Utrecht University

School of Economics

Teacher bias or measurement error bias?

Evidence from track recommendations

Thomas van Huizen

U.S.E. Working Paper Series nr: 21-13

School of Economics

Utrecht University School of Economics (U.S.E.) is part of the faculty of Law, Economics and Governance at Utrecht University. The U.S.E. Research Institute focuses on high quality research in economics and business, with special attention to a multidisciplinary approach. In the working papers series the U.S.E. Research Institute publishes preliminary results of ongoing research for early dissemination, to enhance discussion with the academic community and with society at large.

The research findings reported in this paper are the result of the independent research of the author(s) and do not necessarily reflect the position of U.S.E. or Utrecht University in general.

U.S.E. Research Institute
Kriekenpitplein 21-22, 3584 EC Utrecht, The Netherlands Tel: +31 30 253 9800, e-mail: use.ri@uu.nl www.uu.nl/use/research

Utrecht University School of Economics

U.S.E. Research Institute
Working Paper Series 21-13
ISSN: 2666-8238

Teacher bias or measurement error bias? Evidence from track recommendations

Thomas van Huizen
Utrecht University School of Economics
Utrecht University

November 2021

Abstract
This study examines to what extent measurement error in test scores explains track recommendation bias in the Netherlands. Track recommendations play an important role in allocating children to secondary school tracks. However, track recommendations are subjective evaluations of a child's skills and have been criticized for being biased. Previous studies have shown that children from low socio-economic status (SES) families receive lower track recommendations than their peers from high SES families, conditional on standardized test scores. While it is often argued that this is evidence of teacher bias, such a claim is invalid in the presence of (random) measurement error in test scores. Standardized tests measure the child's true skills with error and the resulting measurement error bias spills over to the estimates of SES differences. This study corrects for measurement error bias in test scores by applying an instrumental variable strategy. The findings show that models that do not address measurement error in test scores substantially overestimate low-high SES differences in track recommendations. Overall, the presumed teacher bias can to a large extent be explained by measurement error in test scores.

Keywords: teacher bias; discrimination; socio-economic inequality; measurement error

JEL classification: I21; I24; J15; J16

Acknowledgements: I gratefully acknowledge valuable comments from Janneke Plantenga, Astrid Sandsør and Jan Skopek on previous versions of this paper.

Comments welcomed to: T.M.vanHuizen@uu.nl

1 Introduction

Teacher bias in grading and expectations may generate inequality in education and labor market outcomes. Previous studies have documented teacher bias by gender (Cornwell et al., 2013; Lavy and Sand, 2018; Terrier, 2020), race and ethnicity (Botelho et al., 2015; Burgess and Greaves, 2013), migration background (Triventi, 2020; Alesina et al., 2018), and social origin (Timmermans et al., 2018). To estimate teacher bias, studies generally rely on comparisons between subjective teacher evaluations and objective measures of student abilities (i.e. performance on standardized tests). When a specific group receives systematically lower teacher evaluations, holding constant the objective measure, there may be an indication of teacher bias. However, evidence of group differences in teacher evaluations conditional on an objective measure of skills is not proof of teacher bias per se. For instance, many studies acknowledge that differences in non-cognitive skills explain part of the group differences conditional on test scores (Burgess and Greaves, 2013; Cornwell et al., 2013; Triventi, 2020).[1]

This study focuses on another, often ignored, explanation for systematic group differences between objective measures and teacher evaluations: measurement error in standardized test scores. Random measurement error in test scores leads to attenuation bias in the test score coefficient and, to the extent that student background characteristics are associated with true skills, this bias spills over to the main parameters of interest. As a consequence, analyses that do not account for measurement error in test scores will overestimate the difference by social origin in subjective assessments. Theoretically, the presumed teacher bias could be merely a statistical artifact. This methodological problem has been coined Kelley's Paradox (Wainer and Brown, 2006). Surprisingly, the issue that measurement error bias spills over to parameters of interest has been largely ignored in the literature on teacher bias (and in empirical studies more general).[2] A notable exception is the study by Botelho et al. (2015) on racial discrimination in Brazil. The authors show that addressing measurement error substantially changes the results, reducing the estimate of the potential teacher bias by half.

This paper examines potential teacher bias in track recommendations in the Nether-

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

[2] As Modalsli and Vosters (2019) note,"[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."

lands, focusing on differences by parental education (a measure of family socio-economic status; SES).[3] A major policy concern is that prejudice plays an important role in formulating track recommendations. The Dutch Inspectorate of Education (2016) demonstrated that children with higher educated parents receive significantly higher track recommendations, even when conditioning on scores from a standardized test taken during the last grade of primary school (see also Timmermans et al. (2018)). These findings and the main explanation provided (i.e. teacher bias) suggest unequal opportunities and have fueled an ongoing public debate on the role of teachers as gatekeepers (Geven et al., 2018). Teacher bias in track recommendations is harmful in terms of equity and efficiency since track recommendations play a key role in allocating students to secondary school tracks. As track placement determines students' learning environment, misallocation to tracks may have negative long-term consequences in terms of attainment and income (Borghans et al., 2019, 2020). Moreover, negatively biased track recommendations signal low teacher expectations, which may by itself negatively affect student outcomes (Carlana, 2019; Papageorge et al., 2020; Hill and Jones, 2021).

A major empirical problem is that (random) measurement error in test scores generates a bias of the parental education coefficients. This bias is in the same direction as the expected teacher bias and therefore it remains unclear whether group differences in track recommendations are due to teacher bias or measurement error bias. The current study demonstrates that correcting for measurement error changes the results both quantitatively and qualitatively. The findings indicate that track recommendation bias can to a large extent be explained by measurement error in test scores.

2 Setting

In the Netherlands, children enter the tracked secondary school system around the age of 12, after the final grade of primary school (6th grade). The secondary school system consists of three main tracks: a) T3: pre-university (*vwo*), the highest and academically most challenging track has a duration of six years and gives students access to university education. b) T2: pre-college (*havo*), the second highest track has a duration of five years and prepares students for college education (higher vocational education); and c) T1: pre-vocational (*vmbo*), the lowest track has a duration of four years and serves

[3]Previous research on this issue indicates no or limited evidence of teacher bias in track recommendations by gender and migration background in the Netherlands (Timmermans et al., 2018). This is consistent with the findings presented in this study.

as a preparation for vocational education. This track consists of several sub-tracks, the highest sub-track within T1 is the most theoretical.

Around half of the students enters the pre-vocational track, the other half the pre-college or pre-university track. The system allows for upstreaming and downstreaming, though this often involves costs (e.g. additional years in school or a transfer to another school). However, especially in the first year or first two years of secondary school, track mobility is rather high as two adjacent tracks are sometimes combined in one class.[4]

The allocation to the secondary school track is determined in 6th grade based on the performance of an objective final (end-of-school) standardized test and the teacher track recommendation.[5] The standardized test score indicates a specific test track recommendation. While schools can choose between several standardized tests, the large majority of children (around 85%) complete the "Cito Eindtoets".

Whereas the final test score provides an objective, independent track recommendation, the teacher track recommendation is more subjective but also aims to be more holistic. The teacher takes into account information on student performance, which includes the final test score but also other objective performance measures. Importantly, throughout the primary school years, schools use standardized (formative) assessments to monitor the learning progress of students. This information is typically used to formulate the track recommendation. While there are no national guidelines or procedures on how to use this information, these are often formulated at the school, local or regional level. Recent comparative evidence demonstrates that Dutch teachers attach relatively high weight to student performance when forming future expectations, which may be explained by the context which "is characterized by an extensive testing culture to assess students' learning potential, and teachers are trained to translate test scores into expectations for students." (Geven et al., 2021). In addition to performance measures, the teacher may take into account non-cognitive dimensions such as motivation, attitude towards school and classroom behavior.

[4]Around a quarter of the pupils switch between tracks in the first three years of secondary school (Feron et al., 2016).

[5]The context description here concerns the period relevant for the sample (2013/2014). Analysis of teacher bias becomes problematic for post 2013/2014 data due to a reform. Since 2014/2015 the teacher track recommendation has become dominant in the track allocation process. Children take a standardized test, but in the new system after the formulation of the track recommendation. The track recommendation can be adjusted upwards (not downwards) if the child perforce better than expected. This implies that for children who are satisfied with the initial track recommendation, the test is a low stakes test. For those who are dissatisfied, the test remains a high stakes test. This process is likely to generate (non-random) measurement error.

3 Data and methodology

3.1 Data

The paper uses data from the last wave of the COOL$^{5\text{-}18}$ study (school year 2013-2014) (Driessen et al., 2015). COOL$^{5\text{-}18}$ includes administrative data from the school information system (e.g. on test scores and parental background) and surveys among teachers, parents and children. The sampling is school-based and the full sample includes both a representative school sample and a smaller sample of schools with a relatively high share of disadvantaged children.

Table 1: Summary statistics

		Share (%)
Track recommendation		
T3		17.32
At least T2		43.32
At least T1 highest sub-track		68.62
SES (parental education)		
Low		21.54
Medium		41.50
High		36.97
Migration background		
Native		76.86
Turkey		4.88
Morocco		6.69
Suriname		2.53
Other		9.03
	Mean	SD
Test scores		
Final test score	533.83	10.45
Reading test score	53.89	18.41
Math test score	110.52	12.42

NOTES: The summary statistics of the test scores concern the original test scores; for the analyses these scores are converted into z-scores.

The data contains information on track recommendations and the scores of the final test (taken in week 5 of 2014) and scores from other standardized tests (taken in other

weeks of the school year). To facilitate the interpretation of the results, test scores are converted into z-scores. The main independent variable is family SES, measured by the highest level education of the parents. I distinguish between three groups: high (at least one parent holds a bachelor's degree or higher), medium (at least one parent completed vocational education) and low (neither parent has completed vocational education). The shares in the sample are around 2/5 (high SES), 2/5 (medium SES) and 1/5 (low SES). In addition, information on gender and migration background (country of birth of the parents) is used. The focus is on the three largest migrant groups: children with a Turkish, Moroccan or Surinamese background. Table 1 provides the summary statistics of the analytical sample. The final sample consists of 4,815 children in 244 schools.

3.2 Strategy

Following the literature on teacher bias, the analyses rely on differences between subjective teacher evaluations (track recommendations) and objective measures of a child's skills (standardized final tests). As shown in this study and previous work (Geven et al., 2018; Timmermans et al., 2018), conditional on final test score, children with high SES parents receive higher track recommendations than their low SES peers. The central aim of this study is to assess to what extent measurement error in test scores explains this difference. The base model is represented by the following equation:

$$TR_{is} = \alpha + \beta_1 MSES_{is} + \beta_2 HSES_{is} + \gamma TS_{is} + X_{is}\delta + e_{is} \quad (1)$$

where the outcome TR_{is} indicates the teacher track recommendation of child i in school s, $MSES$ and $HSES$ are dummies measuring medium and high SES respectively (low SES is the reference category), TS_{is} captures final test scores, the vector X includes gender and migration background dummies and e_{is} is the error term. I present estimates using three alternative binary dependent variables, indicating whether the track recommendation is T3, at least T2 or at least the highest sub-track of T1. The equations are estimated using linear probability models. Standard errors are clustered at the school level. Additional analyses also include school fixed effects (which are in most cases teacher fixed effects) and measures of non-cognitive skills.

The final test captures children's true skills with error. There may be several sources of measurement error in the final test scores: (un)lucky guesses, the child may have

a good (bad) day (i.e. a transitory shock in performance) or the child may by chance be (un)familiar with the test items in such a way that the score is not representative of his/her overall skills.

Measurement error in test scores is a major concern for the interpretation of β_1 and β_2 as parameters of teacher bias. Under the assumption of random measurement error in test scores (classical measurement error), the error generates attenuation bias of γ. Because high SES children have on average higher levels of true (unobserved) skills, this bias spills over to β_1 and β_2, implying an overestimation of the parameters of interest. The intuition is that a high final test score may be the result of a large positive measurement error, and this is statistically more likely to hold for low than for high SES children given the differences in average true skill levels. Addressing this concern and illustrating the extent of the bias is the central objective of this study.

To correct for bias due to measurement error I employ an instrumental variable (IV) strategy, instrumenting the final test score with other standardized test scores. IV provides a solution for classical measurement error (Hyslop and Imbens, 2001) and has been applied to estimate racial discrimination in grading (instrumenting grades using previous test scores) (Botelho et al., 2015), to address measurement error in test scores in longitudinal analyses (Bradbury et al., 2015) and in studies on intergenerational mobility (Modalsli and Vosters, 2019).

For the analyses I use the scores of a standardized reading and math test made in the final year of primary school (most often in calendar week 3, two weeks before the final test). Reading and math are the two main domains of the final test. To construct the instrument, (1) the test scores of both tests are standardized (mean 0; SD 1); (2) the standardized scores are averaged; (3) the average scores are transformed into z-scores.

The central assumption is that measurement error in the final test score is uncorrelated with measurement error in the other tests. Under this assumption, the first stage purges the final test score from measurement error. This assumption is plausible as the reading test and math test are taken on a different day and contain different test items than the final test. The instrument is highly correlated with the final test score ($\rho = 0.85$; first stage F-test statistics are around 3600, see Table A1).

4 Results

Main findings. The main results are presented in Table 2. First, it is clear that the differences between SES groups are substantial. For instance, children from high SES parents are around 23 percentage points more likely of receiving the highest track recommendation (Panel A, column (1)). The inclusion of gender and migration background dummies hardly affects this result (column (2)). Track recommendation differences by SES are to a large extent explained by differences in the final test scores as test scores increase sharply with SES. Nevertheless, in all OLS models, the low-high SES differences remain sizable and significant after controlling for the final test score. The low-high SES gap in receiving the highest track recommendation decreases to 5 percentage points when conditioning on final test scores (column (3)). Including school fixed effects (column (5)) somewhat decreases the gaps, although they remain significant. These gaps in teacher evaluations conditional on an objective measure of skills are often interpreted as teacher bias.

However, the IV models correct for measurement error in test scores and produce quantitatively and qualitatively different results. In fact, when considering higher track recommendations (T3 and T2; panel A and B), low-high SES differences are no longer statistically significant. This finding is important as the two highest tracks give access to higher education. The results indicate a non-monotonic relation between SES and the track recommendation, with low (and high) SES children receiving higher track recommendations than their medium SES peers. Instrumenting the final test score reduces the low-high difference in the probability to receive at least the highest sub-track of the lowest track (Panel C), but the difference remains significant. Hence, within the sub-tracks of the lowest track, teacher bias may exist as here measurement error only explains a relatively small part of the difference between low and high SES children.

Overall, measurement error in test scores appears to matter crucially. As expected, the OLS-IV comparisons indicate that the test score coefficients in the OLS models are attenuated. Moreover, measurement error bias appears to spill over to the SES coefficients, biasing upward the estimates of SES differences in the OLS models.

Finally, there is limited evidence suggesting gender bias or bias towards children with a migration background (see Tables A2-A4), although some of the results indicate that children with a Surinamese background receive higher track recommendations and those with a Moroccan background receive lower track recommendations than their

Table 2: Track recommendations and SES

	Raw gaps	OLS	OLS	IV	FE	IV-FE
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Track recommendation T3						
Medium SES	0.0565***	0.0469***	-0.0283***	-0.0533***	-0.0334***	-0.0558***
	(0.0108)	(0.0111)	(0.0106)	(0.0123)	(0.0109)	(0.0123)
High SES	0.2277***	0.2154***	0.0505***	-0.0043	0.0301**	-0.0183
	(0.0165)	(0.0168)	(0.0143)	(0.0147)	(0.0134)	(0.0146)
Final test			0.1939***	0.2583***	0.2021***	0.2691***
(z-score)			(0.0087)	(0.0116)	(0.0086)	(0.0116)
Panel B: Track recommendation at least T2						
Medium SES	0.1341***	0.1249***	-0.0095	-0.0324**	-0.0117	-0.0316**
	(0.0188)	(0.0179)	(0.0150)	(0.0159)	(0.0140)	(0.0146)
High SES	0.3849***	0.3738***	0.0791***	0.0288	0.0630***	0.0199
	(0.0228)	(0.0222)	(0.0192)	(0.0187)	(0.0157)	(0.0157)
Final test			0.3466***	0.4058***	0.3631***	0.4228***
(z-score)			(0.0074)	(0.0088)	(0.0077)	(0.0086)
Panel C: Track recommendation at least the highest sub-track of T1						
Medium SES	0.1845***	0.1804***	0.0554***	0.0443***	0.0542***	0.0452***
	(0.0195)	(0.0187)	(0.0141)	(0.0143)	(0.0128)	(0.0129)
High SES	0.3787***	0.3735***	0.0995***	0.0753***	0.0830***	0.0635***
	(0.0213)	(0.0217)	(0.0168)	(0.0162)	(0.0150)	(0.0143)
Final test			0.3223***	0.3508***	0.3334***	0.3604***
(z-score)			(0.0068)	(0.0074)	(0.0075)	(0.0077)

NOTES: SES is measured by the highest level of parental education. Standard errors in parentheses are clustered at the school level. The sample consists of 4,815 students in 244 schools. Model (1) does not include any controls, models (2)-(6) control for gender and migration background. Full results are presented in Tables A2-A4.

* $p < 0.1$, ** $p < 0.05$, and *** $p < 0.01$

counterparts with native-born parents. These differences are most pronounced when it concerns the probability to receive the highest track recommendation (see Panel A).

Test track recommendation instead of test score. The final test score implies a specific test track recommendation. Instead of using the (z-)score of the final test, a dummy can be included indicating whether the test score was above the relevant threshold (implying a T3 or at least T2 test score).[6] The results of this alternative specification are found in Table 3. The estimates are largely consistent with the main findings discussed above: the estimates of low-high SES differences are significant and sizable in the OLS models but become insignificant or substantially smaller in the IV models.

Non-cognitive skills. In addition to information from standardized tests, teachers may take into account relevant non-cognitive skills to formulate the track recommendation. The data include teacher-reported measures of attitudes towards school (the average of three 5-point scale items) and behavioral problems (the average of four 5-point scale items). These measures appear to significantly predict the track recommendation in the OLS models, but do not play an important role in the IV models (see Table A5). This suggests that the measurement error in test scores also spills over to the coefficients of these traits. Importantly, including these additional controls hardly affects the SES coefficients in both the OLS and IV models. This shows that, in the Dutch setting, measurement error in test scores rather than differences in non-cognitive skills explains most of the SES differences in track recommendations conditional on test scores.

Perceived parental support. Teachers may believe low SES parents are less actively involved in the learning process of their children or are less able to support their children with school tasks. Additional estimates (see Table A6) show that perceived parental support (a measure based on three 5-point scale items) is significantly associated with track recommendations in the OLS models, but the relations are much weaker (and insignificant in Panel A and B) in the IV models. However, in the models where significant low-high SES differences remain in the IV models (Panel C), including a measure of parental support reduces this difference. As discussed above, teacher bias appears to play mostly a role within the lowest track; these results indicate that this bias is to some extent explained by differences in perceived parental support.

[6] Only the results for the higher track recommendations are presented as the test score cutoffs for the sub-tracks of T1 are less clear.

Table 3: Teacher and test track recommendations

	(1)	(2)	(3)	(4)
Panel A: Track recommendation T3				
Medium SES	0.0140*	-0.0016	0.0108	-0.0007
	(0.0076)	(0.0085)	(0.0076)	(0.0085)
High SES	0.0861***	0.0252**	0.0698***	0.0176
	(0.0127)	(0.0127)	(0.0105)	(0.0111)
Test recommendation: T3	0.6717***	0.9886***	0.6860***	1.0223***
(0/1)	(0.0230)	(0.0331)	(0.0213)	(0.0332)
Panel B: Track recommendation T2 or higher				
Medium SES	0.0183	-0.0206*	0.0184	-0.0166
	(0.0119)	(0.0125)	(0.0123)	(0.0129)
High SES	0.1104***	0.0144	0.0995***	0.0129
	(0.0168)	(0.0167)	(0.0150)	(0.0157)
Test recommendation: T2 or higher	0.7265***	0.9914***	0.7381***	1.0214***
(0/1)	(0.0142)	(0.0114)	(0.0129)	(0.0116)

NOTES: SES is measured by the highest level of parental education. Standard errors in parentheses are clustered at the school level. The sample consists of 4,815 students in 244 schools. All models include controls for gender and migration background.
* $p < 0.1$, ** $p < 0.05$, and *** $p < 0.01$

5 Conclusions

This study shows that measurement error in test scores can to a large extent explain the presumed teacher bias in the Netherlands. Overall, the evidence suggests that teachers do not consistently favor high SES over low SES children. More generally, this study provides a cautionary tale about ignoring measurement error, also when it does not concern the variable of interest.

Standardized tests provide valuable information about a child's skills and are therefore relevant in the track allocation process. However, relying on a single end-of-school test may not be optimal as it is a noisy snapshot of a child's skill level. If a child's performance on the final test deviates substantially from performance on previous standardized tests, it is likely that (bad) luck played a large role in the final test. In that case, the observed test score is a poor reflection of the child's true skills. In the Dutch context, teachers can compare the final test score with scores on previous standardized tests and are essentially able to correct for measurement error. Interestingly, the finding that "the teacher beats the test" as better predictor of track allocation and performance

in secondary school (Feron et al., 2016) is consistent with the idea that teachers correct for measurement error in the final test.

Finally, systematic skill gaps between different socio-economic groups explain (almost) fully differences in track recommendations. To promote equal opportunities, narrowing skill gaps through effective interventions in the school, preschool and home environment should be a key policy priority.

References

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

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

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

Botelho, F., Madeira, R. A., Rangel, M. A., 2015. Racial discrimination in grading: Evidence from brazil. American Economic Journal: Applied Economics 7 (4), 37–52.

Bradbury, B., Corak, M., Waldfogel, J., Washbrook, E., 2015. Too many children left behind: The US achievement gap in comparative perspective. Russell Sage Foundation.

Burgess, S., Greaves, E., 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.

Cornwell, C., Mustard, D. B., Van Parys, J., 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.

Driessen, G., Elshof, D., Mulder, L., Roeleveld, J., 2015. Cohortonderzoek cool5-18 technisch rapport basisonderwijs, derde meting 2013/14.

Feron, E., Schils, T., Ter Weel, B., 2016. Does the teacher beat the test? the value of the teacher's assessment in predicting student ability. De Economist 164 (4), 391–418.

Geven, S., Batruch, A., van de Werfhorst, H., 2018. Inequality in teacher judgements, expectations and track recommendations: A review study. Amsterdam: Universiteit van Amsterdam.

Geven, S., Wiborg, Ø. N., Fish, R. E., van de Werfhorst, H. G., 2021. How teachers form educational expectations for students: a comparative factorial survey experiment in three institutional contexts. Social Science Research, 102599.

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

Hyslop, D. R., Imbens, G. W., 2001. Bias from classical and other forms of measurement error. Journal of Business & Economic Statistics 19 (4), 475–481.

Lavy, V., Sand, E., 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.

Modalsli, J. H., Vosters, K., 2019. Spillover bias in multigenerational income regressions. Discussion Paper 897, Statistics Norway.

Papageorge, N. W., Gershenson, S., Kang, K. M., 2020. Teacher expectations matter. Review of Economics and Statistics 102 (2), 234–251.

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

Timmermans, A. C., de Boer, H., Amsing, H. T., Van Der Werf, M., 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., Brown, L. M., 2006. Three statistical paradoxes in the interpretation of group differences: Illustrated with medical school admission and licensing data. Handbook of statistics 26, 893–918.

Appendix

Table A1: Main IV results (first stage)

	Final test (z-score) (1)	Final test (z-score) (2)
Math & reading test	0.8243***	0.8279***
(z-score)	(0.0136)	(0.0138)
SES (parental education):		
Medium SES	0.0860***	0.0592***
	(0.0231)	(0.0226)
High SES	0.1688***	0.1154***
	(0.0271)	(0.0259)
Migration background:		
Turkey	-0.1353***	-0.0643*
	(0.0479)	(0.0388)
Morocco	-0.0325	0.0645
	(0.0491)	(0.0392)
Surinam	0.0067	0.0192
	(0.0592)	(0.0752)
Other	-0.0548	-0.0066
	(0.0343)	(0.0344)
Girl	0.0303*	0.0241
	(0.0161)	(0.0154)
F-test statistic	3677	3587
School FE	NO	YES

NOTES: Standard errors in parentheses are clustered at the school level. The sample consists of 4,815 students in 244 schools.

* $p < 0.1$, ** $p < 0.05$, and *** $p < 0.01$

Table A2: Track recommendation: T3

	Raw gaps	OLS	OLS	IV	FE	IV-FE
	(1)	(2)	(3)	(4)	(5)	(6)
SES (parental education):						
Medium	0.0565***	0.0469***	-0.0283***	-0.0533***	-0.0334***	-0.0558***
	(0.0108)	(0.0111)	(0.0106)	(0.0123)	(0.0109)	(0.0123)
High	0.2277***	0.2154***	0.0505***	-0.0043	0.0301**	-0.0183
	(0.0165)	(0.0168)	(0.0143)	(0.0147)	(0.0134)	(0.0146)
Migration background:						
Turkey		-0.0628***	-0.0105	0.0069	-0.0295	-0.0206
		(0.0183)	(0.0197)	(0.0232)	(0.0233)	(0.0258)
Morocco		-0.0403*	-0.0393*	-0.0389*	-0.0413*	-0.0479*
		(0.0205)	(0.0200)	(0.0227)	(0.0234)	(0.0256)
Suriname		0.0269	0.0481**	0.0552***	-0.0089	-0.0060
		(0.0338)	(0.0206)	(0.0193)	(0.0278)	(0.0301)
Other		0.0149	0.0277*	0.0319*	0.0024	0.0041
		(0.0182)	(0.0161)	(0.0170)	(0.0185)	(0.0190)
Girl		-0.0091	-0.0077	-0.0073	-0.0054	-0.0047
		(0.0103)	(0.0089)	(0.0091)	(0.0091)	(0.0094)
Final test			0.1939***	0.2583***	0.2021***	0.2691***
(z-score)			(0.0087)	(0.0116)	(0.0086)	(0.0116)

NOTES: Standard errors in parentheses are clustered at the school level. The sample consists of 4,815 students in 244 schools.

* $p < 0.1$, ** $p < 0.05$, and *** $p < 0.01$

Table A3: Track recommendation: at least T2

	Raw gaps	OLS	OLS	IV	FE	IV-FE
	(1)	(2)	(3)	(4)	(5)	(6)
SES (parental education):						
Medium	0.1341***	0.1249***	-0.0095	-0.0324**	-0.0117	-0.0316**
	(0.0188)	(0.0179)	(0.0150)	(0.0159)	(0.0140)	(0.0146)
High	0.3849***	0.3738***	0.0791***	0.0288	0.0630***	0.0199
	(0.0228)	(0.0222)	(0.0192)	(0.0187)	(0.0157)	(0.0157)
Migration background:						
Turkey		-0.0943**	-0.0008	0.0152	-0.0309	-0.0231
		(0.0365)	(0.0286)	(0.0302)	(0.0237)	(0.0246)
Morocco		0.0073	0.0092	0.0095	0.0199	0.0140
		(0.0428)	(0.0290)	(0.0291)	(0.0275)	(0.0278)
Suriname		-0.0158	0.0222	0.0287	-0.0146	-0.0120
		(0.0442)	(0.0317)	(0.0332)	(0.0359)	(0.0380)
Other		-0.0100	0.0129	0.0168	-0.0005	0.0010
		(0.0249)	(0.0198)	(0.0207)	(0.0201)	(0.0208)
Girl		-0.0009	0.0015	0.0020	0.0063	0.0069
		(0.0145)	(0.0104)	(0.0104)	(0.0102)	(0.0103)
Final test			0.3466***	0.4058***	0.3631***	0.4228***
(z-score)			(0.0074)	(0.0088)	(0.0077)	(0.0086)

NOTES: Standard errors in parentheses are clustered at the school level. The sample consists of 4,815 students in 244 schools.

* $p < 0.1$, ** $p < 0.05$, and *** $p < 0.01$

Table A4: Track recommendations: at least the highest sub-track of T1

	Raw gaps	OLS	OLS	IV	FE	IV-FE
	(1)	(2)	(3)	(4)	(5)	(6)
SES (parental education):						
Medium	0.1845***	0.1804***	0.0554***	0.0443***	0.0542***	0.0452***
	(0.0195)	(0.0187)	(0.0141)	(0.0143)	(0.0128)	(0.0129)
High	0.3787***	0.3735***	0.0995***	0.0753***	0.0830***	0.0635***
	(0.0213)	(0.0217)	(0.0168)	(0.0162)	(0.0150)	(0.0143)
Migration background:						
Turkey		-0.0734*	0.0136	0.0213	0.0037	0.0073
		(0.0373)	(0.0251)	(0.0253)	(0.0280)	(0.0278)
Morocco		0.0222	0.0239	0.0240	0.0156	0.0130
		(0.0422)	(0.0280)	(0.0277)	(0.0279)	(0.0278)
Suriname		0.0093	0.0446*	0.0477*	-0.0284	-0.0272
		(0.0472)	(0.0257)	(0.0251)	(0.0344)	(0.0347)
Other		0.0040	0.0253	0.0272	-0.0043	-0.0036
		(0.0246)	(0.0194)	(0.0197)	(0.0194)	(0.0195)
Girl		0.0092	0.0114	0.0116	0.0097	0.0100
		(0.0137)	(0.0096)	(0.0095)	(0.0096)	(0.0097)
Final test			0.3223***	0.3508***	0.3334***	0.3604***
(z-score)			(0.0068)	(0.0074)	(0.0075)	(0.0077)

NOTES: Standard errors in parentheses are clustered at the school level. The sample consists of 4,815 students in 244 schools.

* $p < 0.1$, ** $p < 0.05$, and *** $p < 0.01$

Table A5: Track recommendations and non-cognitive skills

	OLS	OLS	IV	IV
	(1)	(2)	(3)	(4)
Panel A: Track recommendation T3				
Medium SES	-0.0263**	-0.0253**	-0.0513***	-0.0506***
	(0.0110)	(0.0110)	(0.0127)	(0.0127)
High SES	0.0501***	0.0517***	-0.0056	-0.0048
	(0.0146)	(0.0147)	(0.0150)	(0.0150)
Final test	0.1962***	0.1861***	0.2625***	0.2674***
(z-score)	(0.0088)	(0.0087)	(0.0117)	(0.0122)
Attitude towards school		0.0315***		-0.0053
(z-score)		(0.0076)		(0.0078)
Behavioral problems		-0.0187***		-0.0125*
(z-score)		(0.0066)		(0.0066)
Panel B: Track recommendation at least T2				
Medium SES	-0.0043	-0.0039	-0.0267*	-0.0263*
	(0.0148)	(0.0148)	(0.0155)	(0.0155)
High SES	0.0803***	0.0809***	0.0302	0.0311*
	(0.0192)	(0.0192)	(0.0185)	(0.0184)
Final test	0.3494***	0.3340***	0.4090***	0.4057***
(z-score)	(0.0072)	(0.0077)	(0.0087)	(0.0101)
Attitude towards school		0.0408***		0.0084
(z-score)		(0.0079)		(0.0085)
Behavioral problems		-0.0102		-0.0048
(z-score)		(0.0072)		(0.0072)
Panel C: Track recommendation at least the highest sub-track of T1				
Medium SES	0.0557***	0.0567***	0.0448***	0.0457***
	(0.0143)	(0.0141)	(0.0144)	(0.0143)
High SES	0.1033***	0.1047***	0.0789***	0.0804***
	(0.0170)	(0.0169)	(0.0163)	(0.0163)
Final test	0.3204***	0.3131***	0.3496***	0.3481***
(z-score)	(0.0070)	(0.0075)	(0.0077)	(0.0083)
Attitude towards school		0.0236***		0.0078
(z-score)		(0.0080)		(0.0080)
Behavioral problems		-0.0158**		-0.0131*
(z-score)		(0.0067)		(0.0067)

NOTES: SES is measured by the highest level of parental education. Standard errors in parentheses are clustered at the school level. The sample consists of 4,578 students. All models include controls for gender and migration background.
* $p < 0.1$, ** $p < 0.05$, and *** $p < 0.01$

Table A6: Track recommendations and perceived parental support

	OLS	OLS	IV	IV
	(1)	(2)	(3)	(4)
Panel A: Track recommendation T3				
Medium SES	-0.0260**	-0.0367***	-0.0511***	-0.0536***
	(0.0107)	(0.0110)	(0.0125)	(0.0127)
High SES	0.0513***	0.0339**	-0.0045	-0.0086
	(0.0144)	(0.0145)	(0.0148)	(0.0151)
Final test	0.1957***	0.1882***	0.2622***	0.2599***
(z-score)	(0.0087)	(0.0087)	(0.0115)	(0.0117)
Parental support		0.0274***		0.0069
(z-score)		(0.0058)		(0.0061)
Panel B: Track recommendation at least T2				
Medium SES	-0.0064	-0.0172	-0.0292*	-0.0325*
	(0.0150)	(0.0159)	(0.0157)	(0.0167)
High SES	0.0801***	0.0625***	0.0295	0.0241
	(0.0194)	(0.0197)	(0.0187)	(0.0196)
Final test	0.3486***	0.3410***	0.4090***	0.4059***
(z-score)	(0.0072)	(0.0078)	(0.0086)	(0.0093)
Parental support		0.0277***		0.0091
(z-score)		(0.0080)		(0.0082)
Panel C: Track recommendation at least the highest sub-track of T1				
Medium SES	0.0552***	0.0457***	0.0441***	0.0384***
	(0.0144)	(0.0142)	(0.0146)	(0.0145)
High SES	0.0998***	0.0843***	0.0750***	0.0659***
	(0.0171)	(0.0169)	(0.0165)	(0.0167)
Final test	0.3211***	0.3145***	0.3507***	0.3454***
(z-score)	(0.0070)	(0.0073)	(0.0076)	(0.0078)
Parental support		0.0244***		0.0155**
(z-score)		(0.0066)		(0.0064)

NOTES: SES is measured by the highest level of parental education. Standard errors in parentheses are clustered at the school level. The sample consists of 4,624 students. All models include controls for gender and migration background.

* $p < 0.1$, ** $p < 0.05$, and *** $p < 0.01$