Do Test Scores Help Teachers Give Better Track Advice to Students? A Principal Stratification Analysis*

Andrea Ichino Fabrizia Mealli Javier Viviens

February 13, 2026

Abstract

Every year, over one million EU students choose a secondary school track based on their teacher recommendations, yet little evidence supports the optimality of these recommendations. Using Dutch data, we examine whether access to standardized test scores improves recommendation quality. We develop a Principal-Stratification metric in a quasi-randomized setting, conduct a welfare analysis weighting short- and long-term losses, and assess principal fairness across protected attributes. Allowing recommendation upgrades when test scores exceed expectations increases overall welfare and successful placement, while yielding fairer recommendations for immigrant and low-SES students. Our methodology and findings contribute to the literature on algorithm-assisted human decisions.

JEL-Code: I2

Keywords: principal stratification; secondary school track recommendations.

*Ichino: U. Bologna, European University Institute and CEPR (andrea.ichino@unibo.it); Mealli: European University Institute and U. Firenze (fabrizia.mealli@eui.eu); Viviens: European University Institute (javier.viviens@eui.eu). We are grateful to Gustav Axén, Kosuke Imai, Matthijs Osterveen, Lina Segers and Dinand Webbink for insightful discussions.

1 Introduction

Every year, more than one million EU students, upon completing primary education, choose a secondary school track based on recommendations received from their teachers. In some institutional settings—such as the Dutch school system examined in this paper—only under exceptional circumstances students may begin secondary education in a more demanding track than the one recommended to them. Despite the potentially dramatic consequences that bad track recommendations may have, there is no clear evidence that assigning to teachers alone the role of providing these recommendations is the best solution and generates an optimal track assignment.[1] Our goal in this paper is to establish, using a well-defined metric and a quasi-randomized experiment, whether teachers can "improve" the quality of their recommendations after seeing the scores of standardized tests taken by their students.

To the best of our knowledge, there is no consensus on what the goal of secondary school track advice should be or what defines a "better" recommendation. The metric we adopt is the one based on Principal Stratification (Frangakis and Rubin, 2002)[2] that Imai et al. (2023) use to study whether algorithms can help judges in deciding which arrested individuals should be released while waiting for their trial: in this case, a decision is better than another if it avoids releasing subjects at risk of repeating a crime. In the Dutch setting, a reform was introduced that gives teachers the right (but not the obligation) to revise their initial track assignment upwards if the student performs better on a standardized test than the initial recommendation would imply. This reform offers an ideal setting to apply the metric that we adopt[3].

The nature of this metric in the context of track advising can be intuitively described with reference to an education system featuring two tracks: low (e.g., vocational) and high (e.g., academically oriented). We make two assumptions that we maintain throughout the

[1]There is a wide literature on teachers' biases that may affect the quality of secondary school track recommendations. Recent examples are: Driessen et al. (2008), Burgess and Greaves (2013), Gerhenson et al. (2016), Osikominu et al. (2021), van Leest et al. (2021), Geven et al. (2021), Carlana and Fort (2022), Carlana et al. (2022a), Carlana et al. (2022b), Ferman and Fontes (2022), Batruch et al. (2023) Bach (2023), Alesina et al. (2024), van Huizen et al. (2024), and Falk et al. (2025).

[2]For a discussion on Principal Stratification and how it is rooted in the Instrumental Variable literature, see Mealli and Mattei (2012).

[3]The setting offered by the reform was also used in de Ree et al. (2025).

analysis: (1) if a student is able to complete a more difficult track, it is better she is allowed to do so for herself (because she can obtain more desirable lifetime outcomes) and for society (inasmuch as there are collective gains from a more educated community); (2) changing between tracks is costly for students and for the education system. Note that given these two assumptions, even in the absence of negative congestion externalities in the educational process and in the labor market[4] it would be sub-optimal to send all students to the high track because it would be costly for some of them not to complete it and have to change to the low track.

Let's divide the population of students into groups – *principal* strata (Frangakis and Rubin, 2002) – that differ by how the education outcome depends on track recommendations. The first group in which we are interested comprises students for whom the track recommendation determines unequivocally the completed track: if the student is recommended the low track in the first year, she will graduate in the same track at the end of secondary school, while if recommended the high track, this will be her final graduation outcome. These are students for whom the teacher's recommendation is a self-fulfilling prophecy, thus becoming crucially important. In light of this consideration, we call these subjects "Helpable" (H) to signal that they are the ones who can be *helped* by a high first-year track recommendation instead of a low one.[5] Adapting the methodology proposed by Imai et al. (2023) to this context, we can establish whether the information provided by test scores helps teachers improve their recommendations by measuring the fraction of students in the H group that are recommended the high track by teachers and see if it increases when teachers can decide to upgrade based on test score information. In the Principal Stratification framework, this is the Average Principal Causal Effect ($APCE$) of "changing the source" of a recommendation (i.e., giving test score information to the source, in our context) within the principal stratum of H students. If this $APCE$ is positive, test scores help teachers provide better track advice by upgrading to more challenging tracks students who can effectively complete them.

We are also interested in two other groups of students: those who, independently of the track initially recommended to them, always finish secondary education in the low track,

[4]For evidence and a discussion of these externalities see Ichino et al. (2025).

[5]They correspond to the group of cases that, in the context of Imai et al. (2023), are labeled as "Preventable" crimes. Similarly, de Ree et al. (2025) label them as "Trapped-in-Track" (TT).

and those who, instead, always graduate from the high track independently of the initial recommendation. Adapting to our setting the notation of de Ree et al. (2025), we call these students *Always Low* (AL) and *Always High* (AH), respectively. The common characteristic of these two groups is that the final student's outcome is not affected by the teacher's recommendation; however, this outcome can be achieved in different ways, which are more or less costly, depending on the number of track changes they require. In these cases, the goal of a recommendation should be to minimize the number of track changes that these two types of students will experience if they start on a track that differs from the one they will ultimately complete. The best recommendation is then the one that sends all AL students to the low track and all AH students to the high track. Therefore, allowing teachers to upgrade based on test scores produces more desirable outcomes, the larger the $APCE$ is for AH students and the lower it is for their AL peers.[6]

We estimate an $APCE$ for H students which is not lower than 6%, suggesting that when teachers can upgrade on the basis of test scores, the quality of their advice improves by at least 6% as measured by the fraction of these students that are recommended a more difficult track and can complete it successfully, while in the counterfactual case, they would have remained in a lower level track. For AH students, for whom an upgrade is also desirable to reduce the cost of track changes, the $APCE$ is estimated to have approximately the same lower bound.

In the case of AL students, for whom an upgrade is not desirable, the corresponding estimates are also positive with an $APCE$ of at least 7%. This suggests that a significant fraction of students who are unable to complete the more challenging curriculum are upgraded, and we cannot rule out that this fraction is even greater than that estimated for the H and AH strata. A possible reason for this finding is that primary school teachers do not bear any consequence related to track changes in secondary schools and are, therefore, less sensitive to the goal of reducing them in the case of AL students.

Given the opposite desirability of the findings concerning the effect of test scores on the quality of teacher advice for H and AH students on one hand and AL students on the other, it

[6]It is conceivable the existence of a fourth group of students who always finish secondary education in the *opposite* track with respect to the one that was initially recommended to them. This is discussed below.

is natural to seek a method to obtain a comprehensive evaluation. An approach is offered by the classification framework proposed by Ben-Michael et al. (2024). This framework enables us to construct a weighted sum of the losses generated by mistakes in recommendations for the various types of students. We can then test whether this sum decreases when teachers can upgrade based on test scores, depending on the relative weight of the short-term losses suffered by students who must change track because of misplacement versus the lifetime losses of students who are not directed to a more challenging track they can eventually complete. We find that the relative weight of short-term losses would have to be unreasonably high to conclude that teachers should not be allowed to upgrade based on test scores.

Finally, building on Imai and Jiang (2022), a corollary contribution of this analysis is that it makes it possible to evaluate whether being allowed to upgrade based on test scores helps teachers improve the *fairness* of their recommendations with respect to protected attributes such as SES, race, or gender. For example, suppose that gender is a protected attribute with respect to which we want to assess how fair a recommender is. Then, in the stratum of H students, the recommender is fair if the fraction of high recommendations is the same for female and male students. However, the analogous fraction in other strata could be lower or higher as long as it is equal for both genders. Therefore, in the overall population, females may receive a different fraction of high first-year advice from this recommender; however, this would not constitute a violation of fairness, as it would be due to differences in the gender composition of the strata, not to discrimination based on gender within a stratum. We find that allowing teachers to upgrade recommendations based on test scores improves fairness. Specifically, it raises the chances that immigrants and low-SES students, who are otherwise discriminated against, are recommended a more difficult track that they can complete relative to natives and high-SES peers, respectively.

The paper is organized as follows. In Section 2, we describe the Dutch institutional setting and the quasi-randomized experiment that makes our analysis possible. The statistical framework and its assumptions follow in Section 3. Section 4 presents our results, while Section 5 supports their robustness by showing that they also hold when the Exclusion Restriction assumption is substituted by alternative assumptions. In Section 6 we move to the analysis of the fairness of track recommendations, and Section 7 concludes.

2 The Dutch school system and the quasi-experiment

What needs to be known about the Dutch school system for the purpose of this paper is that it features five main secondary school tracks, of which three are vocational (VMBO-BL, VMBO-KL, VMBO-GT ranked in this order by level of difficulty) and two lead to university studies (HAVO, VWO, similarly ranked). In addition, tracks that are contiguous by level of difficulty can be combined in mixed tracks, so that there are in total nine tracks to which a student can be recommended in the first year of secondary school. These nine tracks are listed from the easiest to the hardest in the row headings of Table 1.

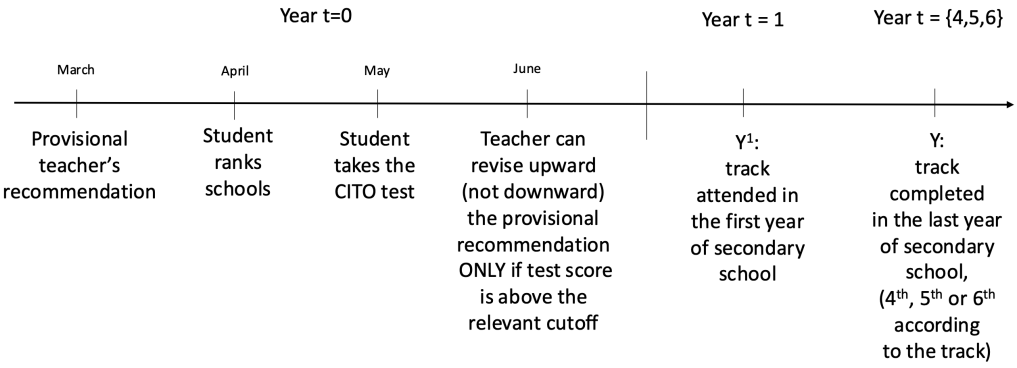
Table 1: Tracks of the Dutch school system

Track	Students initially assigned to the track denoted by the row		Test score cut-off for upgrade to next level	Fraction above the cutoff of students eligible for upgrade to next level	
	1	2	3	4	5
V: BL	9,585	6,929	519	0.45	0.46
V: BL/KL	3,781	3,828	526	0.31	0.30
V: KL	15,698	11,258	529	0.33	0.36
V: KL/GT	3,255	4,003	529	0.49	0.54
V: GT	29,117	21,154	533	0.45	0.46
V: GT/HAVO	8,540	9,620	537	0.40	0.41
A: HAVO	28,226	21,465	540	0.47	0.50
A: HAVO/VWO	9,932	11,136	545	0.29	0.31
A: VWO	27,750	24,469	n.a	n.a.	n.a.
Total	135,884	113,862		0.42	0.43
Cohort	2015	2016	2015 & 2016	2015	2016

Notes: The row headings are the names of the secondary school tracks to which a student can be assigned in the Dutch system, ranked by level of difficulty. The first six tracks are vocational, while the last three lead to university studies (academic). Columns 1 and 2 report the number of students initially assigned to each track by their primary education teachers, in view of starting secondary school in the academic year 2015-16 or 2016-17. Column 3 reports the CITO test score cutoffs above which a student in the track denoted by the corresponding row may be upgraded to the next level by the primary school teacher; the test score takes on integer values $\in \{\underline{S}, \underline{S}+1, \underline{S}+2, \ldots, \overline{S}\}$, with $\underline{S} = 501$ and $\overline{S} = 550$. The upgrade is not mandatory. Columns 4 and 5 report, for each initial track, the fraction of students who, in the years considered, may be upgraded due to a sufficiently high CITO test score.

Using the database of the Dutch National Cohort Study on Education (NCO), we consider the population of students who begin their secondary school in the academic years 2015–16

Figure 1: Time-line for track recommendations and outcomes



and 2016–17 (hereafter, 2015 and 2016 cohorts).[7] The timing implied by these regulations is described in Figure 1. Let $t = 1$ denote the first year of secondary school. In March of the previous year, $t = 0$, the primary education teachers give their students an initial track recommendation. The first and second columns of Table 1 report the number of students in each cohort initially assigned to each track by their primary education teachers as a result of this provisional recommendation.

In April of the same year, $t = 0$, the students rank the schools offering the track that was recommended to them (or a less difficult one if they prefer) and are assigned to one of them.[8] Then, in the following month, they take their end-of-primary education standardized test (CITO), which is graded by a computer with no intervention from primary education teachers. The scores obtained in this test map into track levels on the basis of nationwide, predetermined test score cutoffs. As in de Ree et al. (2025), we refer to the mapping of a CITO test score to a track level as a "test-based" track assignment. If the test-based track assignment is higher than the teacher's track assignment recorded in March of year $t = 0$, the teacher is mandated by the law to consider an upward revision of the provisional recommendation. Such revision is only optional and dependent on the teacher's judgment. Column 3 of Table 1 reports the CITO test score cutoffs above which a student, whose

[7]This database can be accessed following the instructions provided at this link. For documentation in English, see here). The track assignment regulations that generate the quasi-experiment we exploit became effective for students starting secondary school in the 2014-15 academic year, but the data for this cohort do not contain all the necessary information and thus cannot be used.

[8]The more specific method of student allocation differs by municipality. In Amsterdam, for example, a deferred acceptance algorithm is used, with lottery numbers serving as tiebreakers (Ketel et al., 2023).

provisional recommendation is the track of the corresponding row, must be re-evaluated by the teacher for a possible upgrade. Columns 4 and 5 report, for each track, the fraction of students who may be upgraded because of a sufficiently high CITO test score.

Let R_i be a dummy variable equal to 0 if the primary education teacher *finally* recommends the low track to student i. R_i is, instead, equal to 1 if the teacher recommends an upgrade to a higher track. Note that the 52,219 students whose provisional recommendation is the highest track (VWO in the last row of Table 1) cannot be treated with $R_i = 1$ because there is no higher track to which they can be upgraded. Our analysis focuses on the students for whom the track provisionally recommended in March of year $t = 0$ by the teacher is one of those listed in the first 8 rows of Table 1, totaling 197,527 students.

Denote with K_i^τ the track that student i in cohort $\tau \in \{2015, 2016\}$ is provisionally recommended by the teacher; K_i^τ can take on values $\in \{1, 8\}$. Let S_i^τ be the score obtained by student i in cohort τ in the CITO test; S_i^τ can take on integer values $\in \{\underline{S}, \underline{S}+1, \underline{S}+2, \ldots, \overline{S}\}$, with $\underline{S} = 501$ and $\overline{S} = 550$. Let the assignment to treatment be Z_i, defined as follows:

$$Z_i = \begin{cases} 0 & \text{if} \quad S_i^\tau = C_{K_i^\tau}^\tau - 1 \\ 1 & \text{if} \quad S_i^\tau = C_{K_i^\tau}^\tau \\ \text{else} & \text{if} \quad S_i^\tau \neq \{C_{K_i^\tau}^\tau - 1, C_{K_i^\tau}^\tau\} \end{cases} \tag{1}$$

where $C_{K_i^\tau}^\tau$ is the cutoff score that student i in cohort τ provisionally assigned to track K_i^τ has to reach in order to be re-evaluated by her primary school teacher for a possible upgrade; "else" denotes students who do not score close to their respective cutoff score. If we condition on the students in a particular cohort τ, with a specific value of K_i^τ and with a value of $Z_i \in \{0, 1\}$, we can assume that only randomness, possibly conditional on some covariates to capture the heterogeneity of primary schools as illustrated below, determines whether $S_i^\tau = C_{K_i^\tau}^\tau - 1$ or $S_i^\tau = C_{K_i^\tau}^\tau$, that is, whether $Z_i = 0$ or $Z_i = 1$.

This institutional setting generates the quasi-experiment we exploit to answer our research questions. Among students with a realized value of $S_i^\tau \in \{C_{K_i^\tau}^\tau - 1, C_{K_i^\tau}^\tau\}$, Z_i randomly assigns similar subjects in each of the eight initial tracks to two different "sources" of first-year track recommendation: one is the "teacher alone", while the other is the "teacher informed by the CITO test results", who therefore has the possibility to upgrade students

based on such results. This assumption is known in the Regression Discontinuity (RD) literature as "local randomization" (LR).[9] One of its advantages is that it allows the researcher to easily deal with discrete running variables without having to rely on the usual continuity-type assumption[10]. We formalize the LR assumption for our context in Section 3, where we also provide evidence to assess its plausibility.

Next, we denote with $Y_i \in \{0, 1\}$ the outcome that we study. We define $Y_i = 1$ if student i graduates from a higher track than the one provisionally recommended to him/her in March of $t = 0$, before taking the CITO test, and 0 otherwise. We consider a student as graduating in a given track even if he/she does so in more years than the required minimum.[11] Our goal is to evaluate which source of track recommendation ("teacher alone" or "teacher informed by test scores and allowed to upgrade") is "preferable" according to two criteria: a recommendation is better if it (i) induces a student to enroll in the most difficult track that she/he is able to complete successfully and (ii) reduces to a minimum the chances of costly track changes.

An institutional complication of this quasi-experimental setting is that, on rare occasions, also students who score below the cutoff ($Z_i = 0$) may be recommended for an upgrade by their teacher ($R_i = 1$).[12] Moreover, as already mentioned, students with $Z_i = 1$ may have $R_i = 0$ because their primary school teacher does not want to upgrade them even if an

[9]Local randomization has been first formalized by Li et al. (2015) and Cattaneo et al. (2015). The LR approach to the analysis of RD designs has been later discussed in Mattei and Mealli (2017); Branson and Mealli (2019); Cattaneo et al. (2020b,a), and extended to allow for randomization conditional on covariates in Forastiere et al. (2025).

[10]See also Eckles et al. (2025) and Li et al. (2020)

[11]Our conclusions do not change qualitatively if we assume instead that a student graduates in a given track only if he/she does so in the regular possible time (results available from the authors). Four years are typically required for a vocational track, while five and six years are needed for the two academic tracks (HAVO and VWO, respectively). Six years are also needed for students who change from VMBO to HAVO, and seven for those who change from HAVO to VWO. The release of the NCO data at our disposal follows the 2015 cohort for 7 years, and the 2016 one for only 6 years after enrollment. Graduation is not observed for some students in these cohorts. We thus assume that the students of these cohorts who are still enrolled in secondary school in the last year of observation will graduate in the track in which they are enrolled.

[12]As explained by the relevant authorities (Van Po naar Vo, 2025), "[T]here may be situations in which a student does not obtain a higher test recommendation, but a school or teacher still believes that there is more knowledge and skills than expected. In such situations, the school may choose to adjust the provisional school advice. The school's judgment is central here. ... If parents do not agree with the advice, they will consult with the teacher and/or director of the primary school. If they cannot reach an agreement about the level at which the student will enter secondary education, parents can, as a last resort, use the school's complaints procedure."

upgrade is possible. Therefore, as we will show in Section 3, non-compliance may occur in both assignments to treatment conditions.

3 Statistical framework

Consider a sample of N_k^τ units of cohort τ whose provisional recommendation is track k with $Z_i \in \{0,1\}$, $i \in 1, \cdots, N_k^\tau$. Let $\mathbf{Z}$ and $\mathbf{R}$ be the N_k^τ-dimensional vectors of the possible values of Z_i and R_i, with generic element z_i and r_i respectively. Denote with $R_i(\mathbf{Z}) \in \{0,1\}$ the recommendation that student i receives from the teacher as a function of the assignment vector $\mathbf{Z}$. Similarly, denote with $Y_i(\mathbf{Z},\mathbf{R}) \in \{0,1\}$ the potential track (low or high) completed by student i at the end of secondary school as a function of the assignment vector $\mathbf{Z}$ and the vector of recommendations $\mathbf{R}$. Since different primary schools may have different benchmarks to which they compare their students and decide their provisional track assignments, we must consider the possibility that students with different abilities are assigned to the same provisional track k simply because they have more or less lenient primary teachers. We denote with π_i the benchmark against which student i has been evaluated to decide her/his provisional track assignment.

3.1 Assumptions

The first assumption we need is the Stable Unit Treatment Value Assumption (SUTVA, see Rubin, 1980), which restricts the dependence of the potential outcomes for a student only to the assignment and enrolment of that student, ruling out interference:

Assumption 1 *Stable Unit Treatment Value Assumption:*
$R_i(\mathbf{Z}) = R_i(\mathbf{Z}')$ *if* $z_i = z_i'$
$Y_i(\mathbf{Z},\mathbf{R}) = Y_i(\mathbf{Z}',\mathbf{R}')$ *if* $z_i = z_i'$ *and* $r_i = r_i'$.

Assumption 1 postulates the existence of two potential versions of $R_i(z)$ with $z \in \{0,1\}$ and four potential versions of $Y_i(z,r)$ with $z,r \in \{0,1\}$. Given SUTVA, the next assumption is:

Assumption 2 *Local Randomization of the source conditional on primary schools' leniency benchmarks and provisional track assignments: for every cohort τ we have*

$$\{R_i(z), Y_i(z,r), X_i, U_i\} \perp\!\!\!\perp Z_i \mid \pi_i, K_i$$

for $z \in \{0,1\}$ and $r \in \{0,1\}$, where X_i and U_i are observable and unobservable covariates, respectively.

This assumption requires some discussion. It formally says that for students assigned to the same track and whose teachers have a similar benchmark criterion π_i to evaluate beliefs about students' ability, the reasons for them to score just below or at the cutoff are unrelated to either potential outcomes or observed, X, or unobserved, U, covariates. Hence, in their case, $P(Z_i = 1|R_i(z), Y_i(z,r), U_i, X_i, \pi_i, K_i) = P(Z_i = 1|\pi_i, K_i)$. If π_i is observed, this is a testable implication in that if Assumption 2 holds, we can expect, within each provisional track assignment and for a given π_i, the observed characteristics of students to be well-balanced across the score cutoff (Mattei and Mealli, 2017; Forastiere et al., 2025; Angrist and Rokkanen, 2015). Moreover, if Assumption 2 holds for all tracks, we can view the data as coming from a block-randomized experiment, with blocks defined by the tracks, and analyze them accordingly (Imbens and Rubin, 2015).

The problem is that while we observe the initial track assignment and we can condition on scoring just below or at the cutoff, we do not observe the benchmark π_i of a student's primary teachers. A solution would be to condition on primary school fixed effects, assuming that teachers in the same school collectively agree on the same benchmark π. However, we do not have enough students per primary school to conduct such an analysis. What we can do is to proxy π_i with the fraction of students in the primary school of i who score above the cutoffs. This fraction measures the average teacher severity in that school because, as more students score above the cutoffs, the stricter their primary school teachers must have been in their provisional track recommendations. We define this variable as $\tilde{Z}_i = \frac{\sum_{j \in School_i} \mathbb{I}(S_j^\tau \geq C_{K_j^\tau}^\tau)}{\#School_i}$, which takes on the same value for all students of a given school in a given cohort. $\tilde{Z}_i$ will thus be included in all the specifications of our analysis.

Table 2: Balancing of covariates

Panel A: 2015 cohort

Track	Female	Immigrant origin	College mother	Missing college mother	Household income	Urbanity primary school	Primary school confession	Primary school # students	Student's age in months	Obs. for $Z=0$	Obs. for $Z=1$
	1	2	3	4	5	6	7	8	9	10	11
V: BL	1	1	1	1	1	1	1	1	1	519	511
V: BL/KL	1	0.7517	1	1	1	1	1	1	1	173	183
V: KL	1	0.1612	1	0.2872	1	0.1856	1	1	1	1199	860
V: KL/GT	1	1	1	1	1	1	1	1	0.4632	274	219
V: GT	1	1	1	1	1	1	0.6102	1	1	1942	1948
V: GT/HAVO	1	0.8050	1	1	1	1	1	1	1	614	570
A: HAVO	1	1	1	1	1	0.4935	1	1	1	2149	2172
A: HAVO/VWO	0.0787	1	1	1	1	1	1	1	1	843	769

Panel B: 2016 cohort

Track	Female	Immigrant origin	College mother	Missing college mother	Household income	Urbanity primary school	Primary school confession	Primary school # students	Student's age in months	Obs. for $Z=0$	Obs. for $Z=1$
	1	2	3	4	5	6	7	8	9	10	11
V: BL	1	0.7578	0.7132	1	1	1	1	1	1	396	458
V: BL/KL	1	1	1	1	1	1	1	1	1	195	195
V: KL	0.4012	1	1	1	1	1	1	1	0.5522	656	630
V: KL/GT	1	1	0.8141	1	1	1	1	1	1	261	284
V: GT	1	1	0.0193	0.9419	1	1	1	0.8404	1	1386	1376
V: GT/HAVO	1	0.0915	1	1	1	1	1	1	0.3441	923	689
A: HAVO	1	1	1	1	1	1	1	1	1	1643	1728
A: HAVO/VWO	0.0976	0.3389	1	1	1	0.4697	1	1	0.1691	983	952

Notes: The row headings are the names of the secondary school tracks to which a student can be assigned in the Dutch system, ranked by level of difficulty. Columns 1 to 9 report, for each track and covariate, the Holm-Bonferroni p-value of a test for the equality of means in the two assignment-to-treatment groups defined by $Z_i \in \{0,1\}$. The definition of the covariates whose names are not self-explanatory is as follows. Urbanity of a primary school is a categorical variable indicating the population density at the location of the primary school. Primary school confession is a categorical variable indicating the type of confession in primary schools (Protestant, Catholic, Public, or Other). Age in months refers to the age of the student at the time she/he receive the first recommendation. The last two columns report the number of observations in the two assignment groups. Standard errors are always clustered at the primary school level.

Columns 1 to 9 of Table 2 support the validity of Assumption 2 by showing that nine observed characteristics of students are balanced in the immediate vicinity of the cutoff. These covariates are: gender, migrant origin, mother's education, missing mother's education, household income, urban location of the primary school, confession of the primary school, number of students in the primary school, and age in months. Under Assumption 2, our estimates should not be (and in fact are not, as we will show below) affected by the inclusion or exclusion of these balanced covariates because they *are not* correlated with Z_i, conditional on $\tilde{Z}_i$. However, including them may increase efficiency since they *are* likely to be correlated with R_i and Y_i, and for this reason, they will appear in our preferred specification.

It must be noted that the validity of Assumption 2 does not imply that the proportion of students scoring above or below the cutoff is around 50%, as shown in the last two columns of Table 2. The CITO test score is the result of the students responding correctly to a series of questions. To simplify this data-generating process, think about the score being generated (approximately) by a binomial distribution and suppose that units in the subsample that we consider have the same probability χ to respond correctly to a question in the test. If so, these units also have the same probability (propensity score) to score at the cutoff ($S_i^t = C^t_{K_i^\tau}$), yet this probability is not equal to the probability of scoring below ($S_i^\tau = C^\tau_{k_i^\tau} - 1$) because these probabilities follow a binomial distribution. The difference in these two probabilities will depend on χ and on the number of questions in the test, but it is by no means an indication of the failure of Assumption 2. In the data, we may also observe differences in this probability from track to track because the cutoff is different or because the difficulty of the test (i.e., the probability of a correct response to single questions) is different. These findings would not represent violations of Assumption 2.[13]

Imai et al. (2023) invoke two additional assumptions that are necessary to answer our research question.

Assumption 3 *Exclusion restriction:*

$Y_i(z, r) = Y_i(z', r) = Y_i(r)$ *for* $z, z', r \in \{0, 1\}$.

[13]This binomial example is, of course, over-simplistic; most likely, not all the questions have the same difficulty, and the students may not have the exact same probability of responding correctly to all questions. It is only meant to illustrate why, with a discrete score, the probability of being below or above the cutoff may differ even if the mechanism that generated Z_i is random and unrelated to potential outcomes or covariates.

This assumption requires that the potential outcomes $Y_i(z, r)$ depend only on the recommendation r and not on the source z of the recommendation so that they can be written as $Y_i(r)$. Intuitively, this means that the track in which the student graduates should only be affected by scoring above or below the cutoff through its effect on the teacher's recommendation. A violation of this restriction would occur, for example, if scoring above the cutoff increased the probability of graduating in the high track regardless of whether teachers decide to upgrade. In Section 5 we will show that we obtain similar results if we assume a Homogeneity and a Principal Ignorability assumption instead of the Exclusion Restriction (ER) assumption, and therefore that our conclusions are robust and stable across these alternative identifying assumptions.

Assumption 4 *Monotonicity of Y_i with respect to R_i:*
$Y_i(1) \geq Y_i(0)$.

This assumption plausibly excludes the existence of students, whom we call "Rebels", who would not complete the high track if it were the recommended track, but who would complete it if the low track were recommended. These students would be characterized by the following set of potential outcomes: $(Y(1), Y(0)) = (0, 1)$.[14]

Finally, a different type of monotonicity assumption needs to be discussed in our setting.

Assumption 5 *Monotonicity of R_i with respect to Z_i:*
$R_i(1) \geq R_i(0)$.

This assumption is typically made in Instrumental Variable (IV) settings with non-compliance[15] and excludes the existence of "Defiers", that is, students who would be recommended the low track when scoring at or above the cutoff and the high track when scoring below. Imai et al. (2023) do not make this assumption. In their context, a judge who decides with the help of an algorithm may release an arrested subject while the same judge, not helped by the algorithm, keeps the same subject in prison, or vice versa. More generally, every set of counterfactual decisions with and without the help of the algorithm is possible in their

[14] de Ree et al. (2025) allow for the existence of a similar stratum, which they label "Slow Starters". When estimating lower and upper bounds for the corresponding proportion, they cannot reject that this proportion is zero.

[15] See Imbens and Angrist (1994) and Angrist et al. (1996).

context. To put it differently, they do not have (and do not want to have) any prior on which source of a decision is preferable: the fraction of "Preventable" cases kept under arrest may be higher when the judge is helped by the algorithm or when the judge decides alone, and their goal is precisely to assess for which one of the two sources the fraction is higher.

Table 3: Non-compliance statistics

TRACK	$P(R=1\|Z=0)$	$P(R=1\|Z=1)$	$P(R=1\|Z=0)$	$P(R=1\|Z=1)$
	1	2	3	4
V: BL	.0039	.0352	.0025	.0677
V: BL/KL	.0231	.1148	.0205	.1897
V: KL	.0209	.1767	.0122	.1952
V: KL/GT	.0109	.1233	.0153	.1655
V: GT	.0036	.056	.0065	.0574
V: GT/HAVO	.0163	.1684	.0206	.1858
A: HAVO	.0051	.0603	.0085	.0758
A: HAVO/VWO	.0237	.1821	.0193	.188
COHORT	2015	2015	2016	2016

Notes: The row headings are the names of the secondary school tracks to which a student can be assigned in the Dutch system, ranked by level of difficulty. The other columns report the probabilities indicated in the column headings.

In our case, the institutional setting is such that an upgrade to a higher track should be possible only if students score at or above the cutoff. Therefore, Defiers should not exist by construction because no student below the cutoff should be eligible for an upgrade. For the same reasons, also Always Takers should not exist. On the other hand, teachers may decide not to upgrade students who score at or above the cutoff, so Never Takers may exist in our setting.[16] As anticipated in Section 2, however, there are rare instances of non-compliance for students scoring below the cutoff. This is shown in Table 3: in the first and third columns, the probability that a student is upgraded ($R_i = 1$) even if scoring below the cutoff ($Z_i = 0$) is low but positive in all tracks, ranging between 0.3% for BL in the 2016 cohort to 2.4% for HAVO/VWO in the 2015 cohort, and therefore Always Takers do exist in our setting. In the second and fourth columns, the probability that a student is upgraded ($R_i = 1$) when scoring above the cutoff ($Z_i = 1$) is significantly lower than 1 in all tracks, ranging from 3.5%

[16]Table A1 in the Online Appendix to Section 3.1 shows that $\tilde{Z}_i$ correlates positively with the proportion of Never Takers. This finding supports our choice of $\tilde{Z}_i$ as a proxy for the leniency π_i of student i teachers.

for BL in the 2015 cohort to 19.5% for KL in the 2016 cohort. So, and this is no surprise, Never Takers also exist and are frequent. In general, bridge tracks show a significantly higher probability of upgrading than the other tracks, which makes intuitive sense as R_i is defined to be equal to one if the student is upgraded to the higher component of the mixed track.

This evidence indicates that non-compliance is certainly present in our setting. However, we do not see reasons to think that in the rare cases in which students scoring below the cutoff are upgraded (see footnote 12 for the possible reasons), these same students would not be upgraded if scoring above, which makes them Always Takers but not Defiers. Hence, Assumption 5 of Monotonicity of R with respect to Z, excluding the possibility of Defiers, appears plausible. Even if not required to identify causal effects in our setting, we will refer to it for the interpretation of our results.

Table 4: First stage

Track	δ 1	δ 2	δ 3	δ 4
V: BL	.0319	.0322	.07	.0712
	(.0093)	(.0092)	(.0147)	(.0146)
V: BL/KL	.0931	.0872	.1722	.1615
	(.0265)	(.0276)	(.0308)	(.0304)
V: KL	.1639	.1607	.187	.1892
	(.0152)	(.0149)	(.0176)	(.0177)
V: KL/GT	.1152	.11	.1535	.1564
	(.0248)	(.0244)	(.0244)	(.0247)
V: GT	.0564	.0561	.0537	.0537
	(.0058)	(.0058)	(.0074)	(.0074)
V: GT/HAVO	.1579	.1533	.1719	.1675
	(.0186)	(.0182)	(.0162)	(.0162)
A: HAVO	.0581	.058	.0707	.0705
	(.0063)	(.0063)	(.0076)	(.0076)
A: HAVO/VWO	.168	.1667	.1732	.1764
	(.0167)	(.0166)	(.0158)	(.0155)
COVARIATES X_i	NO	YES	NO	YES
COHORT	2015	2015	2016	2016

Notes: The table reports, separately for each track, estimates of parameter δ in the first stage regression (2): $R_i = \gamma + \delta Z_i + \mu X_i + \xi \tilde{Z}_i + \epsilon_i$. The row headings are the names of the secondary school tracks to which a student can be assigned in the Dutch system, ranked by level of difficulty. The covariates X included in columns 2 and 4 are those described in Table 2, respectively, for the two cohorts.

Table 4 reports, separately for each track, estimates of the first stage regression

$$R_i = \gamma + \delta Z_i + \mu X_i + \xi \tilde{Z}_i + \epsilon_i \tag{2}$$

where $\tilde{Z}_i$ is the proxy for π_i on which we need to condition for the validity of Assumption 2 and X_i are the nine balanced covariates described in Table 2. As expected, the estimates of δ confirm the existence of imperfect compliance. Considering the covariates X_i, the stability of the first-stage estimates with and without them supports Assumption 2 that the assignment to treatment is as good as random.

3.2 Students' strata

Under Assumption 1 (SUTVA) and 3 (Exclusion restriction), and according to Frangakis and Rubin (2002), students can be classified in four strata based on the joint values of the two binary potential outcomes: $Y_i(R_i = 1), Y_i(R_i = 0)$.

AH: $(Y_i(1), Y_i(0)) = (1, 1)$

These are *Always High* students who always complete the high track independently of the recommendation they receive.

AL: $(Y_i(1), Y_i(0)) = (0, 0)$

These are *Always Low* students who always complete the low track independently of the recommendation they receive.

H: $(Y_i(1), Y_i(0)) = (1, 0)$

These are *Helpable* students who complete the high track if this is the track recommended to them by primary school teachers, and who would not do the same otherwise.

R: $(Y_i(1), Y_i(0)) = (0, 1)$

These are *Rebel* students who complete the high track if recommended for the low track by primary school teachers, and the low track if recommended for the high track.

These principal strata, which we denote with $J \in \{AL, AH, H, R\}$, will play a crucial role in the definition of the metric we propose in Section 3.3 to measure the quality of track recommendations.

Table 5: The Principal Strata

Strata $R_i \to Y_i$	$\underbrace{Y_i(0,0) = Y_i(1,0) = Y_i(0)}$ Exclusion restriction	$\underbrace{Y_i(0,1) = Y_i(1,1) = Y_i(1)}$ Exclusion restriction	Strata $Z_i \to R_i$	$R_i(0)$	$R_i(1)$	Notes
1	2	3	4	5	6	7
AL	0	0	NT	0	0	
			AT	1	1	
			C	0	1	
			D	1	0	May $\exists$ w/out Monotonicity $Z_i \to R_i$
AH	1	1	NT	0	0	
			AT	1	1	
			C	0	1	
			D	1	0	May $\exists$ w/out Monotonicity $Z_i \to R_i$
H	0	1	NT	0	0	
			AT	1	1	
			C	0	1	
			D	1	0	May $\exists$ w/out Monotonicity $Z_i \to R_i$
R	1	0	NT	0	0	$\nexists$ because Monotonicity $R_i \to Y_i$ holds
			AT	1	1	
			C	0	1	
			D	1	0	

Notes: The four panels refer to the four Principal Strata listed in the first column of the table and defined by the possible values of R_i and Y_i: the Always-Low (AL), the Always-High (AH), the Helpable (H) and the Rebels (R). Columns 2 and 3 describe the values of the potential outcomes Y_i when the treatment R_i is equal to 0 or 1, respectively, and the Exclusion Restriction assumption holds. Each one of these Principal Strata is divided into additional strata defined by the possible values of Z_i and R_i, as listed in the fourth column: the Never Takers (NT), the Always Takers (AT), the Compliers (C) and the Defiers (D). Columns 5 and 6 describe the values of the potential treatments R_i when the assignment to treatment Z_i is equal to 0 or 1, respectively. The Notes in the first three rows of the last column make clear that Defiers may or may not exist depending on whether the Monotonicity Assumption 5 of R_i with respect to Z_i holds. The last note in the fourth row states that given the maintained Monotonicity Assumption 4 of Y_i with respect to R_i, the Rebels do not exist.

Under the same assumptions, the more conventional classification in strata that characterizes compliance types is also possible, based on the joint values of the two binary potential treatment values: $R_i(Z_i = 1), R_i(Z_i = 0)$:

AT: $(R_i(1), R_i(0)) = (1, 1)$

These are *Always Takers* who are always recommended the high track, regardless of their score around the cutoff.

NT: $(R_i(1), R_i(0)) = (0, 0)$

These are *Never Takers* who are never recommended the high track, independently of where they score around the cutoff.

C: $(R_i(1), R_i(0)) = (1, 0)$

These are *Compliers* who are recommended the high track if they score at the cutoff, and the low track if they score below.

D: $(R_i(1), R_i(0)) = (0, 1)$

These are *Defiers* who are recommended the low track if they score at the cutoff, and the high track if they score below.

We denote with $G \in \{AT, NT, C, D\}$ these strata.

Table 5 describes the relationship between these basic principal strata. Under Assumption 4 (Monotonicity of Y_i with respect to R_i), the stratum of the Rebels, $J = R$ in red, does not exist irrespective of the student belonging to any of the strata G, while under Assumption 5 (Monotonicity of R_i with respect to Z_i), the stratum of Defiers, $G = D$ in blue, does not exist in any of the strata J.

3.3 The metric to compare recommendations

Following Imai et al. (2023), the metric that we propose to compare recommendations and establish which one is preferable relies on the classification of students in Principal Strata described in the previous section, combined with the following criteria:

1. If a student is able to complete a more difficult track, it is better that she is allowed to do it for both

   1a) herself, because she can obtain more desirable lifetime outcomes

   1b) and for society, if there are collective gains from a more educated community.

2. Changing tracks is costly for students and for the education system.

The first criterion defines what constitutes a better recommendation for Helpable (H) students. In their case, the goal of teachers should be to recommend the high track to the largest fraction of these students because, as a result of this recommendation, they will complete precisely this track, achieving a better outcome for themselves as well as for society. Therefore, we need to estimate the difference between "the fraction of H students who are recommended a high first-year track by teachers who can upgrade based on students' test scores" and "the analogous fraction when teachers do not see test scores". If this difference is positive, test scores help teachers improve the track advising they provide to their students.

The second criterion defines instead what constitutes a better recommendation for the Always Low (AL) and Always High (AH) students. In their cases, track advising should aim to minimize the number of track changes that these two types of students experience during secondary school, if they start on a track that differs from the one they will ultimately complete. Therefore, the best recommendation is to send all AL students to the low track and all AH students to the high track.

To formalize this intuitive characterization of what constitutes a better recommendation, consider the following Average Principal Causal Effect:

$$APCE_J = E(R_i(1) - R_i(0)| \text{ student } i \text{ is in } J) \tag{3}$$

The $APCE_J$ is the difference between the fraction of students in stratum J who are recommended the high track by teachers who can upgrade based on test scores ($Z_i = 1$) minus the fraction receiving the same recommendation by teachers who have not seen the test scores ($Z_i = 0$). To put it differently, the $APCE_J$ is the Intention To Treat effect of the source of recommendations on their content in stratum J.

Note that each $APCE_J$ is also equal to the difference between the fractions of Compliers and Defiers:

$$APCE_J = P(G_i = \text{C}|\text{ student } i \text{ is in } J) - P(G_i = \text{D}|\text{ student } i \text{ is in } J).$$

If Monotonicity of R_i with respect to Z_i (Assumption 5) holds, Defiers do not exist, and the three $APCE_J$ are all non-negative by construction, being equal to the proportion of Compliers. In this case we could conclude[17] that test scores help teachers improve their recommendations if the $APCE_H$ and the $APCE_{AH}$ are positive while the $APCE_{AL}$ is equal to zero. Having established that the $APCE_J$ are informative about whether test scores help teachers improve the quality of track advising, the next section shows how these population parameters can be identified using data generated by the quasi-experiment under study.

3.4 Identification of the $APCE_J$

Estimands for Helpable (H) students

Adapting Theorem 1 of Imai et al. (2023) to this context, it is possible to show that[18]

$$APCE_H = \frac{E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0)}{Pr(Y_i(1) = 1) - Pr(Y_i(0) = 1)}. \tag{4}$$

where the terms in the numerator do not depend on missing potential outcomes and can be estimated with their sample analogs. The denominator is instead the proportion of Helpable students, given that Rebels are excluded because of the Monotonicity Assumption 4, and can only be partially identified. In some settings, even if principal strata are latent, their proportions can be point identified under specific assumptions. This is the case, for example, of the proportion of compliers in IV settings, because Z is random. In our case, because the principal strata are defined by the joint values of $Y(r)$ and R is confounded, the strata proportions are not point-identified unless we impose further assumptions on the distribution of potential outcomes and the recommendation R.

[17] As already mentioned when we introduced Assumption 5 in the previous section, we emphasize that this assumption allows to interpret $APCE_J$ as the proportion of Compliers in stratum j, but it is not necessary for identification.

[18] For the derivation of this equation and all the others in this section, see the proofs in the Online Appendix to Section 3.4.

To address this problem, Imai et al. (2023) propose to derive non-parametric bounds for the $APCE_H$ by bounding the terms $Pr(Y_i(R) = 1)$ in the denominator of (4). Using Assumption 1 and the Law of Total Probability, it is possible to show that these bounds are:

$$\begin{aligned} \max_z Pr(Y_i = 1, R_i = 0|Z_i = z) \leq \quad & Pr(Y_i(0) = 1) \quad \leq \min_z Pr(Y_i = 1|Z_i = z) \\ \max_z Pr(Y_i = 1|Z_i = z) \leq \quad & Pr(Y_i(1) = 1) \quad \leq 1 - \max_z Pr(Y_i = 0, R_i = 1|Z_i = z) \end{aligned} \tag{5}$$

These results imply that the lower bound of $APCE_H$ is

$$\frac{E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0)}{1 - \max_z Pr(Y_i = 0, R_i = 1|Z_i = z) - \max_z Pr(Y_i = 1, R_i = 0|Z_i = z)} \tag{6}$$

while the upper bound is

$$\frac{E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0)}{\max_z Pr(Y_i = 1|Z_i = z) - \min_z Pr(Y_i = 1|Z_i = z)}. \tag{7}$$

An alternative identification and estimation strategy that we can adopt is instead based on the assumption of unconfoundedness of R with respect to Y:

Assumption 6 *Unconfoundedness of R:*

$Y_i(r) \perp\!\!\!\perp R_i(z) \mid X_i = x, Z_i = z$ *and*

$0 < Pr(R_i(z) = r \mid X_i = x, Z_i = z) < 1 \quad \forall z \in \{0, 1\}, x \in \mathcal{X}, r \in \{0, 1\}$

This assumption, which allows one to point identify the $APCE_J$, implies that, conditioning on observable covariates X, a teacher's decision to upgrade is independent of potential outcomes. In other words, covariates X contain all the information teachers use when making their decision, and the content of the recommendation given this information is random.[19] Identification under this assumption is formally discussed in the Online Appendix to Section 3.4.

[19]This assumption would be violated if, for example, teachers also relied on the student's behavior in class, which we do not observe, when taking their upgrade decisions. As we will see, we achieve similar results under the different identification assumptions we adopt.

Estimands for Always High (AH) students

As for the AH students, we show in the Online Appendix to Section 3.4 that

$$ACPE_{AH} = \frac{Pr(R_i = 0, Y_i = 1 \mid Z_i = 0) - Pr(R_i = 0, Y_i = 1 \mid Z_i = 1)}{Pr(Y_i(0) = 1)}, \tag{8}$$

where, once again, the terms in the numerator do not depend on missing potential outcomes and can be estimated using their sample analogs, whereas further assumptions are needed for the denominator. To bound this denominator, note that:

$$\max_z Pr(Y_i = 1, R_i = 0 \mid Z_i = z) \leq Pr(Y_i(0) = 1) \leq \min_z Pr(Y_i = 1 \mid Z_i = z)$$

Therefore the lower bound of $APCE_{AL}$ is

$$\frac{Pr(R_i = 0, Y_i = 1 \mid Z_i = 0) - Pr(R_i = 0, Y_i = 1 \mid Z_i = 1)}{\min_z Pr(Y_i = 1 \mid Z_i = z)} \tag{9}$$

while the upper bound is

$$\frac{Pr(R_i = 0, Y_i = 1 \mid Z_i = 0) - Pr(R_i = 0, Y_i = 1 \mid Z_i = 1)}{\max_z Pr(R_i = 0, Y_i = 1 \mid Z_i = z)} \tag{10}$$

Also, in the case of AH students, as with their H peers, we will compare the bounds obtained under the above identification strategy with point estimates obtained by assuming the unconfoundedness of R (Assumption 6), as shown in the Online Appendix Appendix to Section 3.4.

Estimands for Always Low (AL) students

Finally, considering the AL students, the Online Appendix to Section 3.4 shows that

$$APCE_{AL} = \frac{Pr(R_i = 1, Y_i = 0 \mid Z_i = 1) - Pr(R_i = 1, Y_i = 0 \mid Z_i = 0)}{1 - Pr(Y_i(1) = 1)}. \tag{11}$$

where, also in this case, the terms in the numerator do not depend on missing potential outcomes and can be estimated with their sample analogs, while the estimation of the denominator requires further assumptions. To bound this denominator, note that:

$$\max_z Pr(Y_i = 1 \mid Z_i = z) \leq Pr(Y_i(1) = 1) \leq 1 - \max_z Pr(Y_i = 0, R_i = 1 \mid Z_i = z)$$

Therefore, the lower bound of $APCE_{AL}$ is

$$\frac{Pr(R_i = 1, Y_i = 0 \mid Z_i = 1) - Pr(R_i = 1, Y_i = 0 \mid Z_i = 0)}{1 - \max_z Pr(Y_i \mid Z_i = z)} \tag{12}$$

while the upper bound is

$$\frac{Pr(R_i = 1, Y_i = 0 \mid Z_i = 1) - Pr(R_i = 1, Y_i = 0 \mid Z_i = 0)}{\max_z Pr(R_i = 1, Y_i = 0 \mid Z_i = z)}. \tag{13}$$

Once again, as for the $APCE_H$ and the $APCE_{AH}$, also the $APCE_{AL}$ can be point identified under unconfoundedness of R (Assumption 6), as shown in the Online Appendix to Section 3.4.

4 Test scores and quality of teachers' advice: evidence

4.1 Estimates for the three strata

Estimates for Helpable (H) students

In the first column of Table 6, we present estimates of the numerator of the $APCE_H$ estimand in the RHS of equation (4), separately for each track and combined for the three aggregate tracks: Vocational, Academic and All.[20] The point estimates of this numerator are strictly positive in all tracks. Students scoring at the cutoff, who can be upgraded by their primary school teachers based on test scores, are 1.8–to–10% more likely to complete a higher track than students who score below the cutoff. These point estimates are statistically different from zero except for the BL track. The estimates for the three aggregate tracks in the last rows of the table are also significant and equal to 4.4%.

Recall that the denominator of the $APCE_H$ estimand in equation (4) is certainly positive

[20]Here and in what follows, each population estimand has been estimated separately for each track and cohort. The estimates reported in the tables for aggregate tracks have been obtained as weighted averages with weights based on the relative sample sizes of each cell. For this reason, the remaining tables report only one set of results, rather than separate results by cohort. Standard errors (reported in parentheses) are bootstrapped (with 1000 repetitions) separately in each track, and then appropriately aggregated with the same weights. The estimates in the tables reported in the text include the covariates $\tilde{Z}_i$ and X_i (see Table 2). The Online Appendix to Section 4.1 reports tables with corresponding estimates obtained without the balanced covariates X_i (but including $\tilde{Z}_i$). The inclusion or exclusion of the balanced covariates does not significantly alter the estimates, as expected. This evidence supports the validity of Assumption 2 (Randomization of the source).

Table 6: Estimates for the $APCE_H$, with covariates

Track	Numerator $APCE_H$	Lower bound denominator $APCE_H$	Upper bound denominator $APCE_H$	Lower bound $APCE_H$	Upper bound $APCE_H$
	1	2	3	4	5
V: BL	.0251	.0251	.6534	.0381	1
	(.0167)	(.0154)	(.0126)	(.0263)	-
V: BL/KL	.0773	.0773	.2877	.2757	1
	(.0294)	(.0279)	(.0222)	(.1029)	-
V: KL	.0996	.0996	.6984	.1426	1
	(.0159)	(.0159)	(.0112)	(.0229)	-
V: KL/GT	.0256	.0256	.3967	.0644	1
	(.0238)	(.0217)	(.0232)	(.0564)	-
V: GT	.0182	.0182	.7607	.0237	1
	(.0077)	(.0077)	(.0076)	(.0104)	-
V: GT/HAVO	.0526	.0526	.5089	.1021	1
	(.0164)	(.0159)	(.0138)	(.0309)	-
A: HAVO	.0349	.0349	.8082	.0432	1
	(.0076)	(.0076)	(.0065)	(.0096)	-
A: HAVO/VWO	.0642	.0642	.5451	.117	1
	(.0156)	(.0154)	(.0126)	(.0276)	-
VOCATIONAL	.0445	.0445	.6486	.0686	1
	(.0058)	(.0059)	(.005)	(.009)	-
ACADEMIC	.0442	.0442	.7251	.0609	1
	(.0072)	(.0071)	(.0059)	(.0099)	-
ALL	.0444	.0444	.6796	.0653	1
	(.0045)	(.0045)	(.0038)	(.0066)	-

Notes: The row headings are the names of the secondary school tracks to which a student can be assigned in the Dutch system, ranked by level of difficulty. The last three rows of the table aggregate the Vocational (VMBO*), the Academic (HAVO*), and All tracks, respectively. With reference to the $APCE_H$ estimand defined in equation (4), the table reports for the outcome Y and separately for each track, estimates of the numerator (in column 1), of the lower and upper bounds of the denominator (in columns 2 and 3), and of the lower bound of the $APCE_H$ (in column 4). Column 5 reports the upper bound of the $APCE_H$ as equal to 1 because it is the ratio between $E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0)$ and $\max_z Pr(Y_i = 1|Z_i = z) - \min_z Pr(Y_i = 1|Z_i = z)$ which are exactly equal under Assumptions 4 and 5. See footnote 21 for the intuition. The definitions of these two last statistics are in equations (6) and (7), respectively. The aggregations for the last three rows are performed as follows: point estimates of the numerator and of the lower and upper bounds of the denominator are created by weighing each track in the corresponding aggregate by the relative population in each track. For the lower and upper bounds of the $APCE_H$, the aggregated estimates for the lower and upper bounds of the denominators are used directly. Standard errors are computed using a block bootstrap at the school level with 1000 repetitions; each quantity in every iteration is calculated in the same way. The specifications in all columns include the covariates $\tilde{Z}_i$ and X_i (see Table 2). The Online Appendix to Section 4.1 reports tables with corresponding estimates obtained without the balanced covariates X_i (but including $\tilde{Z}_i$). The inclusion or exclusion of the balanced covariates does not change the estimates in a relevant way, as expected, supporting the validity of Assumption 2 (Randomization of the source).

and equal to the proportion of Helpable students, given Assumption 4 of Monotonicity of Y_i with respect to R_i. Therefore, a positive numerator supports the conclusion that the information provided by test scores helps recommenders give better advice in the sense that, thanks to this information, they become better at detecting Helpable students and directing them into the most challenging secondary school track that they can successfully complete.

To confirm this conclusion and assess also quantitatively by how much test scores help teachers achieve this goal, estimates of the $APCE_H$ are needed, but as explained in the previous section, the denominator in the RHS of equation (4), which is the proportion of H students, can only be bounded without additional assumptions. Estimates of the lower and upper bounds of this denominator are reported in columns 2 and 3 of Table 6. Focusing on the aggregate tracks in the last three rows, the proportion of H ranges between approximately 4% and 73%.

As a consequence of the large gap between the estimated bounds of the denominator, also the bounds of the $APCE_H$ reported in columns 4 and 5 of Table 6 are substantially different one from the other, ranging between 6% and 100% for the three aggregate tracks.[21] However, what matters most from the viewpoint of our research question is the lower bound of the $APCE_H$, which allows us to conclude that the information provided by test scores improves the quality of recommendations by at least 6% as measured by their capacity to push Helpable students into the high track. This is a remarkable finding, particularly if the proportion of Helpable students in the population can be as high as 80%, as suggested by the estimated upper bound of the $APCE_H$ denominator in column 3 of Table 6.

[21]Note that the upper bound of the $APCE_H$ derived in equation (4) and reported in column 5 of Table 6 is 1 by construction because it is the ratio between $E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0)$ and $\max_z Pr(Y_i = 1|Z_i = z) - \min_z Pr(Y_i = 1|Z_i = z)$, which are exactly equal under the two Monotonicity Assumptions 4 and 5. The intuition is that the intention-to-treat $(E[Y_i \mid Z_i = 1] - E[Y_i \mid Z_i = 0])$ at the numerator of this ratio identifies the proportion of students who are compliers *and* are also helpable. Hence, the minimum plausible share of all Helpable students at the denominator of this ratio cannot be smaller than the share of Helpable who are also compliers at the numerator, and the upper bound of the $APCE_H$ must be 1 by definition. Note also that when the share of Helpable who are also compliers is effectively equal to zero or very small, we may estimate it to be negative because of small sample variability. In tracks where this occurs, we round this proportion to zero. However, even when this proportion is infinitesimally small, it is still the case that the upper bound for the $APCE_H$ is 1 for the reason explained above.

Table 7: Estimates of the $APCE_{AH}$, with covariates

Track	Numerator $APCE_{AH}$	Lower bound denominator $APCE_{AH}$	Upper bound denominator $APCE_{AH}$	Lower bound $APCE_{AH}$	Upper bound $APCE_{AH}$
	1	2	3	4	5
V: BL	.0004	.317	.3189	.0011	.0011
	(.0118)	(.0111)	(.0111)	(.036)	(.036)
V: BL/KL	0227	.6858	.705	.0323	.0332
	(.0243)	(.0209)	(.0214)	(.0336)	(.034)
V: KL	.0083	.2289	.2392	.0368	.0379
	(.0101)	(.0086)	(.0091)	(.0401)	(.0416)
V: KL/GT	.0745	.5681	.581	.1288	.1319
	(.0297)	(.0206)	(.0198)	(.0499)	(.05)
V: GT	.0097	.1938	.1938	.0482	.0482
	(.006)	(.0055)	(.0055)	(.03)	(.03)
V: GT/HAVO	.0439	.4169	.4262	.1023	.1041
	(.0165)	(.0113)	(.0115)	(.0375)	(.0379)
A: HAVO	0	.1283	.1283	0	0
	(.0006)	(.004)	(.004)	(.0047)	(.0047)
A: HAVO/VWO	.0281	.3667	.3777	.074	.0762
	(.0146)	(.0105)	(.0109)	(.037)	(.0379)
VOCATIONAL	.0188	.2988	.3044	.0618	.063
	(.0049)	(.004)	(.0041)	(.0157)	(.0159)
ACADEMIC	.0089	.2035	.207	.0429	.0436
	(.0046)	(.0042)	(.0043)	(.0219)	(.0223)
ALL	.0148	.2601	.2649	.0558	.0568
	(.0034)	(.0029)	(.003)	(.0126)	(.0128)

Notes: The row headings are the names of the secondary school tracks to which a student can be assigned in the Dutch system, ranked by level of difficulty. The last three rows of the table aggregate the Vocational (VMBO*), the Academic (HAVO*), and All tracks, respectively. With reference to the $APCE_{AH}$ estimand defined in equation (8), the table reports for the outcome Y and separately for each track, estimates of the numerator (in column 1), of the lower and upper bounds of the denominator (in columns 2 and 3), and of the lower and upper bounds of the $APCE_{AH}$ (in columns 4 and 5). The definitions of these two last statistics are in equations (9) and (10), respectively. The aggregations for the last three rows are performed as follows: point estimates of the numerator and of the lower and upper bounds of the denominator are created by weighing each track in the corresponding aggregate by the relative population in each track. For the lower and upper bounds of the $APCE_{AH}$, the aggregated estimates for the lower and upper bounds of the denominators are used directly. Standard errors for each quantity are computed using a block bootstrap at the school level with 1000 repetitions; each quantity in every iteration is calculated in the same way. The specifications in all columns include the covariates $\tilde{Z}_i$ and X_i (see Table 2). The Online Appendix to Section 4.1 reports tables with corresponding estimates obtained without the balanced covariates X_i (but including $\tilde{Z}_i$). The inclusion or exclusion of the balanced covariates does not change the estimates in a relevant way, as expected, supporting the validity of Assumption 2 (Randomization of the source).

Estimates for Always High (AH) students

Moving to the stratum of Always High students, the estimates for them are reported in Table 7. In their case, the best source of advice is the one that recommends the high track to the *highest* number of them (possibly to all of them) in order to minimize the cost of track changes they would incur if they started secondary school in the low track. Recall, in fact, that these students always complete the high track no matter where they start.

Focusing for brevity just on Columns 4 and 5 of the table, the gap between the lower and the upper bounds of the $APCE_{AH}$ is remarkably small in almost all tracks, allowing for a very precise estimation of this population parameter.[22] The intuition for why we obtain such a precise result is that we can be confident in the population proportion of AH students (i.e. the denominator) as those with $R = 0$ and $Y = 1$ are for sure AH, and measuring this under the cutoff (i.e. for those with $Z = 0$) gives a precise estimate of the population parameter, showing that the population proportion of AH is approximately 26%.[23] In the eight basic tracks, the point estimates of the $APCE$ range from 0 to 13%. When the tracks are aggregated in the Vocational, Academic, and All tracks (last three rows of the table), the $APCE_{AH}$ is estimated to be about 6.2%, 4.3%, and 5.6%, respectively, and statistically significant. These results indicate that when teachers are allowed to upgrade recommendations based on test scores, they send a significantly higher fraction of AH students to the high track, thereby reducing the costly track changes this group would otherwise incur.

Estimates for Always Low (AL) students

In the case of Always Low students, instead, our estimates of the $APCE_{AL}$ suggest that test score information does not help teachers give better recommendations. These estimates are reported in Table 8. The best source of advice for these students is the one that recommends the high track to the *lowest* number of them (possibly to none), in order to minimize the

[22] In the case of HAVO track, the sample analog of the numerator in the RHS of equation (8) is close to zero but has a negative sign, possibly due to small sample variability. We approximate it with 0.

[23] More specifically, because it is extremely rare to have $Z = 0$ and $R = 1$ (which can only occur due to the institutional complications explained in footnote 12) we know that among those with $Z = 0$, those with $R = 0$ and $Y = 1$ are AH for sure, while those with $R = 0$ and $Y = 0$ are for sure not AH, and these two categories together form virtually the whole population. This means that the fraction of AH students can be estimated very precisely, as Z is randomized and the same fraction should exist above the threshold.

Table 8: Estimates of the $APCE_{AL}$, with covariates

Track	Numerator $APCE_{AL}$	Lower bound denominator $APCE_{AL}$	Upper bound denominator $APCE_{AL}$	Lower bound $APCE_{AL}$	Upper bound $APCE_{AL}$
	1	2	3	4	5
V: BL	.0271	.028	.6576	.0409	.9765
	(.0056)	(.0055)	(.0125)	(.0085)	(.0266)
V: BL/KL	.0239	.0266	.2177	.1126	.8981
	(.0084)	(.008)	(.0208)	(.0404)	.(1399)
V: KL	.0658	.0727	.6612	.0991	.9026
	(.0071)	(.0069)	(.0126)	(.0105)	(.03)
V: KL/GT	.0335	.0352	.3934	.0853	.9569
	(.0084)	(.0083)	(.0194)	(.0212)	(.0599)
V: GT	.0377	.041	.7899	.0477	.9193
	(.0038)	(.0036)	(.0072)	(.0048)	(.0253)
V: GT/HAVO	.0644	.0742	.5213	.1238	.8682
	(.0082)	(.0077)	(.0128)	(.0156)	(.0409)
A: HAVO	.0487	.0539	.8464	.0576	.9035
	(.0042)	(.004)	(.0058)	(.005)	(.0267)
A: HAVO/VWO	.0788	.0883	.5581	.141	.8928
	(.0077)	(.0074)	(.0126)	(.0133)	(.0278)
VOCATIONAL	.0458	.0506	.652	.0703	.9059
	(.0027)	(.0025)	(.005)	(.004)	(.016)
ACADEMIC	.0582	.0648	.7554	.0771	.8988
	(.0037)	(.0035)	(.0055)	.0048	(.0192)
ALL	.0508	.0563	.694	.0733	.9026
	(.0022)	(.0021)	(.0037)	(.0032)	(.0126)

Notes: The row headings are the names of the secondary school tracks to which a student can be assigned in the Dutch system, ranked by level of difficulty. The last three rows of the table aggregate the Vocational (VMBO*), the Academic (HAVO*), and All tracks, respectively. With reference to the $APCE_{AL}$ estimand defined in equation (11), the table reports for the outcome Y and separately for each track, estimates of the numerator (in column 1), of the lower and upper bounds of the denominator (in columns 2 and 3), and of the lower and upper bounds of the $APCE_{AL}$ (in columns 4 and 5). The definitions of these two last statistics are in equations (12) and (13), respectively. The aggregations for the last three rows are performed as follows: point estimates of the numerator and of the lower and upper bounds of the denominator are created by weighing each track in the corresponding aggregate by the relative population in each track. For the lower and upper bounds of the $APCE_{AL}$, the aggregated estimates for the lower and upper bounds of the denominators are used directly. Standard errors for each quantity are computed using a block bootstrap at the school level with 1000 repetitions; each quantity in every iteration is calculated in the same way. The specifications in all columns include the covariates $\tilde{Z}_i$ and X_i (see Table 2). The Online Appendix to Section 4.1 reports tables with corresponding estimates obtained without the balanced covariates X_i (but including $\tilde{Z}_i$). The inclusion or exclusion of the balanced covariates does not change the estimates in a relevant way, as expected, supporting the validity of Assumption 2 (Randomization of the source).

cost of the track changes they would incur if they started secondary school in the high track. Recall that these students can only complete the low track, no matter where they start.

Focusing again for brevity just on the last two columns of the table, Column 4 shows that the estimated lower bound of the $APCE_{AL}$ is positive and statistically different from zero in many tracks, ranging from 4% for BL to 14% for HAVO/VWO. Considering the aggregate tracks in the last three rows of the table, the $APCE_{AL}$ is at least as high as 7% and significantly different from zero in all of them. Therefore, it appears that the information provided by test scores leads teachers to upgrade AL students into the high track. The consequence is a deterioration of the quality of recommendations because AL students cannot complete the high track. A possible reason for this finding is that primary school teachers do not incur any costs resulting from secondary school track changes and are therefore less sensitive to these consequences of the advice they offer to their students.

4.2 A comprehensive evaluation

The evidence in the previous section suggests that, in the case of Helpable and Always High students, test score information helps teachers improve the quality of their recommendations in terms of the two criteria that we have adopted (directing students towards the most difficult track they can successfully complete and reducing track changes). On the contrary, in the case of the Always Low, the positive $APCE_{AL}$ estimate indicates that the same information reduces this quality because it induces AL students to enroll in the high track that they later cannot complete.

The classification framework proposed by Ben-Michael et al. (2024) provides a comprehensive evaluation of whether it is a good idea to offer teachers the possibility to upgrade based on test score information, depending on how much the policy-maker cares about H and AH students compared to their AL peers. Define with l_{ry} the loss deriving from a recommendation $R = r \in \{0, 1\}$ given to a student with potential outcome $Y(1) = y \in \{0, 1\}$. The four possible values of this loss are represented in the following Confusion Matrix:

Confusion Matrix

	Negative ($R = 0$)	Positive ($R = 1$)
Negative ($Y(1) = 0$)	True Negative $l_{00} = 0$	False Positive $l_{10} > 0$
Positive ($Y(1) = 1$)	False Negative $l_{01} > 0$	True positive $l_{11} = 0$

The True Negative is a student whom the teacher correctly identifies as Always Low, while the True Positive is correctly identified as Always High or Helpable. In all these cases, the recommendation of the teacher does not cause any loss: $l_{00} = l_{11} = 0$.[24] The False Negative is a Helpable or Always High student who is not recognized as such by the teacher and thus receives a low recommendation. In her/his case, the loss l_{01} is positive because the student could complete the high track following the opposite advice (in the case of H) and complete the high track regardless, but with a track change (in the case of AH). Similarly positive is the loss for the last type of student, False Positive l_{10}, who is an Always Low not recognized as such by the teacher. Therefore, she/he receives a high recommendation without being later able to complete the high track.

Let's normalize to $l_{01} = 1$ the loss suffered by a False Negative student. It is then reasonable to assume that $l_{10} \leq l_{01} = 1$. This is because l_{10} is the "short-term" loss generated by the cost of changing track that a False Positive AL student incurs if she/he is recommended a too-difficult track. In light of this, if l_{01} is the loss of a False Negative AH, then it is also a short-term cost of having to change track, which can be assumed to be comparable for AL and AH students, and therefore the above inequality holds weakly: $l_{10} = l_{01} = 1$. If, instead, l_{01} is the loss of a False Negative H student, he/she suffers a long-term loss extending well beyond the end of secondary school, because she/he earns, in expectation, the lower lifetime income of a low graduate, while with the opposite advice she/he would earn the higher one of a high graduate. Therefore, in the case of this student, we can assume that $l_{10} < l_{01} = 1$ strictly.

We can then write the overall expected loss of a recommendation given by a source

[24] To simplify the analysis, we abstract here from the possibility of negative losses (gains), which may differ between H, AL, and AH students.

$Z = z \in \{0, 1\}$ as

$$\mathcal{L}(l_{10}, z) = \Pi_{01}(z) + l_{10}\Pi_{10}(z), \tag{14}$$

where $\Pi_{01}(z) = Pr(R(z) = 0; Y(1) = 1)$ is the probability of a False Negative (H or AH) and $\Pi_{10}(z) = Pr(R(z) = 1; Y(1) = 0)$ is the probability of a False Positive (AL). Our goal is to compare the overall loss $\mathcal{L}(l_{10}, z)$ when the source is a "teacher alone" ($z = 0$) or a "teacher informed by test scores and allowed to upgrade"" ($z = 1$), computed at different values of the relative weight l_{10}. This weight is at most equal to 1, when the short-run loss of AL students is considered by the policy-maker as relevant as the long-run loss of H students, so that $l_{10} = 1$, and decreases towards 0 as the former loss becomes less relevant ($l_{10} < 1$).

In the Online Appendix to Section 4.2 we show that

$$\mathcal{L}(l_{01}, 1) - \mathcal{L}(l_{01}, 0) = \Pi_{11}(0) - \Pi_{11}(1) + l_{10}(\Pi_{10}(1) - \Pi_{10}(0)) \tag{15}$$

where the RHS is identified and can be estimated thanks to Assumption 2 (Randomization of the source). The left panel of Figure 2 reports for each track the P-values of the tests of the null that this difference is negative for values of the relative weight l_{01} in the set $\{0, 0.05, 0.1, ..., 1\}$:

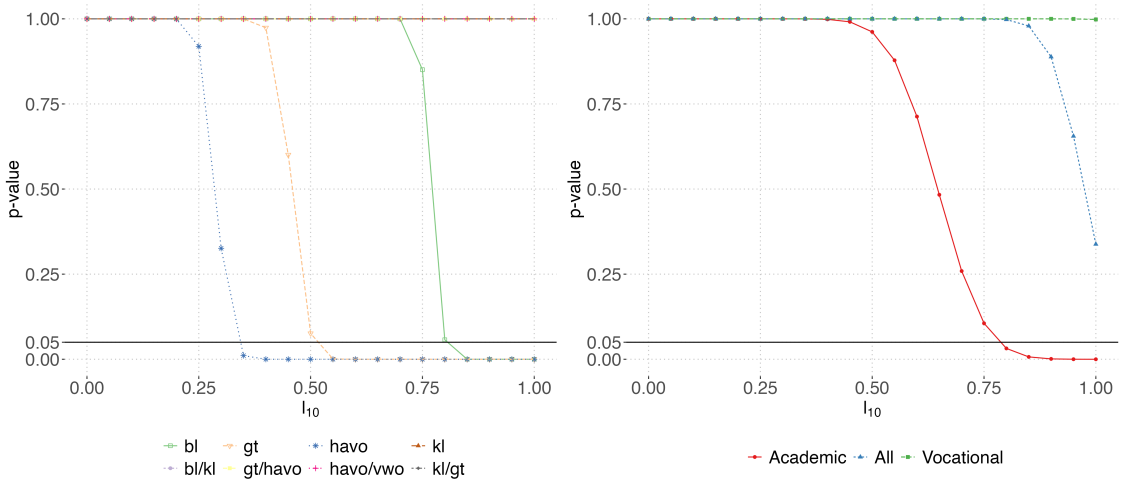
$$H_0 : \mathcal{L}(l_{01}, 1) - \mathcal{L}(l_{01}, 0) \leq 0; \qquad H_1 : \mathcal{L}(l_{01}, 1) - \mathcal{L}(l_{01}, 0) > 0 \tag{16}$$

Failure to reject the null hypothesis of this test (high P-value) indicates that we cannot rule out the possibility that allowing teachers to upgrade their recommendations based on test scores is a better (or equally effective) system because we do not have enough evidence to conclude that the provision of test scores leads to larger losses. Conversely, if we can reject H_0 (with a small P-value), we can rule out the possibility that allowing teachers to upgrade based on test scores generates smaller losses and thus improves recommendations.

With the exception of three tracks (V:BL, V:GT and A:HAVO), we cannot reject that the difference is negative even when the short-run loss of AL students is considered as relevant as the long-run loss suffered by H students ($l_{10} = 1$) and AL students have the maximum relative weight. For the remaining V:BL, V:GT and A:HAVO tracks, we reject that the difference is negative only when the short-term loss suffered by AL students has an (implausibly) high

relative weight – higher than about 0.8, 0.5, and 0.3, respectively, at the 5% significance level. Similar evidence is reported in the right panel of the figure for the three aggregate tracks. We never reject the null at the same significance level for the All and Vocational tracks, even if the relative weight on the short-term loss of AL students is the highest. In the case of the Academic track, we reject only when this weight is higher than about 0.75.

Figure 2: Comprehensive evaluation that test score information improves teachers' recommendations



Notes: The figure reports p-values of the test that the difference $\mathcal{L}(l_{01}, 1) - \mathcal{L}(l_{01}, 0)$ between the overall losses with and without test score information (equation 15) is negative, at different values of the relative weight l_{10} of AL students. Failure to reject the null hypothesis in this test (high P-value) indicates that we cannot rule out that the loss is larger when teachers do not see the test score information. The solid line indicates the 5% significance level of the P-values above which the null cannot be rejected with sufficient statistical confidence.

In light of this evidence, we conclude that, overall, the possibility of upgrading based on test scores helps primary teachers improve the quality of their secondary school track recommendations. For this not to be the case, the short-term loss of AL students (who have to change track because they are recommended too difficult a track that in the end they cannot complete) would have to be considered implausibly almost as relevant as the life-time loss of H students (who are prevented from completing a high track they would be able to complete).

5 An alternative to assuming the Exclusion Restriction

In this section, we explore the identification and estimation of the Average Principal Causal Effects without relying on the Exclusion Restriction (ER) assumption. This assumption

may fail to hold if the causal effect of Z on Y is not mediated solely by R. Consider, for instance, the Never-Takers stratum. Since these students are never upgraded, under the ER their expected value of Y should, in principle, be the same on both sides of the cutoff. However, scoring above the cutoff may boost the confidence of these students and encourage their parents to advocate for an upward track change after initial enrollment in secondary school. More generally, it may have a direct positive effect on these students' performance in secondary school, even in the absence of an upgrade at the moment of the initial track enrollment. In these cases, the Exclusion Restriction would not hold.

We study such a possibility by deriving bounds on the APCEs when a specific form of direct effect of Z on Y for given R is allowed to play a role. Finding that even in the presence of this effect, our results are unchanged would suggest that the ER assumption is not strictly necessary for our conclusions about the preferability of informing teachers about the test score results of their students, or, in other words, that at least this deviation from the Exclusion Restriction assumption is not changing our results (see also Angrist et al., 1996).

To this end, we first assume that the direct effect of Z on Y is additive and equal for all students:

Assumption 7 *Homogeneity of the direct effect of Z on Y*
$\mathbb{E}[Y_i(1,r)] = \mathbb{E}[Y_i(0,r)] + \eta \quad$ *for* $r \in \{0,1\}$,

Under Assumption 7, we derive new expressions for the APCEs and their bounds as a function of η, as described in the Online Appendix to Section 5.[25] These expressions generalize the original framework proposed by Imai et al. (2023). Indeed, it can be shown that as η goes to 0, i.e. the level at which the Exclusion Restriction holds, the new bounds converge to those derived in Section 3 under the ER. More generally, the bounds of the APCEs that we obtain are functions of the unknown parameter η, which serves as a *sensitivity* parameter (e.g., Imbens, 2003), and as such cannot be identified and estimated without invoking further assumptions. In order to gain some insights into the plausible magnitude of η, we can invoke the Principal Ignorability assumption described below that allows to identify and

[25] Mealli and Pacini (2013) study another example of bounds for principal causal effects when Exclusion Restriction assumptions are relaxed.

estimate η. This enables us to establish not only the extent to which the ER assumption is violated but also the stability of our results if the violation indeed occurs and identification is achieved under an alternative assumption.

To understand the nature of this Principal Ignorability assumption, consider the effect of Z on Y for Never-Takers, which under the Homogeneity Assumption 7 is equal to η:

$$\eta = \mathbb{E}[Y \mid Z = 1, R(1) = 0, R(0) = 0] - \mathbb{E}[Y \mid Z = 0, R(1) = 0, R(0) = 0]. \tag{17}$$

The first term on the RHS of equation (17) is identified and can be easily estimated because students who score at the cutoff but are not upgraded by the teacher are certainly Never-Takers and are observed. The second term, however, is not identified because the group of students scoring below the cutoff (and not upgraded) comprises both Compliers and Never Takers, who cannot be distinguished. We solve this problem by invoking the following assumption:

Assumption 8 *Principal Ignorability:*
$Y_i(0) \perp\!\!\!\perp R_i(1) \mid R_i(0) = 0, X_i$

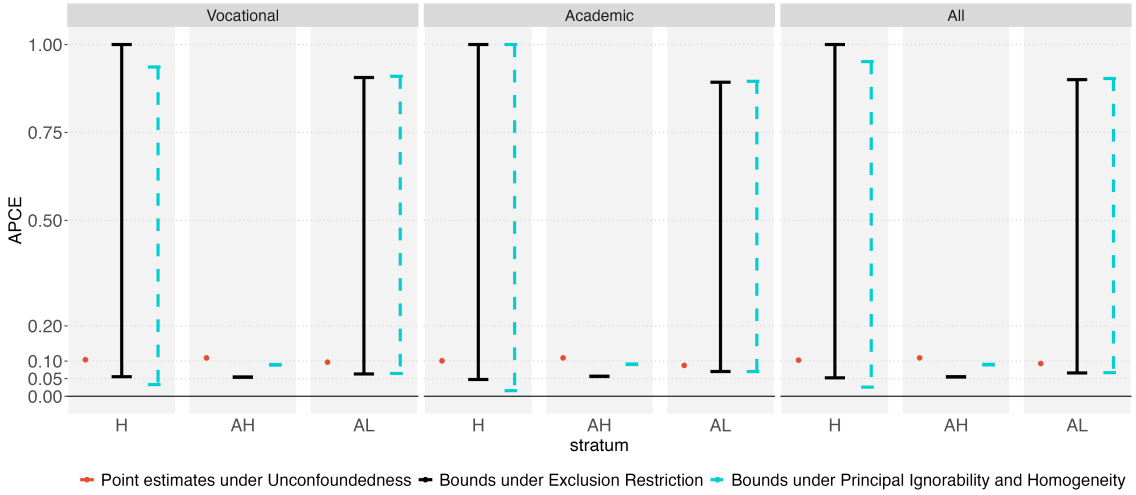
This assumption implies that, conditional on covariates, the distribution of potential outcomes below the cutoff is the same for Compliers and Never-Takers.[26] Therefore, the second term on the RHS of equation (17) is identified and η can be estimated.[27]

Table A5 in the Online Appendix to Section 5 reports estimates of η obtained with the above procedure. Some point estimates of η are quantitatively sizeable. For example, $\hat{\eta}$ for students in track V:BL/KL in the 2016 cohort is 0.1013 (s.e.: 0.0465), which is larger than the estimated numerator of the $APCE_H$ for Helpable students in the same track, i.e. 0.0773. In light of this finding, we use the point estimates $\hat{\eta}$ to estimate the bounds for the APCEs

[26] This assumption is, of course, as debatable as the ER assumption. Moreover, similarly to the unconfoundedness assumption, its plausibility crucially depends on the information contained in the observed covariates (see Feller et al., 2017, Mattei et al., 2023, and Ding and Lu, 2017) However, our goal here is simply to assess how stable our results are when invoking this assumption instead of the ER.

[27] Using observations above the cutoff, where Never-Takers are identified, we can estimate the probability of being a Never-Taker conditional on covariates (the estimated principal score $\hat{p}_{nt}$). Then, $\mathbb{E}[Y_i \mid Z_i = 0, R_i(1) = 0, R_i(0) = 0]$ can be estimated by the average Y_i for units with $Z_i = 0$ and $R_i = 0$, weighted by the principal score. That is, under Assumption 8, $\mathbb{E}[Y_i \mid Z_i = 0, R_i(1) = 0, R_i(0) = 0]$ is identified and can be estimated by $\frac{\sum_{i:Z_i=0,R_i=0} \hat{p}_{i,nt} Y_i}{\sum_{i:Z_i=0,R_i=0} \hat{p}_{i,nt}}$. See Mattei et al. (2023) for further details.

Figure 3: Overview of the $APCE_J$ estimates



Notes: The figure reports, for the three aggregate tracks and for the H, AH and AL strata: the point identified $APCE_J$ estimates obtained under unconfoundedness (Assumption 6, red dot), that are derived in the Online Appendix to Section 3.4; the bounds of the $APCE_J$ obtained under the ER (Assumption 3, solid black), that are derived in Section 3; and the bounds of the $APCE_J$ obtained under PI and Homogeneity (Assumptions 8 and 7, dashed light-blue), that are derived in Section 5.

when the Homogeneity Assumption 7 and the Principal Ignorability Assumption 8 hold instead of the ER.

Figure 3 compares, for the three aggregate tracks and for the H, AH, and AL strata, the point identified $APCE_J$ estimates obtained under unconfoundedness (Assumption 6, red dot), that are derived in the Online Appendix to Section 3.4; the bounds of the $APCE_J$ obtained under the ER (Assumption 3, solid black), that are derived in Section 3; and the bounds of the $APCE_J$ obtained under PI and Homogeneity (Assumptions 8 and 7, dashed light-blue), that are derived in this section. In the case of H and AL students, unconfoundedness delivers estimates that are located near the lower bounds of the partially identified alternative estimates. For the AH students, both with and without unconfoundedness, the $APCE_J$ estimates are precise and close one to the other.

This graphical comparison under the different assumptions indicates that the implications in terms of *lower bounds* of the relevant effects are very similar in all cases. We conclude that our main results are robust with respect to these alternative identification strategies.[28]

[28]The estimates of all the bounds obtained without the Exclusion Restriction and under the alternative PI and Homogeneity assumptions can be found in Online Appendix to Section 5.

6 Fairness of teachers' recommendations

Within the same statistical framework described in the previous sections, Imai and Jiang (2022) and Imai et al. (2023) propose the concept of "Principal Fairness," which can be interestingly used to compare the fairness of two sources of first-year track assignment.

Definition 9 *Principal Fairness:*
A source $Z_i = z \in \{0, 1\}$ of first-year track assignment $R_i(z)$ satisfies Principal Fairness with respect to a protected attribute B_i (e.g., SES, race, gender), if the first-year recommendations deriving from this source are conditionally independent of B_i within each principal stratum $J_i \in \{AH, AL, H\}$

$$Pr(R_i(z)|J_i, B_i) = Pr(R_i(z)|J_i) \tag{18}$$

According to this definition, a recommender is fair as long as her recommendations are independent of the protected attribute among students in the same stratum. The analogous fraction in other strata can be lower or higher if it is equal for both genders within each stratum. A test for Principal Fairness of a source of secondary school track recommendations can then be designed as follows. Let the two sources of recommendations that we would like to compare be: "teacher alone" ($z = 0$) versus "teacher informed by test scores and allowed to upgrade" ($z = 1$). Given two values b and b' of a protected attribute B_i, the Principal Fairness of source z in stratum j is given by:

$$\Delta_j(z) = Pr\{R_i(z) = 1|B_i = b, J_i = j\} - Pr\{R_i(z) = 1|B_i = b', J_i = j\}, \tag{19}$$

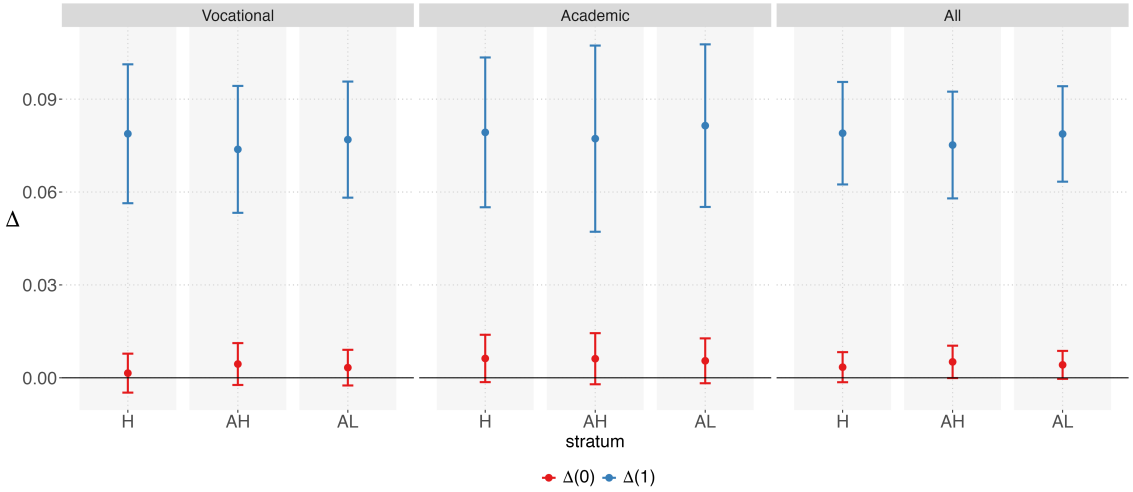
where $R_i(z) = R \in \{0, 1\}$ is the recommendation given by source z to student i. Note that source z is perfectly fair if $\Delta_j(z) = 0$. Otherwise, source z is *unfair* because its probability of recommending the high track to students in stratum j changes with the values of B_i.

It is important to note that, considering the protected attribute B_i without conditioning on the other covariates requires to interpret carefully the reason for a possible lack of fairness. A fairness criterion that does not hold unconditionally may hold conditionally if the covariates correlate with the protected attribute (see Imai and Jiang, 2022, for a discussion). For instance, suppose that teachers do not discriminate based on immigrant status but do

discriminate based on SES, and immigrants have, in general, lower SES. In this case, if the protected attribute $B_i = b$ in equation (19) is immigrant status, this attribute should actually be considered as a proxy for SES. However, from a descriptive viewpoint, it would still be the case that immigrants are treated differently, even if in each stratum defined by SES they are not. As argued by Imai et al. (2023), the choice between marginal principal fairness and conditional principal fairness is not statistical.

The sign of $\Delta_j(z)$ is also important, as it indicates the direction of discrimination. Suppose again that $B_i = b$ denotes immigrant students. Then finding, for example, that $\Delta_j(1) > 0$ while $\Delta_j(0) < 0$, would mean that when recommenders are allowed to upgrade their advice based on test scores, immigrant students are positively discriminated into the high track, while when the recommenders give advice without seeing test scores, immigrants are negatively discriminated. Note that in our setting, where teachers are officially allowed to upgrade only students who score above the cutoff, all the $\Delta_j(0)$ should be equal to 0 by definition. Therefore, a $\Delta_j(0) = 0$ cannot be interpreted as evidence of fairness in the provisional track recommendation made by teachers without the test scores. On the other hand, a $\Delta_j(0)$ different from 0 would indicate different rates of non-compliance below the cutoff across protected attributes in stratum j.

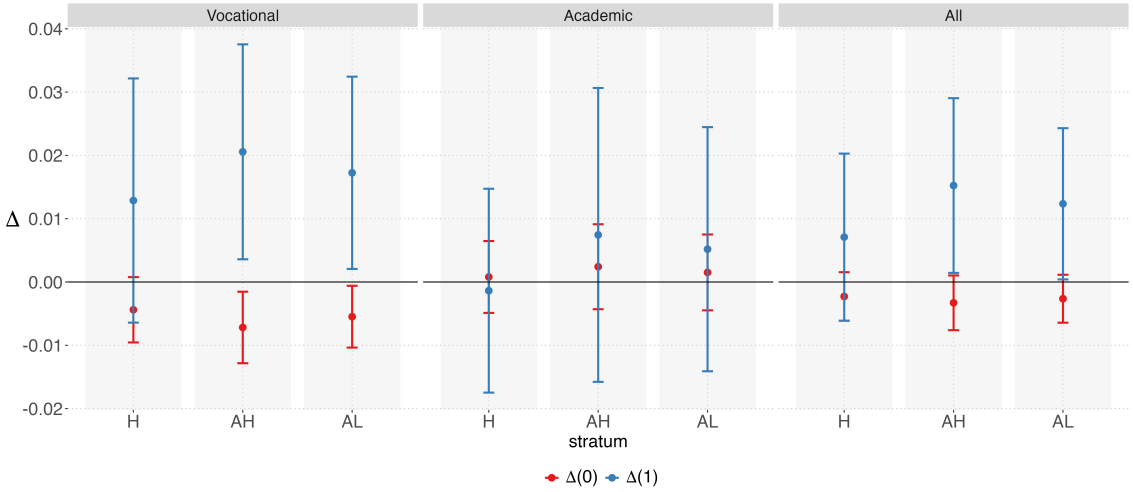
Figure 4: Fairness for immigrants



Notes: the figure plots, for the three aggregate tracks and for the H, AH and AL strata, the point estimates of the statistic $\Delta_j(z) = Pr(R(z) = 1 \mid \text{Immigrant}) - Pr(R(z) = 1 \mid \text{No Immigrant})$ and the correspondent 95% confidence intervals.

Under assumptions 1 (SUTVA), 2 (Local Randomization), 3 (Exclusion Restriction), 4 (Monotonicity of Y with respect to R), and 6 (Unconfoundedness of R with respect to Y),

Figure 5: Fairness by SES



Notes: the figure plots, for the three aggregate tracks and for the H, AH and AL strata, the point estimates of the statistic $\Delta_j(z) = Pr(R(z) = 1 \mid \text{Low SES}) - Pr(R(z) = 1 \mid \text{High SES})$ and the correspondent 95% confidence intervals.

as shown by Imai et al. (2023), we can estimate equation (19). Figure 4 plots, for the three aggregate tracks and for the H, AH, and AL strata, $\Delta_j(1)$ and $\Delta_j(0)$, considering immigrant status as the protected attribute $B_i = b$. In all these cases, $\Delta_j(0)$ is not statistically different from zero and is small in size. $\Delta_j(1)$ is instead clearly positive and significantly larger than $\Delta_j(0)$ in all strata. This evidence suggests that the information provided by test scores considerably increases the chances that immigrant students are upgraded. The interpretation of this finding is that test scores induce a larger revision of the teachers' posterior beliefs regarding the ability of immigrant students compared to those of native students. Note that this outcome is good news for immigrant students in the H and AH strata but not for immigrant students in the AL stratum, for whom this positive discrimination causes track changes.

Figure 5 reports the analogous quantities considering low socioeconomic status (SES) as the protected attribute, where low SES is defined as having a household income below the median. For the academic tracks, none of the $\Delta_j(z)$ is significantly different from 0. In vocational tracks, however, the $\Delta_j(0)$ are negative and the $\Delta_j(1)$ are positive and economically significant, albeit smaller in size than in the case of immigrant students (the scale of the vertical axis changes between Figures 4 and 5). The fact that all the $\Delta_j(0)$ are negative suggests a higher rate of non-compliance below the cutoff among high-SES students. A plausible explanation is that parents of high-SES students may be more likely to pressure teachers to

upgrade their children when they are provisionally recommended for vocational tracks. On the other hand, the positive $\Delta_j(1)$ values indicate that the provision of test scores induces upgrades for low-SES students more often than for high-SES ones. We view this finding as a consequence of test scores inducing a larger shift in teachers' posterior beliefs for low-SES students.

Finally, the analogous evidence for gender is shown in Figure A1 of the Online Appendix to Section 6. We do not find any $\Delta_j(z)$ significantly different from 0, meaning that teacher recommendations are fair when the protected attribute is gender.

7 Conclusions

Using a quasi-experimental setting offered by Dutch educational institutions, we show that when track recommendations given to students at the end of primary education can be upgraded based on standardized test scores, the quality of advice improves by at least 6% as measured by the fraction of students that are recommended a more challenging track they can complete successfully, while in the counterfactual case, they would have remained in a lower level track. An improvement of about the same size is also estimated for students who should be recommended to more challenging tracks, as they are always able to complete them independently of where they start.

We also find, however, that the possibility of upgrading based on test scores results in a significant number of students who are indeed upgraded but are unable to complete a more challenging track. In the case of these students, the goal should be to direct them immediately toward the low track that they can complete successfully, thus minimizing costly track changes. However, an overall evaluation that combines these opposite findings in a weighted objective function indicates that the relative weight of the short-term losses suffered by students who must change track because of misplacement when teachers are given the opportunity to upgrade based on test scores would have to be unreasonably high to conclude that it is better if teachers do not see these scores.

The implications for policy of these results are particularly relevant in countries where tracking is the established way to organize secondary school studies. It is surprising that no

experimental evidence exists to guide educational policymakers of these countries in designing the best system of track advising. We fill in this gap in several ways.

First, we define a reasonable metric to evaluate the quality of teachers' recommendations. This metric values two objectives: reducing costly track changes and directing towards more challenging tracks students who are able to complete them, but who need to be "pushed" towards these tracks and be convinced that they can complete them successfully. Differently from the case of a weather forecast, predicting the school track in which a student would have the best performance and recommending this track may have an effect on the object of the prediction itself, i.e., the choice of the student and her performance in the chosen track, as well as later in life. In other words, track recommendations may be self-fulfilling prophecies, and it is crucial to consider this feature in the optimal design of their implementation. The metric we propose puts this feature at the center stage.

Second, our results show that allowing teachers to upgrade their recommendations based on standardized test scores helps them to improve the quality of track advising. Ideally, one would like to compare recommendations given by teachers with and without the information provided by test scores, which is not what the quasi-experiment we study can offer. However, our evidence suggests that test scores do help teachers provide better recommendations in general, not just to upgrade previous ones. Therefore, appropriate controlled experiments should be implemented to further study and formally test the more general hypothesis.

Third, a large literature (see footnote 1) has demonstrated, in different contexts, that teachers' recommendations do not typically reflect only the previous academic achievement and the future ability potential of a student, but are also highly correlated with characteristics like SES, gender, race, and behavior in class. The bias affecting recommendations that are determined by the conscious or unconscious use of these variables by teachers is typically deemed unacceptable and has reinforced the opposition to school tracking in general. Our evidence suggests that test score information also has an impact on the fairness of recommendations with respect to protected attributes. We find that the provision of test scores induces more frequent upgrades of recommendations for immigrant and low SES students. We interpret these results as indicating that the provision of test scores information induces a larger shift in teachers' posterior beliefs regarding these students. Together with the fact

that high-SES students are more likely to be upgraded when originally recommended a vocational track and scoring below the cutoff, we conclude that test scores are a valuable tool for increasing the fairness of track recommendations.

Independent of track-advising, but no less importantly, our results contribute to the recent and growing literature on algorithm-assisted human decisions (see, for example, Imai and Jiang, 2022, Imai et al., 2023, Ben-Michael et al., 2024, and Rambachan, 2024). It is increasingly common for decision-makers in various fields (justice, medicine, finance, politics, and education, to name a few) to use big data processed by multiple algorithms to make high-stakes decisions in highly uncertain contexts. However, it is still unclear how to make the best use of these algorithmic and data-driven tools in human decision-making. Our results contribute to this literature by providing novel quasi-experimental evidence and by improving existing methods to assess algorithm-assisted human decisions.

References

Alesina, A., M. Carlana, E. La Ferrara, and P. Pinotti (2024). Revealing stereotypes: Evidence from immigrants in schools. *American Economic Review 114*(7), 1916–1948.

Angrist, D. J. and M. Rokkanen (2015). Wanna get away? regression discontinuity estimation of exam school effects away from the cutoff. *Journal of the American Statistical Association 110*(512), 1331–1344.

Angrist, J. D., G. Imbens, and D. Rubin (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association 91*(434), 444–455.

Bach, M. (2023). Heterogeneous responses to school track choice: Evidence from the repeal of binding track recommendations. *Economics of Education Review 95*, 102412.

Batruch, A., S. Geven, E. Kessenich, and H. 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*, 1–18.

Ben-Michael, E., D. J. Greiner, M. Huang, K. Imai, Z. Jiang, and S. Shin (2024). Does ai help humans make better decisions? a methodological framework for experimental evaluation. arXiv preprint.

Ben-Michael, E., K. Imai, and J. Zhichao (2024). Policy learning with asymmetric counterfactual utilities. *Journal of the American Statistical Association 0*(0), 1–14.

Branson, Z. and F. Mealli (2019). The local randomization framework for regression discontinuity designs: A review and some extensions. *arXiv preprint arXiv:1810.02761*.

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. and M. Fort (2022). Hacking gender stereotypes: Girls' participation in coding clubs. *AEA Papers and Proceedings* (112).

Carlana, M., E. La Ferrara, and P. Pinotti (2022a). Goals and gaps: Educational careers of immigrant children. *Econometrica 90*(1), 1–29.

Carlana, M., E. La Ferrara, and P. Pinotti (2022b). Implicit stereotypes in teachers' track recommendations. *AEA Papers and Proceedings* (112), 409–414.

Cattaneo, M., B. R. Frandsen, and R. Titiunik (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the u.s. senate. *Journal of Causal Inference 3*(1), 1–24.

Cattaneo, M. D., N. Idrobo, and R. Titiunik (2020a). *A Practical Introduction to Regression Discontinuity Designs: Extensions*. Cambridge University Press.

Cattaneo, M. D., N. Idrobo, and R. Titiunik (2020b). *A Practical Introduction to Regression Discontinuity Designs: Foundations*. Cambridge University Press.

de Ree, J., M. Oosterveen, and D. Webbink (2025). The quality of school track assignment decisions by teachers. *https://arxiv.org/abs/2304.10636v2*.

Ding, P. and J. Lu (2017). Principal stratification analysis using principal scores. *Journal of the Royal Statistical Society Series B: Statistical Methodology 79*(3), 757–777.

Driessen, G., P. Sleegers, and F. Smit (2008). The transition from primary to secondary education: Meritocracy and ethnicity. *European Sociological Review 24*(4), 527–542.

Eckles, D., N. Ignatiadis, S. Wager, and H. Wu (2025). Noise-induced randomization in regression discontinuity designs. *Biometrika 112*(2).

Falk, A., F. Kosse, and P. Pinger (2025). Mentoring and schooling decisions: Causal evidence. *Journal of Political Economy*. Forthcoming.

Feller, A., F. Mealli, and L. Miratrix (2017). Principal score methods: Assumptions, extensions, and practical considerations. *Journal of Educational and Behavioral Statistics 42*(6), 726–758.

Ferman, B. and L. F. Fontes (2022). Assessing knowledge or classroom behavior? evidence of teachers' grading bias. *Journal of Public Economics 216*, 104773.

Forastiere, L., A. Mattei, J. M. Pescarini, M. L. Barreto, and F. Mealli (2025). Selecting subpopulations for causal inference in regression discontinuity designs. *Annals of Applied Statistics*. Forthcoming.

Frangakis, C. E. and D. B. Rubin (2002). Principal stratification in causal inference. *Biometrics 58*(1), 21–29.

Gerhenson, S., S. B. Holt, and N. Papageorge (2016). Who believes in me? the effect of student-teacher demographic match on teacher expectations. *Economics of Education Review* (52), 209–224.

Geven, S., Øyvind N. Wiborg, R. E. Fish, and H. G. van de Werfhorst (2021). How teachers form educational expectations for students: A comparative factorial survey experiment in three institutional contexts. *Social Science Research 100*, 102–599.

Ichino, A., A. Rustichini, and A. Zanella (2025). College, cognitive ability, and socioeconomic disadvantage: policy lessons from the uk in 1960-2004. *CEPR Discussion Paper 17284*.

Imai, K. and Z. Jiang (2022). Principal fairness for human and algorithmic decision-making. *Statistical Science*.

Imai, K., Z. Jiang, J. Greiner, R. Halen, and S. Shin (2023). Experimental evaluation of algorithm-assisted human decision-making: Application to pretrial public safety assessment. *Journal of the Royal Statistical Society*.

Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review 93*(2), 126–132.

Imbens, G. W. and J. D. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica 62*(2), 467–475.

Imbens, W. and D. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences. An Introduction*. New York, NY, USA: Cambridge University Press.

Ketel, N., H. Oosterbeek, S. Sóvágó, and B. van der Klaauw (2023). The (un)importance of school assignment. *Tinbergen Institute Discussion Paper TI 2023-076/V*.

Li, F., A. Mattei, and F. Mealli (2015). Bayesian inference for regression discontinuity designs with application to the evaluation of italian university grants. *The Annals of Applied Statistics 9*(4), 1906–1931.

Li, F., A. Mercatanti, T. Mäkinein, and A. Silvestrini (2020). A regression discontinuity design for ordinal running variable: Evaluating central bank purchases of corporate bonds. *Annals of Applied Statistics*.

Mattei, A., L. Forastiere, and F. Mealli (2023). Assessing principal causal effects using principal score methods. In *Handbook of Matching and Weighting Adjustments for Causal Inference*, pp. 313–348. Chapman and Hall/CRC.

Mattei, A. and F. Mealli (2017). Regression discontinuity designs as local randomized experiments. *Observational Studies 3*(2), 156–173.

Mealli, F. and A. Mattei (2012). A refreshing account of principal stratification. *The International Journal of Biostatistics 8*(1).

Mealli, F. and B. Pacini (2013). Using secondary outcomes to sharpen inference in randomized experiments with noncompliance. *Journal of the American Statistical Association 108*(503), 1120–1131.

Osikominu, A., G. Pfeifer, and K. Strohmaier (2021). The effects of free secondary school track choice: A disaggregated synthetic control approach. *IZA Discussion Paper Series*.

Rambachan, A. (2024). Identifying prediction mistakes in observational data. *The Quarterly Journal of Economics 139*(3), 1665–1711.

Rubin, D. B. (1980). Discussion of "randomization analysis of experimental data in the fisher randomization test" by basu. *Journal of the American Statistical Association 75*, 591–593.

van Huizen, T., M. Jacobs, and M. Oosterveen (2024). Teacher bias or measurement error? *arXiv preprint arXiv:2401.04200*.

van Leest, A., L. Hornstra, J. van Tartwijk, and J. van de Pol (2021). Test- or judgement based school track recommendations: Equal opportunities for students with different socio-economic backgrounds? *British Journal of Education Psychology 91*, 193–216.

Van Po naar Vo (2025). Veelgestelde vragen. Accessed: 2025-04-07.

Online Appendix

Do Test Scores Help Teachers Give Better Track Advice to Students?
A Principal Stratification Analysis

Andrea Ichino, Fabrizia Mealli, Javier Viviens

February 13, 2026

Appendix to Section 3.1

Table A1: Leniency of teachers and frequency of Never Takers

Track	ξ 1	ξ 2	ξ 3	ξ 4
V: BL	.014	-.0187	.1829	.1482
	(.0535)	(.0527)	(.0702)	(.0734)
V: BL/KL	.1297	.0148	.0582	-.0281
	(.1073)	(.1366)	(.1373)	(.1531)
V: KL	.3561	.2637	.3123	.2355
	(.0846)	(.0866)	(.0763)	(.0776)
V: KL/GT	.1322	.172	.0916	-.0134
	(.121)	(.1195)	(.1359)	(.1507)
V: GT	.1875	.1448	.1712	.1534
	(.0341)	(.0338)	(.0397)	(.0405)
V: GT/HAVO	.2028	.1308	.3215	.2633
	(.096)	(.0981)	(.0854)	(.0898)
A: HAVO	.142	.097	.1365	.0997
	(.0348)	(.0317)	(.0367)	(.0387)
A: HAVO/VWO	.3135	.2067	.1839	.1142
	(.0694)	(.0764)	(.072)	(.0794)
COVARIATES X_i	NO	YES	NO	YES
COHORT	2015	2015	2016	2016

Notes: the table reports, separately for each track and cohort, estimates of the correlation of the proxy for school leniency $\tilde{Z}$ with the probability of being a Never-Taker. Specifically, it reports the coefficient ξ from running the following regression using units with $Z = 1$: $1 - R_i = \gamma + \xi \tilde{Z}_i + \mu X_i + \epsilon_i$. The row headings report the names of the secondary school tracks to which a student can be assigned in the Dutch system, ranked by level of difficulty. The covariates X included in the second and fourth columns are those described in Table 2.

Appendix to Section 3.4

Proof of Theorem 1 of Imai et al. (2023) in our context.

By Assumption 4, we have that:

$$Pr(Y(0)=0, Y(1)=0) = Pr(Y(1)=0) \tag{A1}$$

and

$$Pr(Y(0)=1, Y(1)=1) = Pr(Y(0)=1). \tag{A2}$$

Then, we can write:

$$\begin{aligned} &Pr(R(z)=1, Y(0)=0, Y(1)=0) \\ &\qquad = Pr(R(z)=1, Y(1)=0) \\ &\qquad = Pr(R(z)=1, Y(R(z))=0 \mid Z=z) \\ &\qquad = Pr(R=1, Y=0 \mid Z=z), \end{aligned} \tag{A3}$$

where the second equality comes from Assumption 2. Equivalently, we can write

$$\begin{aligned} &Pr(R(z)=1, Y(0)=1, Y(1)=1) \\ &\qquad = Pr(Y(0)=Y(1)=1) - Pr(R(z)=0, Y(0)=Y(1)=1) \\ &\qquad = Pr(Y(0)=1) - Pr(R(z)=0, Y(0)=1) \\ &\qquad = Pr(Y(0)=1) - Pr(R(z)=0, Y(R(z))=1 \mid Z=z) \\ &\qquad = Pr(Y(0)=1) - Pr(R=0, Y=1 \mid Z=z) \end{aligned} \tag{A4}$$

and

$$\begin{aligned} &Pr(R(z)=1, Y(0)=0, Y(1)=1) = \\ &\qquad Pr(R(z)=1) - Pr(R(z)=1, Y(0)=Y(1)=1) \\ &\qquad - Pr(R(z)=1, Y(0)=Y(1)=0) \\ &\qquad - \underbrace{Pr(R(z)=1, Y(0)=1, Y(1)=0)}_{=0} \end{aligned} \tag{A5}$$

Using equations (A3) and (A4):

$$
\begin{aligned}
&Pr(R(z)=1, Y(0)=0, Y(1)=1) \\
&\qquad = Pr(R(z)=1) - \underbrace{Pr(Y(0)=1) + Pr(R=0, Y=1 \mid Z=z)}_{\text{from equation (A4)}} \\
&\qquad - \underbrace{Pr(R=1, Y=0 \mid Z=z)}_{\text{from equation (A3)}} \\
&\qquad = Pr(R=1 \mid Z=z) - Pr(R=1, Y=0 \mid Z=z) \\
&\qquad + Pr(R=0, Y=1 \mid Z=z) - Pr(Y(0)=1) \\
&\qquad = Pr(R=1, Y=1 \mid Z=z) + Pr(R=0, Y=1 \mid Z=z) \\
&\qquad - Pr(Y(0)=1) \\
&\qquad = Pr(Y=1 \mid Z=z) - Pr(Y(0)=1)
\end{aligned}
\tag{A6}
$$

$$
\begin{aligned}
Pr(Y(0)=0, Y(1)=1) &= Pr(Y(1)=1) - Pr(Y(0)=1, Y(1)=1) = \\
&= Pr(Y(1)=1) - Pr(Y(0)=1)
\end{aligned}
\tag{A7}
$$

Average Principal Causal Effects (APCEs, equations 4, 11 and 8)

$$
\begin{aligned}
APCE_H &= E[R(1) - R(0) \mid \text{student is in H}] \\
&= E[R(1) - R(0) \mid Y(0)=0, Y(1)=1] \\
&= \frac{Pr(R(1)=1, Y(0)=0, Y(1)=1) - Pr(R(0)=1, Y(0)=0, Y(1)=1)}{Pr(Y(0)=0, Y(1)=1)}
\end{aligned}
$$

Using equations (A6) and (A7):

$$
\frac{\overbrace{Pr(Y=1|Z=1)-Pr(Y(0)=1)}^{Pr(Y=1|Z=1)-Pr(Y(0)=1)} - \overbrace{Pr(R(0)=1, Y(0)=0, Y(1)=1)}^{Pr(Y=1|Z=0)-Pr(Y(0)=1)}}{\underbrace{Pr(Y(0)=0, Y(1)=1)}_{Pr(Y(1)=1)-Pr(Y(0)=1)}}
$$

$$
\begin{aligned}
APCE_H &= \frac{Pr(Y=1 \mid Z=1) - Pr(Y=1 \mid Z=0)}{Pr(Y(1)=1) - Pr(Y(0)=1)} = \\
&= \frac{E[Y \mid Z=1] - E[Y \mid Z=0]}{Pr(Y(1)=1) - Pr(Y(0)=1)}
\end{aligned}
$$

$$
\begin{aligned}
APCE_{AL} &= E[R(1) - R(0) \mid \text{student is in AL}] = \\
&= E[R(1) - R(0) \mid Y(0) = 0, Y(1) = 0] = \\
&= \frac{Pr(R(1) = 1, Y(0) = 0, Y(1) = 0) - Pr(R(0) = 1, Y(0) = 0, Y(1) = 0)}{Pr(Y(0) = 0, Y(1) = 0)}
\end{aligned}
$$

From equation (A3):

$$
= \frac{Pr(R = 1, Y = 0 \mid Z = 1) - Pr(R = 1, Y = 0 \mid Z = 0)}{Pr(Y(0) = 0, Y(1) = 0)}
$$

From equation (A1):

$$
\begin{aligned}
&= \frac{Pr(R = 1, Y = 0 \mid Z = 1) - Pr(R = 1, Y = 0 \mid Z = 0)}{Pr(Y(1) = 0)} \\
&= \frac{Pr(R = 1, Y = 0 \mid Z = 1) - Pr(R = 1, Y = 0 \mid Z = 0)}{1 - Pr(Y(1) = 1)}
\end{aligned}
$$

$$
\begin{aligned}
APCE_{AH} &= E[R(1) - R(0) \mid \text{student is in AH}] = \\
&= E[R(1) - R(0) \mid Y(0) = 1, Y(1) = 1] = \\
&= \frac{Pr(R(1) = 1, Y(0) = 1, Y(1) = 1) - Pr(R(0) = 1, Y(0) = 1, Y(1) = 1)}{Pr(Y(0) = 1, Y(1) = 1)}
\end{aligned}
$$

From equation (A4):

$$
= \frac{Pr(Y(0) = 1) - Pr(R = 0, Y = 1 \mid Z = 1) - Pr(Y(0) = 1) + Pr(R = 0, Y = 1 \mid Z = 0)}{Pr(Y(0) = 1, Y(1) = 1)}
$$

From equation (A2):

$$
= \frac{Pr(R = 0, Y = 1 \mid Z = 0) - Pr(R = 0, Y = 1 \mid Z = 1)}{Pr(Y(0) = 1)}
$$

Derivation of the bounds of the $APCE_J$; equation (5)

By assumption 2:

$$
Pr(Y(R) = 1) = Pr(Y(R) = 1 \mid Z = z) \tag{A8}
$$

By the law of total probability:

$$
\begin{aligned}
Pr(Y(R) = 1 \mid Z = z) = & \\
&= Pr(Y = 1 \mid R = R, Z = z) Pr(R = R \mid Z = z) \\
&\quad + Pr(Y(R) = 1 \mid R = 1 - R, Z = z) Pr(R = 1 - R \mid Z = z)
\end{aligned} \tag{A9}
$$

We can rewrite it as:

$$\begin{aligned} Pr(Y(R)=1 \mid Z=z) = \\ &= Pr(Y=1, R=R \mid Z=z) \\ &\quad + \underbrace{Pr(Y(R)=1 \mid R=1-R, Z=z) Pr(R=1-R \mid Z=z)}_{\geq 0} \end{aligned} \tag{A10}$$

By Assumption 4:

$$Pr(Y=1 \mid R=0, Z=z) \leq Pr(Y(1)=1 \mid R=0, Z=z) \leq 1 \tag{A11}$$

and

$$0 \leq Pr(Y(0)=1 \mid R=1, Z=z) \leq Pr(Y=1 \mid R=1, Z=z) \tag{A12}$$

From equation (A10)

$$Pr(Y(0)=1) \geq Pr(Y=1, R=0 \mid Z=z) \tag{A13}$$

and

$$\begin{aligned} Pr(Y(1)=1) = \\ &= Pr(Y=1, R=1 \mid Z=z) \\ &\quad + \underbrace{Pr(Y(1)=1 \mid R=0, Z=z)}_{\leq 1} Pr(R=0 \mid Z=z). \end{aligned} \tag{A14}$$

So that we can write

$$\begin{aligned} Pr(Y(1)=1) \leq & Pr(Y=1, R=1 \mid Z=z) + Pr(R=0 \mid Z=z) \\ = & 1 - Pr(Y=0, R=1 \mid Z=z) \end{aligned} \tag{A15}$$

Then, using equations (A11), (A12), (A13) and (A15), the bounds for $Pr(Y(R)=1)$ are:

$$\begin{aligned} \max_z Pr(Y=1, R=0 \mid Z=z) \leq Pr(Y(0)=1) \leq \min_z Pr(Y=1 \mid Z=z) \\ \max_z Pr(Y=1 \mid Z=z) \leq Pr(Y(1)=1) \leq 1 - \max_z Pr(Y=0, R=1 \mid Z=z) \end{aligned} \tag{A16}$$

Upper and Lower Bounds of the $APCE_J$; equations (6), (7), (9), (10) (12), and (13)

Lower bound of $APCE_H$:

$$\frac{E[Y \mid Z=1] - E[Y \mid Z=0]}{1 - \max_z Pr(Y=0, R=1 \mid Z=z) - \max_z Pr(Y=1, R=0 \mid Z=z)}$$

Upper bound of $APCE_H$:

$$\frac{E[Y \mid Z=1] - E[Y \mid Z=0]}{\max_z Pr(Y=1 \mid Z=z) - \min_z Pr(Y=1 \mid Z=z)}$$

Lower bound of $APCE_{AH}$:

$$\frac{Pr(R=0, Y=1 \mid Z=0) - Pr(R=0, Y=1 \mid Z=1)}{\min_z Pr(Y=1 \mid Z=z)}$$

Upper bound of $APCE_{AH}$:

$$\frac{Pr(R=0, Y=1 \mid Z=0) - Pr(R=0, Y=1 \mid Z=1)}{\max_z Pr(R=0, Y=1 \mid Z=z)}$$

Lower bound of $APCE_{AL}$:

$$\frac{Pr(R=1, Y=0 \mid Z=1) - Pr(R=1, Y=0 \mid Z=0)}{1 - \max_z Pr(Y \mid Z=z)}$$

Upper bound of $APCE_{AL}$:

$$\frac{Pr(R=1, Y=0 \mid Z=1) - Pr(R=1, Y=0 \mid Z=0)}{\max_z Pr(R=1, Y=0 \mid Z=z)}$$

Identification and Estimation of $APCE_J$ and Principal Fairness under Unconfoundedness

Because the goal of fairness analysis is really to understand and compare the size of the Δ_j's, it is crucial to be able to point-identify and estimate these quantities for fairness comparisons. A natural and useful starting point, that would be more or less plausible depending on the information cointained in the baseline covariates, is assuming Assumption 6 Unconfoundedness.

Consider the following principal scores, which represents the population proportions for each of the strata in each cell defined by the covariates:

$$\begin{aligned} e_{AL} &= Pr(Y(1)=0, Y(0)=0 \mid X=x) \\ e_{AH} &= Pr(Y(1)=1, Y(0)=1 \mid X=x) \\ e_H &= Pr(Y(1)=1, Y(0)=0 \mid X=x). \end{aligned}$$

Under assumptions 1, 2, 3, 4 and 6, we can identify these principal strata proportions as:

$$\begin{aligned} e_{AL} =& Pr(Y=0 \mid R=1, X=x) \\ e_{AH} =& Pr(Y=1 \mid R=0, X=x) \\ e_H =& Pr(Y=1 \mid R=1, X=x) - Pr(Y=1 \mid R=0, X=x) \end{aligned}$$

Theorem 10 *Point identification of APCEs under unconfoundedness:*
Under assumptions 1, 2, 3, 4, and 6, $APCE_{AL}$, $APCE_{AH}$ and $APCE_{H}$ are identified as follows:

$$\begin{aligned} APCE_{AL} =& \mathbb{E}[w_{AL}(X)R \mid Z=1] - \mathbb{E}[w_{AL}(X)R \mid Z=0] \\ APCE_{AH} =& \mathbb{E}[w_{AH}(X)R \mid Z=1] - \mathbb{E}[w_{AH}(X)R \mid Z=0] \\ APCE_{H} =& \mathbb{E}[w_{H}(X)R \mid Z=1] - \mathbb{E}[w_{H}(X)R \mid Z=0], \end{aligned}$$

where

$$w_{AL} = \frac{e_{AL}(X)}{\mathbb{E}[e_{AL}(X)]}, \quad w_{AH} = \frac{e_{AH}(X)}{\mathbb{E}[e_{AH}(X)]}, \quad w_{H} = \frac{e_{H}(X)}{\mathbb{E}[e_{H}(X)]}$$

Proof: See Imai et al. (2023).

Similarly, consider the following principal scores for a specific value of a protected attribute $B = b$ with $b \in \{0, 1\}$ (B is in general one of the covariates in X):

$$\begin{aligned} e^{w}_{AL} &= Pr(Y(1) = 0, Y(0) = 0 \mid X = x, B = b) \\ e^{w}_{AH} &= Pr(Y(1) = 1, Y(0) = 1 \mid X = x, B = b) \\ e^{w}_{H} &= Pr(Y(1) = 1, Y(0) = 0 \mid X = x, B = b). \end{aligned}$$

which can be still identified under Assumption 1, 2, 3, 4 and 7.

Theorem 11 *Point identification of fairness under unconfoundedness:*
Under assumptions 1, 2, 3, 4, and 6, $\Delta_{AL}(z)$, $\Delta_{AH}(z)$ and $\Delta_{H}(z)$ are identified as follows:

$$\begin{aligned} \Delta_{AL}(z) =& \mathbb{E}[w^{1}_{AL}(X)R \mid Z=z, B=1] - \mathbb{E}[w^{0}_{AL}(X)R \mid Z=z, B=0] \\ \Delta_{AH}(z) =& \mathbb{E}[w^{1}_{AL}(X)R \mid Z=z, B=1] - \mathbb{E}[w^{0}_{AH}(X)R \mid Z=z, B=0] \\ \Delta_{H}(z) =& \mathbb{E}[w^{1}_{H}(X)R \mid Z=z, B=1] - \mathbb{E}[w^{0}_{H}(X)R \mid Z=z, B=0] \end{aligned}$$

,

where

$$w^{w}_{AL} = \frac{e^{w}_{AL}(X)}{\mathbb{E}[e^{w}_{AL}(X)]}, \quad w^{w}_{AH} = \frac{e^{w}_{AH}(X)}{\mathbb{E}[e^{w}_{AH}(X)]}, \quad w^{w}_{H} = \frac{e^{w}_{H}(X)}{\mathbb{E}[e^{w}_{H}(X)]}$$

Proof: See Imai et al. (2023)

Appendix to Section 4.1

Estimates obtained without covariates

Table A2: Estimates of the $APCE_H$, no covariates

Track	Numerator $APCE_H$	Lower bound denominator $APCE_H$	Upper bound denominator $APCE_H$	Lower bound $APCE_H$	Upper bound $APCE_H$
	1	2	3	4	5
V: BL	.0238	.0238	.651	.036	1
	(.0171)	(.0162)	(.0137)	(.027)	-
V: BL/KL	.0737	.0737	.2872	.2643	1
	(.0298)	(.0288)	(.0226)	(.1019)	-
V: KL	.1049	.1049	.7016	.1496	1
	(.0159)	(.0159)	(.0105)	(.0229)	-
V: KL/GT	.0221	.0221	.3946	.0566	1
	(.0229)	(.0207)	(.0221)	(.0553)	-
V: GT	.0216	.0216	.7624	.0284	1
	(.0089)	(.0087)	(.0075)	(.0118)	-
V: GT/HAVO	.0554	.0554	.5051	.1072	1
	(.017)	(.0165)	(.0146)	(.0323)	-
A: HAVO	.0376	.0376	.8068	.0466	1
	(.0079)	(.0079)	(.0067)	(.0101)	-
A: HAVO/VWO	.062	.062	.5431	.1124	1
	(.016)	(.0156)	(.013)	(.0286)	-
VOCATIONAL	.0469	.0469	.6489	.0723	1
	(.0063)	(.0061)	(.005)	(.0097)	-
ACADEMIC	.0453	.0453	.7236	.0626	1
	(.0076)	(.0075)	(.006)	(.0106)	-
ALL	.0463	.0463	.6792	.0681	1
	(.005)	(.0049)	(.0039)	(.0073)	-

Notes: The row headings reports the names of the secondary school tracks to which a student can be assigned in the Dutch system, ranked by level of difficulty. The last three rows of the table aggregate the Vocational (VMBO*), the Academic (HAVO*), and All tracks, respectively. With reference to the $APCE_H$ estimand defined in equation (4), the table reports for the outcome Y and separately for each track, estimates of the numerator (in column 1), of the lower and upper bounds of the denominator (in columns 2 and 3), and of the lower and upper bounds of the $APCE_H$ (in columns 4 and 5). The definitions of these two last statistics are in equations (6) and (7), respectively. The aggregations for the three last rows are performed as follows: point estimates of the numerator and of the lower and upper bounds of the denominator are created by weighing each track in the corresponding aggregate by the relative population in each track. For the lower and upper bounds of the $APCE_H$, the aggregated estimates for the lower and upper bounds of the denominators are used directly. Standard errors for each quantity are computed using a block bootstrap at the school level with 1000 repetitions; each quantity in every iteration is calculated in the same way.

Table A3: Estimates of the $APCE_{AH}$, no covariates

Track	Numerator $APCE_{AH}$	Lower bound denominator $APCE_{AH}$	Upper bound denominator $APCE_{AH}$	Lower bound $APCE_{AH}$	Upper bound $APCE_{AH}$
	1	2	3	4	5
V: BL	.0033	.3188	.3208	.0097	.0098
	(.0129)	(.0115)	(.0116)	(.0384)	(.0384)
V: BL/KL	.0331	.6843	.7032	.0469	.0482
	(.0275)	(.0208)	(.0211)	(.0378)	(.0383)
V: KL	.0035	.2256	.2367	.0161	.0166
	(.0091)	(.0086)	(.0092)	(.0364)	(.0379)
V: KL/GT	.0792	.5698	.5813	.1369	.1399
	(.0283)	(.0192)	(.0186)	(.0478)	(.0477)
V: GT	.0049	.1923	.193	.0249	.0251
	(.0054)	(.0057)	(.0057)	(.0272)	(.0272)
V: GT/HAVO	.0433	.4178	.4269	.0997	.1014
	(.017)	(.0113)	(.0115)	(.0385)	(.0389)
A: HAVO	0	.1283	.1283	0	0
	(.0004)	(.004)	(.004)	(.0028)	(.0029)
A: HAVO/VWO	.0293	.3684	.3798	.0759	.0783
	(.0146)	(.0108)	(.0113)	(.0368)	(.0375)
VOCATIONAL	.0169	.298	.3038	.0558	.0569
	(.005)	(.004)	(.0041)	(.016)	(.0162)
ACADEMIC	.0092	.2041	.2077	.0445	.0453
	(.0046)	(.0043)	(.0044)	(.0218)	(.0221)
ALL	.0138	.2599	.2648	.0522	.0532
	(.0035)	(.0029)	(.003)	(.013)	(.0132)

Notes: The row headings report the names of the secondary school tracks to which a student can be assigned in the Dutch system, ranked by level of difficulty. The last three rows of the table aggregate the Vocational (VMBO*), the Academic (HAVO*), and All tracks, respectively. With reference to the $APCE_{AH}$ estimand defined in equation (8), the table reports for the outcome Y and separately for each track, estimates of the numerator (in column 1), of the lower and upper bounds of the denominator (in columns 2 and 3), and of the lower and upper bounds of the $APCE_{AH}$ (in columns 4 and 5). The definitions of these two last statistics are in equations (9) and (10), respectively. The aggregations for the last three rows are performed as follows: point estimates of the numerator and of the lower and upper bounds of the denominator are created by weighing each track in the corresponding aggregate by the relative population in each track. For the lower and upper bounds of the $APCE_{AH}$, the aggregated estimates for the lower and upper bounds of the denominators are used directly. Standard errors for each quantity are computed using a block bootstrap at the school level with 1000 repetitions; each quantity in every iteration is calculated in the same way.

Table A4: Estimates of the $APCE_{AL}$, no covariates

Track	Numerator $APCE_{AL}$	Lower bound denominator $APCE_{AL}$	Upper bound denominator $APCE_{AL}$	Lower bound $APCE_{AL}$	Upper bound $APCE_{AL}$
	1	2	3	4	5
V: BL	.0267	.0277	.6579	.0403	.9734
	(.0051)	(.005)	(.0137)	(.0078)	(.0319)
V: BL/KL	.0261	.0285	.223	.1212	.9209
	(.0087)	(.0084)	(.0202)	(.0417)	(.1156)
V: KL	.0663	.0728	.6584	.1005	.91
	(.0071)	(.0068)	(.0128)	(.0105)	(.0276)
V: KL/GT	.0341	.0356	.3966	.0861	.9603
	(.0084)	(.0082)	(.0185)	(.0213)	(.0536)
V: GT	.0376	.0411	.7896	.0477	.9174
	(.0038)	(.0037)	(.0069)	(.0048)	(.0261)
V: GT/HAVO	.0677	.0771	.5177	.131	.8774
	(.0086)	(.0082)	(.0136)	(.0161)	(.0367)
A: HAVO	.0486	.054	.8449	.0575	.8994
	(.0041)	(.0039)	(.0061)	(.0048)	(.027)
A: HAVO/VWO	.0792	.0886	.5583	.1416	.8943
	(.0077)	(.0075)	(.0123)	(.0135)	(.0272)
VOCATIONAL	.0466	.0512	.6512	.0715	.9097
	(.0027)	(.0026)	(.0048)	(.0041)	(.0149)
ACADEMIC	.0583	.0649	.7544	.0772	.8973
	(.0037)	(.0036)	(.0057)	(.0049)	(.0188)
ALL	.0513	.0568	.6931	.074	.9039
	(.0022)	(.0021)	(.0037)	(.0031)	(.0118)

Notes: The row headings report the names of the secondary school tracks to which a student can be assigned in the Dutch system, ranked by level of difficulty. The last three rows of the table aggregate the Vocational (VMBO*), the Academic (HAVO*), and All tracks, respectively. With reference to the $APCE_{AL}$ estimand defined in equation (11), the table reports for the outcome Y and separately for each track, estimates of the numerator (in column 1), of the lower and upper bounds of the denominator (in columns 2 and 3), and of the lower and upper bounds of the $APCE_{AL}$ (in columns 4 and 5). The definitions of these two last statistics are in equations (12) and (13), respectively. The aggregations for the last three rows are performed as follows: point estimates of the numerator and of the lower and upper bounds of the denominator are created by weighing each track in the corresponding aggregate by the relative population in each track. For the lower and upper bounds of the $APCE_{AL}$, the aggregated estimates for the lower and upper bounds of the denominators are used directly. Standard errors for each quantity are computed using a block bootstrap at the school level with 1000 repetitions; each quantity in every iteration is calculated in the same way.

Appendix to Section 4.2

Derivation of the Loss Function

We can estimate

$$\Pi_{11}(z) = Pr(R(z) = 1; Y(1) = 1)$$

and

$$\Pi_{10}(z) = Pr(R(z) = 1; Y(1) = 0),$$

but we cannot identify

$$\Pi_{01}(z) = Pr(R(z) = 0; Y(1) = 1)$$

and

$$\Pi_{00}(z) = Pr(R(z) = 0; Y(1) = 0).$$

From Assumption 2, it follows that the distribution of potential outcomes is the same at both sides of the cutoff:

$$Pr(Y(1) = 1 \mid Z = 1) = Pr(Y(1) = 1 \mid Z = 0)$$

We can rewrite it as:

$$\sum_r Pr(R = r, Y(1) = 1 \mid Z = 1) = \sum_r Pr(R = r, Y(1) = 1 \mid Z = 0)$$

$$\Pi_{11}(1) + \Pi_{01}(1) = \Pi_{11}(0) + \Pi_{01}(0)$$

$$\Pi_{01}(1) - \Pi_{01}(0) = \Pi_{11}(0) - \Pi_{11}(1)$$

From equation (14) we had that:

$$\mathcal{L}(l_{01}, z) = \Pi_{01}(z) + l_{10}\Pi_{10}(z),$$

Consequently,

$$\begin{aligned}\mathcal{L}(l_{01}, 1) - \mathcal{L}(l_{01}, 0) &= \Pi_{01}(1) - \Pi_{01}(0) + l_{10}\Pi_{10}(1) - l_{10}\Pi_{10}(0) \\ &= \Pi_{11}(0) - \Pi_{11}(1) - l_{10}(\Pi_{10}(1) - \Pi_{10}(0))\end{aligned}$$

Appendix to Section 5

Identification of APCEs Without Assuming Exclusion Restriction

Without the exclusion restriction, we have to index the potential outcomes using both Z and R, $Y(z,r)$. Nevertheless, under Assumptions 7 and 8, we can normalize $\mathbb{E}[Y(0,r)]$ to $\mathbb{E}[Y(r)]$ and $\mathbb{E}[Y(1,r)] = \mathbb{E}[Y(r)] + \eta$.

From equations (A3) and (A4), and under the exclusion restriction, we know that:

$$\begin{aligned} Pr(R(z)=1, Y(0)=0, Y(1)=0) &= Pr(R=1, Y=0 \mid Z=z) \\ Pr(R(z)=1, Y(0)=1, Y(1)=1) &= Pr(Y(0)=1) - Pr(R=0, Y=1 \mid Z=z) \end{aligned}$$

Replacing the exclusion restriction with assumptions 7 and 8, this becomes:

$$Pr(R(0)=1, Y(0)=0, Y(1)=0) = Pr(R=1, Y=0 \mid Z=0) \tag{A17}$$

$$Pr(R(1)=1, Y(0)=0, Y(1)=0) = Pr(R=1, Y=0 \mid Z=1) + \eta Pr(R=1 \mid Z=1) \tag{A18}$$

$$Pr(R(0)=1, Y(0)=1, Y(1)=1) = Pr(Y(0)=1) - Pr(R=0, Y=1 \mid Z=0) \tag{A19}$$

$$Pr(R(1)=1, Y(0)=1, Y(1)=1) = Pr(Y(0)=1) - Pr(R=0, Y=1 \mid Z=1) + \eta Pr(R=0 \mid Z=1) \tag{A20}$$

Similarly, from the proof in Appendix to Section 3.4, we known that

$$\begin{aligned} Pr(R(z)=1, Y(0)=0, Y(1)=1) & \\ &= Pr(R(z)=1) - Pr(R(z)=1, Y(0)=0, Y(1)=0) && \text{(A21)} \\ &- Pr(R(z)=1, Y(0)=1, Y(1)=1). && \text{(A22)} \end{aligned}$$

Therefore, using the results from (A17-A20), it follows that

$$\begin{aligned} Pr(R(0)=1, Y(0)=0, Y(1)=1) = & \\ &= Pr(R=1 \mid Z=0) - Pr(R=1, Y=0 \mid Z=0) \\ &- Pr(Y(0)=1) + Pr(R=0, Y=1 \mid Z=0) \\ &= Pr(Y=1 \mid Z=0) - Pr(Y(0)=1) && \text{(A23)} \end{aligned}$$

and

$$
\begin{aligned}
Pr(R(1)=1, Y(0)=0, Y(1)=1) = & \\
& = Pr(R=1 \mid Z=1) - Pr(R=1, Y=0 \mid Z=1) \qquad \text{(A24)} \\
& - \eta Pr(R=1 \mid Z=1) \\
& - Pr(Y(0)=1) + Pr(R=0, Y=1 \mid Z=1) - \eta Pr(R=0 \mid Z=1) \\
& = Pr(Y=1 \mid Z=0) - Pr(Y(0)=1) - \eta \qquad \text{(A25)}
\end{aligned}
$$

From equations (A22) and (A14), we have that:

$$
\begin{aligned}
APCE_H &= E[R(1) - R(0) \mid \text{student is in H}] \\
&= \frac{Pr(R(1)=1, Y(0)=0, Y(1)=1) - Pr(R(0)=1, Y(0)=0, Y(1)=1)}{Pr(Y(0)=0, Y(1)=1)} \\
&= \frac{E[Y \mid Z=1] - E[Y \mid Z=0] - \eta}{Pr(Y(1)=1) - Pr(Y(0)=1)}.
\end{aligned}
$$

Similarly, from equations (A17) - (A20),

$$
\begin{aligned}
APCE_{AL} &= E[R(1) - R(0) \mid \text{student is in AL}] = \\
&= \frac{Pr(R(1)=1, Y(0)=0, Y(1)=0) - Pr(R(0)=1, Y(0)=0, Y(1)=0)}{Pr(Y(0)=0, Y(1)=0)} \\
&= \frac{Pr(R=1, Y=0 \mid Z=1) + \eta Pr(R=1 \mid Z=1) - Pr(R=1, Y=0 \mid Z=0)}{1 - Pr(Y(1)=1)}
\end{aligned}
$$

$$
\begin{aligned}
APCE_{AH} &= E[R(1) - R(0) \mid \text{student is in AH}] = \\
&= \frac{Pr(R(1)=1, Y(0)=1, Y(1)=1) - Pr(R(0)=1, Y(0)=1, Y(1)=1)}{Pr(Y(0)=1, Y(1)=1)} \\
&= \frac{Pr(R=0, Y=1 \mid Z=0) - Pr(R=0, Y=1 \mid Z=1) + \eta Pr(R=0 \mid Z=1)}{Pr(Y(0)=1)}
\end{aligned}
$$

Finally, we derive the bounds on $Pr(Y(0)=1)$ and $Pr(Y(1)=1)$ without the exclusion restriction. Following the same intuition, and using the results from the section with the exclusion restriction, we have that:

$$
\begin{aligned}
\max\{Pr(Y=1, R=0 \mid Z=0), Pr(Y=1, R=0 \mid Z=1) - \eta Pr(R=0 \mid Z=1)\} & \\
\leq Pr(Y(0)=1) \leq \min\{Pr(Y=1 \mid Z=0), Pr(Y=1 \mid Z=1) - \eta\} &
\end{aligned}
$$

$$
\begin{aligned}
\max\{Pr(Y=1 \mid Z=0), Pr(Y=1 \mid Z=1) - \eta\} \leq Pr(Y(1)=1) & \\
\leq 1 - \max\{Pr(Y=0, R=1 \mid Z=0), Pr(Y=0, R=1 \mid Z=1) + \eta Pr(R=1 \mid Z=1)\} &
\end{aligned}
$$

All in all, the estimators of the bounds of the APCEs are given by the sample analogs of the following expressions:

Lower bound of $APCE_H$:

$$\frac{Pr[Y \mid Z=1]-Pr[Y \mid Z=0]-\eta}{1-\max\left\{Pr(Y=0,R=1 \mid Z=1)+\eta Pr(R=1 \mid Z=1), Pr(Y=0,R=1 \mid Z=0)\right\}-\max\left\{Pr(Y=1,R=0 \mid Z=1)-\eta Pr(R=0 \mid Z=1), Pr(Y=1,R=0 \mid Z=0)\right\}} \tag{A26}$$

Upper bound of $APCE_H$:

$$\frac{Pr[Y \mid Z=1]-Pr[Y \mid Z=0]-\eta}{\max\left\{Pr(Y=1 \mid Z=1)-\eta, Pr(Y=1 \mid Z=0)\right\}-\min\left\{Pr(Y=1 \mid Z=1)-\eta, Pr(Y=1 \mid Z=0)\right\}} \tag{A27}$$

Lower bound of $APCE_{AH}$:

$$\frac{Pr(R=0,Y=1 \mid Z=0)-Pr(R=0,Y=1 \mid Z=1)+\eta Pr(R=0 \mid Z=1)}{\min\left\{Pr(Y=1 \mid Z=1)-\eta, Pr(Y=1 \mid Z=0)\right\}} \tag{A28}$$

Upper bound of $APCE_{AH}$:

$$\frac{Pr(R=0,Y=1 \mid Z=0)-Pr(R=0,Y=1 \mid Z=1)+\eta Pr(R=0 \mid Z=1)}{\max\left\{Pr(R=0,Y=1 \mid Z=1)-\eta Pr(R=0 \mid Z=1), Pr(R=0,Y=1 \mid Z=0)\right\}} \tag{A29}$$

Lower bound of $APCE_{AL}$:

$$\frac{Pr(R=1,Y=0 \mid Z=1)-Pr(R=1,Y=0 \mid Z=0)+\eta Pr(R=1 \mid Z=1)}{1-\max\left\{Pr(Y=1 \mid Z=1)-\eta, Pr(Y=1 \mid Z=0)\right\}} \tag{A30}$$

Upper bound of $APCE_{AL}$:

$$\frac{Pr(R=1,Y=0 \mid Z=1)-Pr(R=1,Y=0 \mid Z=0)+\eta Pr(R=1 \mid Z=1)}{\max\left\{Pr(R=1,Y=0 \mid Z=1)+\eta Pr(R=1 \mid Z=1), Pr(R=1,Y=0 \mid Z=0)\right\}} \tag{A31}$$

Estimates of Scoring Above Cutoff for Never-Takers (η)

Table A5 shows the estimated values of η, as outlined in Section 5.

Table A5: Effect of scoring above the cutoff for never-takers.

Track	η 1	η 2
V: BL	-.0097	.0258
	(.0299)	(.0332)
V: BL/KL	.0336	.1013
	(.0521)	(.0465)
V: KL	.0492	.036
	(.0216)	(.025)
V: KL/GT	.0269	-.0338
	(.044)	(.0444)
V: GT	.0195	-.0057
	(.0128)	(.0154)
V: GT/HAVO	-.0302	.0639
	(.0305)	(.0265)
A: HAVO	.0242	.0278
	(.0106)	(.0122)
A: HAVO/VWO	.0064	.0551
	(.0268)	(.0226)
VOCATIONAL	.0278	.0338
	(.0092)	(.009)
ACADEMIC	.0196	-.0215
	(.0106)	(.0099)
ALL	.0245	.0108
	(.007)	(.0068)
COHORT	2015	2016
COVARIATES	YES	YES

Notes: The row headings report the names of the secondary school tracks to which a student can be assigned in the Dutch system, ranked by level of difficulty. The last three rows of the table aggregate the Vocational (VMBO*), the Academic (HAVO*), and All tracks, respectively. Columns 1 and 2 show the estimates of η for each track and cohort, as explained in section 5.

Estimates of APCEs Without Assuming Exclusion Restriction

Table A6: Estimates of the $APCE_H$, without exclusion restriction

Track	Numerator $APCE_H$	Lower bound denominator $APCE_H$	Upper bound denominator $APCE_H$	Lower bound $APCE_H$	Upper bound $APCE_H$
	1	2	3	4	5
V: BL	.0187	.0187	.6513	.0289	1
	(.0094)	(.0076)	(.0164)	(.0147)	(.4627)
V: BL/KL	.0083	.0083	.2764	.0318	1
	(.019)	(.0128)	(.0239)	(.069)	(.6564)
V: KL	.0555	.0555	.6907	.0803	1
	(.0088)	(.0088)	(.0112)	(.0125)	(0)
V: KL/GT	.0306	.0344	.3982	.0731	.8895
	(.0163)	(.0135)	(.021)	(.0414)	(.4774)
V: GT	.0064	.0072	.7677	.0082	.8888
	(.0064)	(.0061)	(.009)	(.0124)	(.4177)
V: GT/HAVO	.0285	.0285	.5041	.0566	1
	(.0098)	(.0091)	(.0135)	(.0191)	(.337)
A: HAVO	.0091	.0091	.8266	.0111	1
	(.0029)	(.0028)	(.006)	(.0034)	(.1736)
A: HAVO/VWO	.0313	.0313	.5387	.0583	1
	(.0087)	(.0086)	(.0128)	(.0157)	(.1429)
VOCATIONAL	.0232	.0237	.6484	.0357	.9772
	(.004)	(.0037)	(.0054)	(.0064)	(.0761)
ACADEMIC	.0161	.0161	.7357	.0219	1
	(.0034)	(.0033)	(.0059)	(.0046)	(.0193)
ALL	.0203	.0206	.6838	.0297	.9844
	(.0027)	(.0025)	(.004)	(.004)	(.0538)
COVARIATES	YES	YES	YES	YES	YES

Notes: The row headings report the names of the secondary school tracks to which a student can be assigned in the Dutch system, ranked by level of difficulty. The last three rows of the table aggregate the Vocational (VMBO*), the Academic (HAVO*), and All tracks, respectively. With reference to the $APCE_H$ estimand defined in equation (4), the table reports for the outcome Y and separately for each track, estimates of the numerator (in column 1), of the lower and upper bounds of the denominator (in columns 2 and 3), and of the lower and upper bounds of the $APCE_H$ (in columns 4 and 5), without assuming the exclusion restriction, as given by equations equations (A26) and (A27), respectively. The aggregations for the three last rows are performed as follows: point estimates of the numerator and of the lower and upper bounds of the denominator are created by weighing each track in the corresponding aggregate by the relative population in each track. For the lower and upper bounds of the $APCE_H$, the aggregated estimates for the lower and upper bounds of the denominators are used directly. Standard errors for each quantity are bootstrapped with 1000 repetitions; each quantity in every iteration is calculated in the same way.

Table A7: Estimates of the $APCE_{AH}$, without exclusion restriction

Track	Numerator $APCE_{AH}$	Lower bound denominator $APCE_{AH}$	Upper bound denominator $APCE_{AH}$	Lower bound $APCE_{AH}$	Upper bound $APCE_{AH}$
	1	2	3	4	5
V: BL	.0028	.3187	.3187	.0125	.0125
	(.0088)	(.0159)	(.0158)	(.0283)	(.0283)
V: BL/KL	.0805	.6858	.705	.1138	.1167
	(.0213)	(.0248)	(.0247)	(.0304)	(.0303)
V: KL	.0433	.2282	.2392	.1827	.191
	(.0055)	(.0101)	(.0101)	(.0216)	(.0224)
V: KL/GT	.0709	.5681	.5791	.1224	.1248
	(.0169)	(.022)	(.0225)	(.0292)	(.0293)
V: GT	.0106	.1908	.1916	.054	.0542
	(.0063)	(.0077)	(.0077)	(.024)	(.024)
V: GT/HAVO	.0632	.4169	.4262	.1486	.1522
	(.0081)	(.0131)	(.0132)	(.0183)	(.0185)
A: HAVO	.0037	.1176	.1186	.0315	.0318
	(.0022)	(.0052)	(.0052)	(.018)	(.0181)
A: HAVO/VWO	.0546	.3667	.3777	.1455	.1498
	(.0076)	(.0116)	(.0117)	(.0192)	(.0196)
VOCATIONAL	.0323	.2977	.3034	.1064	.1084
	(.0036)	(.005)	(.005)	(.0117)	(.0119)
ACADEMIC	.0198	.1962	.2004	.0988	.1009
	(.0028)	(.0052)	(.0052)	(.0136)	(.0138)
ALL	.0272	.2565	.2616	.104	.1061
	(.0024)	(.0036)	(.0037)	(.0091)	(.0092)
COVARIATES	YES	YES	YES	YES	YES

Notes: The row headings report the names of the secondary school tracks to which a student can be assigned in the Dutch system, ranked by level of difficulty. The last three rows of the table aggregate the Vocational (VMBO*), the Academic (HAVO*), and All tracks, respectively. With reference to the $APCE_H$ estimand defined in equation (4), the table reports for the outcome Y and separately for each track, estimates of the numerator (in column 1), of the lower and upper bounds of the denominator (in columns 2 and 3), and of the lower and upper bounds of the $APCE_H$ (in columns 4 and 5), without assuming the exclusion restriction, as given by equations equations (A30) and (A31), respectively. The aggregations for the three last rows are performed as follows: point estimates of the numerator and of the lower and upper bounds of the denominator are created by weighing each track in the corresponding aggregate by the relative population in each track. For the lower and upper bounds of the $APCE_H$, the aggregated estimates for the lower and upper bounds of the denominators are used directly. Standard errors for each quantity are bootstrapped with 1000 repetitions; each quantity in every iteration is calculated in the same way.

Table A8: Estimates of the $APCE_{AL}$, without exclusion restriction

Track	Numerator $APCE_{AL}$	Lower bound denominator $APCE_{AL}$	Upper bound denominator $APCE_{AL}$	Lower bound $APCE_{AL}$	Upper bound $APCE_{AL}$
	1	2	3	4	5
V: BL	.0277	.0286	.664	.0411	.9776
	(.0053)	(.0053)	(.0172)	(.0078)	(.0266)
V: BL/KL	.0351	.0378	.2867	.1237	.9394
	(.0103)	(.0099)	(.0253)	(.0338)	(.0884)
V: KL	.0741	.0811	.7053	.1049	.9132
	(.0078)	(.0076)	(.0131)	(.0105)	(.0266)
V: KL/GT	.0321	.0337	.3865	.0825	.9603
	(.0098)	(.0095)	(.0244)	(.0236)	(.0584)
V: GT	.0382	.0416	.8012	.0477	.9197
	(.0039)	(.0038)	(.0095)	(.0048)	(.0256)
V: GT/HAVO	.0691	.0789	.5453	.1264	.8734
	(.0091)	(.0086)	(.0157)	(.0152)	(.0395)
A: HAVO	.0506	.0558	.8722	.0579	.9066
	(.0042)	(.004)	(.0059)	(.0047)	(.0249)
A: HAVO/VWO	.0851	.0946	.591	.1442	.8999
	(.0087)	(.0085)	(.0144)	(.0134)	(.0258)
VOCATIONAL	.049	.0538	.6731	.0728	.9115
	(.0029)	(.0027)	(.0062)	(.0041)	(.0148)
ACADEMIC	.0615	.068	.7835	.0785	.9036
	(.004)	(.0039)	(.0063)	(.0049)	(.0176)
ALL	.0541	.0596	.7179	.0753	.9079
	(.0024)	(.0023)	(.0044)	(.0032)	(.0113)
COVARIATES	YES	YES	YES	YES	YES

Notes: The row headings report the names of the secondary school tracks to which a student can be assigned in the Dutch system, ranked by level of difficulty. The last three rows of the table aggregate the Vocational (VMBO*), the Academic (HAVO*), and All tracks, respectively. With reference to the $APCE_H$ estimand defined in equation (4), the table reports for the outcome Y and separately for each track, estimates of the numerator (in column 1), of the lower and upper bounds of the denominator (in columns 2 and 3), and of the lower and upper bounds of the $APCE_H$ (in columns 4 and 5), without assuming the exclusion restriction, as given by equations equations (A28) and (A29), respectively. The aggregations for the three last rows are performed as follows: point estimates of the numerator and of the lower and upper bounds of the denominator are created by weighing each track in the corresponding aggregate by the relative population in each track. For the lower and upper bounds of the $APCE_H$, the aggregated estimates for the lower and upper bounds of the denominators are used directly. Standard errors for each quantity are bootstrapped with 1000 repetitions; each quantity in every iteration is calculated in the same way.

Estimates of the $APCE_J$ based on unconfoundedness

Table A9 shows the point estimates of $APCE_J$ from assuming unconfoundedness. As can be seen, the results show that the estimates for the three strata are similar in all tracks.

Table A9: Estimates of the $APCE_J$ based on unconfoundedness assumption

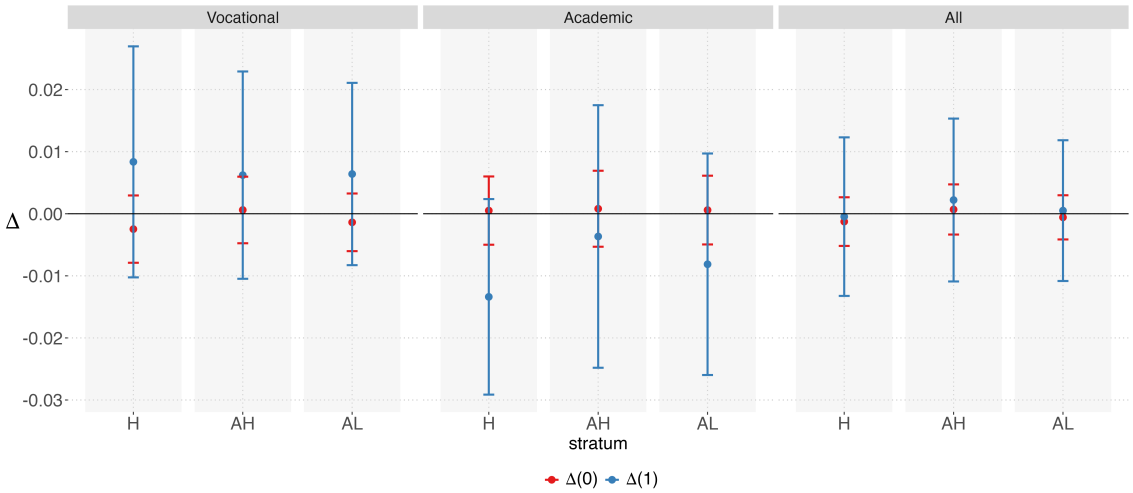
Track	$APCE_{AL}$ 1	$APCE_H$ 2	$APCE_{AH}$ 3
V: BL	0.0444	0.0481	0.0508
	(0.0071)	(0.0076)	(0.0089)
V: BL/KL	0.1360	0.1357	0.1311
	(0.0306)	(0.0236)	(0.0207)
V: KL	0.1560	0.1682	0.1807
	(0.0107)	(0.0108)	(0.0129)
V: KL/GT	0.1247	0.1288	0.1379
	(0.0206)	(0.0175)	(0.0182)
V: GT	0.0497	0.0556	0.0584
	(0.0040)	(0.0047)	(0.0057)
V: GT/HAVO	0.1530	0.1607	0.1659
	(0.0122)	(0.0117)	(0.0126)
A: HAVO	0.0574	0.0700	0.0776
	(0.0041)	(0.0051)	(0.0063)
A: HAVO/VWO	0.1531	0.1674	0.1770
	(0.0099)	(0.0103)	(0.0116)
VOCATIONAL	0.0969	0.1037	0.1090
	(0.0040)	(0.0039)	(0.0045)
ACADEMIC	0.0876	0.1007	0.1090
	(0.0042)	(0.0048)	(0.0057)
ALL	0.0931	0.1025	0.1090
	(0.0029)	(0.0030)	(0.0035)

Notes: The row headings report the names of the secondary school tracks to which a student can be assigned in the Dutch system, ranked by level of difficulty. The last three rows of the table aggregate the Vocational (VMBO*), the Academic (HAVO*), and All tracks, respectively. The columns show point estimates obtained by assuming unconfoundedness. Standard errors for each quantity are bootstrapped with 1000 repetitions; each quantity in every iteration is calculated in the same way.

Appendix to Section 6

Estimates of fairness by gender

Figure A1: Fairness by gender



Notes: the figure plots, for the three aggregate tracks and for the H, AH and AL strata, the point estimates of the statistic $\Delta_j(z) = Pr(R(z) = 1 \mid \text{Girl}) - Pr(R(z) = 1 \mid \text{Boy})$ and the correspondent 95% confidence intervals.