Should we sort it out later? The effect of tracking age on long-run outcomes

Economics of Education Review 75 (2020) 101973

Contents lists available at ScienceDirect

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev

Should we sort it out later? The effect of tracking age on long-run outcomes

By Lex Borghans[a], Ron Diris[*,a], Wendy Smits[b,c], Jannes de Vries[c]

[a] Department of Economics, Maastricht University, 6200 MD Maastricht, the Netherlands
[b] Research Centre for Education and the Labour Market (ROA), Maastricht University, 6200 MD Maastricht, the Netherlands
[c] Statistics Netherlands,6401 CZ Heerlen, the Netherlands

ARTICLE INFO

Keywords:
Educational economics
Efficiency
Wage differentials

JEL classification:
I21
J24

ABSTRACT

This study estimates the long-run effect of the school tracking age on educational attainment and labour market outcomes. We exploit within-country variation in tracking ages for students in the highest two tracks in the Netherlands, using the supply of early tracking schools at the municipal level as an instrument for early tracking (tracking at age 12–13 vs. age 14). Combining several data sources, we find that early tracking leads to a decrease in higher education completion, and that it lowers earnings for both low-ability and medium-ability students in the sample. Estimates for high-ability students are positive but imprecisely estimated. The negative effects appear largely driven by higher misallocation of students to tracks when they are sorted early. Robustness analyses strongly suggest that the results are not driven by sorting between municipalities.

1. Introduction

There is an ongoing debate on the merits of tracked versus comprehensive education. Tracking can enhance teaching efficiency because instruction is targeted towards a homogeneous group of students, but can also enhance inequality by sorting peer and school quality. Additionally, assessments that determine track assignment cannot perfectly forecast student potential, thereby allocating some students to a track that is suboptimal for their learning process. Among countries that track students, there is substantial variation in the age at which tracking takes place. The implications of even small differences in the tracking age can be large in the long run. For one, skill formation is a dynamic process in which early learning begets future learning, and therefore early differences in skill development can expand over time (Cunha & Heckman, 2007). Additionally, different tracking ages also imply differences in the accuracy of sending students to their optimal track, which can influence educational pathways throughout secondary school as well as student choices towards post-secondary education and subsequent job opportunities.

This paper estimates the effect of early tracking on educational attainment and labour market outcomes, for students in the intermediate and academic track of Dutch education. We exploit within-country variation, as tracking in the Netherlands can take place in grade 7, 8 or 9, depending on school policy. We define early tracking as taking place in grade 8 (when students are typically 13 years old) or before. We estimate an instrumental variable model that uses the relative supply of early tracking schools in the municipality as an instrument for early tracking. This study uses a rich dataset that matches students from a secondary school cohort study with administrative data on educational careers and labour market outcomes, and to data on the school tracking policy for every school in the Netherlands to deduce the supply of early tracking schools that each student faces.

We find that early tracking initially leads to a higher likelihood of being sorted into the academic track, but ultimately to lower levels of educational attainment and lower earnings. Both low-ability and medium-ability students experience negative earnings effects from early tracking, but with a different dynamic across outcomes. Those (initially) perceived as low-ability students appear negatively affected because earlier tracking leads to a less accurate assessment of their abilities. Medium-ability students appear negatively affected through a too strong tendency to put them in the more demanding academic track when tracking is earlier, which puts them on a downward trajectory during the remainder of their educational career and early labour market years. The point estimates of the effect of early tracking for students of high ability are positive and substantial, but imprecisely estimated. The negative earnings effects for the sample as a whole, which are around 14%, are largely attributed to a decrease in hours worked. This, in turn, is partly mediated by field of study.

Students in the Netherlands enter secondary school in grade 7, when they are typically 12 years old, but track assignment can still be postponed for one or two grades. The exact tracking age in lower secondary education is up to the discretion of school leadership. Literature

* Corresponding author.
E-mail address: r.diris@maastrichtuniversity.nl (R. Diris).

https://doi.org/10.1016/j.econedurev.2020.101973
Received 18 January 2019; Received in revised form 31 January 2020; Accepted 7 February 2020
Available online 19 February 2020
0272-7757/

suggests that this school policy is predominantly driven by the pedagogical and didactic views of school directors (Korpershoek, Naaijer, & Bosker, 2016). While our estimation approach corrects for sorting within municipalities, the estimates can still be biased through sorting of students between municipalities. Our results provide evidence against this. First, while early tracking correlates strongly with a wide range of individual background variables and baseline test scores, the instrument is orthogonal to all observable characteristics, suggesting that sorting is concentrated within municipalities. Second, the estimates remain highly similar when including geographical controls such as level of urbanization, and we also observe the same pattern of results within low urbanised areas as within high urbanised areas. Hence, the results are not driven by unobserved differences between students/parents living in cities and students/parents living in rural areas. Third, we find no evidence of a relationship between the instrument and parental attitudes and investments, derived from the survey data. Fourth, we find that the different didactic views that are behind the different tracking decisions are not reflected in (other) aspects of school quality.

The literature on tracking is rich and expanding. Several studies exploit between-country variation through difference-in-difference designs to estimate the relation between tracking and student achievement; see, e.g., Hanushek and Zhang (2006). A rare instance of experimental variation is exploited by Duflo, Dupas, and Kremer (2011), who identify positive effects on achievement in Kenya. Most recent studies exploit tracking policy changes in European countries. Guyon, Maurin, and McNally (2012) find that an expansion of the elite track in Northern Ireland led to increases in educational attainment, while Piopiunik (2014) identifies a negative effect on the educational achievement of low-ability students from a shift of tracking from grade 6 to grade 4 in Germany. Pekkala Kerr, Pekkarinen, and Uusitalo (2013) similarly find that a postponement of the tracking age in Finland improved test scores of low SES students.

In general, these studies conclude that (earlier) tracking does not benefit mean performance and harms low-ability students.[1] However, these findings are limited to educational achievement and attainment. One of the motivations for tracking (earlier) is that not all students are deemed fit to pursue a track with an academic focus. It might therefore not be surprising that earlier sorting leads to lower performance on academic achievement tests and fewer academic degrees for low-ability students. It would arguably be more valuable to evaluate such practices by looking at how students fare in the labour market. Studies on the labour market effects of early tracking are more scarce but notable exceptions exist. Their findings are mixed and underline the importance of estimating long-run effects. Hall (2012) finds that a policy change that made the vocational track in Sweden more academic increased educational attainment in secondary school, but did not lead to higher university enrolment or higher earnings in later life. Malamud and Pop-Eleches (2010, 2011) identify a similar dynamic across outcomes for a postponement of the start of vocational education in Romania. On the other hand, Meghir and Palme (2005) (for a joint reform that increased both the tracking age and the compulsory schooling age in Sweden) and Pekkarinen, Uusitalo, and Kerr (2009) (for Finland) identify positive wage effects for low-ability students from attending (more) comprehensive schooling.

All of these long-run studies exploit tracking policy changes. This provides a robust way around the selection issues that plague estimation relying on cross-country or (static) within-country variation. On the other hand, identification through policy changes could pick up on transitional effects or unobserved changes in related policies. For example, teachers have to adjust to teaching either more or less homogeneous groups and new curricula have to take shape over time. This study, in contrast, provides evidence for an educational system in which tracking policies have been stable for a long time, relying on within-country differences at the same point in time. Additionally, previous studies have examined situations in which at least part of the student population shifts from an academic to a vocational curriculum or vice versa. In contrast, we study a setting with two non-vocational tracks for which the formal curricula are similar (i.e. students follow the same set of school subjects). Hence, our study is not about the effects of (more) academic versus vocational education. The setting allows us to zoom in on the "efficiency arguments" of early tracking, in terms of having a more homogeneous set of peers versus having a lower accuracy of sending students to their optimal track.[2] Van Elk, Van der Steeg, and Webbink (2011) provide evidence in the same Dutch setting, also exploiting geographical variation, to estimate the effect of early tracking for a subgroup of low-ability students. They find that it leads to fewer higher education diplomas for this subgroup. Our study differs by also estimating labour market effects, focusing on the full ability distribution within the top two tracks, and relying on a different instrument to correct for the endogeneity of tracking age.

To sum up, this study provides three main contributions to the tracking literature. First, we contribute to the expanding but still scarce literature on the long-run effects of tracking age. Second, by focusing on a settled tracking system and on students in two non-vocational tracks, our setting allows us to identify effects that are not driven by adjustments to new policies or by differences in the formal curriculum, but rather capture the allocative aspects of earlier versus later tracking. We also develop a theoretical model that specifies these different mechanisms and how they interact with ability. Third, by combining administrative and survey data, we can uncover more of the mechanisms and heterogeneity that are behind tracking age effects. Previous studies typically rely on registered data on final educational attainment. Our longitudinal data register the position in the educational system for every year of the educational career as well as post-secondary study choices, which allows us to estimate the complete dynamics of the early tracking effect. These dynamics turn out to be highly important in explaining the long-run effects. The availability of pre-treatment achievement data also allows for a rich heterogeneity analysis across student ability.

This paper is organized as follows. Section 2 presents our theoretical framework. Section 3 gives an overview of the Dutch educational system. Section 4 discusses data. The methodological approach is explained in Section 5. Section 6 presents the main results, after which robustness analyses are discussed in Section 7. Section 8 concludes.

2. Theory

The effect of the age of tracking on future outcomes can operate through multiple mechanisms. Brunello, Giannini, and Ariga (2007) develop a theoretical model that pits peer sorting (which favours earlier selection) against uncertainty about the student's type (which favours later selection). In this section, we expand on this model, by also incorporating instruction effects and heterogeneity by ability.

2.1. Instruction effect

In the empirical setting of this paper, two tracks are considered and

[1] Conversely, Galindo-Rueda and Vignoles (2005) find that the shift in the United Kingdom from a tracked to a comprehensive system had no benefits for low-ability students and harmed high-ability students, but Manning and Pischke (2006) show that the approach of the study does not solve the endogeneity issue.

[2] The setting therefore also relates to the US-based literature on ability-grouping; see, e.g., Betts and Shkolnik (2000); Rees, Brewer, and Argys (2000). There are two crucial distinctions: students are separated from the other track for all school subjects, and the tracks provide different eligibility towards post-secondary education.

sorting is based on average student ability. The theoretical model is also built from this perspective. Students can be tracked, or held together in a comprehensive class. This results in three possible investment paths, pertaining to either the low track L, the high track H, or the comprehensive track C. The return on these investment paths depends on the ability type θ_i, and is reflected in a future outcome variable Y_i. For each track T, we have:

$$Y_i = \alpha^T + \beta^T \theta_i \quad (1)$$

We specify that, also in absence of peer effects, α^T and β^T will differ between tracks because instruction will be adjusted to the ability distribution in class. We label this as the 'instruction effect' of tracking. In general, this can comprise both differences in formal curricula and 'informal' differences that reflect that teachers adjust pace and level of instruction to the ability level of the class for a given subject. In the empirical setting of this study, students follow the same school subjects in each track, and only the latter difference applies. The same principles apply.[3]

We assume that instruction is adjusted to the level of the median student in each track. As a result, lower (higher) tracks provide better outcomes for students of lower (higher) ability. This implies that $\alpha^L > \alpha^C > \alpha^H$ and $\beta^L < \beta^C < \beta^H$. This situation is depicted in Fig. 1. If the allocation of students to tracks maximizes individual payoffs, outcomes under tracking will equal:

$$Y_i = max(\alpha^L + \beta^L \theta_i, \alpha^H + \beta^H \theta_i) \quad (2)$$

Alternatively, the payoff under a comprehensive system will equal:

$$Y_i = \alpha^C + \beta^C \theta_i \quad (3)$$

In Fig. 1, the tracked system is favourable for students to the left of type θ_a and to the right of type θ_b, while the comprehensive system is favourable for those in between. It could also be argued that heterogeneity in class is detrimental to the efficiency of instruction in general, such that comprehensive classes also harm students of medium-ability. Values for α^T and β^T can be chosen accordingly. Still, one may assume that C is *relatively* more favourable for medium-ability types compared to low and high-ability types, leading to a similar pattern across θ_i. Additionally, α^T and β^T can be different for different outcome variables. This underlines the importance of estimating both the short-run and the long-run impacts of different tracking policies.

2.2. Noise effect

The model so far assumes that there is perfect information and each student is allocated to the track that maximizes Y_i. In reality, ability is measured with error and students can be allocated to a suboptimal track. This is labelled by Brunello et al. (2007) as the "noise effect of tracking". The later tracking takes place, the larger the information set and the lower the risk of misallocation. This noise effect is a key aspect of tracking, since misallocation influences the learning environment throughout secondary schooling, and can significantly impact trajectories after secondary education as well. We label the noise penalty of earlier tracking as γ_i. How this effect differs across ability depends on two opposite forces. The risk of misallocation is highest at the ability threshold θ_c, but the size of the penalty increases when we move away from θ_c. The relation between γ_i and θ_i thus depends on the assumed functional forms. The negative effect of misallocation is assumed to be linearly related to $\theta_i - \theta_c$ (by $\beta^H - \beta^L$). Additionally, we (reasonably) assume that the measurement error on the ability signal is normally distributed around the true ability. The resulting simulation is presented on the right side of Fig. 1. The parameter γ equals zero at θ_c, where there is no loss from misallocation, and at the extremes, where there is no probability of misallocation, while it is positive in between.

To avoid being too restrictive, we simply specify that γ_i depends on the distance from the cut-off point, without assuming any functional form:

$$\gamma_i = f(|\theta_i - \theta_c|) \quad (4)$$

It is likely that γ_i also depends on the type of misallocation. Being assigned to a track that is above ones capabilities involves different mechanisms and likely leads to a different impact than being assigned to a track that is below ones capabilities (also depending on the possibilities for retracking in each direction). As such, the payoffs from early tracking are equal to:

$$Y_i = max(\alpha^L + \beta^L \theta_i - \gamma_i^L, \alpha^H + \beta^H \theta_i - \gamma_i^H) \quad (5)$$

The payoff from later tracking is the same as under Eq. (3). Later tracking is also based on a noisy ability signal, but the noise is lower than under early tracking. As such, γ_i^T reflects the *increase* in the noise parameter from tracking earlier. The noise effect shifts the cut-off types θ_a and θ_b further to the left and right, respectively, thereby increasing the total set of students that favours later tracking.[4]

2.3. Peer effects

In Fig. 1, the effect of early versus late tracking is symmetric around θ_c, leading to a similar optimal size of tracks L and H. However, tracking also reallocates peers. Peer quality decreases in Track L and increases in Track H, relative to Track C.[5] Additionally, dispersion in peer quality reduces through tracking. The literature identifies substantial positive effects of peer quality on school achievement, and mixed evidence for the effect of peer dispersion (Epple & Romano, 2011; Sacerdote, 2011). Following these findings, panel (a) of Appendix Figure A1 incorporates peer effects by decreasing α^L and increasing α^H relative to α^M. Fig. 2 directly shows the resulting effect of early versus late tracking across the ability distribution. Peer effects decrease the optimal size of track L and increase the optimal size of track H. Hence, (stronger) peer effects shift the cut-off types θ_a, θ_b and θ_c to the left.

2.4. Heterogeneity and thresholds

The empirical analysis of this study estimates the effects of early tracking separately for students of low, medium and high ability. While the model states that students between θ_a and θ_b are hurt by earlier tracking, we cannot predict how this translates to these ability types, as the location of θ_a, θ_b and θ_c depends on the true values of α^T, β^T and γ^T. Strong peer effects will lead to a high α^H and will shift the cut-off types to the left. The student at the median of the ability distribution can then still prefer early tracking. Although we cannot unambiguously say that those with high θ_i benefit from early tracking, it is highly likely that

[3] One may consider 'curriculum effect' (comprising both formal and informal aspects of the curriculum) to be a more appropriate term for this mechanism. However, since it is a particular feature of this study that the empirical setting involves no differences in the formal curriculum, we consider this terminology to be confusing here.

[4] While there is no noise effect for students at θ_c, the negative effect of early tracking is still maximized at this point. This is a result of the chosen parameters; β^C is exactly in between β^L and β^H. Since the probability of assignment to the right track is at least 50%, the noise effect when moving away from θ_c does not compensate for the stronger instruction effect at θ_c. Panel (d) of Appendix Figures A1 and A2 shows that the effect of early tracking is minimized at a different point for alternative values for α^T and β^T (panels (b) and (c) show that this can also occur when the threshold is not at the efficient point). Even for the rather extreme parameters chosen here, the minimum point is not far from θ_c.

[5] Teacher and school quality can differ as well between tracks. For simplicity of presentation, we model these under the same umbrella, assuming they all gradually improve across the distribution and have a linear positive effect. The model can easily be expanded to incorporate non-linear effects.

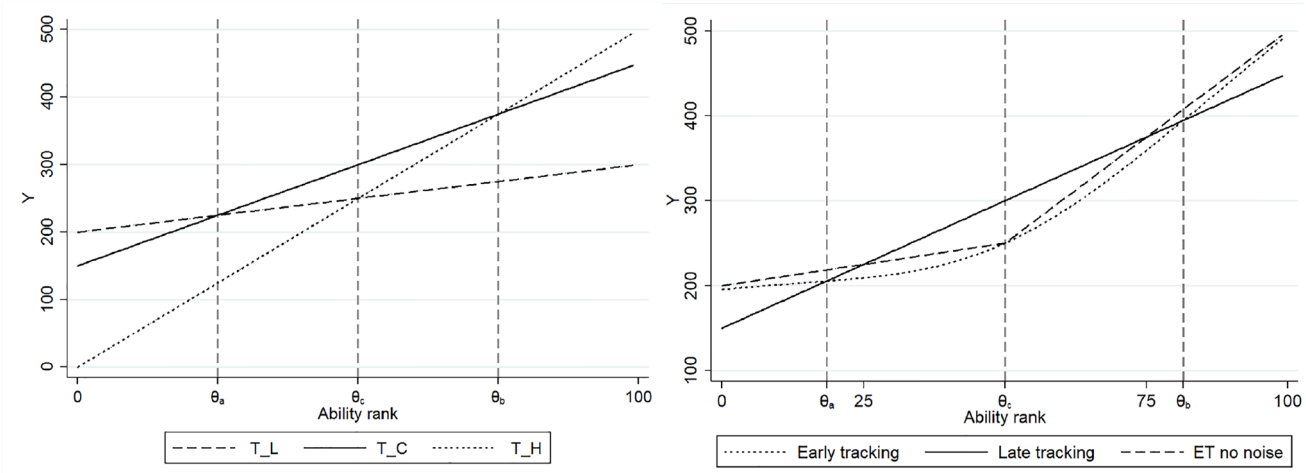

Fig. 1. Tracking vs. comprehensive grade (simulation). **Notes:** The figures presents a simulation of the effect of attending a low track (L), comprehensive track (C) or a high track (H), on an arbitrary outcome variable Y, across the ability ranking. The simulation is conducted for 1,000,000 observations. The chosen parameters for α^T are 200, 150 and 0; the chosen parameters for β^T are 1, 3 and 5. The effect of early tracking is zero at θ_a and θ_b. θ_c is the cut-off ability type for attending the higher track. The right figure adds noise in the ability signal to the simulation. The noise is assumed to be normally distributed.

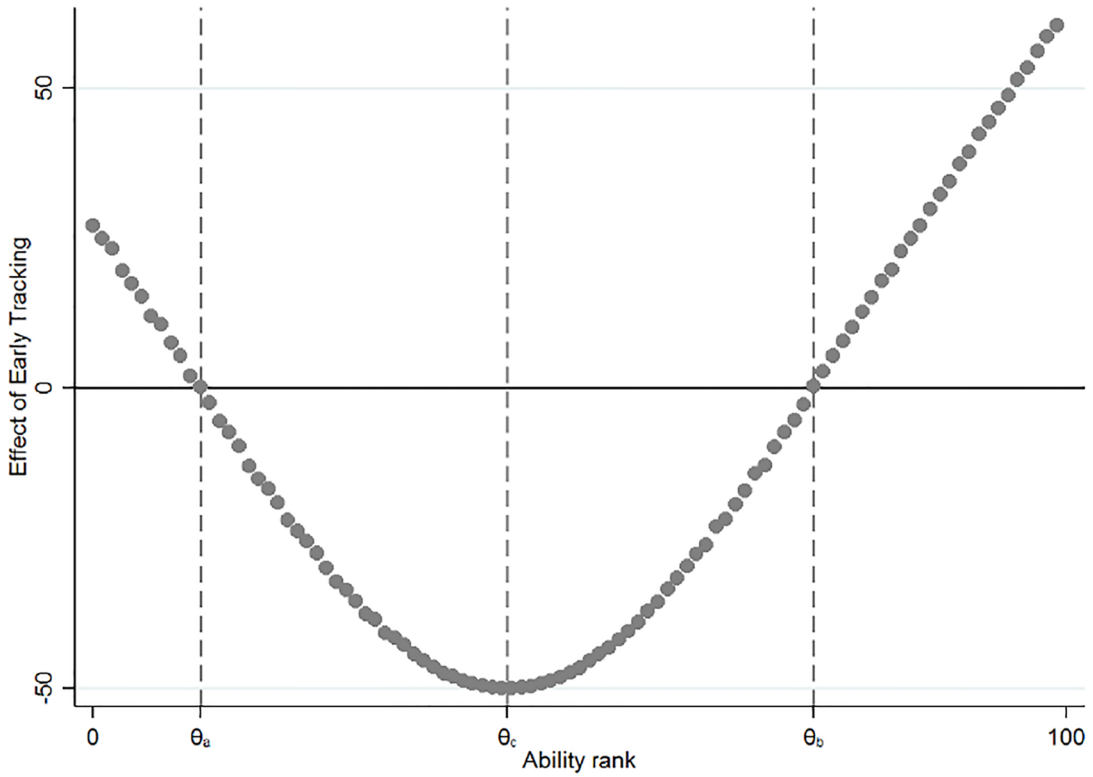

Fig. 2. Simulated effect of early tracking across θ_i. **Notes:** The figure shows the simulated effect of early tracking, i.e. the difference in payoff from early vs. late tracking, incorporating curriculum, noise and peer effects (situation depicted in panel (a) of Figure A1).

they do. Only a very high noise penalty can compensate for the favourable instruction and peer effect, which is unlikely given the low probability of misallocation at this distance from θ_c. It is comparatively more likely that low-ability students are hurt by early tracking, as the peer effect works in the opposite direction.

While recognizing that the ability signal is noisy, we have still assumed that the *targeted* ability margin for going to the higher track is θ_c, as is efficient (i.e. θ_c is at the student that is indifferent between tracks L and H). The effective ability margin may either be more lenient or more strict than that efficient margin. For example, schools that offer track H may lower the ability threshold to attract more students, or may set a higher ability threshold to increase average student ability. Additionally, overconfidence by students and parents may expand track H. Panels (b) and (c) of Appendix Figures A1 and A2 show the effects of early tracking when thresholds are set either too low or too high. Too lenient assignment will shift θ_a and θ_b to the left, while too strict assignment will shift these cut-off types to the right. The effective location of the threshold may also differ between early and late tracking. Under higher uncertainty (i.e. earlier tracking), schools may be risk-averse towards letting marginal students enter track H, as failure at the higher track presents costs for the school (e.g. through grade retention). Overconfidence of parents and students might make them more eager to

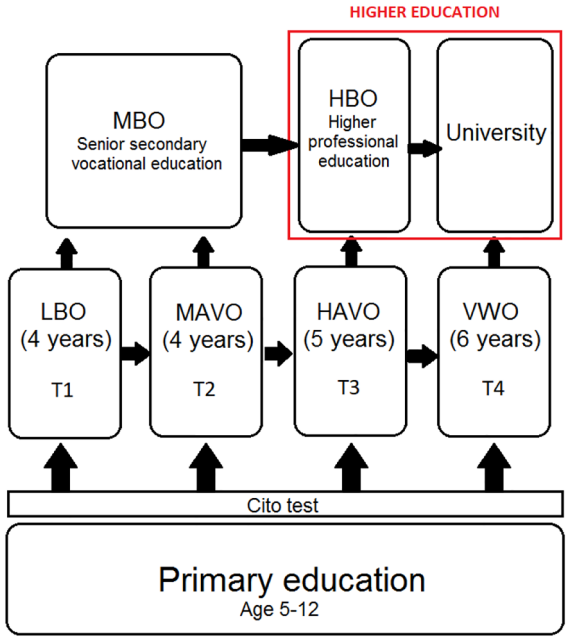

Fig. 3. Dutch educational system. **Notes:** Source: Center of International Education Benchmarking.

push for track H when uncertainty is high. Assuming that more noise would also increase the probability of setting a "suboptimal" threshold, this would be an additional argument against early tracking.

3. Dutch education

3.1. Educational system

This study analyses the effect of the tracking age within the context of Dutch education. A schematic overview of the Dutch educational system is provided in Fig. 3. Students in the Netherlands attend six years of primary school. In secondary education, they can be subsequently sorted into four tracks. These are *lbo* (lower vocational), *mavo* (higher vocational), *havo* (higher general) and *vwo* (pre-university). We label these tracks as T1, T2, T3 and T4, respectively, for the remainder of this paper.[6] After being tracked, students can still drop to lower tracks when their achievement is low. Moving to higher tracks is generally only possible when the current track has been completed.

After completing secondary school, students can enrol in three types of post-secondary education. The lowest level is *mbo*, providing post-secondary vocational education. Students of both T1 and T2 are eligible for *mbo*. Higher education can be divided into two categories; *hbo* (higher professional education) and university. Students with a diploma from the T3 track or higher can enter *hbo*, while only those with a T4 diploma can (directly) enter university. Students can also reach higher levels of post-secondary education through "horizontal" pathways: they are eligible for *hbo* after completing the highest level of *mbo*, and for university after completing *hbo*. As such, T3 students can still ultimately complete university, but these alternative routes require more effective years of education.

3.2. Assignment to tracks

A particular feature of Dutch education is that the assignment of students to tracks can happen at different stages. Initial sorting of students upon entry of secondary education (grade 7) is based on two instruments. One is the standardized high-stakes exit test taken at the end of primary education (grade 6), the so-called CITO test. The obtained score on the test is tied to a recommendation for a secondary school track. Additionally, the 6th grade teacher provides a track recommendation for each student. This recommendation is provided after the test score is known, and correlates highly (around 0.85) with the recommendation from the test. Students can receive "mixed" recommendations (e.g. T3/T4) when considered to be at the margin of the required level for a certain track. It is possible for students and parents to not comply with the track recommendation, if the secondary school allows this. It is common practice for secondary schools to set a certain minimum CITO-score and/or a minimum teacher recommendation as an entrance requirement.[7] Appendix Figure A3 shows the allocation to tracks by teacher recommendation. Students are more often (initially) sorted to a track above their recommendation than to a track below their recommendation. This is most prominent for students with T3 or T3/T4 recommendations.

While tracking can occur from the start of secondary education, it can also be postponed to grade 8 or grade 9. This occurs through the existence of comprehensive "bridge classes", where students of two or more tracks are still kept together. This is most common for students in the highest two tracks. In our sample, 90% of T3 and T4 students is in a comprehensive grade for at least one year, and 35% for 2 years. For T1 and T2, 70% of students are already tracked in grade 7. Bridge classes can have different mixtures of (prospective) tracks, with T3/T4 being the most prevalent combination.[8]

Later tracking follows a similar approach as tracking at the point of secondary school entry: schools set a minimum threshold, often based on average grades, but there still exists a grey area in which students' and parents' preferences can be decisive. Both the length of the bridge class and the achievement threshold are up to the discretion of each individual school (federal policy only prescribes that students need to be tracked at the start of grade 10). The use of bridge classes induces variation in the age at which students are being tracked. Students who do not attend a bridge class are tracked at age 12 (grade 7), students who attend a one-year bridge class are tracked at age 13 (grade 8) and students who attend a two-year bridge class are tracked at age 14 (grade 9). It is this variation in tracking age that we exploit in the empirical analysis.

Because bridge classes predominantly occur for T3 and T4, our empirical analysis focuses on these two tracks. T4 is typically labelled as the classical "academic" track and T3 as the "intermediate" track, but they both provide theoretical education and are both non-vocational.[9] The formal curricula for these tracks, which specify the set of subjects schools need to teach and the educational goals students should master at the end of each year, are the same in grades 7 through 9 ("junior high"). However, schools are free to differentiate subject matters and, since students in T4 are of higher ability, the material in this track is typically more advanced (comprising the "instruction effect" specified in Section 2). In senior high school (grades 10 and above), students still follow the same subjects in each track but exact subject matter differs more strongly. This is mainly because T4 lasts until grade 12 while T3 lasts until grade 11. As all students are tracked when in senior high, the early tracking effect involves differences in instructional difficulty, but

[6] Since 1999, the lowest two tracks are merged into the *vmbo* track, but the cohorts we analyse are from before this transition.

[7] Korthals (2012) reports that around 88% of students in the Netherlands attend secondary schools that always consider entrance requirements, which is among the highest rates worldwide. Dutch schools are lawfully obliged to consider at least one of the two instruments when allowing and sorting students, but are free to determine their exact assignment rule.

[8] Of all T3 and T4 students that are not tracked in grade 7, 59% is in a T3/T4 class, 34% in a T2/T3/T4 class, 5% in a T2/T3 class and 2% in a T1/T2/T3/T4 class. For grade 8, these numbers are 81%, 13%, 5% and 2%, respectively.

[9] The T3 track is therefore more theoretical than the German intermediate *realschule*.

no difference in formal curricula. This is in contrast to most literature in this area, in which the different treatment conditions typically involve different degrees of academic versus vocational education.

4. Data

This study relies on several data sources. First, we use data from the Dutch Secondary Education Cohort Studies (VOCL), conducted by Statistics Netherlands and the Groningen Institute for Educational Research (Driessen & Van der Werf, 1991; Statistics Netherlands, 1991). These are longitudinal surveys of Dutch students that are followed across secondary education. We use data from the 1989 and 1993 cohorts of VOCL. The name of the cohort refers to the year the students entered secondary education (grade 7). VOCL registers the track recommendation that students received at the end of 6th grade, and takes baseline tests at the start of grade 7. The track recommendation data are the basis of the subgroup analysis by ability that is a central part of the empirical analysis. For the remainder of this paper, we refer to students with T3, T3/T4 and T4 recommendation as students of respectively low, medium and high ability.[10]

VOCL further records the track that students attend each year, including any combination of mixed tracks in junior high. Additional background information is collected from student and parental questionnaires. VOCL also contains test scores from grade 9, for math and language. There are concerns about the reliability of these 9th grade test scores, and we therefore report results for these outcomes in Appendix E and focus on educational attainment and labour market outcomes in the main paper.

These data are matched with administrative information from the System of Social Statistical Datasets (SSD) on educational attainment and labour market outcomes. SSD contains information on attendance and completion of any post-secondary study program. Registered educational attainment data are available until 2008, when the 1989 cohort is 31 and the 1993 cohort is 27 years old. The share of (T3 and T4) students still in education in 2008 equals 3.1% for the 1989 cohort and 9.3% for the 1993 cohort. Labour market information was made available for the period 2001–2007, implying that wages can be observed from the early 20s until age 30 (the 1989 sample in 2007). Earnings are registered for the month of September in that particular year. September earnings are seen as most representative, because they are not affected by end-of-year bonuses or vacation pay. The level of attrition in the data is very low; 98% of students participating in VOCL 1989 and 99% of students participating in VOCL 1993 are retrieved in the registered data. Sample sizes per cohort equal 4709 and 5644 (taking only T3 and T4 students).

We further collect data on school tracking policies from the Educational Inspection Office. These data describe how students are allocated to tracks for every grade in every school in the Netherlands, which includes every combination of mixed tracks in junior high (e.g. T3/T4, T2/T3/T4 etc.). This is measured for the year 1997.[11] We deduce the relative supply of early tracking schools for every municipality in the Netherlands, based on all secondary schools (see Section 5.2).

We assess the impact of early tracking across several measures of educational attainment, to capture the complete dynamic of treatment effects across the educational career. The main outcomes are: placement into the academic track (T4), completion of T4, completion of higher education (comprising both professional higher education and university) and completion of university. We only include the 1989 cohort when estimating labour market effects, as individuals in the 1993 cohort are at most 26 years old in the labour market data. As many are still in education or at the very start of working life, their earnings potential is underestimated (results are reported in the appendix). We focus on earnings data from 2004 to 2007, when students from the 1989 cohort are 27–30 years old. We take these years since over 90% of the 1989 cohort has left full-time education by 2004. We use the administrative gross monthly earnings as a main outcome, as well as a recoded 'mean wage' variable. The latter correct for the full-time equivalent (FTE) of the main job. As all earnings are measured in logs, they exclude those with zero earnings (the estimate of early tracking on the probability of having no earnings is low and statistically insignificant).

Table 1 reports summary statistics for students in the two highest tracks, which forms the estimation sample for the empirical analysis. The two tracks are roughly of similar size and track assignment is stable between the two cohorts. There are small differences in the background characteristics of students across cohorts. The 1993 cohort performs slightly better at baseline tests, has less individuals at either extreme of the socio-economic background indicators and has fewer students living in urbanized areas. Early tracking is more common in the 1993 sample than in the 1989 sample. This difference is due to sampling, since there are very few changes in tracking policy among schools that appear in both cohorts. Rates of higher education completion are higher in 1993, which is part of a general trend, while earnings are naturally lower for the younger 1993 cohort. Observations are spread across 443 municipalities in the Netherlands.

Appendix Figure A4 shows the final educational attainment across teacher recommendations. The prevalence of higher education completion is around 60% for students with a T3 recommendation and 80% for students with a T4 recommendation. Looking at university completion only, the respective shares are 20% and 60%.[12]

5. Methodology

5.1. Selection bias

Attending an early tracking school depends on decisions made by students and parents. Appendix Table B1 shows the mean values of several background characteristics and outcomes, separately for those that are tracked early and those that are tracked late. Early tracked students have more favourable characteristics. This is reflected by statistically significant differences in baseline test scores, parental education and social class.[13] This result likely reflects that (parents of) students of high ability have stronger preferences to be selected early into an elite track. Assessing observables across ability groups reveals some heterogeneity in selection (Appendix Table B2). Low-ability students are rather balanced across both types of schools while early tracked students of both medium and high ability are positively selected.

Table B1 also lists outcome variables. Early tracking is associated with more frequent assignment to the academic track. Interestingly, early tracked students have lower completion rates of higher education, even though they have more favourable baseline characteristics. Assuming that conditional selection on unobservables is of the same

[10] From the perspective of the total student population, students with T3 recommendation are at or above the median of the ability distribution, but the terminology is used here to make the *relative* distinction with those with T3/T4 recommendations and T4 recommendation.

[11] The data for the instrument being from a different year naturally involves a loss in first stage power, but is otherwise not a threat to instrument validity (students in schools that switch policy are non-compliers). Moreover, tracking policies by school are very consistent over time; of the 86 schools that appear in both the 1989 and 1993 cohort, only 5 changed their tracking policy.

[12] The share of students with a higher education degree is around 34% for the full sample (all tracks). Population data show that the share of students aged 25–34 with a higher educational degree in 2007 is also 34% (Ministerie van Onderwijs Cultuur en Wetenschap, 2011). Hence, the overall sample appears representative of the total Dutch population in this age group.

[13] Interestingly, the difference is opposite for the intelligence test. The predictive value of this test towards later outcomes is, however, markedly lower than that of the other test scores.

Table 1
Summary statistics.

	1989		1993	
	Mean	Std. dev.	Mean	Std. dev.
Track 3 (*havo*)	0.471	0.499	0.472	0.499
Track 4 (*vwo*)	0.529	0.499	0.528	0.499
Language test	0.731	0.678	0.769	0.647
Math test	0.763	0.696	0.774	0.663
Study skills test	0.712	0.722	0.770	0.686
Intelligence test	0.311	0.840	0.389	0.828
Female	0.509	0.500	0.523	0.499
Urbanisation	3.66	1.28	2.87	1.26
Non-Dutch	0.111	0.315	0.133	0.339
Lowest social class	0.145	0.353	0.131	0.337
Highest social class	0.280	0.449	0.251	0.433
Parent high educ.	0.410	0.492	0.381	0.486
Parent low educ.	0.055	0.228	0.035	0.184
Early Tracking	0.512	0.500	0.653	0.476
Higher education	0.674	0.469	0.719	0.450
University	0.324	0.468	0.325	0.468
Earnings 2007	2979.69	1392.73	2237.30	991.01
Earnings 2006	2748.76	1287.63	1910.49	978.89
Earnings 2005	2419.56	1035.35	1523.20	858.19
Earnings 2004	2239.32	880.24	1281.28	801.84
Wage 2007	3335.57	1332.30	2634.47	935.02
Full-time equivalent 2007	0.887	0.194	0.841	0.245
N	4709		5644	

Notes: The table shows means and standard deviations of all main variables, by cohort. Social class is based on the occupational status of the parents. Degree of urbanisation is measured in five categories. Higher education jointly comprises higher professional education (*hbo*) and university education. Earnings are per month, gross and in euro's. 'Wage' corrects the earnings measure for the full-time equivalent of the main job.

sign as selection on observables, the unbiased ATE is expected to be even more negative. The difference between early tracked students and late tracked students is of opposite sign for university completion, while there are no statistically significant differences for labour market outcomes. The sign of the difference in earnings outcomes is in favour of the later tracked students which again is suggestive of a negative ATE for early tracking.

5.2. Instrument construction

Given the established selection issues, an alternative to OLS is needed to obtain unbiased estimates. We employ an Instrumental Variable (IV) approach. Our instrument exploits variation in the share of early tracking schools across municipalities. This share represents the local supply of early tracking and as such the choice set of each student. It is expected to be a strong positive predictor of the actual age at tracking. At the same time, we assume that it does not correlate with other determinants of future outcomes. In other words, it is assumed that the choice for an early or a late tracking school within the local choice set is selective, but the choice set itself is not.

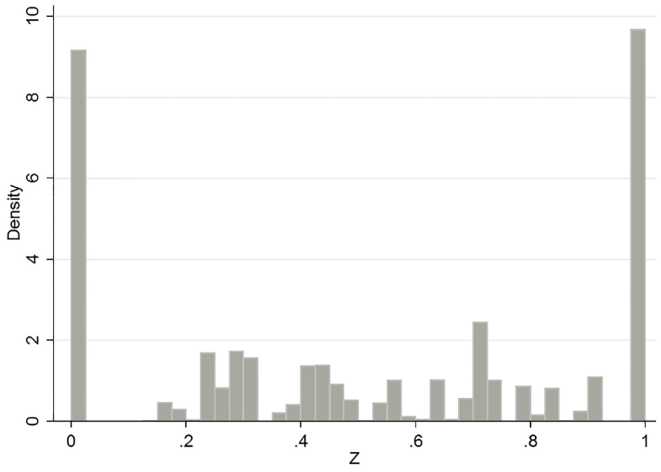

Fig. 4. Distribution of early tracking instrument. **Notes:** The figure shows the distribution of the instrumental variable Z_i. The instrument represents the relative supply of schools that track early (in grade 7 or 8) in the municipality where the student resides.

Using the aforementioned national school-level data, we construct instrument Z_i as the student-weighted share of early tracking schools in the municipality. This involves two steps: (i) We categorize each school as either an early or a late tracking school, depending on whether they separate T3 and T4 students in grade 8 or not (ii) per municipality, we divide the number of 9th grade T3 and T4 students in ET schools by the total number of 9th grade T3 and T4 students (as everyone is tracked in grade 9).

There are two issues in the construction of this instrument. One, some schools track some students early and others late. This applies to 13 out of the 204 schools in the estimation sample (comprising 11.7% of the sample). We choose to categorize these schools as early tracking (i.e. the condition for being an early tracking school is to have at least one class in which T3 students are separated from T4 students). As the majority of students in these 13 schools are tracked in grade 8 and 70% of those that are tracked later are sorted to T3, the earlier tracking moment appears more decisive in these schools.[14] A second issue is that some students do not have a school that offers T3 or T4 in their municipality (around 30% of the sample). In this case, we take the relative supply of the municipality that most students in these municipalities commute to, implicitly assuming that this is their effective choice set. Section 7.1 assesses sensitivity to how students in schools with mixed policies and students in municipalities without a T3/T4 school are dealt with.

Fig. 4 shows the distribution of the instrument. For 47% of the sample, it takes either value 0 (no school in the municipality of residence tracks early) or 1 (all schools in the municipality of residence track early). The remaining 53% has both policies in effect in their municipality. Figure A6 in the appendix shows the number of schools (having T3/T4 students) per municipality. It shows that 77% of students attend a school in a municipality where they can choose between at least two schools that offer T3 and/or T4.

5.3. The IV model

Using the defined instrument Z_i, we specify the following two-stage model to estimate the effect of Early Tracking (ET_i) on outcome Y_i:

$$ET_i = \delta_0 + \delta_1 Z_i + \delta_2 X_i' + \eta_i$$

$$Y_i = \rho_0 + \rho_1 \hat{ET_i} + \rho_2 X_i' + \epsilon_i \quad (6)$$

Y_i can represent educational and labour market outcomes. X_i' is a vector of controls and ϵ_i represents a classical error term. As individual-level controls, we include: month and year of birth, gender, ethnic origin, social class (6 categories), parental education (three categories), and baseline scores for an IQ test and for three test domains (language, math and study skills) of an achievement test. These tests are taken upon entry of secondary education, and can therefore be considered as

[14] We choose a dichotomous classification of schools (rather than using shares for schools with mixed policies) because we want to focus on whether the policy is offered or not. Additionally, the school-level data are for one particular year and therefore do not allow us to follow a cohort of students over time. As such, we do not know the exact share of students within a cohort that is tracked early and therefore classify each school as a whole. We take an alternative approach with (approximate) shares in the robustness analysis.

pre-treatment. We employ robust standard errors, which are clustered at the level of the municipality. As we estimate the effect of ET between the top two tracks in Dutch education, we also restrict the sample to students that are (initially) sorted to either T3 or T4.[15] We also estimate effects by ability subgroup, depending on their 6th grade track recommendation.

5.4. Instrument validity

A valid instrument is both relevant and exogenous. We first assess instrument relevance. The first stage coefficients for Z_i are portrayed in the first row of Tables 3 and 4, for the total sample and subgroups respectively. The final rows provide the Kleibergen–Paap test statistics. As we only use the 1989 cohort for the labour market outcomes, first stage power is lower there. For the T4 subgroup in the 1989 cohort, it is just above the conventional critical value of 10 (Staiger & Stock, 1997) but below the Stock–Yogo threshold of 16.38 (Stock & Yogo, 2005). Earnings estimates for this subgroup are therefore less precise. The instrumentation for the T3 and T3/T4 recommendation subgroups is strong.[16]

Z_i relies on cross-sectional variation within the Netherlands, driven by school-specific policies. This induces potential identification threats. Before we discuss these, the broader question is where the variation in ET policy comes from. The decision when to track students is fully up to school leadership, who operate independently from national and local governments in the Netherlands. Korpershoek et al. (2016) execute a case study among 97 secondary schools in the Netherlands to investigate the motivations for their tracking policy. Schools report that their own pedagogical and didactic beliefs are the main driver of this decision. Capacity constraints and parental preferences are presented as far less important determinants. We interpret this as an indication that variation in tracking age is not predominantly driven by demand-side variables, but rather by what each school leader beliefs is the better choice for the learning process of students.

Given the approach, there are several potential identification threats in this empirical approach. These fall into four main categories: (1) Selective commuting of students to schools, (2) parents selecting their place of residence based on the tracking policy of nearby schools, (3) schools adjusting tracking policies to the preferences of parents and students in their area, and (4) tracking policies of schools being correlated with other important aspects of school policy and school quality.

The first identification threat is largely addressed by relying on the supply of ET schools in the municipality of residence, rather than the municipality of the school. Still, the imputation of the instrument for students that live in a municipality without any secondary school can be partially driven by selective commuting. Section 7.1 addresses the sensitivity of the results to how we deal with this particular group of students.

Identification threats (2) and (3) are inherently similar; they concern the assumption that students and parents living in areas with high Z_i have the same baseline characteristics as students and parents living in areas with low Z_i. This cannot be formally tested, but several arguments validate this assumption in the context of our study. First of all, this is partly mitigated by the institutional setting. In Dutch education, there are no catchment areas that only allow students that live nearby to attend, as is common in e.g. the United States. Parents and students are free to select the primary or secondary school of their choice (provided they pass eligibility thresholds for achievement, in the case of secondary schools). Moreover, as the barriers to start a new school are low and population density is high, there are typically multiple schools available within a reasonable travel distance (see Appendix Figure A5). This lowers the incentive of parents to select their residence based on characteristics of the nearest school. This is corroborated by recent research. Borghans, Langen, Meshcheriakova, and Palacios Temprano (2017) find that the housing premium from living close to a high quality primary school is about ten times smaller in the Netherlands compared to countries in which catchment areas are in effect, while Borghans, Golsteyn, and Zölitz (2015) identify a very weak relation between primary school quality and a range of neighbourhood characteristics. Additionally, the variation in tracking age is predominantly between the second year (grade 8) and the third year (grade 9) of secondary education. Hence, schools typically provides both tracks, and late tracking schools are not compared to strictly categorical schools.[17] Their tracking policy is therefore less salient and less likely to be taken into account when parents choose a region of residence.

Schools may still adjust tracking policies towards characteristics of the students and parents in their area. The data contain a wide range of background characteristics, as well as baseline tests taken at the beginning of secondary school. We run a regression of the instrument on this set of controls, for the sample as a whole and for the ability subgroups. Results are reported in Table 2. They show no evidence of a relation between Z_i and observable characteristics.[18] A related issue is that Z_i may affect how primary school teachers give track recommendations, which would imply that the specified ability subgroups are not comparable across municipalities. Table 2 shows that there is also no relation between Z_i and the 6th grade track recommendation. The robustness analysis will assess instrument validity towards additional parental indicators.

While the instrument does not correlate with any observed individual characteristic, we identify a small but statistically significant correlation with the degree of urbanisation of the area.[19] Relatedly, Z_i differs across provinces of residence. Geographical indicators can have an independent effect on the outcomes, for example through differences in the availability of higher educational institutions or local labour market conditions. Some evidence against the latter issue is already provided by the fact that Z_i does not correlate with any of the social status dummies (which are based on parental occupation). Nonetheless, to assess whether the estimates of ET operate through urbanisation, we additionally control for urbanization dummies (five categories) and province of residence (twelve provinces exist).

Finally, schools with different tracking policies might differ in other characteristics. In particular, the different pedagogical beliefs that are behind the tracking policy choice can be reflected in other policies. We analyse the relation between the instrument and a range of school characteristics, relying on questionnaire data from VOCL and administrative data. The results of a regression of the instrument on these characteristics are reported in Appendix Table C1. We do not identify statistically significant correlations between Z_i and school denomination, student-reported school quality, degree of ability-grouping *within*

[15] A potential concern with limiting the sample is that ET can also affect the probability of being at least in T3. We find that there is no relation between the instrument and the probability of attending T3 or higher. The share of students that ends up in tracks below T3 is also very limited for the recommendation subgroups we focus on in the heterogeneity analysis (see Appendix Figure A3).

[16] Some students in the top two tracks have a track recommendation below T3 (2049 observations), but the first stage power is consistently too low for this subgroup. Additionally the track recommendation is missing for 402 students. As such, the sample sizes for the three portrayed subgroups do not add up to that for the total sample. Appendix Table G1 reports results when including recommendations below T3 in the low-ability subgroup.

[17] Around 9% of schools are categorical, representing around 5% of the student population; we assess to what extent this group drives results in Section 7.1.

[18] This is in contrast with other studies that use an IV approach to estimate (early) tracking effects; e.g. the instrument from both Van Elk et al. (2011) and Galindo-Rueda and Vignoles (2005) correlates with baseline test scores.

[19] The correlation equals 0.035. It is not strictly linear across the five urbanization categories, which is why we include dummy variables.

Table 2
Instrument validity: Correlation with observable characteristics.

	Full sample	LA	MA	HA
Age (in years)	0.001	0.004	0.000	0.004*
	(0.001)	(0.002)	(0.002)	(0.002)
Age (in months)	-0.013	-0.032	0.000	-0.010
	(0.009)	(0.020)	(0.016)	(0.016)
Female	-0.001	-0.007	0.002	-0.023
	(0.008)	(0.013)	(0.013)	(0.015)
Non-Dutch	0.006	-0.008	0.038*	0.021
	(0.012)	(0.020)	(0.020)	(0.020)
Parent high education	0.005	0.002	0.0090	0.000
	(0.011)	(0.016)	(0.016)	(0.015)
Parent low education	0.030	0.049	0.0068	-0.043
	(0.020)	(0.030)	(0.038)	(0.045)
Social class cat. II	0.024	0.013	0.028	-0.010
	(0.017)	(0.022)	(0.029)	(0.042)
Social class cat. III	-0.013	-0.021	0.026	0.002
	(0.017)	(0.025)	(0.026)	(0.034)
Social class cat. IV	0.003	0.007	0.019	0.026
	(0.015)	(0.023)	(0.024)	(0.031)
Social class cat. V	0.009	-0.033	0.030	0.004
	(0.015)	(0.022)	(0.025)	(0.036)
Social class cat. VI	0.011	-0.005	0.027	-0.043
	(0.014)	(0.023)	(0.031)	(0.039)
Intelligence test score	-0.011	-0.002	-0.008	0.012
	(0.009)	(0.010)	(0.012)	(0.010)
Language subscore	0.005	-0.002	0.000	0.017
	(0.008)	(0.013)	(0.011)	(0.013)
Math subscore	0.004	0.001	0.008	-0.007
	(0.008)	(0.011)	(0.013)	(0.017)
Study skills subscore	-0.015*	-0.022	-0.017	-0.020
	(0.009)	(0.013)	(0.010)	(0.018)
Track rec. MA	0.018	-	-	-
	(0.024)			
Track rec. HA	0.029	-	-	-
	(0.037)			
Joint significance	0.438	0.434	0.331	0.252
Joint significance with geo	0.000	0.000	0.000	0.000

Notes: *Significant at 10% level **Significant at 5% level ***Significant at 1% level The table shows results of a regression of the instrument Z_i on the referred list of control variables. Separate regressions are run for the sample as a whole and for each track recommendation subgroup. 'Joint significance' gives the p-value of a joint significance test on all individual controls, and separately on all reported variables plus urbanisation and province dummies ('geo'). For social class, I = blue collar, II = self-employed, III = low skilled, IV = medium-skilled, V = high-skilled, VI = not employed. Standard errors are between parentheses and are robust and corrected for clustering at the municipal level.

size. Schools that track early have a total student population that is 15% larger than that of schools that track late. There is no correlation with class size. It appears unlikely that school size has a large independent effect on outcomes that biases our estimates. Including school size as an additional control has no impact on the estimated coefficients, as it has no explanatory power towards any of our outcomes.

6. Results

We estimate the effect of Early Tracking (ET) on educational and labour market outcomes. Estimates using OLS are portrayed in Appendix Table B3, for comparative purposes. We expect these to be positively biased given the results from Table B1. OLS and IV estimates are not fully comparable since the former are ATE and the latter are LATE, but the OLS estimates are still informative as they likely represent upper bounds of the ATE. OLS results indicate a negative relationship between ET and completion of higher education, which is concentrated in the lower recommendation subgroups. The estimated relation between ET and university completion is positive for the subgroups with the highest track recommendation. For earnings, we observe negative coefficients for low-ability and, especially, medium-ability students, but a strong positive estimate for high-ability students.

We now discuss results for the main IV model, separately for the sample as a whole and for the ability subgroups.

6.1. Full sample

The IV estimates of the effect of ET for the full sample can be observed in Table 3. The table presents three different models; one without control variables, one adding individual-level controls and one further adding geographical controls (urbanization and province dummies). The results are highly consistent across the three models, for all outcomes. The consistency between model I and II confirm results from the previous section that the instrument is not related to individual background characteristics. As discussed before, the instrument correlates with urbanization and provincial dummies, which also implies a slight reduction in first stage power in model III. However, the results indicate that the estimates do not operate through this channel.

Table 3 shows that ET increases the probability of being selected in the highest track, by around 12 percentage points (p.p.). Relative to an average incidence of T4 assignment of 53%, this is a sizable effect. However, it does not increase the probability of ultimately completing

Table 3
Long-run effects of early tracking: main analysis (IV model).

	I	II	III		I	II	III
First stage	0.475***	0.477***	0.491***	First stage	0.563***	0.573***	0.579***
	(0.060)	(0.059)	(0.062)		(0.095)	(0.093)	(0.114)
T4 assignment	0.138***	0.118**	0.115***	Mean earnings	-0.128***	-0.123***	-0.127***
	(0.048)	(0.048)	(0.043)		(0.041)	(0.039)	(0.047)
T4 diploma	0.025	0.016	0.002	Mean wage	-0.033	-0.029	-0.027
	(0.041)	(0.035)	(0.035)		(0.025)	(0.023)	(0.028)
High education	-0.089***	-0.085***	-0.105***	2007 earnings	-0.131***	-0.136***	-0.136***
	(0.029)	(0.024)	(0.027)		(0.039)	(0.034)	(0.044)
University	0.038	0.032	0.019	2007 FTE	-0.043***	-0.048***	-0.048***
	(0.036)	(0.032)	(0.029)		(0.014)	(0.013)	(0.017)
KP stat	119.15	120.76	102.73		75.44	83.54	51.91
Ind. controls		yes	yes			yes	yes
Geo. controls			yes				yes

Notes: *Significant at 10% level **Significant at 5% level ***Significant at 1% level 'First stage' estimates are for Z_i. Labour market effects are estimated for the 1989 cohort only. 'High education' jointly comprises higher professional education and university. See Table 2 for the list of controls. Standard errors, between parentheses, are robust and corrected for clustering at the municipal level.

class, quality and quantity of school counselling, frequency of group work, heterogeneity in teaching styles, or use of low-stakes versus high-stakes testing. There is a statistically significant correlation with school

T4. Hence, many early tracked students fall back to T3 after initial assignment to T4. This is confirmed by Appendix Table E2, which gives a more detailed overview of educational attainment results. Moreover,

Table 4
Long-run effects of early tracking by ability: Subgroup analysis (IV model).

	LA	MA	HA		LA	MA	HA
First stage	0.545***	0.504***	0.435***	First stage	0.616***	0.671***	0.390***
	(0.070)	(0.077)	(0.078)		(0.103)	(0.103)	(0.116)
T4 assignment	0.014	0.140***	0.176***	Mean earnings	-0.145**	-0.192***	0.086
	(0.061)	(0.053)	(0.059)		(0.060)	(0.057)	(0.146)
T4 diploma	-0.123***	0.058	0.147**	Mean wage	-0.023	-0.098***	0.084
	(0.047)	(0.063)	(0.071)		(0.036)	(0.037)	(0.096)
High education	-0.136***	-0.102**	0.053	2007 Earnings	-0.130**	-0.237***	0.106
	(0.044)	(0.042)	(0.062)		(0.060)	(0.059)	(0.166)
University	-0.030	-0.034	0.175**	2007 FTE	-0.041	-0.076***	0.064
	(0.038)	(0.054)	(0.076)		(0.025)	(0.021)	(0.055)
KP stat	90.45	61.00	39.20		76.85	82.78	14.14
N	2954	3004	1944		1420	1404	915

Notes: *Significant at 10% level **Significant at 5% level ***Significant at 1% level Estimates are shown for those of low ability (T3 recommendation; LA), medium ability (T3/T4 recommendation; MA) and high ability (T4 recommendation; HA). Standard errors, between parentheses, are robust and corrected for clustering at the municipal level.

ET leads to the attainment of fewer higher education diplomas, by around 10 p.p. (compared to an average incidence of 70%). The estimate for university completion is positive but statistically insignificant.[20]

The right side of Table 3 presents results for the effect of ET on labour market outcomes. These are estimated for the 1989 cohort only.[21] ET reduces the average monthly earnings between age 27 and age 30 by around 14%. Subsequent rows show that the effect of ET on the mean wage is near zero, while ET significantly reduces the FTE of the job (by around 0.05, relative to a mean of 0.86). Hence, earnings effects occur because, given employment, early tracked students work fewer hours. The results indicate that labour market effects mainly operate at the lower margin, especially given that average full-time equivalents are high in this sample. Similarly, we identify no effects when we exclude the lowest quantile of earnings (not shown).

6.2. Heterogeneity by ability

Table 4 reports estimates of the effect of ET for the three ability subgroups separately. This reveals a different dynamic in the results for low-ability students versus medium-ability students. The low-ability subgroup experiences no increase in T4 assignment, but a substantially higher probability of retracking to T3 after initial assignment to T4 (see Appendix Table E2). These results are suggestive of "inefficient" sorting. ET sends a substantial amount of students to T4 for which the track is too demanding, as suggested by the increase in retracking to T3. At the same time, there likely exists a separate group of T3 students that would have attended and completed T4 under later tracking.[22] Since it is much more common in Dutch secondary education to downgrade students to a lower track than to promote them to a higher track, the result is a relative increase in T3 diplomas at the expense of T4 diplomas (12 p.p.), which also translates into less frequent enrolment in higher education (14 p.p.).

Medium-ability students experience a strong increase in T4 assignment when tracked early (14 p.p.), but no significant increase in T4 diplomas (although the point estimate of 6 p.p. is positive and non-negligible). They subsequently experience a significant decrease in the share of higher education degrees of 10 p.p., while there is no effect on higher education enrolment. For these students, it appears that ET leads to overambitious sorting into T4, which in subsequent years leads to frequent downgrading to T3, and lower persistence in higher education studies.

For both the low-ability and the medium-ability students, ET also translates into lower earnings in early adulthood. The point estimates are stronger for the medium-ability group (19%) compared to the low-ability group (15%). The former also experience a negative impact on their wage, of around 10%. Hence, not all of the effect for this group operates through hours worked. The earnings effect for the medium-ability subgroup continues a pattern of deteriorating effects across age, as the point estimates for this group are positive but insignificant for T4 completion, negative but relatively small for higher education completion and strongly negative for earnings.

The estimates for the highest ability students are consistently positive though have low precision. Similar to the medium-ability subgroup, ET increases the probability of attending T4, by around 18 p.p. This also translates into an increase in T4 diplomas, by 15 p.p. High-ability students do not obtain more higher education diplomas from ET, but do substitute higher professional education (*hbo*) diplomas for university diplomas. The latter increase by a substantial 18 p.p. (the average incidence for this subgroup is 57%). Earnings estimates are positive but statistically insignificant for high-ability students. The point estimates are high, ranging between 8% and 11%, but suffer from high imprecision due to low first stage power. Hence, we cannot draw any definite conclusions for the earnings effects of ET for high-ability students, but the estimates are suggestive of a positive effect. Hence, in addition to inducing a (local) average negative effect, ET also leads to an increase in the inequality of educational attainment and earnings across ability.

6.2.1. Additional outcomes

We further analyze additional outcome variables that may act as mechanisms toward the long-run effects identified before. First of all, we analyse whether ET affects achievement in grade 9. To deal with limitations of the achievement data, we employ several sensitivity analyses. A more elaborate explanation is provided in Appendix E, as are the results. In short, we do not find any strong evidence that ET affects achievement, or acts as a mechanism for long-run effects within any of the ability subgroups.

Secondly, we look at study field choice. Earnings effects of ET are

[20] Table E2 further shows that the effect of ET on higher education diplomas is mainly the result of lower enrolment, rather than completion given enrolment. ET also leads to a (marginally significant) increase in retention rates. This likely relates to the higher probability of misallocation.

[21] Appendix Table G2 shows that the educational outcomes for the 1989 cohort are highly similar to those for the full sample. There is a small difference in the estimates for the full sample but the subgroup estimates are very similar. The former is mainly driven by the stronger first stage for low-ability students in 1989, leading to a stronger weight of this subgroup in the overall coefficient.

[22] An alternative explanation is that late tracked students are downgraded less often from T4 to T3 through better learning during the extra year of comprehensive education. However, estimates appear to be too large to be attributed to differences in the learning environment that only last one year, also given that peer effects should work in favour of early tracked students in T4 in that one year.

large relative to the effects on educational attainment, especially for the medium-ability subgroup.[23] Study choice may be a mediator. We distinguish eight different study fields; results are portrayed in Table E3 in the Appendix. The negative effect of ET on higher education completion for low-ability students mainly comes at the expense of economics degrees. For the medium-ability subgroup, the point estimate is largest for economics as well, and just shy of statistical significance thresholds. For the high-ability subgroup, math and technical degrees increase, at the expense of mainly humanities. As shown at the bottom of the table, average earnings are highest in economics. This is partly because FTE's are comparatively high in economics. As such, part of the earnings and labour supply effect of ET operates via study choice. We have seen that ET leads to a lower inclination of entering higher education, also when students are eligible to do so. The results for study choice may indicate that ET also leads to a lower selection of majors that can be perceived as more challenging.

6.3. Additional heterogeneity analysis

We have also analysed effect heterogeneity by student background (gender, parental education and relative age). We do not identify strong evidence of heterogeneity but estimates are imprecise. Results are reported in Appendix F. There is some indication that negative wage effects are stronger for (low-ability) women. This likely relates to the observation that, while boys obtain slightly better CITO scores than girls, there are substantially more girls in T4 by grade 9. As girls are more likely to outperform their initial ability assessments, they are more at risk of being sorted to a "too low" track when tracking is done early.

Finally, we conduct heterogeneity analysis by endogenous subgroups, to further understand the mechanisms behind the long-run effects. The main results show that ET leads to high initial assignment to T4 but frequent downgrading to T3 in later grades. It remains unclear whether the latter completely drives the negative long-run effects (e.g. through detrimental effects on motivation and ability beliefs). We assess this by estimating effects for subsamples that select on earlier outcomes. This sample selection naturally leads to biased estimates, but the direction of the bias can be predicted and lower or upper bounds can be established as a result. For example, ET has a negative impact on completing higher education. When we estimate the model only for those with a higher education degree, ET students are positively selected (as they need to overcome the 'penalty' from being tracked earlier). The results then represent a conservative estimate of the negative impact of ET on earnings, net of its effect on higher education completion. Similarly, we can select a subsample of students who obtained a T4 diploma. As there is no (average) effect of ET on T4 diplomas for the full sample, the bias from selecting on this outcome is likely small.

Selecting only students with a T4 diploma reduces the coefficient for higher education completion substantially, but it remains negative and just shy of statistical significance (see Appendix Table G3). This suggests that track switches explain a substantial part of the relation between ET and higher education completion, but does not preclude a negative impact for those that remain in the academic track. Labour market effects are smaller but still negative and statistically significant. When we select only students with a higher education diploma, labour market results similarly show smaller but still substantially negative point estimates. This suggests that the effect of ET on earnings operates through more than just lowering higher education diplomas. Results for the subgroups are also reported, but they are more difficult to interpret, as the expected biases have different directions and sizes. Nonetheless, they confirm the main result that these intermediary outcomes are mediators but do not explain away the complete effect.

One reason for this result could be that the more lenient track assignment under early tracking creates a group of students that become less motivated to persist in post-secondary education after following the demanding track T4, also when they complete T4.[24] This also relates to recent findings from Elsner and Isphording (2017) that a student's rank in class impacts post-secondary education decisions, conditional on own ability. Such effects could potentially also explain why early tracked students select into majors with lower expected wages and jobs with fewer working hours.

6.4. Discussion

Coming back to the theoretical framework in Section 2, the results confirm the prediction of especially large gains from later tracking for medium-ability students. This goes against the traditional linear depiction of tracking as a loss for the low-achieving and a gain for the high-achieving. The high negative effect for the medium-ability subgroup could arise through the high probability of misallocation and through instruction effects. While we cannot disentangle these, the strong effects on track switching and the subsequent pattern of results suggest that misallocation is a strong factor. Scenario (b) from Figures A1 and A2, in which assignment is noisy and too lenient, mimics the actual findings most closely.

We find negative effects for low-ability students as well. Negative impacts of ET are more likely at the bottom of the distribution than at the top, because peer quality is negatively affected in the former and positively affected in the latter case.[25] Aside from peer effects, the results for low-ability students also suggest that the ability signal is especially noisy, leading to misallocation also for students whose perceived ability level is further away from the threshold. That this does not lead to negative effects for the high-ability group can be because peer effects operate in the opposite direction there. Moreover, this is a relatively smaller group at the very top of the ability distribution. Retracking even occurs within this group, further highlighting the ability signal's high noise. This is not surprising in light of recent insights from neuroscience and educational investment analysis, which indicate that early adolescence is the period in which cognitive skills are most malleable (Hoxby, 2018). As such, a difference in the tracking age of even one year can lead to very different conclusions about what is a student's optimal track. As these track choices in turn determine secondary and post-secondary career paths, they can potentially have major effects throughout the life cycle.

We emphasize that although the results suggest that a substantial share of the negative effects of ET is driven by putting students in too demanding tracks, one should not interpret this as evidence in favour of keeping the academic track small. The LATE suggests that placement in a high track can have negative consequences for students that are selected in a higher track *because* of early tracking. In other words, these students would not have been in the higher track if they had attended a late tracking school (i.e. under a larger information set). This does not preclude that other students could have benefited from being sent to the high track. In fact, results suggest that a sizable group of students with a T3 recommendation would have.

[23] When measured in years of schooling (coded to the obtained diploma using common standards for the Netherlands), the educational attainment effects are around -0.9 for the low-ability subgroup and -0.5 for the medium-ability subgroup.

[24] Alternatively, this same group of students could effectively learn less in the academic track because of the instruction effect. While we do not identify strong achievement effects in grade 9, these might still arise in upper secondary education.

[25] We note that T3 recommendation students are not the very lowest ability students in the sample, as 15% has a lower recommendation. When including these students within the low-ability subgroup, point estimates become even slightly stronger (see Appendix Table G1).

Additionally, it is somewhat surprising that the decrease in T4 diplomas translates into a decrease in diplomas for higher professional education rather than university education. The eligibility threshold for university lies between tracks T3 and T4, while both T3 and T4 students are eligible for higher professional education. However, even though students that complete T3 are eligible for (part of) higher education, they enrol to a significantly lower extent than T4 students (68% versus 90% in the full sample and 58% versus 79% for LA). The effect of ET on higher education for the low-ability subgroup might therefore operate through conforming to peer behaviour rather than through formal eligibility.[26]

7. Robustness

We conduct several robustness checks that centre around the threats to the validity of the instrument that have been specified in Section 5.4.

7.1. Construction of the instrument

We first assess sensitivity to the exact construction of Z_i. Results for the full sample are reported in Table 5. Robustness results by subgroup are provided in Appendix D.

The main analysis imputes values for students without a T3 or T4 school in their municipality (around 30% of the sample) by relying on the municipality that most students commute to. This imputation may be a threat to instrument validity, as it is based on choices of parents and students where to commute to, and thus prone to selective mobility. As a robustness check, we exclude all students for which the instrument is thusly imputed. The results for this alternative approach are highly similar to those of the base model. We also report results from a model in which we strictly rely on the supply of ET of the municipality where the school resides. The long-run estimates are slightly less negative in this case, suggesting that there is some selective mobility of better students to municipalities with a high supply of ET schools, but the differences are small in magnitude.

Another potential issue is that Z_i is based on student counts. We choose this approach because the supply of ET places should also reflect school size. Moreover, using student shares increases variation and therefore first stage power. However, school size is co-determined by students' enrolment decisions. We alternatively construct Z_i using the share of ET schools, not weighted by school size. Table 5 shows that the pattern of results is very similar. The estimate for T4 assignment is statistically insignificant in the full sample, but effects for this outcome are still positive and strongly statistically significant for the medium-ability and high-ability subgroups, while remaining statistically insignificant (now with negative sign) for the low-ability subgroup. Results are also similar when we replace the dichotomous classification of ET schools with (approximated) shares for the 13 schools that have mixed policies ('student share' column).

Additionally, we have generally classified the effect of ET in our sample as the effect of tracking in grade 8 versus tracking in grade 9, while around 10% of the sample is tracked in grade 7 already. These students could be especially important as they are more constrained by their track recommendation and more at risk of misallocation. Moreover, one might be more concerned about instrument validity within this subgroup, as it also contains students that are in categorical schools. When we exclude those tracked in grade 7 from the estimation, results are highly similar to those for the full sample (Appendix Table G4).

[26] Several studies show the substantial effect that peers can have on enrolment decisions; see, e.g., Bobonis and Finan (2009) and De Giorgi, Pellizzari, and Redaelli (2009)

7.2. Estimation by degree of urbanization

Analysis has shown that there is a correlation between Z_i and the degree of urbanization, but that the estimated effects do not operate through this channel. We further assess to what extent the effects of ET are consistent across rural and urban areas. The nature of the variation in Z_i in each of these areas differs. Rural municipalities typically have only one school and the instrument will equal either 0 or 1, while urban municipalities typically offer both options but with variation in the relative shares between 0 and 1. One might consider the former source of variation as 'cleaner'. The single school has to decide one of the two options and that single decision leads to large variation in Z_i. On the other hand, when there are two municipalities with 5 schools and one has 4 ET schools and the other has 1 ET school, the threat of systematic differences between these two municipalities likely is larger. If urban areas completely drive the results, this would therefore be a concern. Conversely, when rural areas completely drive the results this could be problematic given that the lower supply of schools in the area (see Figure A5) increases the threat of parents choosing their place of residence based on ET policy.

The final columns of Table 5 portray results separately for municipalities with low urbanization (highest two categories) and high urbanization (lowest three categories). The cut-off is chosen so that the first stage power is roughly equal in each case. There are some differences in exact magnitudes, but we identify the same pattern of results in both cases. The fact that we obtain the same overall conclusions, relying on variation in Z_i that is of different nature and is differently vulnerable to the identification threats specified before, provides further evidence in favour of the validity of our results.

7.3. Parental attitudes and investments

Earlier analysis shows that individual background characteristics are not correlated with Z_i. Parents can still differ in attitudes and beliefs. When parental attitudes towards tracking ages influence ET policies of the schools in their area of residence, and when these attitudes also independently affect long-run outcomes, the estimates will be biased.

We use data from VOCL parental and student questionnaires to assess whether the instrument correlates with parental attitudes and investments. Results are provided in Appendix Table C2, for the full sample and the ability subgroups. The data contain parental involvement measures on homework help, talking about school at home and providing encouragement to work hard in school, for both the mother and the father. They further measure attendance of PTA meetings, and museum and library visits. None of these indicators correlates significantly with Z_i. We do identify a statistically significant negative correlation with the size of the parental social network. Given the number of variables we assess, this could simply be the result of multiple hypothesis testing. Moreover, the variable does not correlate with any outcome variable, so it appears unlikely that this drives the pattern of results. We find no associations with parental gender attitudes (which are potentially important in light of the labour supply effects for women), parental educational aspirations, and authoritarian parenting styles.

7.4. School quality

ET schools may differ in other aspects than just their tracking policy. We have already shown that Z_i does not correlate with a range of school characteristics, other than school size. A specific issue in light of the identified results is that schools that keep students of different ability together for a longer time might be more concerned in general about inequality and therefore especially invest in improving outcomes of low achieving students. We therefore assess whether ET affects the variation in 9th grade test scores but identify no effects. Moreover, if later

Table 5
Robustness to alternative specifications.

	Base model	Exclude no school	School municipality	School share	Student share	Low urbaniz.	High urbaniz.
T4 assignment	0.115***	0.139**	0.112***	0.065	0.113***	0.171***	0.087
	(0.043)	(0.054)	(0.039)	(0.044)	(0.041)	(0.054)	(0.060)
T4 diploma	0.002	0.029	0.006	-0.029	0.010	0.013	-0.013
	(0.035)	(0.044)	(0.030)	(0.035)	(0.034)	(0.049)	(0.046)
High education	-0.105***	-0.111***	-0.064**	-0.079***	-0.099***	-0.069*	-0.131***
	(0.027)	(0.033)	(0.026)	(0.027)	(0.026)	(0.036)	(0.039)
University	0.019	0.021	0.029	0.013	0.016	0.015	0.0037
	(0.029)	(0.037)	(0.024)	(0.029)	(0.027)	(0.037)	(0.038)
Mean earnings	-0.127***	-0.117**	-0.099***	-0.083*	-0.113***	-0.153*	-0.103*
	(0.047)	(0.053)	(0.036)	(0.049)	(0.044)	(0.080)	(0.054)
Mean wage	-0.027	-0.023	-0.011	0.002	-0.024	-0.003	-0.034
	(0.028)	(0.032)	(0.025)	(0.030)	(0.027)	(0.042)	(0.033)
2007 earnings	-0.136***	-0.110**	-0.095***	-0.110**	-0.130***	-0.159*	-0.120**
	(0.044)	(0.046)	(0.031)	(0.046)	(0.041)	(0.082)	(0.053)
2007 FTE	-0.048***	-0.040**	-0.051***	-0.061***	-0.044***	-0.084***	-0.032
	(0.017)	(0.019)	(0.014)	(0.019)	(0.016)	(0.028)	(0.022)

Notes: *Significant at 10% level **Significant at 5% level ***Significant at 1% level The columns report, in order: estimates from the main IV model; estimates when excluding students without a school in their municipality; estimates when Z_i is based on municipality where the school resides; estimates where Z_i is based on school shares rather than student shares; estimates where Z_i is based on student shares for schools with mixed policies; estimates on a sample from low urbanized areas; estimates on a sample from high urbanized areas. All estimations use the full set of controls. Standard errors are between parentheses and are robust and corrected for clustering at the municipal level.

tracking schools would indeed invest more in low-ability students, one would also have expected a negative effect of ET on achievement of low-ability students, and this would also not explain the fact that the long-run effects are largest for medium-ability students.

Given that ET policies are largely based on pedagogical beliefs of the principal or school board, we acknowledge that it is unlikely that ET schools and late tracking schools are completely identical in how they otherwise teach students. One might therefore claim that we rather estimate the effect of being in a (self-determined) ET school rather than the effect of ET per se. Nonetheless, data show that this is not reflected in a range of key objective measures of school policy and quality. Moreover, the hypothesis that school quality differences are driving the results would be inconsistent with the lack of an achievement effect, and with the heterogeneity in effects across ability.

7.5. Students still in education

Some students have not completed their education yet in the final year for which the data are available. This applies to 3.1% of those from the 1989 cohort and 9.3% of the 1993 cohort. Table G5 in the appendix shows that the estimates are not sensitive to this limitation. Most importantly, the probability of still being in education at this point is not affected by ET, across cohorts and ability groups.

The scope of the data is also relevant for the interpretation of the earnings effects, as these are only estimated in young adulthood. Still, effects for 2007 earnings are very similar to those for average earnings in 2004–2007. We find that negative wage effects are statistically significant from age 27 onwards, and do not increase afterwards (see Appendix Table E4). This also explains the largely insignificant results for the 1993 cohort, as these students are 26 years old in 2007. The signs of the estimates for the 1993 cohort point in the same direction.[27]

One would preferably estimate earnings effects over the complete lifetime. Given that ET induces lower average educational attainment, which typically implies flatter age-earnings profiles (Borjas, 2015), we may underestimate the negative impact of ET on earnings. Still, the fact that estimates are rather stable between ages 27 and 30 while average earnings are strongly increasing for these ages is suggestive evidence that the effect is constant once people are settled in the labour market. Moreover, Bhuller, Mogstad, and Salvanes (2011) show that the lifetime earnings impacts of an extra year of schooling are closest approximated by earnings impacts in the early 30's. The 2007 earnings effect may therefore be a reasonable approximation of the average lifetime earnings effect of ET.

8. Conclusion

This study estimates the impact of the age of tracking on long-run outcomes, for students in the intermediate and academic track in the Netherlands. We use the relative supply of early tracking schools at the municipal level as an instrument for being tracked early (grade 7 or 8 versus grade 9). Results show that early tracking negatively affects the probability of obtaining a higher education degree (by 10 p.p) as well as earnings at age 30 (by 14%). Earnings effects are largely driven by decreases in hours worked, and are larger for women. Negative earnings effects are present for students of both low ability and medium ability, but are the result of different dynamics. Point estimates of the effect of early tracking on high-ability students are imprecisely estimated, but strongly suggestive of a positive effect. Hence, early tracking in the Netherlands negatively affects both efficiency and equality in long-run outcomes. The pattern of results appears largely driven by strong misallocation of students to tracks under early tracking, when the information set on the student is smaller. These effects arise not only from "late bloomers" who are put in tracks below their potential, but also from initial assignment to a track that is too demanding. The results suggest that misallocation (partly mediated by subsequent downgrading in later years) affects decisions that students make in post-secondary education and thereafter, in terms of how long they persist in education, what study field they pursue and how many hours they work. Higher peer quality (for the low-ability students) and better-targeted instruction (for the medium-ability students) could further contribute to the favourable outcomes for later tracking.

Our reliance on (static) within-country variation in tracking ages for a settled tracking system contrasts our study with those that rely on policy changes for identification. The downside of the IV approach is that the validity of an instrument cannot be formally tested. While different robustness tests consistently provide evidence in favour of instrument validity, we cannot rule out that unobserved differences in

[27] The estimates appear slightly lower for the low-ability subgroup in 1993, compared to the same ages in 1989. Additional analysis shows that ET increases the probability of having no earnings in the 1993 cohort (in contrast to the 1989 cohort), which are naturally excluded in the log earnings estimates. We identify a marginally significant negative effect of ET on age 26 earnings for low-ability students when including zero-earners (using a level specification).

parental and student characteristics exist between areas with few and areas with many early tracking schools, or that early and late tracking schools differ in other (unobserved) dimensions than their tracking policy alone. We emphasize, however, that if early tracking policies are shaped by the preferences of parents and students in the area, then we would expect the bias to be of the same direction as in the OLS estimation. Correlations show that better achieving students, from more affluent backgrounds, prefer to attend early tracking schools. If these preferences are also reflected in the instrument, any potential bias would be positive in the IV as well, implying even an underestimation of the negative effect of early tracking. Additionally, any substantial bias through differences in school policy or quality appears inconsistent with the specific dynamic and heterogeneity in the identified results.

It is difficult to assess how representative our findings are for other countries. The negative effects we identify appear partly mediated by more lenient initial sorting under early tracking (for medium-ability students) and a lack of upward retracking opportunities (for low-ability students), which may both be particular of Dutch education. On the other hand, sorting more students into a higher track under earlier tracking could be a more general result of higher uncertainty, while upward retracking is also limited in other countries that track early. Germany, in contrast, provides more opportunities for upward retracking (Dustmann, Puhani, & Schönberg, 2017). The effects of ET might be less harmful in that context, especially for students of (initially perceived) low ability. As previous literature is focused on a different setting (academic versus vocational education) any definite conclusions on external validity can only be made after future research for other countries in a similar setting. Nonetheless, our results further emphasize that the effect of tracking policy can be different between the short and the long run, as shown before by Hall (2012) and Malamud and Pop-Eleches (2010, 2011), while they confirm the negative long-run effects of earlier tracking for low-ability students, as identified by Meghir and Palme (2005) and Pekkarinen et al. (2009). They also confirm the more general finding that (earlier) tracking increases inequality, which is not offset by gains in efficiency. In our study, this increase in inequality arises when looking across baseline ability, but is also reflected by negative effects that are concentrated at the bottom of the earnings distribution within these ability groups.

We emphasize that our study strictly looks at tracking for students in the academic and the intermediate track. While this makes the analysis complementary to the existing literature that is mainly focused on academic versus vocational students, these results cannot be extrapolated towards vocational education. In the context of the theoretical model, one may expect that both the positive and negative aspects of early tracking are larger in a setting that sorts to both academic and vocational education, as the differences in peer ability and curriculum (and thereby also the cost of misallocation) are bigger as well. On the other hand, allocation may be less noisy when separating between vocational and academic education than when sorting between two non-vocational tracks. Nonetheless, our study shows that even when there are no differences in (academic vs. vocational) curricular focus and the difference in the tracking age is only one year, tracking ages can have substantial effects on students' educational career paths and labour market success.

CRediT authorship contribution statement

By Lex Borghans: Conceptualization, Methodology, Formal analysis, Data curation, Writing - review & editing. **Ron Diris:** Conceptualization, Methodology, Formal analysis, Data curation, Writing - review & editing. **Wendy Smits:** Conceptualization, Methodology, Formal analysis, Data curation, Writing - review & editing. **Jannes de Vries:** Conceptualization, Methodology, Formal analysis, Data curation, Writing - review & editing.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.econedurev.2020.101973 .

References

Betts, J. R., & Shkolnik, J. L. (2000). The effects of ability grouping on student achievement and resource allocation in secondary schools. *Economics of Education Review, 19*(1), 1–15.

Bhuller, M., Mogstad, M., & Salvanes, K. G. (2011). Life-cycle bias and the returns to schooling in current and lifetime earnings. NHH Dept. of Economics Discussion Paper no. 4.

Bobonis, G. J., & Finan, F. (2009). Neighborhood peer effects in secondary school enrollment decisions. *The Review of Economics and Statistics, 91*(4), 695–716. https://doi.org/10.1162/rest.91.4.695.

Borghans, L., Golsteyn, B. H., & Zölitz, U. (2015). School quality and the development of cognitive skills between age four and six. *PloS one, 10*(7), e0129700.

Borghans, L., Langen, M., Meshcheriakova, O., & Palacios Temprano, J. (2017). *Is it more expensive to live next to a good school? Comparing externalities of good and bad primary schoolsPaper presented at the 2017 EALE conference in Sankt Gallen*.

Borjas, G. J. (2015). *Labor economics (seventh edition)*. McGraw-Hill/Irwin Boston.

Brunello, G., Giannini, M., & Ariga, K. (2007). The optimal timing of school tracking: A general model with calibration for germany. In L. Woessmann, & P. E. Peterson (Eds.). *Schools and the equal opportunity problem* (pp. 129–156). MIT Press.

Cunha, F., & Heckman, J. J. (2007). The technology of skill formation. *American Economic Review, 97*(2), 31–47. https://doi.org/10.1257/aer.97.2.31.

De Giorgi, G., Pellizzari, M., & Redaelli, S. (2009). Be as careful of the company you keep as of the books you read: Peer effects in education and on the labor market. National Bureau of Economic Research, no. 14948. 10.3386/w14948.

Driessen, G., & Van der Werf, G. (1991). *Het functioneren van het voortgezet onderwijs. Beschrijving steekproef en psychometrische kwaliteit instrumententenTechnical report.* RION/ITS, Groningen/Nijmegen.

Duflo, E., Dupas, P., & Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in kenya. *American Economic Review, 101*(5), 1739–1774. https://doi.org/10.1257/aer.101.5.1739.

Dustmann, C., Puhani, P. A., & Schönberg, U. (2017). The long-term effects of early track choice. *The Economic Journal, 127*(603), 1348–1380.

Elsner, B., & Isphording, I. E. (2017). A big fish in a small pond: Ability rank and human capital investment. *Journal of Labor Economics, 35*(3), 787–828.

Epple, D., & Romano, R. E. (2011). *Peer effects in education: A survey of the theory and evidence. Handbook of social economics1. Handbook of social economics* Elsevier1053–1163.

Galindo-Rueda, F., & Vignoles, A. (2005). *The heterogeneous effect of selection in secondary schools: Understanding the changing role of abilityCEP Discussion Paper*. Centre for Economic Performance, London School of Economics.

Guyon, N., Maurin, E., & McNally, S. (2012). The effect of tracking students by ability into different schools: A natural experiment. *Journal of Human Resources, 47*(3), 684–721. https://doi.org/10.1353/jhr.2012.0022.

Hall, C. (2012). The effects of tracking in upper secondary school: Evidence from a large-scale pilot scheme. *Journal of Human Resources, 47*(1), 237–269. https://doi.org/10.3368/jhr.47.1.237.

Hanushek, E. A., & Zhang, L. (2006). *Quality-consistent estimates of international returns to skillWorking Paper*. National Bureau of Economic Research.

Hoxby, C. M. (2018). Taking productivity in education seriously: Insights from primary and secondary education. Alfred Marshall lecture. University of Cambridge.

Korpershoek, H., Naaijer, H., & Bosker, R. (2016). *De inrichting van de onderbouw: Onderzoek naar de motieven en de beweegredenen van vo-scholen naar soort brugklasTechnical Report.* Groningen: GION onderzoek/onderwijs.

Korthals, R. (2012). Selection and tracking in secondary education: A cross country analysis of student performance and educational opportunities. ROA Research Memorandum, ROA-RM-2012/14.

Malamud, O., & Pop-Eleches, C. (2010). General education versus vocational training: evidence from an economy in transition. *The Review of Economics and Statistics, 92*(1), 43–60.

Malamud, O., & Pop-Eleches, C. (2011). School tracking and access to higher education among disadvantaged groups. *Journal of Public Economics, 95*(11–12), 1538–1549.

Manning, A., & Pischke, J.-S. (2006). *Comprehensive Versus Selective Schooling in England and Wales: What Do We Know?CEE Discussion Paper 66*. Centre for Economics of Education, London School of Economics and Political Science.

Meghir, C., & Palme, M. (2005). Educational reform, ability, and family background. *American Economic Review, 95*(1), 414–424.

Ministerie van Onderwijs Cultuur en Wetenschap (2011). *Kerncijfers 2006–2010Technical Report*.

Pekkala Kerr, S., Pekkarinen, T., & Uusitalo, R. (2013). School tracking and development of cognitive skills. *Journal of Labor Economics, 31*(3), 577–602.

Pekkarinen, T., Uusitalo, R., & Kerr, S. (2009). School tracking and intergenerational income mobility: Evidence from the finnish comprehensive school reform. *Journal of Public Economics, 93*(7–8), 965–973. https://doi.org/10.1016/j.jpubeco.2009.04.006.

Piopiunik, M. (2014). The effects of early tracking on student performance: evidence from a school reform in bavaria. *Economics of Education Review, 42*, 12–33. https://doi.org/10.1016/j.econedurev.2014.06.002.

Rees, D. I., Brewer, D. J., & Argys, L. M. (2000). How should we measure the effect of

ability grouping on student performance? *Economics of Education Review, 19*(1), 17–20.
Sacerdote, B. (2011). Peer effects in education: How might they work, how big are they and how much do we know thus far? In E. Hanushek, S. Machin, & L. Woessman (Vol. Eds.), *Handbook of the economics of education. 3. Handbook of the economics of education* (pp. 249–277). Amsterdam: North-Holland.
Staiger, D., & Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica, 65*(3), 557–586.
Statistics Netherlands (1991). *Schoolloopbanen en achtergrond van leerlingen: cohort 1989. Deel 1: instroomTechnical report.* The Hague: Statistics Netherlands.
Stock, J. H., & Yogo, M. (2005). Testing for weak instruments in linear ivregression. In J. H. Stock, & D. W. Andrews (Eds.). *Identification and Inference for Econometric Models: Essays in Honor of Thomas J. Rothenberg.* Cambridge University Press.
Van Elk, R., Van der Steeg, M., & Webbink, D. (2011). Does the timing of tracking affect higher education completion? *Economics of Education Review, 30*(5), 1009–1021. https://doi.org/10.1016/j.econedurev.2011.04.014.