TI 2019-038/V
Tinbergen Institute Discussion Paper

The Heterogeneous Effects of Early Track Assignment on Cognitive and Non-cognitive Skills

Maria Cotofan[1]
Ron Diris[2]
Trudie Schils[2]

[1] Erasmus School of Economics
[2] Maastricht University

The Heterogeneous Effects of Early Track Assignment on Cognitive and Non-cognitive Skills*

Maria Cotofan[a], Ron Diris[b] and Trudie Schils[b]

[a]*Department of Economics, Erasmus School of Economics, 3062 PA Rotterdam, the Netherlands*

[b]*Department of Economics, Maastricht University, 6200 MD Maastricht, the Netherlands.*

Abstract

Previous findings on (fleeting) relative age effects in school suggest that, given innate ability, too few younger and too many older students attend academic tracks. Using a regression discontinuity design around school-specific admission thresholds, we estimate the cognitive and non-cognitive effects of track assignment at the achievement margin, across relative age. We find that attending the higher track does not affect cognitive outcomes at any relative age. For older students, attending the higher track increases perseverance, need for achievement, and emotional stability. The results suggest that older students compensate lower ability (given high track attendance) with higher effort.

Keywords: educational economics, school tracking, relative age, non-cognitive skills

JEL Classification: I21, J24

*We would like to thank Robert Dur, Marjolein Muskens, Bart Golsteyn, Roxanne Korthals, Sergio Parra Cely and seminar participants at Maastricht university for their helpful comments and feedback. Data collection has been funded by the government of the province of Limburg (Provincie Limburg), Maastricht University, school boards in primary, secondary and vocational education, and institutes for higher education in Limburg, within the program Educatieve Agenda Limburg

1 Introduction

The use of educational tracking to sort students into different learning environments is common practice worldwide. Tracking has traditionally been criticized because it is argued to enhance inequality by concentrating peer and school quality at the top of the achievement distribution; see, e.g. Hanushek and Woessmann (2006). Another point of critique towards (early) tracking is that students can be misallocated to tracks that do not fit with their abilities, as the ability signals that tracking is based on are noisy (Brunello et al., 2007). Several studies have particularly focused on relatively young students in class as a group that is especially harmed by (early) tracking. It has been well-documented that there are differences in student achievement by relative age, and that these decrease as students grow older; see, e.g., Bedard and Dhuey (2006); Elder and Lubotsky (2009); Mühlenweg and Puhani (2010). As tracking is based on achievement, younger students are sorted more often to tracks that are below their academic potential, especially when they are tracked at early ages. It has been recognized to a much lesser extent that there is an equivalent issue on the other end of the relative age distribution; relatively older students are at risk of being placed above their potential, which can be harmful for their learning development as well. Hence, the effects of tracking can be highly heterogeneous by relative age, which could thereby be an important source of misallocation of students to optimal tracks. Moreover, such potential misallocation can be expected to not only affect cognitive learning development but also non-cognitive skills.

This study analyses the cognitive and non-cognitive effects of secondary school tracking, with a particular focus on heterogeneity across relative ages. The empirical analysis uses data from the Netherlands, where track allocation is strongly based on a high-stakes exit test taken near the end of primary school (age 12). We apply a regression discontinuity design that estimates school-specific thresholds for this exit test from the data, to identify the causal effect of track allocation for students who are at the achievement margin of the top track. We estimate overall effects, as well as interaction terms with relative age to assess heterogeneity across older and younger students. We use data from the OnderwijsMonitor Limburg (OML), which collects longitudinal data on primary

and secondary school students from both administrative sources and surveys, in the Dutch province of Limburg.

We find that attending the higher track has no effects on math and reading achievement, across relative age, but do identify heterogeneous treatment effects for non-cognitive skills. Attending the higher track benefits older students especially in terms of perseverance, need for achievement and emotional stability. These gains in non-cognitive skills are absent for the relatively young in class. We further find that older students who go to the higher track are not more likely to fall back to lower tracks in subsequent grades despite the fact that their cognitive abilities are lower on average (contingent on being in the top track), which could be explained by these compensating spillovers on non-cognitive skills.

Previous studies have examined the interplay between relative age and tracking. Korthals et al. (2016) find that the relation between relative age and academic track attendance is stronger in early tracking countries, but simultaneously identify that the relation between relative age and achievement at the end of compulsory education is smaller in such countries. Moreover, they find that younger students have *higher* wages than their older peers when tracking is done early. Dustmann et al. (2017) provide a key contribution to the field by directly estimating the causal effect of track attendance at the achievement margin, and doing so for both educational and labour market outcomes. The authors use the variation in academic track attendance by month of birth to identify causal effects. They find that attending a higher track does not lead to more favorable long-run outcomes.[1] These studies suggest that the underestimation of the potential of the relatively young in early tracking systems is not necessarily detrimental for their future attainment. However, these studies do not elicit the direct effect of track attendance for either relatively young or relatively old students, because they rely on a comparison between each group. What is typically neglected in discussions around tracking and relative age is that the relatively older students can also be harmed from being sent to a track that is above their potential. For example, the lack of a tracking effect

[1] A similar zero long-run treatment effect of attending a more academic track is found by Malamud and Pop-Eleches (2010, 2011) and Hall (2012), exploiting policy changes in Romania and Sweden, respectively. In contrast, Guyon et al. (2012) identify positive effects in the short and medium run from expanding the elite track in Northern Ireland.

in Dustmann et al. (2017) could be the result of the negative effect of the young being in a too low track being canceled out by the negative effect of the old being in a too high track. If this would be the case, the near-zero effect sizes would hide that there can be considerable welfare gains by reallocating students on each side of the relative age distribution. To assess whether allocation that is partly based on fleeting relative age effects is harmful or not, it is required to directly identify the effects of track allocation across relative age.

Additionally, our study estimates the effects of track assignment for both cognitive and non-cognitive skills. The importance of non-cognitive skills is increasingly recognized in economic literature, see, e.g., Almlund et al. (2011) and Kautz et al. (2014). These studies show that non-cognitive skills are especially malleable in adolescence, more so than cognitive skills. As such, the track environment in secondary education can have potentially important consequences for such skills. Moreover, Mühlenweg et al. (2012) show that relative age affects the development of non-cognitive skills or personality in childhood. Hence, students of different relative ages enter secondary school tracks with a different set of non-cognitive skills. If there exists complementarity between non-cognitive skills and investments, one would expect these skills to be differently affected across relative age by tracking. In order to assess the efficiency of track assignment across relative age, it is therefore essential to look at both types of skills.

Analyzing treatment effects for different types of skills also provides insights into the exact mechanisms driving the effects of track allocation, on which little empirical evidence still exists. Class rank effects are often mentioned as a possible explanation in studies that identify no or weak effects of attending the higher over the lower track (Korthals et al., 2016). In a different context, Elsner and Isphording (2017) show that, conditional on ability, class rank has a strong impact on student expectations, perceived ability and, subsequently, on educational attainment. Such relative rank effects can be especially important in the context of track allocation. On the one hand, being in the high track in itself can lead students to perceive themselves more favorably, but moving from the lower to the higher track at the ability margin simultaneously moves a student from the top to the

bottom of the ability ranking within each track. In that light, it is especially valuable to look into the impact of tracking on non-cognitive outcomes, such as self-esteem, motivation, and perseverance. Estimation of non-cognitive effects of tracking is rare in the literature, let alone exploring their heterogeneity across relative age.

This remainder of this study is organized as follows. Section 2 provides an overview of relevant aspects of the Dutch educational system. A description of data sources is given in Section 3, while Section 4 discusses the methodological approach. Section 5 presents and discusses the main results. Robustness analysis is provided in Section 6. Section 7 concludes.

2 The Dutch Education System

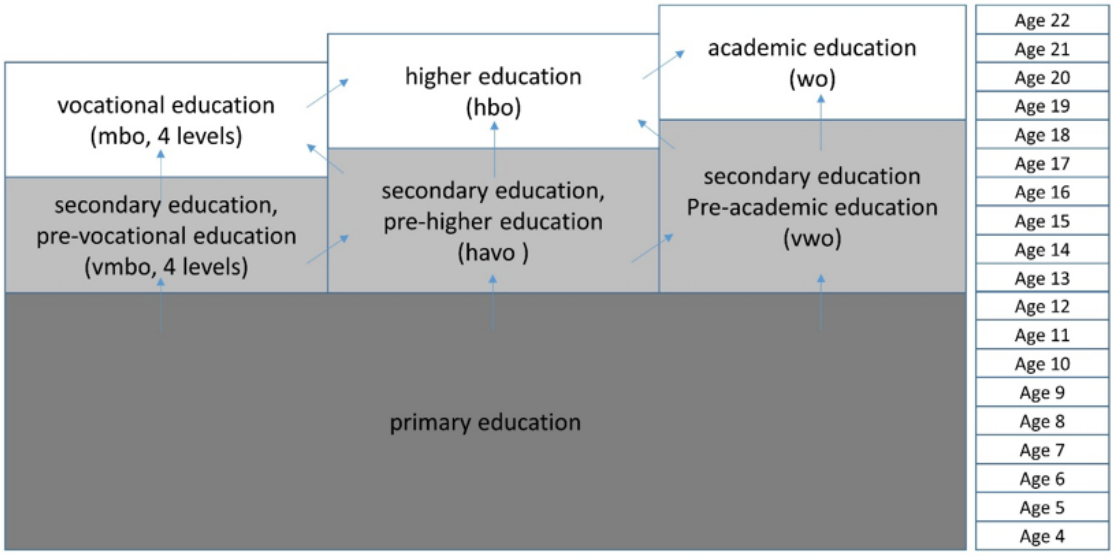
Figure 1 shows an overview of the main stages of the Dutch educational system. Primary education consists of eight years of which the first two are spent in kindergarten. As of the third year of primary school (1st grade), children formally learn how to read and write. Most children start kindergarten at the age of 4, enter 1st grade at the age of 6, and finish primary school at the age of 12. When entering secondary school (grade 7), students are sorted into different school tracks. There are two underlying indicators for the sorting of students to tracks:

1. A standardized exit test: In 6th grade, students take a national achievement test. At the time of data collection, three separate tests were officially approved by the government, and schools are free to choose which to employ, but 95% of schools in the Netherlands administered the so-called CITO test, including all of the schools that appear in our data sample. The testing bureau reports for each range of final test scores the appropriate track pertaining to that score. This can also take the form of a 'mixed' recommendation of two adjacent tracks.

2. A teacher recommendation: In 6th grade, teachers provide a subjective assessment of the child's ability and the associated track in secondary school. In the recommendation, the teacher is supposed to summarize the history of achievement of the student, but also an assessment of

the broader (cognitive and non-cognitive) development, based on the teachers' classroom observations throughout the year. As with the 'test recommendation', the teacher recommendations can be towards a single track, or a mixed recommendation for two adjacent tracks. The teacher recommendation is granted after the results on the exit test and may thus be strongly affected by the test. The correlation between teacher recommendation and test recommendation equals 0.82.

Figure 1: Dutch education system



Note: The age column on the right of the figure shows the average age at each educational stage, not taking into account retention or acceleration. The scheme excludes special needs education.

The Dutch secondary education system is hierarchically structured by ability and consists of three main tracks that differ in duration and qualification. The four-year track (*vmbo*) qualifies children for vocational education, the five-year track (*havo*) qualifies children for higher (professional) education and the six-year track (*vwo*) qualifies children for academic education/university. On average, 55 percent of the children end up in the pre-vocational track, 25 percent in the pre-higher education track and 20 percent in the pre-academic track. The pre-vocational track is further divided into four sub-tracks (*vmbo-b*, *vmbo-k*, *vmbo-g*, *vmbo-t*) that differ in their focus on practical versus theoretical content in the curriculum. Time spent on more theoretically oriented courses increases from *vmbo* b (25% of time) to *vmbo* t (100% of time). The *vmbo-b* and *vmbo-g* tracks have reduced considerably in recent years and in many schools have effectively been integrated

Table 1: Share of students per track recommendation given by 6th grade teacher

Teacher recommendation	% Students	Mean score
vmbo-b	8.8%	522
vmbo-b/k	3.3%	524
vmbo-k	9.8%	527
vmbo-gt	19.4%	533
vmbo-gt/havo	13.4%	537
havo	16.1%	540
havo/vwo	12.9%	543
vwo	16.3%	547

Note: The score on the exit test ranges from 501 to 550.

within the *vmbo-k* and *vmbo-t* tracks respectively. Track recommendations also do not distinguish between the *vmbo-g* and *vmbo-t* track, which are jointly put under the label *vmbo-gt*. Recommendations do distinguish between *vmbo-b* and *vmbo-k* but the former track is relatively small (around 8% of the total student population). Hence, in practice the distinction is between one (largely) practical and one theoretical subtrack within *vmbo*, and the Dutch system as a whole can be argued to have four main tracks.

Table 1 below shows for each track the average score on the primary school exit test and the percentage of students attending it (for all students in the OML data). Roughly half the students in the sample receive mixed recommendations, which are given when the teacher believes a student to be at the margin of two tracks. As shown, a mixed recommendation for *vmbo-k* and *vmbo-gt* is not given (by rule).

Final tracking decisions can still be postponed for one year. Schools have the discretion to keep two adjacent tracks together in grade 7. Around 50% of students are tracked directly in grade 7 and the remainder only in grade 8.[2] These temporary comprehensive classes are most common for the combination of the practical and theoretical subtracks in *vmbo* and for the combination of *havo* and *vwo*.

[2]A small minority of students is only tracked at the start of grade 9. This applies to two schools in our sample, comprising around 4% of all students.

3 Data

The data we use stem from the Dutch Onderwijsmonitor Limburg (OML). This is a cooperative project between Maastricht University and schools, schools boards and government bodies in Limburg, a province in the South of the Netherlands.[3] The province has about 1.1 million inhabitants, and a population density of 520 inhabitants per square kilometer, which is slightly above the average population density in the Netherlands of 502 inhabitants per km^2 (Statistics Netherlands, 2017). The average disposable household income in the region is about 34000 Euro per year, which is somewhat below the national average of 36200 Euro (Statistics Netherlands, 2017). Despite this, average scores on the standardized 6th grade exit test are above the national average.

The OML aims to collect and analyze information about the educational development of students in this region to provide feedback to schools and policy-makers about students' performance. The OML supplements administrative data with surveys among students, parents, and teachers at different stages in the education career, i.e. in kindergarten, grade 6 (final year of primary school), and grade 9 (secondary school). The administrative data include the exit test score and teacher recommendation from grade 6, track placement from 7th to 9th grade. Questionnaires collect additional information on demographic indicators, socio-economic status, and non-cognitive skills of students. Additionally, the OML administers an IQ test in grade 6 and 9th grade test s in math and reading. Parents can decide to withdraw from the survey, based on the passive consent principle. A legal cooperation agreement states that the data collection is allowed to reach the overall goal of the program and is signed by all partners. The data collection is approved by the local ethical committee (ERCIC-092-12-07-2018). Researchers get fully anonymous data files.

The OML collects data for schools across the province in secondary school and for schools in the South of Limburg in secondary school. The coverage of schools is around 90 percent of the full population in that area in each case. For students that attended primary school in the North of the province, data on teacher recommendations and exit test scores are still collected, as they are

[3]For more information, see `http://www.educatieve-agenda.nl/onderwijsmonitor-p/english`.

also registered in the administrative data of secondary schools. Non-participating schools include special education schools, schools with a philosophy not to test children, and schools unable to plan the survey activities (participation takes ones hour per class in primary school and two hours per class in secondary school). Since most schools in the region participate in the project, sample selection problems are expected to be small.

Data collections for the OML have taken place yearly from 2009 onward in primary education and biannually in secondary education from 2010 onward. The basis of our data sample are the 2012, 2014, and 2016 9th grade cohorts, as the 2010 cohort largely lacks information on attended track in grades 7 and 8, as well as a linked 6th grade data collection from three years before. Each cohort contains around 9.000 observations.

Below we describe the data and variables in more detail.

3.1 Demographic variables

A key variable in our empirical analysis is (relative) age. We observe exact birth dates of all students. The cutoff date for formal education in the Netherlands is the 1st of October. Cutoff dates are not strictly enforced but compliance is relatively strong. Due to this threshold, the youngest students born just before the cut-off will be 12 months younger than classmates born at or just after the 1st of October, provided they comply with the cutoff date and do not retain or skip grades. We construct a measure of relative age that equals 12 on the first of October and 0 on the 30th of September, with daily increments in between. Additional background variables are gender, parental education (based on the highest completed level between both parents), family structure (a dummy variable that equals 1 if the student lives with both biological parents), ethnicity, language spoken at home, and the working status of mother and father.

3.2 Cognitive measures

The OML contains data on the high-stakes exit test that students take at the end of grade 6 (see Section 2). The test is standardized for all students and externally graded. It contains 200 multiple-choice questions testing the students on three main domains: Dutch language, mathematics, and study skills.

Data are also available for low-stakes tests on math and language in 9th grade. These tests were developed for the research project and administered digitally.[4] As the time for testing students was limited, not all students were administered both tests. Using a random algorithm, one third of the students were assigned to take only the language test, one third to take only the math test, and one third to take (shorter versions of) both tests. While this reduces the sample size for estimating effects on cognitive outcomes, the randomization should ensure that this does not lead to any bias in our estimates. In addition, there are some missing test data on the school-cohort level, indicating that schools decided not to administer the test in class.[5] As we include school and year fixed effects, this is not a major concern for the internal validity of the point estimates. Exit test scores and background characteristics are not correlated with having missed the test, suggesting that external validity concerns are minor as well. Students from different tracks received items with different difficulty level, but overlap in the test items ensures that we can construct a comparable scale using item response theory (IRT).

The data additionally contain information of the attended track in secondary education in grades 7, 8 and 9, which are all derived from the school administrative records. This includes any indication of mixed tracks that can still occur in grade 7 (and in some rare exceptions in grade 8 as well). Data on 7th and 8th grade placement are retrieved retrospectively from the school administration

[4]The language test contains items on comprehensive reading taken from PISA 2000-2006 tests (Organisation for Economic Co-operation and Development, 2011) and items on spelling and word knowledge taken from a Dutch Cohort Study (COOL5-18, see (Zijsling et al., 2009). The math test contained a number of items taken from the PISA 2000-2006 tests, additional items from COOL5-18, and some items from a Belgian Study on School Feedback (Verhaeghe and Van Damme, 2007).

[5]This is comparatively most frequent in the 2014 cohort, as the digital test was made available relatively late in the academic year. Apart from the missing test data at the school-cohort level, around 7% of testing data is missing at the individual level, in all likelihood because students were absent at the testing moment.

systems. Not all schools register this in an identifiable manner, and therefore this information is missing for around 20% of the students, who are excluded from the sample. As these missing observations are predominantly at the school-level rather than the individual level, this type of attrition is unlikely to be a major concern for our identification.[6] In addition, school switchers are likely to be overrepresented among these missing values. Since all of the schools in the sample offer both of the top tracks, there is no direct concern that school switching is linked to being re-tracked to the lower of the two tracks. A comparison further shows that the sample with missing data is slightly less affluent in terms of exit test score, teacher recommendation and parental background, but this disappears when we control for school fixed effects. In other words, schools with more low-ability students are more likely not to report this information. As we control for school fixed effects, this is corrected for in the analysis.

3.3 Non-cognitive measures

The aim of the empirical analysis is to estimate track effects for both cognitive and non-cognitive skills. The OML data contain a wide array of measures of non-cognitive measures. To prevent identifying statistically significant results simply due to multiple hypothesis testing, we select those measures that appear to be most relevant in the context of the literature. Among personality traits, we look at conscientiousness and neuroticism, which have been shown as the most predictive of the Big Five traits with respect to educational outcomes (Poropat, 2009; Kautz et al., 2014). Additionally, we use need for achievement and perseverance/grit, which have been shown as highly predictive traits for educational outcomes in, e.g., Duckworth et al. (2007). The items on which the personality factors are based are not fully consistent across cohorts but do have substantial overlap. On average, there are around six items per facet in each cohort. We use factor analysis to create factor variables for each of these indicators.

In addition to personality traits, we estimate effects for motivation, self-confidence and expecta-

[6]The grade 7 data are generally missing in one particular year. There is no school within the dataset for which this information is missing in all years.

tions for future educational attainment. Attending the higher track versus the lower track at the margin implies being at the bottom rather than the top of the ability distribution. One could expect that this has marked effects on students' self-perception and their motivation for school. On the other hand, being in the higher track is connected to higher levels of post-secondary education and therefore would be expected to lead to higher educational aspirations. Additionally, the attenuation of relative age effects over time would also imply that rank differences would be different for students depending on their month of birth, and therefore the effects of tracking on these outcomes may differ as well across relative age.

Motivation is measured through 20 student-reported items (1-5 Likert scale) and self-confidence through 25 student-reported items (1-4 Likert scale). We again create factor variables for each measure. There are two measures of educational aspirations, either on the secondary school track students expect to complete or on the highest level of post-secondary education they expect to complete. As the empirical analysis is focused on the top tracks in Dutch education, we create two dummy indicators taking value 1 if students expect to obtain a diploma from the top track (Exp. Track) or from university (Exp. Univ.). Expectations are not measured in the 2012 cohort of the OML. Appendix A11 provides an overview of the items that the non-cognitive outcomes are based upon.

4 Methodology

Our analysis is motivated by the common finding in the literature that younger students in class obtain lower scores on achievement tests, especially early in life. When tracking occurs relatively early, such differences in maturity are expected to be more prominent in the allocation of students to tracks. In Appendix A1 we use primary school data on student achievement on the high-stakes exit test and on the tracking recommendation and estimate how these relate to age of testing (predicted by the month and day of birth). We show that our results are in line with the literature on relative age and performance, such that older students perform better on cognitive tests. Addition-

ally, older students are recommended to higher tracks by their teacher, even when controlling for test performance. Hence, teachers appear to weigh maturity as an additional factor in the recommendation. This in turn also induces a positive relation between attendance of the high track in grade 7 and relative age: the predicted probability of high track attendance is around 8.1 percentage points higher for the very oldest student compared to the very youngest student in class, within the sample used in the empirical analysis.[7] We also find evidence that these relative age effects substantially reduce as students grow older, which confirms the findings of previous studies. These results motivate our expectation that students who are older at the time of assignment are also more likely to struggle academically in higher tracks when at the achievement margin.

In light of these dynamics between relative age and school achievement, the goal of this study is to estimate the causal effect of track assignment on cognitive and non-cognitive skills, across relative age in class. We restrict this analysis to the highest two tracks in Dutch secondary education, *havo* and *vwo*. The assignment of students to *vmbo* and its subtracks is not transparent. As mentioned before, schools can differ in the manner in which they organize subtracks. Additionally, correlations suggest that they are markedly more likely to consider the teacher recommendation than the exit test score in their sorting decisions, compared to sorting in the higher tracks. Given the high fuzzyness of this allocation, we therefore restrict the sample in our analysis to students that are assigned to either a *havo*, *havo-vwo* or *vwo* class in grade 7.

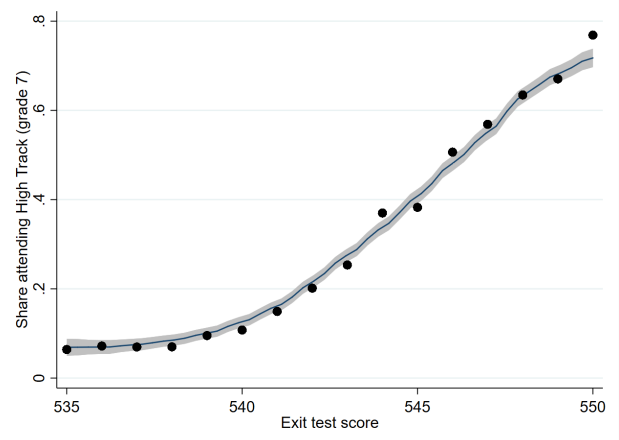
We define a student as being assigned to a high track if they were placed in *vwo* in 7th grade, as opposed to those placed in the next available track, namely *havo* or *havo-vwo*, depending on the school. In other words, those students who are placed at the *vwo* level are in the 'high track', while those who are placed in the second highest track that the school offers, are considered to be placed in the 'low track'. If the school has (next to a *vwo* track) a *havo-vwo* track, that will be the lower track. If the school only has *vwo* and *havo* tracks, then *vwo* will be the high track

[7] There is no relation anymore between relative age and high track attendance in grade 7 once controlled for exit test score and track recommendation, indicating that secondary schools do not additionally consider relative age when sorting students in the first year.

and *havo* will be the low track.[8] This implies that our treatment effect is essentially compared to two different counterfactuals jointly. We choose this approach because the alternative would imply endogenously excluding schools based on their sorting policies. Those in the mixed *havo-vwo* track in grade 7 comprise 60% of the group that is in the 'lower track'. Hence, this counterfactual carries more weight towards our results. The group in the mixed track still has a substantial chance of getting into the top track one year later, in which case their treatment would only consist of the one year being in a mixed grade. Section 6.4 will assess to what extent this may affect the results of the analysis.

To assess the treatment effect of being in the highest track on cognitive and non-cognitive outcomes, we employ a regression discontinuity design. While the testing bureau links each of their final test scores to a recommended track, these score thresholds are not formally enforced for secondary schools in the Netherlands. Figure 2 below plots the the primary school exit test score against the probability of being assigned to the highest track. As shown, the probability of being assigned to the highest track increases smoothly with the exit test score, between 0 and 1. There is no clear discontinuity for any particular score, when all schools in our data are pooled together. The same applies when we drop one of the two counterfactual 'low track' situations.

Figure 2: Share of students attending high track across exit test score



[8] When a school only has a *havo-vwo* class in 7th grade, there is no school threshold to estimate and the observations are excluded. This applies to 12 out of the 54 school-cohort observations.

As we will show in this section, empirical evidence supports the notion that secondary schools do use school-specific thresholds around the exit test scores to determine the allocation of students to the highest track. We will show that the probability of being assigned to a higher track jumps at the empirically-estimated school-specific thresholds. As this probability of being accepted to the high track does not jump from 0 to 1, we make use of a fuzzy regression discontinuity design.

4.1 Fuzzy RD design

Our fuzzy RD design instruments attendance of the high track in grade 7 by a dummy variable that takes value 1 if the student scored above the school threshold. The first stage is given by:

$$HT_{ic} = \lambda_0 + \lambda_1 I_{ic}(S_{ic} \geq \overline{S_s}) + \lambda_2 RelAge_{ic} + f^k(S_{ic}) + X'_{ic}\lambda_4 + \tau_c + \sigma_s + \mu_{ic} \quad (1)$$

where HT_{ic} is an indicator for being in the high track, for student i in cohort c, S_{ic} represents the exit test score of each individual and $\overline{S_s}$ is the (school-specific) threshold score for eligibility for the high track. $f^k(S_{ic})$ is a polynomial control function of order k for the test score S_{ic}. In all of the main analyses, we use separate linear specifications on each side of the threshold, as this provides the best fit. $RelAge_{ic}$ is the relative age of each student, measured by the month and day of birth of the student.[9] X_{ic} is a vector of individual-level controls including gender, parental education, ethnicity, living with both parents, language spoken at home and the employment status of each parent. We further include fixed effects for each cohort c (τ) and each secondary school s (σ). The second stage becomes:

$$Y_{ic} = \kappa_0 + \kappa_1 \widehat{HT_{ic}} + \kappa_2 RelAge_{ic} + f^k(S_{ic}) + X'_{ic}\kappa_5 + \tau_c + \sigma_s + \varepsilon_{ic} \quad (2)$$

Y_{ic} represents the outcome variable 3 years after track placement in the beginning of secondary school. These outcome variables can be classified in two categories: (i) cognitive and (ii) non-

[9] See Appendix A1 for a detailed account on constructing the relative age variable.

cognitive. Cognitive outcomes include the track level, math, and reading scores. Non-cognitive outcome variables include the personality traits of conscientiousness, neuroticism, need for achievement and persistence, measures of school motivation and self-confidence, as well as expectations students have with respect to future diplomas.

To analyze whether there are any heterogeneous treatment effects with respect to age, the model is expanded such that the interaction between being in the high track and relative age is instrumented with the interaction between treatment eligibility and relative age, adding a second first stage regression. The estimated equation becomes:[10]

$$HT_{ic} = \lambda_0 + \lambda_1 I_{ic}(S_{ic} \geq \overline{S_s}) + \lambda_2 RelAge_{ic} + f^k(S_{ic}) + X'_{ic}\lambda_3 + \tau_c + \sigma_s + \mu_{ic}$$

$$HT_{ic} * RelAge_{ic} = \rho_0 + \rho_1 I_{ic}(S_{ic} \geq \overline{S_s}) * RelAge + \rho_2 RelAge_{ic} + f^k(S_{ic}) + X'_{ic}\rho_3 + \tau_c + \sigma_s + \nu_{ic}$$

$$O_{ic} = \eta_0 + \eta_1 \widehat{HT_{ic}} + \eta_2 \widehat{HT_{ic} * RelAge_{ic}} + \eta_3 RelAge_{ic} + f^k(S_{ic}) + X'_{ic}\eta_5 + \tau_c + \sigma_s + \varepsilon_{ic} \qquad (3)$$

4.2 Estimating school-specific thresholds

Since there is no nationally-determined cut-off point for being admitted to the high track, we investigate whether school-specific thresholds are indeed used when placing students into the *vwo* track. This approach is partly motivated by the fact that schools are lawfully obliged to use at least one of the instruments (teacher recommendation or exit test score) in determining their admission

[10] We note that the reported standard errors in the main analysis are robust against heteroskedasticity but not corrected for clustering, as the number of school or school-cohort clusters falls below the informal threshold of 50. The inclusion of school fixed effects will partially, but not fully, correct for potentially clustered errors at the school level (Cameron and Miller, 2015). To further assess the possible threat of underestimated standard errors, we have performed wild bootstrap tests. Additionally, we have executed the model with corrections for clustering at the class level instead (in a few instances the number of clusters is still too low in this case; mainly for the expectation variables which are not available for one cohort). All these analyses lead to highly similar standard errors and p-values, indicating that we are not at risk of overrejecting the null through not correcting for clustering.

or sorting policy, although they are free to decide *how* to use that information.[11] The latter reflects the very high autonomy that schools have in decision-making within the Dutch educational system; see, e.g., Hanushek et al. (2013) who show that school autonomy across the OECD is highest in the Netherlands.

Little is known about the exact decision process of schools in setting thresholds, but evidence shows that it is common to have agreements about admission and sorting policies between different schools in the region. Reports indicate that around 25% of schools have an agreement with at least one other secondary school, 25% have agreements with all schools in the region, and 10-15% indicate that there are formal agreements at the municipal level (Inspectie van het Onderwijs, 2014). Agreements at the municipal level are especially common in bigger cities. For example, all schools in Amsterdam have been subject to a formal agreement which states that all schools have to allow students to *vwo* above a certain exit score threshold, and indicating another range of scores just below for 'further consideration'.[12] To the best of our knowledge, such written agreements are lacking for the region of Limburg. Nonetheless, the school-specific thresholds that we will discuss in this section show substantial within-region correlation, which suggests that schools in this area operate in a similar way, though less formally.

We use secondary-school-level data on students admitted to the highest track and their exit test scores, to assess the prevalence of school-specific discontinuities. Porter and Yu (2015) show that program effects in an RD design can also be identified using a two-stage procedure where the cutoff is estimated from the data. Furthermore, they show that estimating the cutoff in the first stage in this way does not affect the efficiency of the estimate in the second stage. In large samples, the fact that the threshold is unknown therefore does not affect the treatment estimates.

[11] The Education Inspectorate reviews whether individual schools indeed adhere to this lawful obligation. Schools may combine the two instruments with other considerations. The presence of a sibling at the same school or lotteries can be used if students have the same score/track recommendation and the number of places is limited. The legal obligation has changed since the year 2014/2015, as now schools are only allowed to consider the teacher recommendation. All of the cohorts we include in out empirical analysis have conducted the exit test before this point.

[12] See `http://www.onderwijsconsument.nl/citoscores-en-schooladviezen-citobandbreedtes/` for an example. Other cities have similar agreements, or alternatively set standards based on the teacher recommendation only. Note that this pertains to an older agreement, as setting exit test score requirements has been abandoned nationwide since 2014/2015.

An application of this approach that is similar in nature to ours is provided by Booij et al. (2016). They investigate the effects of gifted secondary education programs in the Netherlands, where the acceptance thresholds are not known in every period. By comparing estimated thresholds with the observed thresholds, they find that this methodology indeed leads to accurate predictions of the true thresholds. This approach differs slightly from that of Porter and Yu (2015), because the approach of the latter assumes that there is a treatment effect, which may be violated in the setting of Booij et al. (2016), as well as ours.

We estimate school-specific exit test threshold scores, using the same approach as Booij et al. (2016). This implies that, for each school and cohort, we regress a dummy indicator for attending the high track in grade 7 on a threshold dummy and the exit test score, excluding other covariates. We do this for all possible threshold scores and select the threshold that maximizes the R^2 (excluding control variables) . In other words, we select the threshold score at which the discontinuity is strongest in that school.

Based on the threshold of each school, those students scoring above are considered treatment eligible, while those scoring below the threshold are not. In a strict regression discontinuity design, 100% of the students should be assigned to tracks in line with their eligibility. In our sample, 83% of students starting secondary school are placed into tracks in line with their treatment eligibility, indicating that the estimated school thresholds have significant predictive power for track placement:

Table 2: Treatment eligibility and track assignment

	Scores above threshold	Scores below threshold
Attends high track	2,685 (44.2%)	513 (8.4%)
Attends low track	517 (8.5%)	2,365 (38.9%)

As using a threshold to assign students to the highest track is not a strictly enforced method, it

could be the case that some schools simply do not employ any threshold.[13] The potential problem of our approach is that we would then estimate the threshold where the noise in the data is strongest. This would elicit uncommonly good draws on the right hand side of the 'fake' threshold, or uncommonly bad draws on the left hand side of the 'fake' threshold, which could lead to biased results.

Figure A1 in Appendix A4 separately plots the jumps around empirically estimated thresholds for each secondary school. For a number of schools, the discontinuity is very fuzzy and most likely driven by random variation. As such, for our main estimates, we restrict the sample to the schools for which the jump at the threshold is statistically significant. After imposing this additional restriction, we capture 70% of the relevant schools and students in the region. This indicates that the majority of schools in our data use a threshold to assign students to the highest track.[14]

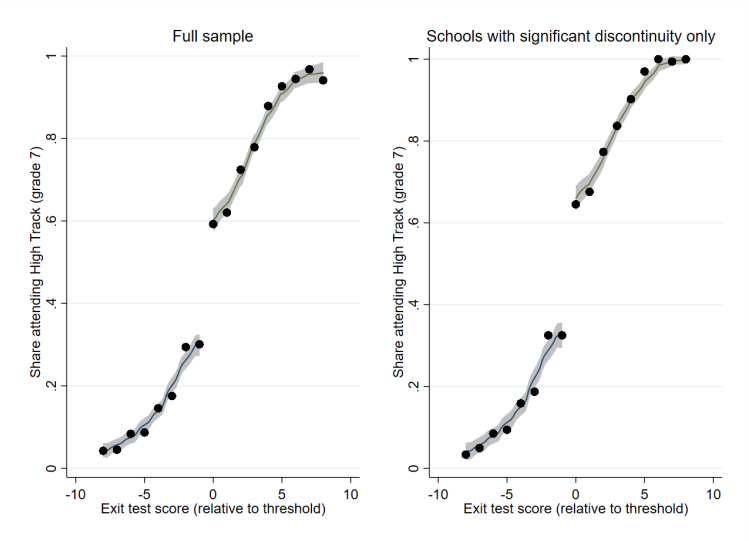
Figure 3 below plots the discontinuity around the estimated thresholds, for both the complete and the 'strict' sample. There is a precisely estimated jump in the probability of being assigned to the high track at the threshold: the probability of being assigned to the high track approximately doubles, from 30% to 60%.

Note that the thresholds are estimated here per school and per cohort. We conduct a robustness analysis that uses thresholds that are estimated per school across the three cohorts as well (see Appendix A5), but choose the former as the main approach. For one, this approach increases first stage power. More importantly, we believe that this reflects that thresholds can be updated from year to year. In fact, the score bands that the testing bureau reports also change every few years, and the formal agreements that exist in bigger cities in the Netherlands also show change over time. At

[13] Schools are lawfully obliged to consider either the test or the teacher recommendation in some way, but if they, for example, consider a very broad band of test scores for the top track or if many students forego the opportunity of attending the higher track, a true discontinuity will be absent.

[14] Including the full set of schools in the analysis leads to similar results as in our main analysis. Point estimates are slightly more favourable towards attending the higher track, but significance levels are the same as in the main analysis across outcomes. While the discarded 'non-strict' schools comprise a non-negligible share of the sample, their naturally higher degree of non-compliance means they do not receive a strong weight in the estimation of the LATE.

Figure 3: Discontinuity around the school-specific threshold



the same time, we should recognize that a school-cohort threshold is potentially more vulnerable to schools endogenously adjusting the threshold to the composition of the student population in that year. The robustness test using thresholds estimated across years can show us whether this is a potential concern.

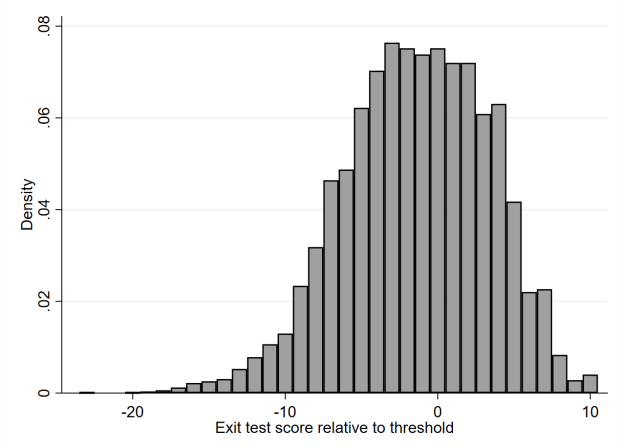
4.3 Assumptions

The crucial assumption for the validity of our methodology is that students are not able to self-select on any side of the threshold. This assumption is particularly important when thresholds are not strictly enforced. In particular, one can be concerned that students 'shop around', by repeatedly applying to different schools in their region until they are placed in the desired track. This type of selection is a threat to the validity of our results and, if present, prevents us from interpreting the results causally.

A first test to see whether students would sort themselves to be just above the threshold is to examine the distribution of the exit test score, centered around the threshold. If sorting would occur, one would expect bunching in the distribution at 0. Figure 4 shows that the distribution is very smooth. A McCrary density test confirms this (McCrary, 2008). This provides first evidence

against this concern. This also implies that schools do not adjust the threshold to maximize the inflow of students. There is also little incentive for them to do so in this setting, as there is no school in our sample that only offers the top track.

Figure 4: Distribution of test score around school threshold



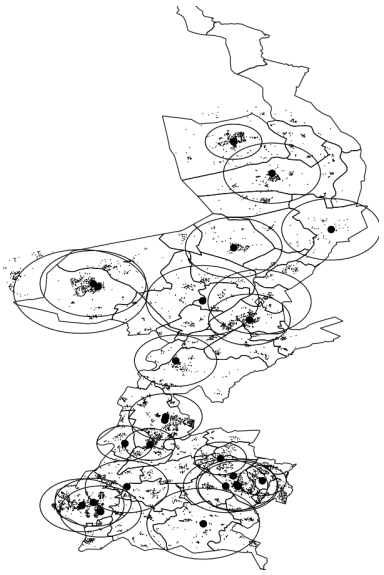
To deal with potential sorting more formally, we conduct tests that allow us to relax the no-sorting assumption in favor of two conditions which we believe to be much less restrictive.

First, we argue that *everything else equal, students prefer to attend the school which is closest to their home.*

It appears unlikely that students/parents choose their place of residence based on how lenient the threshold score of the nearest secondary school is, especially since this is not public information. Hence, the main concern lies in endogenous commuting of students to schools. Figure 4 below plots the population of students, location of schools, and the effective school catchment areas, in the region of Limburg. In our sample, 74% of students attend the school closest to their home. Many schools in the center and north of the region are far apart with little overlap in the living places of their students. The degree of urbanization is high in the South, and thus schools are frequently placed within short distances of each other.

Table 3 shows the results from a logit regression, where the dependent variable is a dummy indicating whether a student attends the closest school or not, on a number of control variables (gender,

Figure 5: Distribution of schools and students across the province of Limburg



Note: This figure draws upon Borghans et al. (2018), and has been updated to include all of the waves of the OML. Large dots represent schools, small dots represent students and circles represent 75% catchment areas.

parental education, ethnicity, language spoken at home, living with both parents, employment status of both the mother and the father, and exit test score). The students who attend the closest school are less likely to speak Dutch at home, as opposed to the regional dialect. This is likely a consequence of the fact that there is more school choice in the southeast, where the speaking of a dialect at home is more common. They do not differ on other characteristics.

In general, most students not attending the closest school live in an urbanized area with a high concentration of schools close to each other. Around 40% of students that do not attend their closest school travel less than one kilometer extra. In those cases, it is less likely that students select the attended school based on specific characteristics (such as sorting policy), as they are essentially indifferent in terms of commuting distance.

This observation leads to our second underlying assumption, namely that *student mobility becomes a problem if students are not treatment eligible for the closest school, and travel further away to be in a higher track.*

This is founded in the idea that endogenous self-selection of students to schools based on the school

Table 3: Balancing tests for students not attending closest school

	Coefficient	Standard error
Female	-0.043	0.069
Parental education		
Lower secondary	0.256	0.337
Upper secondary/lower vocational	-0.022	0.270
Higher professional education	0.142	0.270
University	0.040	0.272
Lives with both parents	-0.015	0.111
Ethnicity		
(Other) Dutch	-0.034	0.101
Non-Dutch	-0.181	0.131
Language at home		
Dutch	-0.246***	0.084
Other	-0.297	0.186
Mother is working	-0.228	0.146
Father is working	0.276	0.278

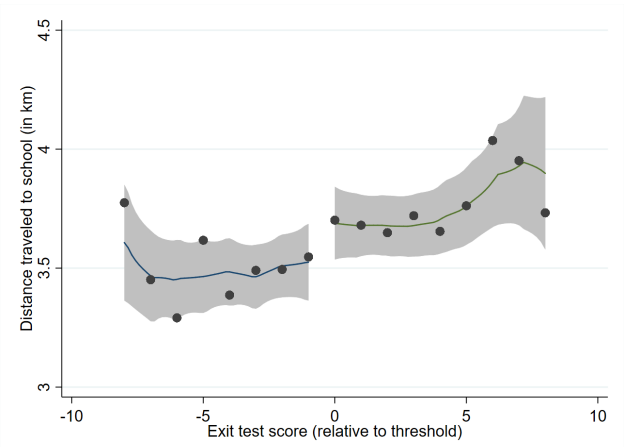
Note: Estimates are from a logit regression.
$*p < 0.10, **p < 0.05, ***p < 0.01$

threshold would only occur when wanting to enter a higher track, since going to a lower track is always possible. As such, students not attending their closest school is an issue if students travel further away to be in a higher track, thereby self-sorting on the right-hand side of the threshold. If this is the case, we should observe those not attending the closest school to be especially traveling towards schools where the acceptance threshold to the highest track is lower than that of their closest school. While 28% of students in our sample do not comply with going to the school closest to their home, half of these students attend the lower track and therefore by definition do not search out a distant school in order to be a in a higher track. Among the remaining 14%, there are also students that attend a school with a similar or lower threshold as their closest school. As such, we define student assignment as potentially problematic, if a student travels further away, to attend a school with a lower threshold than the school that is closest to their home. Only 5% of students in our sample fall in that category. These students originate disproportionately from the region in the South-East of Limburg (3% of the total 5%). Again, this is likely due to the higher availability of schools in this area. In additional robustness checks (see Appendix 8.6), we

show that excluding this region yields very similar estimates, confirming that this is not driving our results.

Figure 6 plots the student scores relative to the school threshold against the distance students travel to school. There is only a minor increase in travel distance at the threshold, which is statistically not significant. This is further indicative evidence that students do not consistently choose to travel further in order to select into the higher track.

Figure 6: Average distance traveled to school relative to school threshold



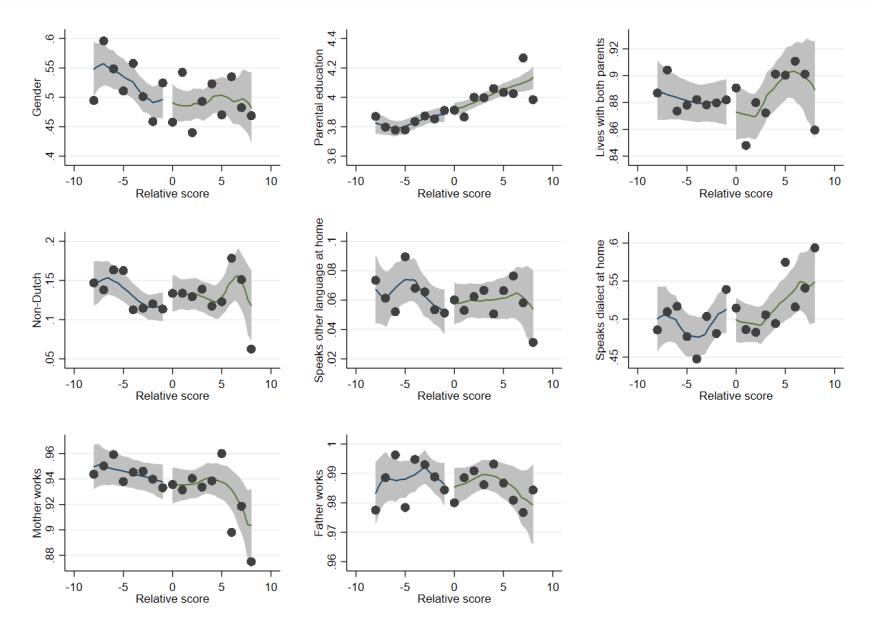
While potential sorting issues thus seem limited, we will further address this by additional robustness checks. In Section 6.1, we show that our results are robust to an alternative specification where we instrument track assignment with treatment eligibility as determined by the threshold of the closest school.

4.4 Balance tests

We now test whether important student characteristics and potential confounders are balanced around the threshold. As our dataset is very comprehensive as well as longitudinal, we can test this across a wide set of dimensions. We first look at our set of control variables. These are parental education, student gender, ethnicity (Dutch or other), whether student is living with both parents, language spoken at home, and whether the parents are employed. In Figure 7, we plot the exit test scores around the threshold for each of these control variables. No jumps are observed at the

threshold, confirming that the instrument is indeed exogenous to these characteristics. A regression of the treatment instrument on all controls gives no statistically significant coefficient, and the joint significance tests has a p-value of 0.925 (0.674 when school fixed effects are also included).

Figure 7: Balancing tests around the threshold



Because our data are longitudinal, we can also run similar balancing tests across lagged measures of our outcome variables or other pre-treatment measures of cognitive and non-cognitive skills. In Appendix A8 we show that the track assignment is also orthogonal to 6th grade measures of need for achievement, persistence, conscientiousness and neuroticism, several teacher-reported measures of non-cognitive skills, and the score on a non-verbal IQ test.

5 Results

We estimate the fuzzy regression discontinuity model described in equations (1) to (3). To do so we make use of the empirically estimated school-specific thresholds, restricting the sample to the schools for which the discontinuity is statistically significant.

Before exploring the regression results and the heterogeneous treatment effects with respect to

relative age, we first plot the second-stage for each individual outcome variable. Figure 7 below plots the second-stage for each of the three cognitive outcomes in 9th grade, while Figure 8 plots the reduced forms for the non-cognitive outcomes measured in 9th grade.

Figure 8: Reduced form graphs cognitive outcomes (9th grade)

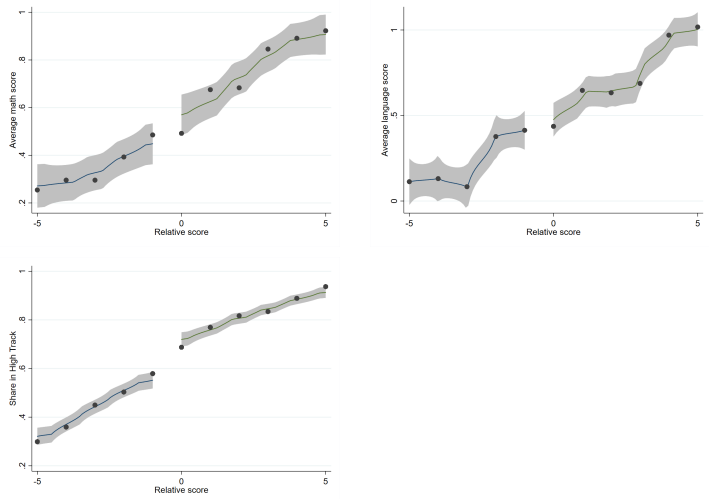
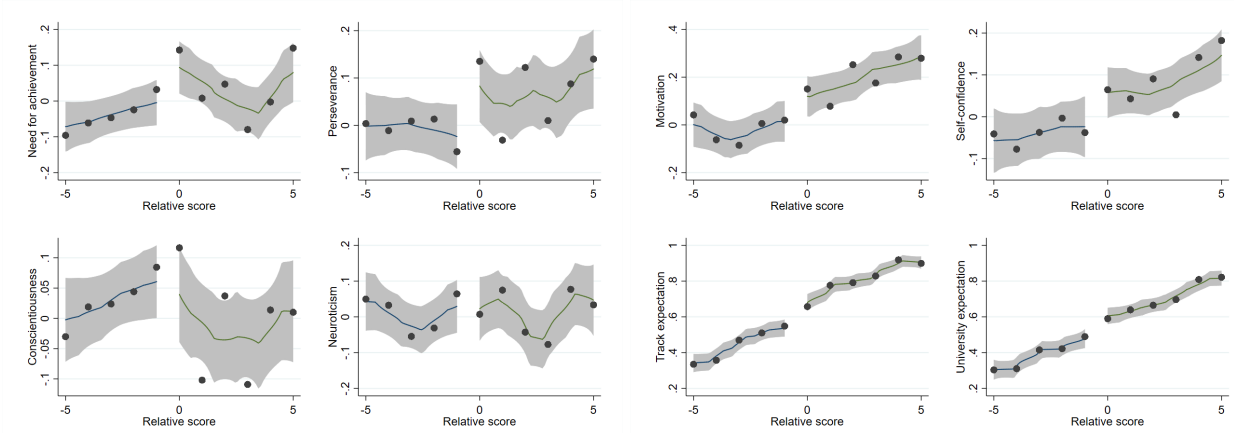


Figure 9: Reduced form graphs non-cognitive outcomes (9th grade)



Three years after track placement, both math and reading performance increase quite smoothly with the relative exit test score. However, there is no jump at the estimated threshold, suggesting that track assignment is efficient with respect to cognitive performance (i.e. no students can switch and be better off). On the other hand, we see that being assigned to a higher track in grade 7 is sustained by also being in a higher track in grade 9.

Likely as a result of being locked into specific tracks, students' expectations for finished track and post-secondary education also jump at the assignment threshold. For the remaining non-cognitive outcomes, there is no clear and precisely estimated jump at the threshold, although the patterns for mainly need for achievement and perseverance point somewhat in that direction. Effects for the latter two seem especially concentrated for those that just obtained the threshold score, suggesting a possible need to compensate lower cognitive ability by better non-cognitive development to stay on track for these academically marginal students. Regression analysis is needed for conclusive results here (as well as for any indications of heterogeneity across relative age).

We present and discuss these results in turn for both cognitive and non-cognitive outcomes. The optimal bandwidths are calculated using the approach developed by Imbens and Kalyanaraman (2011), and their sensitivity is discussed in section 6.2.

5.1 Cognitive outcomes

Table 4 below presents the second stage results using equation (3), where the dependent variables are (i) track position 3 years after track placement[15], (ii) math scores 3 years after track placement and (iii) reading scores 3 years after track placement. Appendix A10 shows the first stage results of these estimations.[16] All results reported below include the full set of controls and school and time fixed effects. A comparison to results without controls and school fixed effects is provided in Appendix Table A7. These estimates are highly similar, confirming that our instrument is orthogonal to observed characteristics.

As expected, students that are placed in the high track at the age of 12, are significantly more likely to still be in the high track by the age of 15. The coefficient size equals 0.386. Naturally, the difference in high track assignment in grade 7 equals 1, so the difference has narrowed. However,

[15]Defined as a dummy variable which takes value 1 if the student is in the high *vwo* track, and value 0 if the student is in a track below *vwo*.

[16]We report Kleibergen-Paap test statistics on the relevance of the instrument. The Stock-Yogo critical values for weak instruments equal 16.38 for the case with one instrument (which applies to the results reported in Appendix A3) and 7.03 for the case with two instruments (which applies to the main results portrayed below).

Table 4: Cognitive Skills

	Track 9th grade	Math Score	Reading Score
High Track	0.386**	-0.032	-0.294
	(0.164)	(0.393)	(0.427)
High Track * Age	-0.008	0.015	0.014
	(0.010)	(0.022)	(0.029)
Age	0.004	-0.022	-0.011
	(0.006)	(0.014)	(0.017)
N	3,001	1,908	1,518
KP stat	17.60	13.13	10.03
Optimal BW	± 3	± 4	± 3

Note: The regressions include a control function for the exit test score on each side of the threshold, and controls for gender, parental education, ethnicity, family structure, language spoken at home, parent employment status, cohort and school fixed effects. Optimal bandwidths are calculated (throughout the analysis) using the method by Imbens and Kalyanaraman (2011). Standard errors are robust. The Kleibergen-Paap (KP) statistic is provided to asses the strength of the identification.
$*p < 0.10, **p < 0.05, ***p < 0.01$

this is to be expected because students in the mixed *havo*/*vwo* track in grade 7, which make up 60% of those not in the high track, still need to be allocated. Part of this also reflects downgrading of students that are in the high track in grade 7; descriptive statistics show that this occurs for around 13% of these students, and for around 18% of those at the school-specific threshold.

There are no differences in terms of cognitive skills at the achievement margin, although the results should be interpreted with caution due to relatively low power. Nonetheless, the point estimates suggest that the higher track environment does not translate into more efficient learning in these areas in lower secondary education. In terms of both math and reading, those marginal students who were allocated to the higher track do not appear to be performing significantly different than their low track counterparts. This seems to indicate that in terms of cognitive achievement, track allocation is generally efficient, such that no marginal students would benefit from switching track.

The results for the age interaction show that older students are not more likely to fall back to a lower track in the 3 years following track assignment. This is somewhat surprising, given that the older students in the top track are expected to be of lower ability, and given that retracking to

lower tracks happens rather frequently in Dutch education. The point estimate of the interaction is in the expected direction (i.e. the treatment effect is lower for older students) but statistically insignificant. It predicts a difference in treatment effects of around 0.09 between the very oldest and the very youngest student in class, suggesting a very minor treatment heterogeneity at best. In terms of math and reading achievement, treatment effects do not differ significantly between the older and the younger students.

Coefficient for math and language are especially imprecise because not all students take both tests. Since this is determined randomly, it may be reasonable to impute missing test scores from one test from the items completed in the other test (using the fact that some students completed both tests). Once we do this, standard errors indeed reduce considerably, but coefficient remain low and statistically insignificant (availabe on request).

5.2 Non-cognitive outcomes

Table 5 presents the results for the 9th grade non-cognitive outcomes.[17] Appendix A10 shows the first stage results of these estimations.

The columns in Table 5 estimate equation (3) for: (1) need for achievement, (2) perseverance, (3) neuroticism, (4) conscientiousness. (5) confidence, (6) school motivation, (7) track expected to finish, (8) post-high school education expected to finish. Note that the baseline effects for being in the high track reflect treatment effects for the youngest student in class.

The first two columns show that the attending the high track benefits older students (more) in terms of need for achievement and perseverance. The effects are sizable and economically significant. Assignment to the high track contributes 0.040 and 0.045 of a standard deviation (sd) more on need for achievement and perseverance, respectively, for every month that the student is older. This means that the positive treatment effect of HT is 0.48 and 0.54 sd larger for the oldest student

[17] For a number of students, self reported responses are missing on some of the non-cognitive outcomes. When this is the case, we impute responses from a parent questionnaire, administered in the same period. While this imputation marginally increases the precision of the estimates, it does not change the overall conclusions.

Table 5: Non-Cognitive Skills

	Need Ach.	Persev.	Conscient.	Neurot.	Confid.	Motiv.	Exp. Track	Exp. Univ.
High Track	-0.048	0.517	-0.418	0.256	0.084	0.175	0.379*	0.317*
	(0.204)	(0.452)	(0.333)	(0.414)	(0.201)	(0.225)	(0.228)	(0.191)
High Track * Age	0.040***	0.045*	0.027	-0.084***	0.017	0.007	-0.003	-0.005
	(0.013)	(0.026)	(0.019)	(0.024)	(0.012)	(0.014)	(0.010)	(0.009)
Age	-0.030***	-0.022	-0.010	0.039***	-0.003	-0.001	0.003	0.002
	(0.008)	(0.015)	(0.011)	(0.015)	(0.008)	(0.008)	(0.006)	(0.005)
N	4,366	2,433	3,114	2,440	3,255	3,378	2,050	2,401
KP stat	66.24	15.59	23.48	15.78	45.85	50.56	11.02	17.86
Optimal BW	± 7	± 3	± 4	± 3	± 7	± 7	± 4	± 5

Outcome variables are, in order, need for achievement, perseverance, conscientiousness, neuroticism, self-confidence, school motivation, expected to finish high track and expected to finish university. The latter two are dichotomous, all other variables are standardized. The second stage regressions include a control function for the exit test score on each side of the threshold, and the control variables discussed in Table 4. Standard errors are robust. The Kleibergen-Paap (KP) statistic is provided to asses the strength of the identification. $*p < 0.10, **p < 0.05, ***p < 0.01$

in class, compared to the youngest student in class. Attending the high track also adds an additional reduction in neuroticism of 0.084 sd for every month that the student is older. This leads to a sizable treatment heterogeneity of 1 sd between the oldest and youngest students in the sample. Appendix A3 reports results for the model without the age interaction, which shows that the favourable effect of the high track on perseverance is also present for the (local) average student, but statistically insignificant in case of need for achievement and neuroticism. In other words, high track assignment improves perseverance for the average student but more so for older students, while it improves need for achievement and neuroticism only for relatively older students. We do not find any treatment effect, across relative age, for conscientiousness.[18]

The magnitude of these non-cognitive treatment effects is sizable. When we center the baseline effect on the oldest (rather than the youngest) student, the estimates for need for achievement, perseverance and neuroticism are 0.44, 1.06 and 0.75 respectively. To provide some comparison, the non-cognitive effects of the Perry Preschool Program center around 0.5 sd (Heckman et al., 2010). Heckman et al. (2006) identify a causal effect of completing high school on Locus of Control of 0.4 sd and a causal effect of (some) college on self-esteem of 0.6 sd. The estimates in this study for the oldest students are slightly above this (at least for the non-cognitive outcomes with a statistically significant effect). This could reflect the high malleability of these skills in early adolescence. At the same time, one has to keep in mind that this elicits the specific group that benefits the most. Moreover, these effects may be particularly large at the achievement margin, as high track assignment involves an allocation from the very bottom to the very top of the achievement distribution for these students. In any case, the effects we identify are of substantial size and as such represent important improvements in non-cognitive skills, which in turn have been shown to have strong links to multiple later-life outcomes.

The non-cognitive results could explain why older students are not more likely to be retracked to a lower track: being assigned to a track that may be 'too high' for their cognitive ability may

[18] As shown in the appendix, the interaction estimate with respect to conscientiousness is statistically significant in some of the robustness tests, suggesting that older students may also benefit more from high track attendance .

challenge the older students more. In order to keep up with a more challenging environment, older students compensate by working harder, as reflected by the coefficients relating to need for achievement and perseverance. By being more challenged, older students assigned to the highest track appear to compensate a relatively lower ability with higher effort and a better ability to deal with psychological stress. Thus, these positive spill-overs on non-cognitive skills appear to help the older students to remain in the highest track and perform as well as the younger on achievement tests, despite the latter being more cognitively able on average when in the top track.

These (heterogeneous) estimates of the effect of high track attendance potentially comprise a wide range of mechanisms. Being in a different track environment can represent a different curriculum, different teachers, different peers, a different class rank and different expectations about future educational attainment. While we cannot disentangle all these potential forces, the setting and set of outcomes allow us to provide some insights. First of all, the formal curricula in these two tracks are virtually identical in terms of the followed set of school subjects, but there can be differences in how advanced the level of instruction is.[19] All schools in our sample offer both tracks and thus students in each track are subject to the same teacher staff, making it highly unlikely that teacher (or school) quality effects partially drive our results.

It does not appear that rank effects have operated through self-confidence, given the insignificant estimate in column 5 of Table 5. This may be a result of the negative effect of a lower rank within class being offset by the positive effect of being in a high track in itself. As such, the setting differs from studies that show negative impact of class rank on self-image, expectations and achievement in comprehensive schooling settings (Elsner and Isphording (2017) for the United States and Murphy and Weinhardt (2018) for the United Kingdom). In fact, students have higher expectations in the top track, although this effect does not differ by age and therefore does not appear to explain the favourable non-cognitive impacts of high track attendance for older students. We identify no treatment effects, across relative age, for motivation. In summary, it appears that

[19]This can operate through having more advanced material in the higher track for a specific course, or more informally by teachers adjusting the context of their classes to the level of the students.

having different peers and the (informal) effects that this has on the level of instruction in class is the main driver of the results, as this appears to induce higher levels of effort from especially those students that are at the lower end of the ability distribution within the higher track (i.e. academically marginal students with a high relative age).

6 Robustness

6.1 Student mobility

We further address the potential selection issue of students sorting to schools in our sample. In section 4.3, we have argued that, under some mild assumptions, student mobility is only an issue if students travel further away from home to potentially select into the track of their choice. We have shown that 74% of our sample attends the closest school, and that the ones that do not are not strongly different with respect to background characteristics. Additionally, only 5% of the sample attends a school further away that also has a more lenient threshold and is in the high track.

In this section we provide additional evidence that this sort of selection is not driving our results. We redefine our instrumental variable to measure treatment eligibility for the closest school, rather than eligibility with respect to the attended school. In this alternative estimation, the students who drive the results are those that comply with assignment, as defined by attending the closest school (and complying with eligibility).

Tables 6 and 7 below present the results for cognitive and non-cognitive outcomes, based on this alternative approach. The findings are largely consistent with the main estimation results, presented in section 5. For the cognitive outcomes, we find a very similar effect for 9th grade track. The coefficients for math and language are difficult to interpret as the KP-statistics are markedly below acceptable thresholds.

In Table 7, we find the same pattern of baseline effects as well as treatment heterogeneity as in the main analysis in terms of magnitude, sign and significance. Namely, we capture a positive

Table 6: Closest school instrument: Cognitive Skills

	Track 9th grade	Math Score	Reading Score
High Track	0.345*	-0.769	-1.255
	(0.195)	(0.746)	(1.067)
High Track * Age	-0.006	-0.012	0.020
	(0.008)	(0.032)	(0.049)
Age	0.004	-0.005	-0.009
	(0.005)	(0.018)	(0.026)
N	3646	1626	1282
KP stat	12.58	3.28	2.61
Optimal BW	± 5	± 4	± 3

The second stage regressions include a control function for the exit test score on each side of the threshold, and the control variables discussed in Table 4. Standard errors are robust. The Kleibergen-Paap (KP) statistic is provided to asses the strength of the identification. Optimal bandwidths are re-estimated for this alternative approach. $*p < 0.10, **p < 0.05, ***p < 0.01$

interaction between high track attendance and relative age for need for achievement and perseverance, and a negative interaction for neuroticism. Again, expectations regarding track completion and university degree are markedly higher in the upper track, although the estimate for university expectations is statistically insignificant because of higher imprecision. This confirms that our conclusions are not driven by a small sample of potentially self-selecting students.

6.2 Bandwidth sensitivity

The bandwidths for Tables 4 and 5 have been optimally selected for each outcome variable, using the Imbens-Kalyanaraman method. To check the sensitivity of our results to variations in the bandwidth, we provide additional results where the bandwidth used is larger or smaller than the optimal one.

In the tables below, we extend or restrict each bandwidth by 1 point on the exit test. As expected, wider bandwidths add more precision to the estimates, while smaller bandwidths decrease power. However, all our results remain very similar. The only true difference is that the estimate for track in 9th grade is statistically insignificant for the very smallest bandwidth of 2. The estimate is still

Table 7: Closest school instrument: Non-Cognitive Skills

	Need Ach.	Persev.	Conscient.	Neurot.	Confid.	Motiv.	Exp. Track	Exp. Univ.
High Track	0.486	0.672	-0.433	-0.421	0.602	0.256	0.917**	0.634
	(0.323)	(0.578)	(0.327)	(0.567)	(0.419)	(0.309)	(0.417)	(0.422)
High Track * Age	0.044***	0.049*	0.025	-0.076***	0.019	0.011	-0.023	-0.024
	(0.017)	(0.025)	(0.017)	(0.026)	(0.019)	(0.016)	(0.016)	(0.019)
Age	-0.024**	-0.021	-0.006	0.035**	-0.001	0.000	0.022**	0.012
	(0.0097)	(0.014)	(0.009)	(0.014)	(0.011)	(0.009)	(0.010)	(0.010)
N	3550	2553	3556	2560	2260	2943	2375	2365
KP stat	25.72	8.37	25.95	8.69	8.97	22.99	12.76	13.87
Optimal BW	± 7	± 4	± 7	± 4	± 5	± 8	± 7	± 7

The second stage regressions include a control function for the exit test score on each side of the threshold, and the control variables discussed in Table 4. Standard errors are robust. The Kleibergen-Paap (KP) statistic is provided to asses the strength of the identification. Optimal bandwidths are re-estimated for this alternative approach. $*p < 0.10, **p < 0.05, ***p < 0.01$

clearly positive, and results for any other bandwidth are statistically significant. Moreover, the estimate for the age interaction is low and statistically insignificant across all bandwidths.

In Appendix A9 we show results of stronger changes in bandwidth, for a subset of relevant outcome variables. The main results remain consistent.[20] In particular, Table A16 shows that, while the optimal bandwidth with respect to Need for Achievement is broad in the main estimation, the favourable track effect for older students is also present for more narrow bandwidths. Hence, our findings are not dependent on the employed bandwidth. Results are also robust to the inclusion of higher-order polynomials in the control function (available on request).

Table 8: Bandwidth sensitivity: Cognitive Skills

	Track 9th grade	Math Score	Reading Score
BW+1			
High Track	0.310**	0.291	-0.255
	(0.131)	(0.342)	(0.346)
High Track * age	-0.007	-0.000	0.010
	(0.007)	(0.018)	(0.022)
N	3,818	2,254	1,959
KP stat	28.01	18.12	16.87
BW	± 4	± 5	± 4
BW-1			
High Track	0.236	-0.167	-0.023
	(0.174)	(0.519)	(0.411)
High Track * Age	-0.005	0.010	0.027
	(0.014)	(0.029)	(0.040)
N	2,214	1,468	1,132
KP stat	17.75	7.71	9.61
Optimal BW	± 2	± 3	± 2

The second stage regressions include a control function for the exit test score on each side of the threshold, and the control variables discussed in Table 4. Standard errors are robust. The Kleibergen-Paap (KP) statistic is provided to asses the strength of the identification. $*p < 0.10, **p < 0.05, ***p < 0.01$

[20] Bandwidth changes for the other outcomes also show little sensitivity; a minor exception is that the interaction term is statistically significant (and positive) for conscientiousness for bandwidths of 6 and larger.

Table 9: Bandwidth sensitivity: Non-Cognitive Skills

	Need Ach.	Persev.	Conscient.	Neurot.	Confid.	Motiv.	Exp. Track	Exp. Univ.
High Track	0.008	0.344	-0.409	0.228	0.120	0.266	0.434**	0.328**
	(0.186)	(0.363)	(0.295)	(0.348)	(0.182)	(0.205)	(0.186)	(0.155)
High Track * Age	0.036***	0.035*	0.024	-0.064***	0.018	0.006	-0.008	-0.007
	(0.013)	(0.019)	(0.016)	(0.019)	(0.012)	(0.013)	(0.009)	(0.008)
N	4,554	3,107	3,662	3,115	3,402	3,525	2,409	2,634
KP stat	82.30	22.70	32.35	22.83	58.00	63.03	17.55	27.76
Optimal BW	±8	±4	±5	±4	±8	±8	±5	±6
BW-1								
High Track	-0.064	0.319	-0.267	0.227	0.072	0.241	0.505*	0.284
	(0.243)	(0.473)	(0.390)	(0.444)	(0.229)	(0.257)	(0.259)	(0.233)
High Track * Age	0.043***	0.054	0.022	-0.080**	0.017	0.006	0.004	-0.002
	(0.015)	(0.036)	(0.024)	(0.036)	(0.013)	(0.015)	(0.014)	(0.011)
N	4,004	1,795	2,442	1,800	3,004	3,121	1,634	2,042
KP stat	46.59	16.18	16.60	16.40	34.19	37.78	8.77	11.38
Optimal BW	±6	±2	±3	±2	±6	±6	±3	±4

The second stage regressions include a control function for the exit test score on each side of the threshold, and the control variables discussed in Table 4. Standard errors are robust. The Kleibergen-Paap (KP) statistic is provided to asses the strength of the identification. $*p < 0.10, **p < 0.05, ***p < 0.01$

6.3 Controlling for lagged outcomes

As we can rely on longitudinal data, we have pre-tracking (i.e. grade 6) measures for some of the non-cognitive outcome variables. This applies to need for achievement, perseverance, conscientiousness and neuroticism. These measures are only available for around two thirds of the sample (as primary school questionnaires are not administered in the North of Limburg). As losing the other third would further reduce statistical power, the lagged outcomes are not included in the main specification. However, they do serve as a valuable robustness test.

Table 10 below shows results when we control for these lagged outcomes (III), which are compared to the main results (I) and the results for the main specification run on the (smaller) sample for which the 6th grade measures are available (II). The table shows that results are highly similar when lagged outcomes are controlled for. The age interaction for perseverance even slightly increases and is now statistically significant at the 5% level, although this is primarily due to the sample change. These results solidify our main conclusion. They show that the favorable non-cognitive impacts for older students truly arise during the tracking period and that our estimates are not affected by baseline differences in these non-cognitive skills between those on each side of the threshold.

This further confirms the results for the balance tests of 6th grade measures discussed in Section 4.4, and available in Appendix Figure A2. Importantly, Figure A2 also shows that there is no particularly high value for the first data point above the threshold. The reduced form graphs for especially perseverance and need for achievement in grade 9 (Figure 9) suggested that the treatment effects are mainly concentrated at this truly marginal student. The different analyses have shown that the same marginal student is not different from those at the other side of the threshold in background characteristics or a wide set of pre-treatment non-cognitive skills.

Table 10: Non-Cognitive Skills: lagged outcomes

	Need Ach.			Persev.			Conscient.			Neurot.		
	I	II	III	I	II	III	I	II	III	I	II	III
High Track	-0.048	-0.037	0.045	0.517	-0.170	-0.049	-0.418	-0.053	-0.054	0.256	0.191	0.076
	(0.204)	(0.220)	(0.217)	(0.452)	(0.438)	(0.424)	(0.333)	(0.348)	(0.330)	(0.414)	(0.457)	(0.432)
HT * Age	0.040***	0.040***	0.036**	0.045*	0.065**	0.063**	0.027	0.018	0.008	-0.084***	-0.075***	-0.066**
	(0.013)	(0.015)	(0.015)	(0.026)	(0.027)	(0.027)	(0.019)	(0.022)	(0.020)	(0.024)	(0.027)	(0.026)
N	4366	2988	2988	2433	1256	1256	3114	2117	2117	2440	1587	1587
KP-stat	66.24	58.77	59.89	15.59	14.52	15.39	23.48	21.99	21.81	15.78	12.53	12.55
BW	± 7	± 7	± 7	± 3	± 4	± 4	± 4	± 4	± 4	± 3	± 3	± 3

Column I provides the main results, column II for the main model on the sample with available lagged outcomes and column III for the model including the lagged outcome as additional control. The second stage regressions include a control function for the exit test score on each side of the threshold, and the control variables discussed in Table 4. Standard errors are robust. The Kleibergen-Paap (KP) statistic is provided to asses the strength of the identification. Optimal bandwidths are re-estimated for the reduced. $*p < 0.10, **p < 0.05, ***p < 0.01$

6.4 Different counterfactuals

As mentioned before, the 'low track' condition jointly comprises the *havo*-track and the *havo-vwo* mixed track in grade 7. These two counterfactual settings differ in several dimensions, most prominently the fact that *havo-vwo* students still have a considerable probability of assignment to the top track in case of good performance in grade 7. In this section, we separately analyze treatment effects with respect to each counterfactual situation, by dropping students in the other counterfactual from the analysis. These results, presented in Tables A12 through A15, should be interpreted with care as the estimates are naturally less precise and there may be issues of sample selection bias as well. Nonetheless, these results can indicate to what extent the existence of a mixed counterfactual can influence the main estimates and shed further light on some of the mechanisms.

Tables A12 and A13 show one particular difference for the cognitive results, namely a substantially stronger effect on *vwo*-attendance in grade 9 for the *havo* counterfactual. This is not surprising, since these students are effectively locked in the lower track, while the *havo-vwo* students still have a direct path to the top track.[21] The effect on 9th grade track is statistically insignificant for the *havo-vwo* counterfactual, although the point estimate is positive and does not preclude a meaningful effect. Interestingly, we observe in Tables A14 and A15 that the non-cognitive gains for the older students are present with respect to both counterfactual situations.

This can be suggestive evidence that the higher probability of being in the more demanding track in grades 8 and 9 when assigned to the top track in grade 7 is not the main driver behind the non-cognitive treatment effects. On the other hand, we do find suggestive evidence of an effect on 9th grade track for the very oldest in class (while the interaction effect is relatively low, we find point estimates 0.086 and 0.325 when we split the sample in the younger and the older half). Hence, the non-cognitive gains for this group of older students may still operate through a higher probability

[21] Taken together, students in the mixed *havo-vwo* class are assigned to the *vwo* track in around 45% of all cases, but this is naturally higher for students who are at the achievement margin for the top track. For those just below the school threshold and in a *havo/vwo* class in grade 7, around 60% ends up in *vwo*.

of being in the more demanding track, inducing compensating gains in non-cognitive skills. Still, the results could suggest that other mechanisms may be important as well. The different peer environment in grade 7 may create a different reference point that has lasting effects on non-cognitive skills.[22] It could also be that, because *havo-vwo* students only get into the top track when they show high achievement in grade 7, these are students that are not likely to struggle academically in *vwo*, thereby reducing the need for compensation through effort or non-cognitive skills. We can only speculate on the relative importance of each of these potential mechanisms, and also need to be careful in interpreting these estimates that are rather imprecise and potentially subject to sample selection bias. In any case, this analysis shows that the identified gains in non-cognitive skills for especially older students are not dependent on the counterfactual that we employ.

6.5 Additional robustness tests

We have performed several exercises that rely on a slightly different construction of the instrument. As mentioned before, there appears to be some (regional) coordination in the setting of thresholds. We therefore have alternatively calculated region-specific thresholds (splitting the province of Limburg into five sub-regions) and school-board-specific thresholds. As the threat of endogenously setting thresholds towards the student population may be more relevant for schools that deviate from the 'regional norm', this approach could be considered as less susceptible to bias. Naturally, this also reduces first stage power, as those that comply to their school-specific threshold but not to the region-specific threshold are added to the group of non-compliers. While the estimates are therefore less precise, they show the same pattern of results and identify a similar gain in non-cognitive skills for older students from attending the high track (these results are available on request).

The robustness tests described above are aimed at assessing potential endogeneity of the threshold, or sorting around the thresholds. A separate concern may lie in the relative age measure, which is

[22] E.g., even when getting into the top track, (especially younger) *havo-vwo* students may still stick with their grade 7 classroom peers, and are therefore less induced to catch up to their new higher-ability peers in the *vwo*-track. Alternatively, being among high-quality peers from the beginning as an academically marginal student could make one's own relatively low ability more salient (especially for older students, who are of lower ability given top track attendance), inducing the need for higher effort and persistence.

based on date of birth. While commonly used as instrument for relative age in class, some studies have questioned the exogeneity of birth dates; see, e.g., Buckles and Hungerman (2013). We note that the inclusion of a control for relative age in our model corrects for any general effect from potential non-randomness of birth dates. Moreover, relative age does not correlate with the control variables in our model, indicating that we do not pick up on potential interactions between high track attendance and, for example, socio-economic background. A related concern is that relative age does correlate with having repeated a grade. However, controlling for retention or excluding the 'later' birth dates (where retention is concentrated) does not affect our overall results. The same applies to the exclusion of the very earliest birth dates, as one may be concerned about deliberate planning of parents to let their child be the oldest in class (results are available on request). Hence, we believe that the heterogeneous effects we identify across relative age are not driven by other factors that may correlate with birth dates.

7 Conclusion

This study has analyzed the effect of academic track attendance on cognitive and non-cognitive outcomes and its interaction with relative age. Our results show that assignment to the high track has no effect on cognitive outcomes for students at the achievement margin, irrespective of relative age. Track assignment does affect non-cognitive skills, and heterogeneously across relative age. Relatively older students benefit from attending the higher track in terms of higher perseverance, higher need for achievement and higher emotional stability. These effects are identified using a regression discontinuity design that estimates school-specific thresholds from the data. We show that results are not driven by selective mobility of students to schools, and also robust to alternative bandwidths, alternative approaches for threshold estimation, and the inclusion of lagged outcomes.

Earlier research on relative age has documented that relatively younger students are tracked lower than is warranted given their academic potential. This automatically implies that relatively older students are at-risk of being tracked to too demanding tracks. In light of the evidence on the de-

creasing nature of relative age effects as students grow older, one would expect this group to be especially at risk of falling to lower tracks in later grades and losing motivation for school. The results from our study appear to indicate an opposite story, namely that the more demanding environment of the higher track induces positive spillovers on non-cognitive skills for these students, which mainly appear in areas related to applied effort and motivation to achieve. Hence, while these students might fall short of the cognitive ability level that is believed to be necessary for the top track (i.e. they would fall short of achievement thresholds if tests would be corrected for relative age), they appear to compensate by working harder. This could also explain why older students do not relegate more often to lower tracks after initial track selection, despite their higher susceptibility to being tracked above their ability level. Put differently, non-cognitive spillovers appear to mitigate the expected complementarity between ability and attending the higher track.

While our results are based on a different estimation approach and elicit a different treatment margin than other studies in this area, they confirm the overall finding that the 'undertracking' of those with low relative ages does not directly harm their educational development; see, e.g., Dustmann et al. (2017); Korthals et al. (2016). While those studies have shown that this holds when comparing younger students in a lower track with older students in a higher track of similar ability, we show that this also holds when comparing younger students in a low track with other younger students in a higher track. Moreover, we have shown that the 'overtracking' of older students actually has benefits in terms of non-cognitive development, when we compare older students in the high track with older students in the low track.

In this light, one could conclude that high-stakes tests for track selection should not be corrected for relative age effects, but rather that a higher relative age is a positive signal for achieving a better non-cognitive development from attending the higher track, conditional on test performance. We emphasize, however, that these treatment effects by relative age may be highly dependent on the location of the threshold. If the lack of non-cognitive gains for younger students indeed occurs because they are of higher ability given top track attendance, then there may be a group

of younger students with equivalent treatment effects on non-cognitive skills further below the threshold (i.e. those with similar *age-corrected* test scores as the older students with high non-cognitive treatment effects). The results can therefore as well be interpreted in a more general perspective that demanding learning environments can benefit non-cognitive skills.

The results further emphasize that the effects of educational decisions and policies should be evaluated with respect to both cognitive and non-cognitive skills. Besides the general importance of non-cognitive skills for later-life outcomes, they are also shown to be especially malleable in early adolescence, and therefore highly relevant for educational decisions in early secondary school, of which tracking is one of the most prominent examples.

We have analyzed the effects of track selection specifically for the top track in the Netherlands, which gives access to university education. Analysis for other choice margins in the Dutch tracking system was not feasible, as the relation between track selection and achievement is too fuzzy for the application of an RD design. Estimation of the effects of track selection for vocational tracks provides an interesting avenue for future research. Moreover, it would be valuable to also look at the long-run implications of track selection across relative age in future studies. While we identify favorable effects on non-cognitive skills, the question arises to what extent these effects, which predominantly seem to reflect higher effort levels, are sustainable in the long run.

References

Almlund, M., A. L. Duckworth, J. Heckman, and T. Kautz (2011). Personality psychology and economics. In *Handbook of the Economics of Education*, Volume 4, pp. 1–181. Elsevier.

Bedard, K. and E. Dhuey (2006). The persistence of early childhood maturity: International evidence of long-run age effects. *Quarterly Journal of Economics 121*(4), 1437–1472.

Booij, A., F. Haan, and E. Plug (2016). Enriching students pays off: Evidence from an individualized gifted and talented program in secondary education. IZA discussion paper, no. 9757.

Borghans, L., R. Korthals, and T. Schils (2018). Track placement and the development of cognitive and noncognitive skills. Unpublished manuscript.

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

Buckles, K. and D. M. Hungerman (2013). Season of birth and later outcomes: Old questions, new answers. *The Review of Economics and Statistics 95*(3), 711–724.

Cameron, A. C. and D. L. Miller (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources 50*(2), 317–372.

Duckworth, A. L., C. Peterson, M. D. Matthews, and D. R. Kelly (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology 92*(6), 1087–1101.

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

Elder, T. E. and D. H. Lubotsky (2009). Kindergarten entrance age and children's achievement impacts of state policies, family background, and peers. *Journal of Human Resources 44*(3), 641–683.

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

Goldberg, L. R. (1992). The development of markers for the big-five factor structure. *Psychological assessment 4*(1), 26.

Guyon, N., E. Maurin, and S. McNally (2012). The effect of tracking students by ability into different schools: A natural experiment. *Journal of Human Resources 47*(3), 684–721.

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.

Hanushek, E. A., S. Link, and L. Woessmann (2013). Does school autonomy make sense everywhere? panel estimates from pisa. *Journal of Development Economics 104*, 212–232.

Hanushek, E. A. and L. Woessmann (2006). Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. *Economic Journal 116*(510), 63–76.

Heckman, J. J., S. H. Moon, R. Pinto, P. A. Savelyev, and A. Q. Yavitz (2010). Analyzing social experiments as implemented: A reexamination of the evidence from the HighScope Perry Preschool Program. *Quantitative Economics 1*(1), 1–46.

Heckman, J. J., J. Stixrud, and S. Urzua (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics 24*(3), 411–482.

Imbens, G. and K. Kalyanaraman (2011). Optimal bandwidth choice for the regression discontinuity estimator. *Review of Economic Studies 79*(3), 933–959.

Inspectie van het Onderwijs (2014). De kwaliteit van het baissschooladvies. Technical report, Utrecht: Inspectie van het Onderwijs.

Kautz, T., J. J. Heckman, R. Diris, B. Ter Weel, and L. Borghans (2014). Fostering and measuring skills: Improving cognitive and non-cognitive skills to promote lifetime success. Technical Report 19656, National Bureau of Economic Research.

Korthals, R., O. Marie, and D. Webbink (2016). Does early educational tracking increase inequality? Short and long term international evidence. Paper presented at ESPE 2016 Berlin.

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

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

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics 142*(2), 698–714.

Mühlenweg, A., D. Blomeyer, H. Stichnoth, and M. Laucht (2012). Effects of age at school entry (ase) on the development of non-cognitive skills: Evidence from psychometric data. *Economics of Education Review 31*(3), 68–76.

Mühlenweg, A. M. and P. A. Puhani (2010). The evolution of the school-entry age effect in a school tracking system. *Journal of Human Resources 45*(2), 407–438.

Murphy, R. and F. Weinhardt (2018). Top of the class: The importance of ordinal rank. Working Paper 24958, NBER.

Organisation for Economic Co-operation and Development (2011). *PISA take the test: Sample questions from OECD's PISA assessments*. Paris, France: OECD Publishing.

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological bulletin 135*(2), 322.

Porter, J. and P. Yu (2015). Regression discontinuity designs with unknown discontinuity points: Testing and estimation. *Journal of Econometrics 189*(1), 132–147.

Verhaeghe, J. and J. Van Damme (2007). Leerwinst en toegevoegde waarde voor wiskunde, technisch lezen en spelling in eerste en tweede leerjaar. Technical report, Leuven, Steunpunt Studie- en Schoolloopbanen, Report No. OD1/05.

Zijsling, D., J. Keuning, H. Kuyper, T. van Batenburg, and B. Hemker (2009). Cohortonderzoek cool5 18. technisch rapport eerste meting in het derde leerjaar van het voortgezet onderwijs. Technical report, Groningen/Arnhem, the Netherlands: GION/Cito.

Appendix

A1 Relative age effects

We investigate whether (i) age influences cognitive achievement on high stake tests at the end of primary school and (ii) whether the teacher recommendation that students receive at the end of primary school accounts for the cognitive immaturity of the younger students at the time of the test.

The main measure of cognitive achievement in primary school is the exit test score (CITO) at the end of 6th grade. We standardize the exit test with mean 0 and standard deviation 1. To have a comparable scale for the teacher recommendation outcome, we assign to each category the average exit test score of students with that recommendation.

A well-known difficulty in investigating the relative age effect is that assignment into grades is non-random such that the weaker students who often are younger are also more likely to be retained or sent to special education. In order to solve this issue, Bedard and Dhuey (2006) suggest using the assigned relative age as an instrument for observed age since the distribution of birth dates is exogenous, estimating the effect of age net of grade retention or late entry.

As in the main analysis, we identify relative age from 0 (October 1st) to 12 (September 30th), with daily increments in between. We use an instrumental variable approach where the first stage is given by:

$$A_{ic} = \beta_1 + \beta_2 RelA_{ic} + X'_{ic}\beta_3 + \tau_c + \varepsilon_{ic} \tag{4}$$

where A_{ic} is the age of each student i in each cohort c, $RelA_{ic}$ is the relative age of the student as defined above, X_{ic} is a vector of controls, τ_c are cohort fixed effects and ε_{ic} is an individual-specific (robust) error term clustered at the primary school level.

The second stage equation then equals:

$$Y_{ic} = \alpha_1 + \alpha_2 \widehat{A_{ic}} + X'_{ic}\alpha_3 + \tau_c + \nu_{ic} \tag{5}$$

where the outcome variable Y_{ic} can represent the exit test score, teacher recommendation, or 9th grade scores for math and language. The vector of controls includes the same set as in the main analysis of the paper. For consistency, we run the analysis on the same sample as used for our main analysis, hence only including students from the top two tracks.

Table A1 presents the results from both OLS estimates and IV coefficients where the dependent variable is the exit test score. Following the IV results, we identify that being one month older leads to an increase in the test score of 0.033 of a standard deviation higher. This means that the oldest in class score 0.4 standard deviations higher that the youngest students. This is a sizable difference and equal to around 40% of the difference in average scores between the *havo* and *vwo* tracks.

Table A2 shows results when we use the teacher recommendation as an outcome. Older students receive higher recommendations, which is not surprising as the recommendation follows the test, which is characterized by large gaps by relative age. More importantly, we still identify a positive relation between age and teacher recommendation when the exit test score is controlled for. This indicates that teachers do not compensate the youngest based on their cognitive immaturity at the time of the test. In fact, they penalize the younger students additionally, possibly by associating their immaturity with a lower ability that is not captured by the high stake test, or valuing relative maturity in itself as an important quality for being successful in future education.

Finally, Table A3 reports results for 9th grade achievement. These estimates need to be interpreted with care, because the relation between relative age and attendance of the high track (through the established gaps in exit test scores and teacher recommendations) may impact these estimates. Nonetheless, they can provide indicative results for the development of relative age effects over time. Table A3 shows that relative age effects have virtually disappeared by grade 9. Moreover,

estimates become negative once we control for the 6th grade ability indicators. The latter suggests that younger students indeed experience faster achievement growth between grades 6 and 9. While differences in high track attendance by age may endogenously contribute to these estimates, they are unlikely to be behind the narrowing of the relative age gap. For one, we identify no effects of attending the higher track on achievement in the main analysis. Second, other literature suggests that if there is any effect of tracking on relative age gaps, it is in favour of the older students (Bedard and Dhuey, 2006; Korthals et al., 2016), implying that we would even underestimate the catching up of younger students. Therefore, we conclude that the expected narrowing of relative age gaps as students grow older, which is one of the motivations for the main analysis conducted in this study, indeed appears to be present within our sample.

We emphasize that the narrowing of relative age gaps by grade 9 is not at odds with our finding of no effect of attending the high track on 9th grade achievement, as identified in the main analysis. The main analysis implies that younger students do not gain more when attending the higher track than older students (in fact, none experience any cognitive gains from attending the high track). In other words, the stronger achievement gains of younger students between grades 6 and 9 are present (and of equal size) irrespective of the track they attend.

Table A1: The effect of relative age on 6th grade achievement

	OLS	IV	N
Exit test	-0.013***	0.033***	10.122
	(0.0013)	(0.0038)	
Math subscore	-0.0094***	0.018***	5.677
	(0.0025)	(0.0059)	
Language subscore	-0.013***	0.030***	5.677
	(0.0025)	(0.0067)	
N	10.122	10.122	

All regressions include the vector of control variables X'_{ic}. Subscores for the exit test are only available for a subset of students. Standard errors are robust and clustered at the primary school level. $*p < 0.10, **p < 0.05, ***p < 0.01$

Table A2: The effect of relative age on teacher recommendation

	OLS	OLS	IV	IV
RelAge	-0.0080***	-0.0029***	0.026***	0.013***
	(0.0014)	(0.011)	(0.0026)	(0.0019)
Exit test		0.412***		0.417***
		(0.0096)		(0.0095)

All regressions include the vector of control variables X'_{ic}. Sample size equals 10,122 in all instances. Standard errors are robust and clustered at the primary school level. $*p < 0.10, **p < 0.05, ***p < 0.01$

Table A3: The effect of relative age on 9th grade achievement

	OLS	OLS	OLS	IV	IV	IV
Math						
RelAge	-0.021***	-0.017***	-0.016***	-0.0052	-0.012*	-0.015**
	(0.0017)	(0.0016)	(0.0016)	(0.0065)	(0.0064)	(0.0064)
Exit test		0.257***	0.153***		0.259***	0.153***
		(0.015)	(0.011)		(0.016)	(0.011)
Teacher rec.			0.253***			0.254***
			(0.026)			(0.027)
Language						
RelAge	-0.019***	-0.014***	-0.013***	0.0022	-0.0078	-0.012
	(0.0028)	(0.0020)	(0.0019)	(0.0088)	(0.0082)	(0.0080)
Exit test		0.334***	0.203***		0.337***	0.204***
		(0.025	(0.019)		(0.026)	(0.019)
Teacher rec.			0.312***			0.313***
			(0.036)			(0.037)

All regressions include the vector of control variables X'_{ic}. Sample size equals 5,450 for math scores and 5,082 for language scores. Standard errors are robust and clustered at the secondary school level. $*p < 0.10, **p < 0.05, ***p < 0.01$

A2 Descriptives

Table A4: Descriptive Statistics

	Share (%)	N
High track 9th grade	52.6	6,080
Female	50.8	6,056
Parental education		6,056
Primary	1.6	95
Lower Secondary	2.6	157
Upper Secondary	21.0	1,269
Higher Professional	52.2	3,160
University	22.7	1,375
Lives with both parents	88.4	6,056
Ethnicity		6,056
Limburg	66.0	3,999
Other Dutch	20.7	1,255
Non-Dutch	13.2	802
Language at home		6,056
Dialect	43.4	2,626
Dutch	50.5	3,058
Other	6.1	372
Mother employed	93.9	6,080
Father employed	98.8	6,080
Region		5,938
Central/North Limburg	13.7	811
Western Limburg	30.2	1,791
South-West Limburg	35.5	2,105
South-East Limburg	20.7	1,231

	Mean	Standard deviation	N
CITO score	543.3	4.60	6,080
Math score	0.54	0.92	2,995
Reading score	0.49	0.98	3,053
Need for achievement	0.01	1.01	4,911
Perseverance	0.04	0.99	4,910
Conscientiousness	0.01	1.00	4,917
Neuroticism	0.02	1.03	4,920
Confidence	0.04	0.800	3,663
School motivation	0.10	1.03	3,789
Expectation Track	0.59	0.49	3,216
Expectation University	0.51	0.50	3,217

Note: All summary statistics are reported for the main estimation sample. Standardization of outcomes has been conducted on the sample of all students in a *havo*, *havo/vwo* and *vwo* class in grade 7 (including non-strict schools). 'Confidence' takes the mean of two separate standardized self-confidence measures (instrumental and social self-confidence).

A3 Alternative model specifications

Below, we report results for two changes to the main model specification. Tables A5 and A6 show estimates for the RD specification that only includes the high track indicator, without the age interaction. As such, estimates represent the (local) effect for a student of average relative age, while the baseline effects in the main tables reflect the effect for the very youngest students in class. Table A7 shows a comparison between results with and without the control vector and school fixed effects.

Table A5: Without age interaction: Cognitive Skills

	Track 9th grade	Math Score	Reading Score
High track	0.342***	0.054	-0.226
	(0.158)	(0.365)	(0.426)
Age	-0.000	-0.014**	-0.004
	(0.003)	(0.006)	(0.007)
N	3001	1908	1518
KP stat	35.36	26.28	20.71
Optimal BW	± 3	± 4	± 3

The second stage regressions include a control function for the exit test score on each side of the threshold, and the control variables discussed in Table 4. Standard errors are robust. The Kleibergen-Paap (KP) statistic is provided to asses the strength of the identification. $*p < 0.10, **p < 0.05, ***p < 0.01$

Table A6: Without age interaction: Non-Cognitive Skills

	Need Ach.	Persev.	Conscient.	Neurot.	Confid.	Motiv.	Exp. Track	Exp. Univ.
High track	0.200	0.774*	-0.255	-0.218	0.183	0.218	0.359*	0.285*
	(0.188)	(0.428)	(0.309)	(0.389)	(0.192)	(0.207)	(0.203)	(0.168)
Age	-0.009**	0.003	0.005	-0.007	0.007*	0.003	0.001	-0.001
	(0.004)	(0.006)	(0.005)	(0.006)	(0.004)	(0.004)	(0.003)	(0.003)
N	4366	2433	3114	2440	3255	3378	2050	2401
KP stat	132.56	31.22	46.96	31.66	91.83	101.26	22.70	36.82
Optimal BW	± 7	± 3	± 4	± 3	± 7	± 7	± 4	± 5

The second stage regressions include a control function for the exit test score on each side of the threshold, and the control variables discussed in Table 4. Standard errors are robust. The Kleibergen-Paap (KP) statistic is provided to asses the strength of the identification. $*p < 0.10, **p < 0.05, ***p < 0.01$

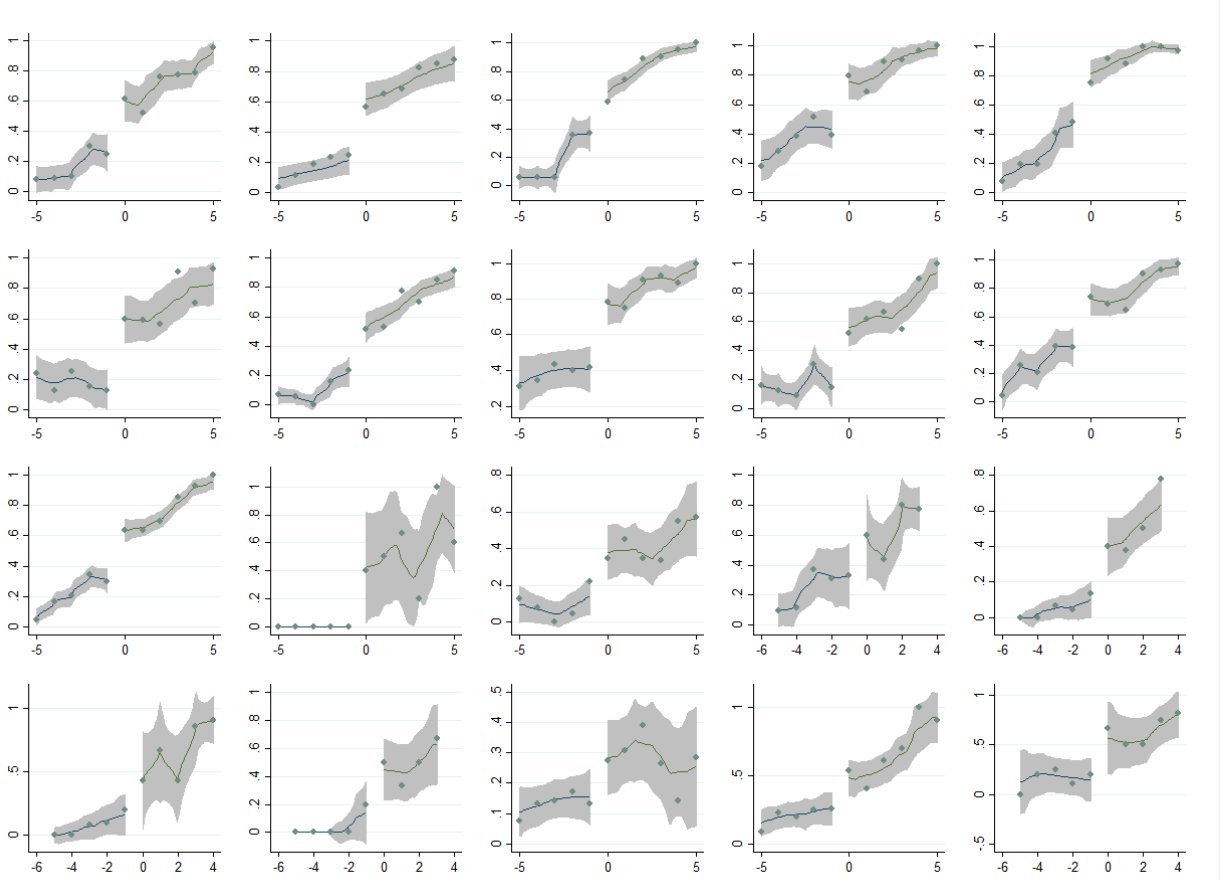
Table A7: Estimates with and without control variables

	Track G9		Math		Language			
High Track	0.309*	0.386**	0.019	-0.032	-0.584	-0.294		
	(0.182)	(0.164)	(0.431)	(0.393)	(0.502)	(0.427)		
HT * Age	-0.005	-0.008	0.019	0.015	0.019	0.014		
	(0.010)	(0.009)	(0.022)	(0.022)	(0.032)	(0.029)		
	Need Ach.		Persev.		Conscient.		Neurot.	
High Track	-0.056	-0.048	0.490	0.517	-0.431	-0.418	0.171	0.256
	(0.211)	(0.204)	(0.455)	(0.452)	(0.340)	(0.333)	(0.429)	(0.414)
HT * Age	0.042***	0.040***	0.041	0.045*	0.025	0.027	-0.075***	-0.084***
	(0.013)	(0.013)	(0.026)	(0.026)	(0.019)	(0.019)	(0.025)	(0.024)
	Confid.		Motiv.		Exp. Track		Exp. Univ.	
High Track	0.075	0.084	0.103	0.175	0.288	0.379*	0.254	0.317*
	(0.209)	(0.201)	(0.232)	(0.225)	(0.230)	(0.228)	(0.200)	(0.191)
HT * Age	0.018	0.017	0.009	0.007	-0.001	-0.003	-0.005	-0.005
	(0.012)	(0.012)	(0.014)	(0.014)	(0.011)	(0.010)	(0.010)	(0.009)
Controls		X		X		X		X

The table shows estimates for all outcome, either with or without the control vector and school fixed effects. $*p < 0.10, **p < 0.05, ***p < 0.01$

A4 School-specific thresholds

Figure A1: Discontinuity at threshold: school-specific figures



Note: The figure shows discontinuities in high track attendance, separately for all the schools in our sample. Both strict and non-strict schools are portrayed in the figure, where the latter group is excluded for the final estimation sample.

A5 Estimates when thresholds are estimated across cohorts

Table A8: Constant thresholds: Cognitive Skills

	Track 9th grade	Math Score	Reading Score
High Track	0.421*	-0.421	0.093
	(0.216)	(0.528)	(0.285)
High Track * Age	-0.009	-0.010	0.000
	(0.008)	(0.021)	(0.016)
Age	0.005	-0.001	-0.003
	(0.005)	(0.013)	(0.010)
N	3,980	2,207	2,799
KP stat	9.59	7.74	28.51
Optimal BW	± 4	± 5	± 8

The second stage regressions include a control function for the exit test score on each side of the threshold, and the control variables discussed in Table 4. Standard errors are robust. The Kleibergen-Paap (KP) statistic is provided to asses the strength of the identification. Optimal bandwidths are re-estimated for this alternative approach. $*p < 0.10, **p < 0.05, ***p < 0.01$

Table A9: Constant thresholds: Non-Cognitive Skills

	Need Ach.	Persev.	Conscient.	Neurot.	Confid.	Motiv.	Exp. Track	Exp. Univ.
High Track	-0.541	0.117	-0.242	0.110	0.365	0.150	0.296	0.236
	(0.334)	(0.348)	(0.340)	(0.550)	(0.496)	(0.840)	(0.294)	(0.242)
High Track * Age	0.045***	0.019	0.019	-0.053**	-0.009	-0.033	0.007	-0.005
	(0.017)	(0.017)	(0.017)	(0.022)	(0.019)	(0.033)	(0.013)	(0.011)
Age	-0.029***	-0.010	-0.008	0.026**	0.014	0.032*	-0.001	0.002
	(0.010)	(0.009)	(0.010)	(0.012)	(0.012)	(0.018)	(0.007)	(0.006)
N	4,033	4,034	4,038	3,195	2,326	1,904	2,161	2,491
KP stat	24.50	24.57	24.42	8.50	6.22	3.08	7.13	11.51
Optimal BW	± 6	± 6	± 6	± 4	± 4	± 3	± 4	± 5

The second stage regressions include a control function for the exit test score on each side of the threshold, and the control variables discussed in Table 4. Standard errors are robust. The Kleibergen-Paap (KP) statistic is provided to asses the strength of the identification. Optimal bandwidths are re-estimated for this alternative approach. $*p < 0.10, **p < 0.05, ***p < 0.01$

A6 Results excluding the South-East Limburg province

Table A10: Cognitive Skills: Excluding South-East

	Track 9th grade	Math Score	Reading Score
High Track	0.313**	-0.322	-0.115
	(0.142)	(0.482)	(0.383)
High Track * Age	-0.003	0.034	0.016
	(0.008)	(0.025)	(0.022)
Age	0.003	-0.030**	-0.007
	(0.005)	(0.015)	(0.014)
N	3,005	1,497	1,572
KP stat	24.13	9.31	12.89
Optimal BW	± 4	± 4	± 4

The second stage regressions include a control function for the exit test score on each side of the threshold, and the control variables discussed in Table 4. Standard errors are robust. The Kleibergen-Paap (KP) statistic is provided to asses the strength of the identification. Optimal bandwidths are re-estimated for this alternative sample. $*p < 0.10, **p < 0.05, ***p < 0.01$

Table A11: Non-Cognitive Skills: Excluding South-East

	Need Ach.	Persev.	Conscient.	Neurot.	Confid.	Motiv.	Exp. Track	Exp. Univ.
High Track	-0.164	0.473	-0.534*	0.189	0.212	0.300	0.399	0.408
	(0.226)	(0.492)	(0.289)	(0.457)	(0.382)	(0.293)	(0.257)	(0.264)
High Track * Age	0.037**	0.050*	0.036**	-0.074***	0.003	-0.002	-0.005	-0.005
	(0.015)	(0.029)	(0.017)	(0.027)	(0.021)	(0.016)	(0.012)	(0.013)
Age	-0.025***	-0.022	-0.017*	0.034**	0.002	0.011	0.004	-0.000
	(0.009)	(0.017)	(0.009)	(0.016)	(0.012)	(0.010)	(0.007)	(0.007)
N	3,486	1,952	3,199	1,954	1,995	2,234	1,643	1,637
KP stat	51.33	12.79	37.05	13.16	14.69	19.22	9.40	9.63
Optimal BW	± 7	± 3	± 6	± 3	± 5	± 4	± 4	± 4

The second stage regressions include a control function for the exit test score on each side of the threshold, and the control variables discussed in Table 4. Standard errors are robust. The Kleibergen-Paap (KP) statistic is provided to asses the strength of the identification. Optimal bandwidths are re-estimated for this alternative sample. $*p < 0.10, **p < 0.05, ***p < 0.01$

A7 Separating the counterfactuals

Table A12: Cognitive Skills: vwo vs havo-vwo

	Track 9th grade	Math Score	Reading Score
High Track	0.181	-0.553	-0.017
	(0.200)	(0.499)	(0.446)
High Track * Age	0.005	0.045	0.006
	(0.014)	(0.031)	(0.031)
Age	-0.001	-0.044**	-0.011
	(0.007)	(0.018)	(0.027)
N	1,642	1,025	1,057
KP stat	13.26	9.10	11.88
Optimal BW	± 3	± 4	± 4

The second stage regressions include a control function for the exit test score on each side of the threshold, and the control variables discussed in Table 4. Standard errors are robust. The Kleibergen-Paap (KP) statistic is provided to asses the strength of the identification. Optimal bandwidths are re-estimated for this alternative sample. $*p < 0.10, **p < 0.05, ***p < 0.01$

Table A13: Cognitive Skills: vwo vs havo

	Track 9th grade	Math Score	Reading Score
High Track	0.616***	0.806*	0.096
	(0.136)	(0.486)	(0.386)
High Track * Age	-0.010	-0.009	0.004
	(0.008)	(0.027)	(0.027)
Age	0.005	-0.004	-0.003
	(0.005)	(0.019)	(0.019)
N	2046	1095	1125
Kleibergen-Paap tat	17.29	7.47	9.69
Optimal BW	± 4	± 4	± 4

The second stage regressions include a control function for the exit test score on each side of the threshold, and the control variables discussed in Table 4. Standard errors are robust. The Kleibergen-Paap (KP) statistic is provided to asses the strength of the identification. Optimal bandwidths are re-estimated for this alternative sample. $*p < 0.10, **p < 0.05, ***p < 0.01$

Table A14: Non-Cognitive Skills: vwo vs havo-vwo

	Need Ach.	Persev.	Conscient.	Neurot.	Confid.	Motiv.	Exp. Track	Exp. Univ.
High Track	0.262	-0.064	-0.565	0.383	-0.082	0.103	0.192	0.139
	(0.302)	(0.485)	(0.417)	(0.496)	(0.300)	(0.327)	(0.261)	(0.247)
High Track * Age	0.051**	0.072**	0.061**	-0.063*	0.026	0.009	-0.000	0.007
	(0.020)	(0.034)	(0.027)	(0.035)	(0.019)	(0.021)	(0.014)	(0.013)
Age	-0.029***	-0.024	-0.017	0.025	-0.001	0.000	0.001	-0.007
	(0.011)	(0.018)	(0.014)	(0.018)	(0.011)	(0.011)	(0.008)	(0.007)
N	2,254	1,350	1,753	1,351	1,656	1,761	1,091	1,240
widstat	33.16	14.18	16.97	14.40	24.81	22.83	10.12	12.45
Optimal BW	± 7	± 3	± 4	± 3	± 7	± 7	± 4	± 5

The second stage regressions include a control function for the exit test score on each side of the threshold, and the control variables discussed in Table 4. Standard errors are robust. The Kleibergen-Paap (KP) statistic is provided to asses the strength of the identification. Optimal bandwidths are re-estimated for this alternative sample. $*p < 0.10, **p < 0.05, ***p < 0.01$

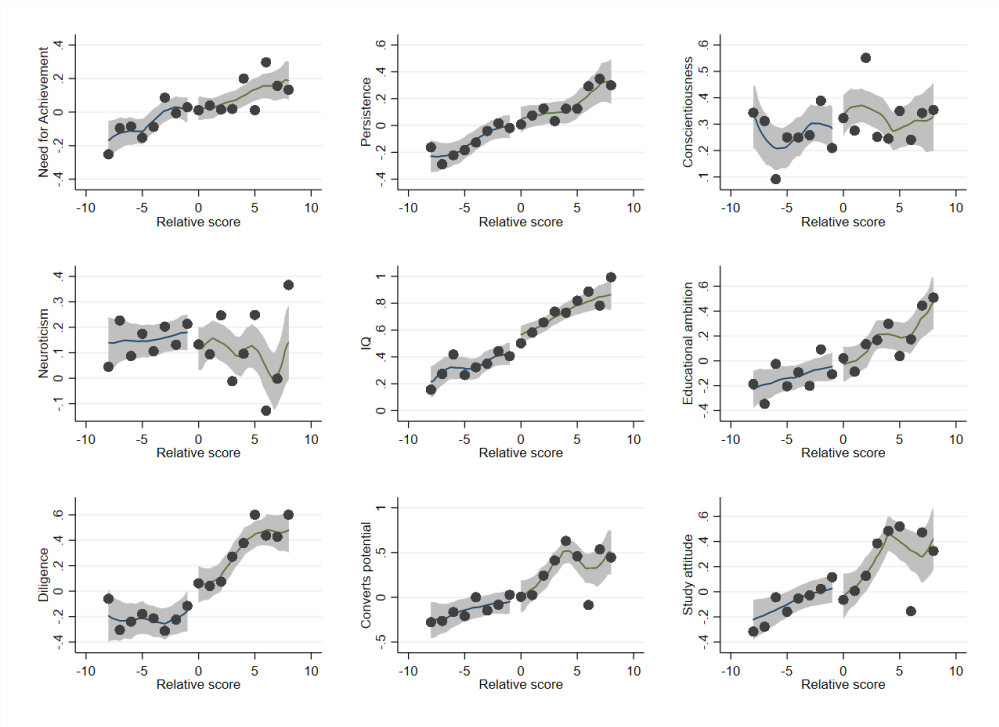
Table A15: Non-Cognitive Skills: vwo vs havo

	Need Ach.	Persev.	Conscient.	Neurot.	Confid.	Motiv.	Exp. Track	Exp. Univ.
High Track	-0.093	0.430	0.177	0.157	0.288	0.131	0.471**	0.471**
(0.237)	(0.460)	(0.397)	(0.390)	(0.219)	(0.249)	(0.240)	(0.197)	
High Track * Age	0.035**	0.045*	0.014	-0.048**	0.021	0.010	0.001	-0.014
	(0.016)	(0.025)	(0.023)	(0.022)	(0.016)	(0.016)	(0.012)	(0.011)
Age	-0.030**	-0.033*	-0.015	0.021	-0.012	-0.002	-0.000	0.012
	(0.012)	(0.018)	(0.016)	(0.015)	(0.012)	(0.012)	(0.008)	(0.008)
N	2655	1667	1673	1673	2044	2064	1188	1448
KP stat	45.80	13.82	14.24	13.47	36.06	33.53	7.18	13.35
Optimal BW	± 7	± 4	± 4	± 4	± 7	± 7	± 4	± 5

The second stage regressions include a control function for the exit test score on each side of the threshold, and the control variables discussed in Table 4. Standard errors are robust. The Kleibergen-Paap (KP) statistic is provided to asses the strength of the identification. Optimal bandwidths are re-estimated for this alternative sample. $*p < 0.10, **p < 0.05, ***p < 0.01$

A8 Balance tests for 6th grade non-cognitive skills and IQ

Figure A2: Balance tests 6th grade outcomes



Need for achievement, persistence, conscientiousness, neuroticism and educational ambition are student-reported. IQ is based on a non-verbal test. The final three outcomes are reported by the 6th grade teacher.

A9 Additional bandwidth sensitivity

Table A16: Non-Cognitive Skills: bandwidth sensitivity

	BW-4	BW-3	BW-2	BW+2	BW+3	BW+4
Track 9th grade						
High Track				0.282**	0.269***	0.325***
				(0.115)	(0.098)	(0.083)
High Track * Age				-0.002	-0.003	-0.003
				(0.006)	(0.006)	(0.005)
N				4,488	4,914	5,347
KP stat				37.92	53.67	77.07
Need for Achievement						
High Track	0.064	0.024	-0.040	0.052	0.027	-0.013
	(0.407)	(0.335)	(0.289)	(0.173)	(0.166)	(0.160)
High Track * Age	0.059**	0.047**	0.044***	0.032***	0.029**	0.031***
	(0.025)	(0.019)	(0.016)	(0.012)	(0.012)	(0.012)
N	2,433	3,107	3,655	4,675	4,765	4,809
KP stat	15.59	22.70	31.69	96.63	112.02	121.67
Perseverance						
High Track				0.284	0.110	-0.010
				(0.313)	(0.260)	(0.219)
High Track * Age				0.027*	0.034**	0.032**
				(0.016)	(0.014)	(0.013)
N				3,655	4,004	4,367
KP stat				31.692	46.591	66.245
Neuroticism						
High Track				0.168	0.163	0.207
				(0.304)	(0.253)	(0.215)
High Track * Age				-0.049***	-0.041***	-0.042**
				(0.016)	(0.014)	(0.013)
N				3,663	4,012	4,376
KP stat				31.87	46.85	66.61

For Track, Perseverance and Neuroticism, the optimal bandwidth equals 3 and a reduction of more than 1 is not feasible. The second stage regressions include a control function for the exit test score on each side of the threshold, and the control variables discussed in Table 4. Standard errors are robust. The Kleibergen-Paap (KP) statistic is provided to asses the strength of the identification. $*p < 0.10, **p < 0.05, ***p < 0.01$

A10 First stage coefficients

Tables 26 to 29 below present the first stage coefficients for our two instruments, namely track eligibility, and the interaction between track eligibility and age. The first stage coefficients corresponds to the our main second stage results, presented in Section 5.

Table A17: Cognitive Skills: First stage

Second stage outcome	Track 9th grade	Math Score	Reading Score
Track eligibility	0.225***	0.222***	0.257***
	(0.046)	(0.056)	(0.064)
F-value	17.73	13.18	10.57
Track eligibility*age	0.443***	0.534***	0.447***
	(0.036)	(0.042)	(0.348)
F-value	77.98	76.75	39.82
N	3,001	1,908	1,518

The table reports the coefficient for high track eligibility when predicting high track attendance and the coefficient for the eligibility*age interaction when predicting the attendance*age interaction. $*p < 0.10, **p < 0.05, ***p < 0.01$

Table A18: Non-Cognitive Skills: First stage

Second stage outcome	Need Ach.	Persev.	Conscient.	Neurot.	Confid.	Motiv.	Exp. Track	Exp. Univ.
Track eligibility	0.284***	0.233***	0.237***	0.236***	0.281***	0.287***	0.164***	0.172***
	(0.039)	(0.051)	(0.043)	(0.051)	(0.037)	(0.036)	(0.053)	(0.046)
F-value	66.28	34.07	23.48	15.84	45.90	50.62	11.93	19.75
Track eligibility*age	0.638***	0.460***	0.528***	0.459***	0.623***	0.625***	0.555***	0.613***
	(0.025)	(0.040)	(0.033)	(0.040)	(0.029)	(0.028)	(0.040)	(0.035)
F-value	349.40	67.75	127.41	67.34	249.70	262.81	98.53	164.30
N	4,366	2,433	3,114	2,440	3,255	3,378	2,050	2,401

The table reports the coefficient for high track eligibility when predicting high track attendance and the coefficient for the eligibility*age interaction when predicting the attendance*age interaction. $*p < 0.10, **p < 0.05, ***p < 0.01$

A11 Items non-cognitive skills

Below are the English translations of all the items of the non-cognitive outcome measures. Items for self-confidence ask students to rate how good they consider themselves at the various tasks, on a 4-point scale. All other items are measured on a 5-point Likert scale. Personality items are Dutch translations of the 50 item IPIP Big Five scale based on Goldberg (1992).

Table A19: Items Need for Achievement

	2012	2014	2016
I want to get high grades	X	X	X
I want to be good in my job	X	X	X
Trying your best is important to me	X		
Success is made too important	X		

Table A20: Items Perseverance

	2012	2014	2016
I work hard		X	X
If I start something, I finish it	X	X	X
If something becomes too difficult, I quit	X	X	X
If something is disappointing, I lose motivation	X	X	X

Table A21: Items Conscientiousness

	2012	2014	2016
I do chores right away	X		
I often leave my things around	X	X	X
I always keep my appointments	X		
Sometimes I forget that I need to do something	X		
I am accurate	X		
I do things without planning		X	X
I like order and regularity		X	X
I like working according to a scheme		X	X
I do things at the last moment		X	X
I finish my work in time		X	X
I neglect my work		X	X
I respond quickly		X	X
I prepare things well		X	X
I neglect tasks		X	X

Table A22: Items Neuroticism

	2012	2014	2016
I am stressed easily	X		
I easily get upset	X		
I am often in a sad mood	X		
My mood often changes	X		
I am pessimistic about the future		X	X
I always fear the worst		X	X
I burst into tears		X	X
I look at this with a positive view		X	X
I can put problems aside		X	X
I am self-confident		X	X
I panic		X	X
I am depressed		X	X
I ponder about something		X	X
I remain calm		X	X

Table A23: Items self-confidence

	2012	2014	2016
Writing without mistakes	X	X	X
Writing an essay	X	X	X
Calculation by head	X	X	X
Concentrating	X	X	X
Reading out loud		X	X
Drawing and painting		X	X
Making music		X	X
Drawing, painting or making music	X		
Finding something on the computer	X	X	X
Comforting somebody	X	X	X
Giving your own opinion	X	X	X
Winning at a fight	X	X	X
Getting my way	X	X	X
Getting along with my classmates	X	X	X
Listening to somebody who has a difficult time	X	X	X
Dressing nicely and looking good	X	X	X
Discussing	X	X	X
Doing sports		X	X
Keeping track of the news	X	X	X
Taking the lead	X	X	X
Being stronger than others		X	X
Making new friends		X	X
Giving a presentation in class		X	X
Writing nicely		X	X
Working in an orderly fashion		X	X
Control myself		X	X

Table A24: Items Motivation

	2012	2014	2016
I quit this school without finishing it	X		
As soon as I can, I stop learning	X		
I am gonna learn a job, but outside school	X		
I am very motivated to continue learning	X		
I am gonna learn interesting things	X		
I am gonna continue learning because I like to	X		
I am gonna continue learning for a long time	X		
As soon as I find a job, I quit learning	X		
I am gonna continue learning after this school		X	X
As soon as I can, I quit this school		X	X
When I get up in the morning I look forward to going to school		X	X
I have the feeling that I have too much schoolwork		X	X
I feel fit and strong when being at school		X	X
I am enthusiastic about what I learn at school		X	X
I often think about quitting school		X	X
I often feel I cannot handle the schoolwork		X	X
School inspires me		X	X
If I am learning intensively, I feel happy		X	X
I often sleep badly due to things that have to do with schoolwork		X	X
I am proud of going to school		X	X
I lose my interest in school		X	X
I often question whether schoolwork makes sense		X	X
I lose myself in schoolwork		X	X
I can motivate myself less and less for school		X	X
At school I have a lot of energy		X	X
During leisure time I worry about schoolwork		X	X
When I am learning, I can get carried away by the content		X	X
I used to be able to do more for school than I nowadays can		X	X