e_t, \Delta_{s_t}]}{\text{Cov}[y_t, u_t]}\right) \\ =& 1 - \left(\frac{\text{Cov}[e_t, \Delta_{s_t}]}{\text{Cov}[y_t, u_t]}\right).\end{aligned} \tag{53}$$

The results above allow us to write (44) as follows,

$$\theta_1 = \gamma\lambda\left(1 - \left(\frac{\text{Cov}[e_t, \Delta_{s_t}]}{\text{Cov}[y_t, u_t]}\right)\right)\left(\frac{\sigma^2_{u_t} + \sigma^2_{m_t}}{\sigma^2_{u_t} + \sigma^2_{\Delta_{s_t}} + \sigma^2_{m_{t-1}}}\right). \tag{54}$$

We finish with the second stage estimates γ^{iv} and β^{iv}. Combining (52) and (54) results in

$$\gamma^{iv} = \frac{\theta_1}{\pi} = \gamma\left(1 - \left(\frac{\text{Cov}[e_t, \Delta_{s_t}]}{\text{Cov}[y_t, u_t]}\right)\right). \tag{55}$$

Using the results above, it follows from (49) that,

$$\beta^{iv} = \beta + \gamma\delta\left(\frac{\text{Cov}[e_t, \Delta_{s_t}]}{\text{Cov}[y_t, u_t]}\right). \tag{56}$$

B.3 Case (iii)

First we use Assumption 3 to find a link between u_{t-1} and u_t. Similar to (50), we write

$$\delta = \delta_{t-1} + \alpha, \tag{57}$$

with $\alpha = \left(\frac{\text{Cov}[\Delta_{s_t}, SES]}{\sigma^2_{SES}}\right)$. Assumption 3 guarantees again that $\delta = \delta_{t-1}$. Moreover, we can write (4) and (10) as

$$\begin{pmatrix} s_t \\ s_{t-1} \end{pmatrix} = \begin{pmatrix} s_t \\ s_t - \Delta s_t \end{pmatrix} = \delta SES + \begin{pmatrix} u_t \\ u_t \end{pmatrix}, \tag{58}$$

where the constant Δs_t is completely absorbed by the intercept of (10), so that $u_{t-1} = u_t$.

Case (i) assumes Assumption 1 and case (iii) assumes both Assumption 1 and 3. Hence, we use the results above derived under Assumption 3 to simplify the expressions in case (i).

We start with the first stage coefficient π derived in (43). We substitute for $u_{t-1} = u_t$ in $\left(\frac{\text{Cov}[u_t, u_{t-1}]}{\sigma^2_{u_t}}\right)$, so that this term equals 1. We further substitute for $\sigma^2_{u_{t-1}} = \sigma^2_{u_t}$ in the term

$\left(\frac{\sigma^2_{u_t}+\sigma^2_{m_t}}{\sigma^2_{u_{t-1}}+\sigma^2_{m_{t-1}}}\right)$. This results in

$$\pi = \lambda\left(\frac{\sigma^2_{u_t}+\sigma^2_{m_t}}{\sigma^2_{u_t}+\sigma^2_{m_{t-1}}}\right). \tag{59}$$

We continue with the reduced form coefficient θ_1 derived in (44). We substitute for $u_{t-1} = u_t$ in the term $\left(\frac{\text{Cov}[y_t,u_{t-1}]}{\text{Cov}[y_t,u_t]}\right)$, so that this term equals 1. This results in,

$$\theta_1 = \gamma\lambda\left(\frac{\sigma^2_{u_t}+\sigma^2_{m_t}}{\sigma^2_{u_t}+\sigma^2_{m_{t-1}}}\right). \tag{60}$$

We finish with the second stage estimates γ^{iv} and β^{iv}. Combining (59) and (60) results in

$$\gamma^{iv} = \frac{\theta_1}{\pi} = \gamma. \tag{61}$$

Using the results above, it follows from (49) that

$$\beta^{iv} = \beta. \tag{62}$$

C Proof Proposition 3

C.1 Case (i)

Case (i) of Proposition 2 shows that under Assumption 1 we have that $\widetilde{\pi} = \lambda\left(\frac{1}{B^1_\pi}\right)$. We substitute $\widetilde{\pi}$ for λ in (21) to arrive at,

$$\begin{aligned}\gamma^{fs} &= \frac{\gamma^m}{\widetilde{\pi}} = \left(\frac{\gamma\lambda}{\lambda}\right)B^1_\pi \\ &= \gamma B^1_\pi.\end{aligned} \tag{63}$$

Next we substitute $\widetilde{\pi}$ for λ in (22) to arrive at,

$$\begin{aligned}\beta^{fs} &= \beta^m - \gamma^m\delta\left(\frac{1-\widetilde{\pi}}{\widetilde{\pi}}\right) \\ &= \beta^m - \gamma^m\delta\left(\frac{1-\lambda}{\lambda}\right) + \gamma^m\delta\left(\frac{1-\lambda}{\lambda}\right) - \gamma^m\delta\left(\frac{1-\widetilde{\pi}}{\widetilde{\pi}}\right) \\ &= \beta + \gamma^m\delta\left(\frac{1-\lambda}{\lambda} - \frac{1-\widetilde{\pi}}{\widetilde{\pi}}\right) \\ &= \beta + \gamma\delta\left(1 - B^1_\pi\right).\end{aligned} \tag{64}$$

C.2 Case (ii)

Case (ii) of Proposition 2 shows that under Assumption 1 and 2 we have that $B_{\pi}^{1} = 1$, so that $\widetilde{\pi} = \lambda$. It follows from case (i) of Proposition 3 that

$$\gamma^{fs} = \gamma, \tag{65}$$

$$\beta^{fs} = \beta. \tag{66}$$

D Additional empirical results

Table 5: Descriptive statistics on the test scores across all grades

	(1) Obs	(2) Mean	(3) SD	(4) Min	(5) 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. Due to confidentiality reasons we cannot show 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 6: OLS estimates with control variables

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

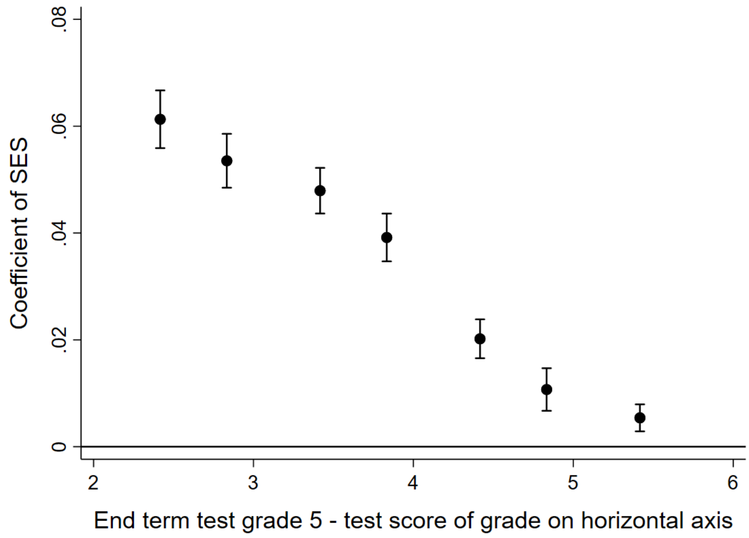
Notes: ***, **, * refers to statistical significance at the 1, 5, and 10% level. The numbers represent the estimated coefficients and the numbers in parentheses the corresponding standard errors. Coefficients are estimated with OLS. Standard errors are clustered on the school id.

Table 7: 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***				
	(0.001)				
$\mathbb{1}$(female)		0.007***			
		(0.002)			
$\mathbb{1}$(Western imm.)			0.018***		
			(0.004)		
$\mathbb{1}$(non-Western imm.)				0.012***	
				(0.004)	
Age					-0.005**
					(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. The numbers represent the estimated coefficients and the numbers in parentheses the corresponding standard errors. Coefficients are estimated with OLS. Standard errors are clustered on the school id.

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

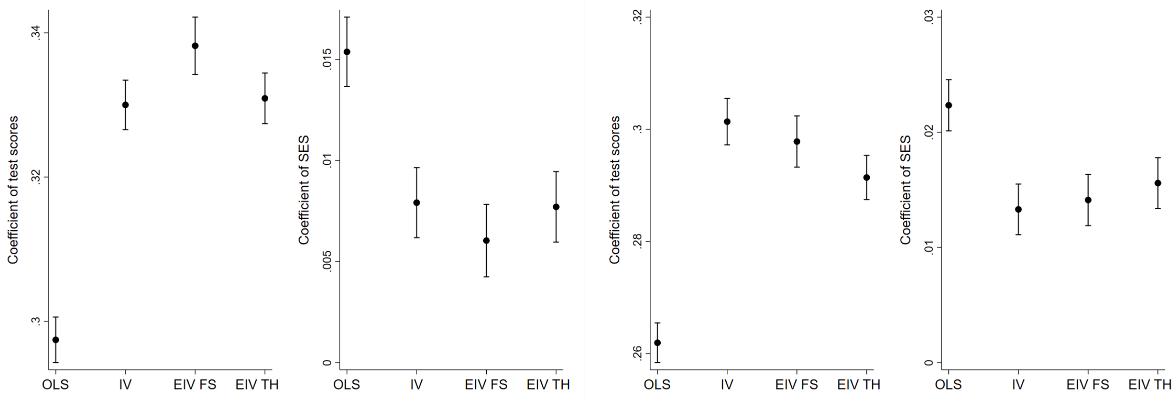


Notes: this figure plots the estimates for the coefficients of SES from regressions of the change in test scores ($s_t^m - s_{t-\tau}^m$) upon SES over time τ. Coefficients are estimated with OLS. Standard errors are clustered on the school id.

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

(a) Theory-oriented vocational track

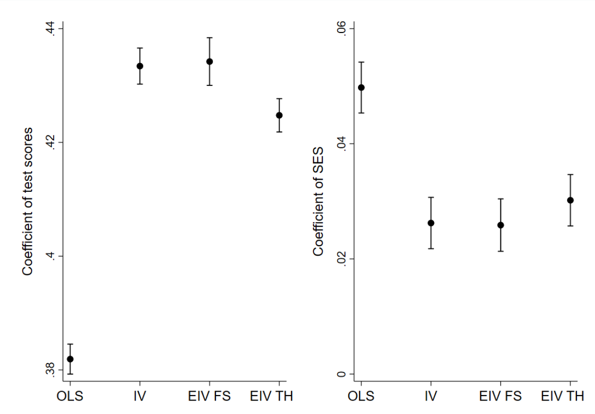
(b) University track



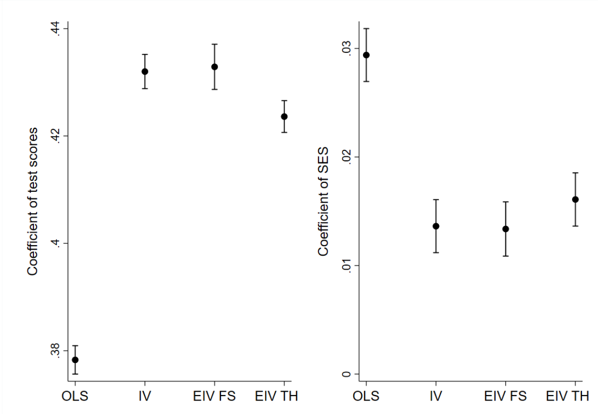
Notes: this figure shows the estimated coefficients for the OLS, IV, and EIV strategy with two alternative measures of teacher recommendation. The lines are 95% confidence intervals. The two alternative measures are a dummy variable that is equal to one if the recommendation is equal to or higher than 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

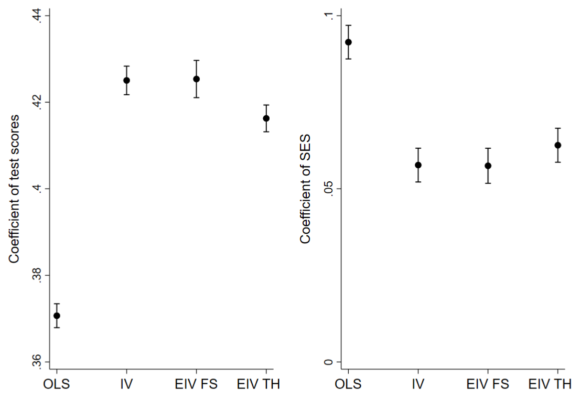
(a) Binarized income



(b) Years of schooling

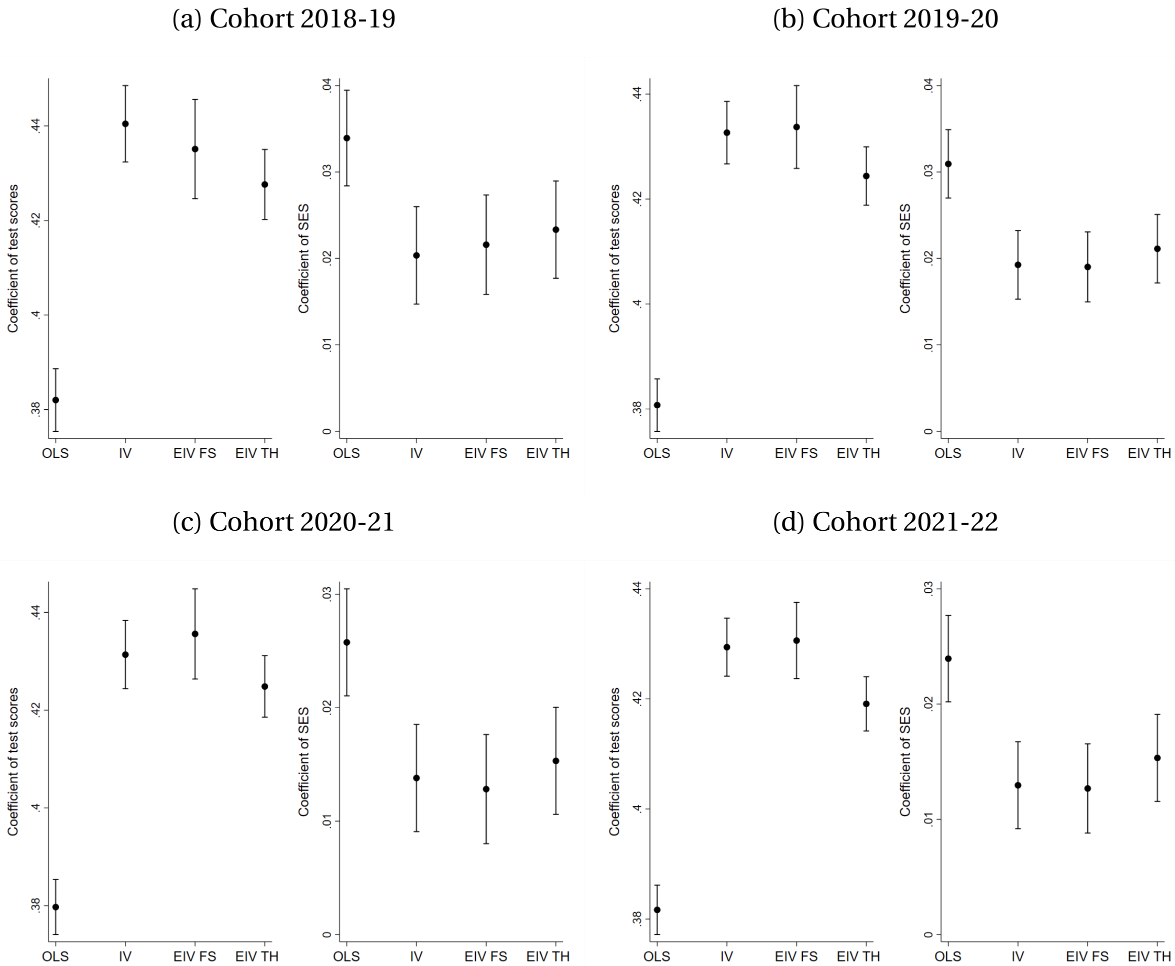


(c) At least college education



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.

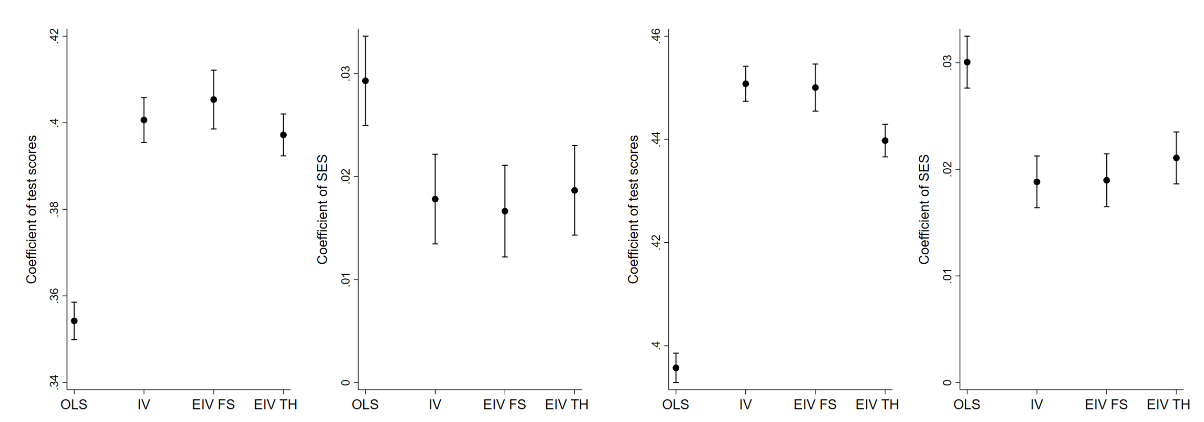
Figure 6: OLS, IV, and EIV estimates per cohort



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

(a) Immigrants (b) 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.