Who benefits from attending the higher track? The effect of track assignment on skill development and the role of relative age*

Maria Cotofan
Centre for Economic Performance, London School of Economics

Ron Diris[1]
Department of Economics, Leiden University

Trudie Schils
Department of Macro, International and Labour Economics, Maastricht University

Abstract

Previous findings on (gradually decreasing) relative age effects in school suggest that 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 thank Robert Dur, Bart Golsteyn, Roxanne Korthals, Marjolein Muskens, Sergio Parra Cely, participants at the ESPE 2019 (Bath) and EALE 2019 (Uppsala) conferences, seminar participants at Maastricht University and two anonymous referees for their helpful comments and feedback.

[1]Corresponding author, r.e.m.diris@law.leidenuniv.nl

1 Introduction

The use of tracking to sort students to different learning environments is common practice worldwide. Empirical studies on tracking have typically focused on the effects of tracked educational systems versus comprehensive systems, or on earlier versus later tracking. (Early) tracking is argued to foster inequality by concentrating peer and school quality at the top of the achievement distribution; see, e.g. Hanushek and Woessmann (2006); Pekkala Kerr et al. (2013). A separate policy question is how students, given that they are tracked at a certain age, should be sorted to tracks. This sorting process is complicated by the fact that ability signals that determine track assignment are noisy (Brunello et al., 2007), and that students may develop differently in secondary school than their primary school achievement suggests. With respect to the latter issue, several studies have focused on relatively young students in class as a group that is harmed by (early) tracking. It has been well-documented that student achievement differs by relative age in class, and that such differences decrease as students grow older; see, e.g., Bedard and Dhuey (2006); Elder and Lubotsky (2009); Mühlenweg and Puhani (2010); Cascio and Schanzenbach (2016). As tracking is based on achievement, younger students are sorted more often to tracks that are below their academic potential, especially when tracking takes place early (Bedard and Dhuey, 2006). At the other end of the relative age distribution, older students are at risk of being placed above their cognitive potential.[2] Relative age could thereby be an important source of misallocation of students to optimal tracks. Decision-makers thus face two key choices in determining an efficient sorting: where to set the achievement threshold, and whether the threshold should take into account that achievement reflects both ability and relative age.

This study estimates the cognitive and non-cognitive effects of track assignment at the achievement margin, and how these interact with relative age. We use data from the Netherlands, where track allocation is strongly based on a high-stakes exit test taken at the end of primary school (age 12). We apply a regression discontinuity design that relies on school-specific thresholds for the exit test, to

[2] Borghans et al. (2020) show that both "too high" and "too low" misallocation are important drivers of the negative long-run effects of early tracking.

identify the causal effect of track allocation for students who are at the achievement margin of the highest track. We further estimate how these marginal effects differ between older and younger students, by estimating interactions with relative age as well as a difference-in-discontinuity design. We use longitudinal data on primary and secondary school students from administrative sources and surveys, covering almost the full student population 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 identify positive non-cognitive effects for the oldest in class. Attending the higher track benefits older students in terms of perseverance, need for achievement and emotional stability. These gains in non-cognitive skills are absent for the 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 highest track), which could be explained by these compensating spillovers on non-cognitive skills.

Other studies have examined the interplay between relative age and attended track. Korthals et al. (2016) find that this relation is stronger in early tracking countries, but that younger students have similar achievement by age 15 and higher wages when tracking takes place early. Dustmann et al. (2017) estimate the long-run effects of attended track, using 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 higher wages.[3] These studies suggest that the underestimation of the potential of the relatively young in early tracking systems is not necessarily detrimental to their future attainment. However, this answers a different question than whether achievement thresholds are optimally set, such that no student can switch and be better off.[4] Additionally, it needs to be considered that those achievement thresholds imply different levels of ability for students of different relative age.

[3] 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. Guyon et al. (2012) find positive effects in the short and medium run from expanding the elite track in Northern Ireland, while Borghans et al. (2019) identify small positive wage effects from attending the highest track in the Netherlands.

[4] For example, these results could reflect that sending too few young students to the higher track is canceled out by sending too many old students to the higher track. Comparing students by month of birth only identifies the net effect of these two forces.

This study analyses the efficiency of track assignment by assessing whether students can benefit at the achievement margin from switching tracks, and whether such benefits are different between old and young students. Our RDD setup thus differs from much of the relative age literature, as it employs the exit score as running variable rather than the relative age of the student. This reflects that our focus is on the effect of switching tracks at the exit score threshold (either for the average marginal student, the young marginal student, or the old marginal student). Our approach is a direct test of whether the achievement threshold is indeed set at the "indifferent" student. By exploring differences between the old and young in class, it also answers the question of whether the threshold should be corrected for relative age in order to produce a more efficient sorting of students to tracks. We focus on this particular heterogeneity because the dynamics of relative age effects highlight a high potential for misallocation. As such effects gradually decrease, this implies that younger students move up in the achievement distribution and older students move down. As a result, students may find themselves in the 'wrong track' as they grow older. Moreover, addressing such misallocation is under direct policy control. The relation between relative age and track assignment is an artifact of the setup of educational systems and their use of calendar cut-off dates. If relative age is a source of misallocation, then age-correcting threshold scores could lead to welfare gains, at no direct expense. Policies that aim to remedy misallocation by relative age would also be less susceptible to practical and ethical difficulties than policies that would differentiate threshold scores by e.g. gender or parental education.

The educational context of our analysis comprises a low track and a high track for which curricula differ in difficulty. Assignment to tracks is largely based on an achievement test score, which is not corrected for relative age. In this context, the expected heterogeneities by relative age may differ by type of skill. For cognitive skills, one may assume that the benefits of attending the high track are increasing in ability.[5] Low ability students learn little in the most demanding track, as will high ability students in the low track. There will exist a cut-off ability level at which a student

[5]Classical theoretical models typically assume complementarity between ability and level of investment, such as in the context of college entry decisions; see, e.g., (Arrow, 1971).

is indifferent between attending the low or the high track. Everyone to the left of this threshold in the ability distribution would benefit from attending the low track and everyone to the right will benefit from attending the high track. Students are assigned to tracks based on an ability signal that is affected by relative age. This means that a young marginal student is of higher ability than an old marginal student, as they need to overcome the relative age penalty to pass the score threshold. Assuming complementarity between ability and the gains from attending a higher track, younger students are expected to benefit more from the higher track in terms of cognitive skills.

For non-cognitive skills, expected effects similarly depend on whether there are differences in such skills by relative age, and how these develop over time. Evidence from Germany (Mühlenweg et al., 2012) and Japan (Yamaguchi et al., 2020) shows that older students have better developed non-cognitive skills at the time of track assignment. This may imply that the higher baseline non-cognitive skills of older students lead to better non-cognitive skill development in the high track. However, if such differences fade out, non-cognitive gains from being in the higher track may be similar between younger and older students. A potential mediator for the non-cognitive effects of track assignment is the fact that switching tracks at the margin shifts students between the top and the bottom of the class rank (see, e.g., Elsner and Isphording (2017) for evidence on the importance of class rank). This may induce non-cognitive effects that are dependent on the baseline levels of non-cognitive skills.

In other words, the marginal young student is expected to enter the high track with higher cognitive and lower non-cognitive skills than the marginal old student. Consequently, the gains from attending the higher track at the margin may not only differ by relative age but also by type of skill. Despite the increased recognition of the importance of non-cognitive skills in education (Cunha et al., 2010; Jackson, 2018), estimation of non-cognitive (causal) effects of tracking is absent in the literature, let alone exploring their heterogeneity across relative age. Causal evidence shows that secondary education does enhance non-cognitive skill development (Heckman et al., 2006), although it remains largely unclear how such gains arise.

The remainder of this study is organized as follows. Section 2 provides an overview of relevant aspects of the Dutch educational system. Section 3 describes the data, while Section 4 discusses the methodological approach. Section 5 presents the main results and Section 6 the robustness analysis. 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. Education is compulsory until age 16, or age 18 when without a degree. When entering secondary school (grade 7/age 12), students are sorted to tracks. There are two main determinants for sorting students to tracks in secondary education:

1. A standardized exit test: In 6th grade, students take an achievement test. At the time of data collection, 95% of schools in the Netherlands administered the so-called CITO test, including all of the schools that appear in our data sample. The test is standardized and externally graded. It contains 200 multiple-choice questions testing the students in Dutch language, mathematics, and study skills. The testing bureau provides the score, and what they consider the appropriate track pertaining to that score. This can also be a 'mixed' recommendation of two adjacent tracks.

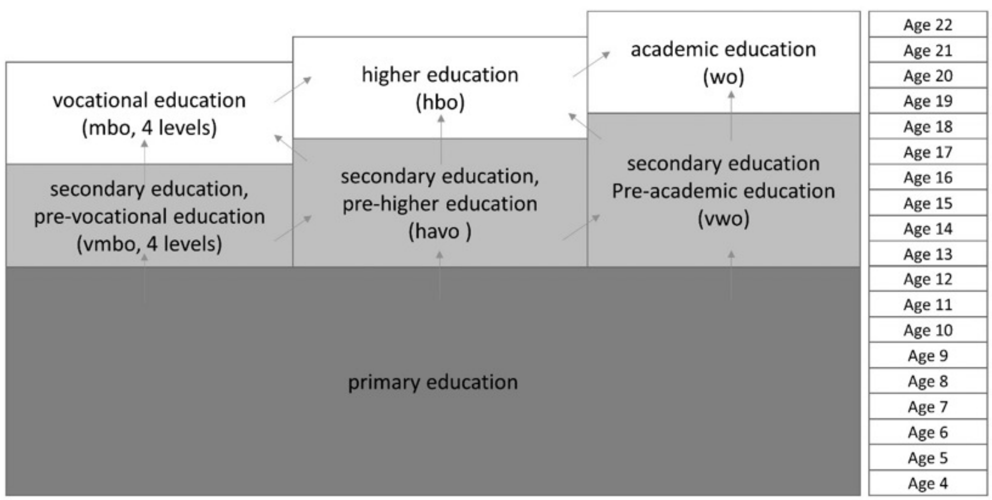
2. A teacher recommendation: In 6th grade, teachers provide a track recommendation for each student. In this assessment, the teacher is supposed to incorporate the history of achievement of the student, but also the broader cognitive and non-cognitive development, based on the teachers' classroom observations throughout the year.[6] As with the recommendation from the test, the teacher recommendation can be towards a single track, or a mixed recommendation for two adjacent tracks. The teacher recommendation is given after the results on the exit test and thus takes the test into account. The correlation between teacher recommendation and test recommendation,

[6] Such motivations are not expressed within the recommendation, which simply consists of a form that reports the recommended track(s).

when expressed on an ordinal scale, equals 0.82.

Within our sample, sorting is decentralized to the school level and students may apply to any school, after which the school rejects or accepts based on test score and/or teacher recommendation. Schools are legally obliged to consider at least one of the two instruments in this decision, but are free to decide how to combine these into an assignment rule.[7] Schools may consider additional criteria if there are capacity constraints.[8] If not, schools are not allowed to reject students with a certain test score and/or recommendation, if they admit other students with the same or lower score or recommendation.

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.

There are three main tracks in secondary education. 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 university. Around 55 percent of students are in the pre-vocational track, 25 percent in the pre-higher education track and 20 percent in the pre-academic track. While sorting starts in grade 7, final tracking decisions can still be

[7]This freedom is in line with the very high school autonomy in educational decisions in the Netherlands, which is highest across the OECD (Hanushek et al., 2013).

[8]If capacity constraints exist, it is most common to give preference to a student with a sibling or working parent at the school, living within a certain postal code and/or to hold a lottery. Students may apply to multiple schools simultaneously, and choose their preferred option when admitted to more than one.

postponed for one year. Schools have the discretion to keep two adjacent tracks together in grade 7 and 8. Around 50% of students are tracked directly in grade 7 and the remainder only in grade 8.[9] Secondary schools can choose to offer only one track, or multiple tracks. The most common options are to offer all tracks (45 percent), both the medium and the high track (23 percent), or only the lower track (22 percent).[10] Within the province of Limburg, all schools that offer the high track also offer the medium track. Table 1 shows student shares and average scores on the exit test for each teacher recommendation, across all the students in our data.[11] Our empirical analysis focuses on the *havo* and *vwo* tracks. We label these as T^M and T^H respectively, for the remainder of this study. Students in these two tracks follow the same set of subjects in lower secondary school, although typically with a higher level of difficulty in T^H. As such, the treatment effect of assignment to T^H over T^M comprises having better peers (but a lower class rank), and getting more advanced class material. Since all schools in our sample offer both tracks, differences in school quality are absent. Differences in teacher quality can still occur through within-school sorting of teachers to tracks.

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

Teacher recommendation	% Students	Mean score
vmbo	39.8%	529
vmbo/havo	9.9%	536
havo	19.0%	540
havo/vwo	12.8%	543
vwo	18.6%	547

Note: All vmbo recommendations that distinguish between different subtracks have been grouped. The score on the exit test ranges from 501 to 550.

[9] 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.

[10] Source: https://www.onderwijsincijfers.nl/kengetallen/vo/instellingen-vo/aantallen-aantal-vo-scholen.

[11] There are separate recommendations within the *vmbo* track that distinguish between practical and theoretical subtracks but we group these together as our focus is on the highest two tracks.

3 Data

We use data from the Onderwijsmonitor Limburg (OML). This is a cooperative project between Maastricht University, schools, schools boards and government bodies in Limburg, a province in the South of the Netherlands.[12] The OML supplements administrative data with surveys among students, parents, and teachers taken in kindergarten, grade 6, and grade 9. The administrative data include the exit test score and teacher recommendation from grade 6, as well as track placement from 7th to 9th grade. Surveys collect additional information on demographic indicators, socio-economic status, cognitive and non-cognitive skills of students.[13]

The OML collects data from secondary schools across Limburg and for primary schools in the south of Limburg. The coverage of schools is around 90 percent within each respective area. For students that were not part of the primary school data collection, teacher recommendations and exit test scores were retrieved from the administrative data of secondary schools.[14] Secondary school data collection started in 2010. The basis of our data sample are the 2012, 2014, and 2016 9th grade cohorts, as the 2010 cohort lacks information on key variables. Each cohort contains around 9,000 observations. Below we describe the key variables in more detail.

3.1 Secondary school tracks

Administrative data register the track students attend in grades 7, 8 and 9. This includes any indication of mixed tracks that can still occur in grade 7.[15] We focus on tracking in the highest two tracks in Dutch secondary education. The assignment of students to the vocational track and

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

[13]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. Parents can decide to withdraw from the survey, based on the passive consent principle.

[14]Non-participating schools include special education schools, schools with a philosophy not to test children, and schools unable to plan the survey activities.

[15]Mixed grades are not clearly identifiable for 15% of students, which are excluded from the sample. As missings are predominantly at the school-level, this attrition is unlikely to be a major concern. Additionally, school switchers may be overrepresented among missing values, but since all schools in the sample offer both T^M and T^H, this is not directly connected to being retracked to the lower of the two tracks. The sample with missing data is slightly less affluent, but this disappears when we control for school fixed effects (which are also included in the main analysis).

its subtracks is not transparent. Schools can differ in the manner in which they organize subtracks, and correlations indicate that they are much 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 restrict the sample to students that are assigned to either a T^M class, a T^H class or a mixed T^M/T^H class in grade 7.

We define a student as being assigned to a high track if they were placed in a T^H class in 7th grade, as opposed to T^M or T^M/T^H, depending on the school. If the school has a T^M/T^H and a T^H track, the former will be the low track counterfactual. If the school only has T^M and T^H tracks, then T^M will be the low track counterfactual.[16] This implies that our treatment effect is compared to two joint counterfactuals. We choose this approach because the alternative would imply endogenously excluding schools based on their sorting policy. The mixed T^M/T^H track in grade 7 comprise 60% of the group that is in the 'lower track', and thus carries more weight. The group in the mixed track still has a substantial chance of getting into T^H one year later, in which case their treatment would only consist of the one year being in a mixed grade. Robustness analysis will assess to what extent this may affect the results of the analysis.

3.2 Cognitive variables

The data contain scores for the high-stakes exit test taken in grade 6, and for low-stakes tests on math and language taken in grade 9. The 9th grade tests are developed for the research project and administered digitally.[17] Because the time for testing students was limited, students are randomly assigned to take either the math test, the language test, or (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

[16] When a school only has a T^M/T^H 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.

[17] The language test contains items on comprehensive reading taken from PISA 2000-2006 tests (OECD, 2011) and items on spelling and word knowledge taken from a Dutch Cohort Study (COOL5-18, see (Zijsling et al., 2009). The math test contains items taken from the PISA 2000-2006 tests, items from COOL5-18, and items from a Belgian Study on School Feedback (Verhaeghe and Van Damme, 2007).

data at the school-cohort level, indicating that schools decided not to administer the test in class.[18] As we include school and year fixed effects, this is not a major concern for the internal validity of the estimates. Students in the T^M and T^H tracks were provided with the exact same test, allowing us to directly compare scores. Final scores are constructed using Item Response Theory (IRT).

3.3 Non-cognitive variables

The data contain a wide array of non-cognitive measures. To prevent identifying statistically significant results due to multiple hypothesis testing, we select those that the educational literature deems most relevant. 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, the two facets of 'grit' which have been shown as highly predictive traits for educational outcomes in, e.g., Duckworth et al. (2007).

We additionally use motivation, self-confidence and educational aspirations as non-cognitive outcomes, in light of the impact of track assignment on class rank. Attending the higher track versus the lower track at the margin implies being at the bottom rather than the top of the ability distribution, which can affect motivation and self-assessment. On the other hand, being in the higher track is connected to higher levels of post-secondary education and could thereby lead to higher educational aspirations. Additionally, the attenuation of relative age effects over time implies that class rank develops differently for (marginal) students of different relative age, The effects of tracking on these outcomes may therefore differ as well across relative age. For educational aspirations, we create two dummy outcome variables taking value 1 if students expect to obtain a diploma from T^H (Exp. Track), or expect to complete university (Exp. Univ.), respectively. Expectations are not measured in the 2012 cohort. All other non-cognitive outcomes are constructed using factor

[18] Around 7% of testing data is missing at the individual level, in all likelihood because students were absent at the testing moment. Exit test scores and background characteristics are not correlated with having missed the test.

analysis. Appendix A9 provides individual items.[19]

3.4 Demographic variables

We use exact birth dates to construct a relative age measure. The cutoff date for formal education in the Netherlands is the 1st of October, which is not strictly enforced but has high compliance. Our measure of relative age 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 among 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.

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, and consequently have lower probabilities of attending an academic track. We shortly review whether this also applies to our sample, in Appendix A1. Our results are highly consistent with previous literature. Older students perform better on cognitive tests, and are more often recommended to higher tracks by their teacher (also when controlling for test performance). This translates into higher odds of high track attendance: the predicted probability of T^H attendance is around 8.1 percentage points higher for the very oldest student compared to the very youngest student in class.[20] We further identify that relative age effects for achievement reduce as students grow older, and confirm that older students have better developed non-cognitive skills at age 12.

[19] For a small number of students, self-reported items are missing for some non-cognitive outcomes. We impute responses from the parental questionnaire, administered in the same period. While this imputation marginally increases the precision of the estimates, it does not change the overall conclusions.

[20] There is no relationship 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.

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 final test score to a recommended track, these thresholds are not formally enforced. Figure A2.1 in the Appendix plots the primary school exit test score against the probability of being assigned to the highest track. The latter increases smoothly with the exit test score. There is no clear discontinuity for any particular score, when all schools in our data are pooled together.

4.1 Estimating school-specific thresholds

As there is no nationally enforced cut-off point for being admitted to the high track, we assess whether school-specific thresholds are used when placing students into T^H. This approach is motivated by the fact that schools are lawfully obliged to base their admission or sorting policy on at least one of the instruments (teacher recommendation or exit test score), although they are free to decide *how* to use that information.[21] Given the institutional and legal context, the exit test score should therefore be a leading determinant of student sorting at the school level. This should hold even more so in this sample, as all schools offer both T^M and T^H. As such, thresholds are not about entry to the school as a whole, and are therefore less likely to be driven by capacity constraints.

Public information on the exact decision process of schools in setting thresholds is limited, 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% have formal agreements at the municipal level (Inspectie van het Onderwijs, 2014). Agreements at the municipal level are especially common in bigger cities.[22] To the best of our

[21]The Education Inspectorate reviews whether individual schools adhere to this obligation, and do not consider other criteria (apart from those allowed under capacity constraints). The legal obligation has changed since 2015, after which schools are only allowed to consider the teacher recommendation. All of the cohorts in the empirical analysis have taken the exit test before this point.

[22]For example, all schools in Amsterdam have been subject to a formal agreement that obliges them to allow all students to T^H above a certain exit score threshold; see `http://www.onderwijsconsument.nl/citoscores-en-schooladviezen-citobandbreedtes/`. Other cities have similar agreements, or 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.

knowledge, such written agreements are lacking for the region of Limburg. The school-specific thresholds discussed in this section show substantial within-region correlation, suggesting that schools in Limburg operate similarly, though less formally.

Porter and Yu (2015) show that program effects in an RD design can 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. Therefore, the fact that the threshold is unknown does not affect the treatment estimates in large samples. An application of this approach in a setting that is similar to ours is provided by Booij et al. (2016). They investigate the effects of a secondary education program for gifted children, wherein the acceptance threshold is not known in every period. By comparing estimated thresholds with the observed thresholds, they find that this methodology leads to accurate predictions of the true thresholds. This approach differs slightly from that of Porter and Yu (2015) who assume that there is a treatment effect, which may not hold in the setting of Booij et al. (2016), nor in ours.

We estimate school-specific threshold scores, using the same approach as Booij et al. (2016). 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. Students scoring above the school-specific threshold are considered treatment eligible, while those scoring below are not. Table 2 shows that in our sample, 83% of students starting secondary school are placed into tracks in line with their treatment eligibility.

Table 2: Treatment eligibility and track assignment

	Scores above threshold	Scores below threshold
Attends T^H	2,685 (44.2%)	513 (8.4%)
Attends T^M	517 (8.5%)	2,365 (38.9%)

As using score thresholds is not strictly enforced, it is possible that some schools do not employ an effective threshold.[23] 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 'threshold', or uncommonly bad draws on the left hand side of the 'threshold', leading to biased results. Figure A2.2 in the Appendix separately plots the jumps around empirically estimated thresholds for each secondary school. For a number of schools, the discontinuity is highly fuzzy. As such, for our main estimates, we restrict the sample to the schools for which the jump at the threshold is statistically significant at the 5% level.

After imposing this additional restriction, 70% of the sample remains. This indicates that the majority of schools in our data use a threshold to assign students to the highest track. Moreover, the exercise confirms the use of a unique threshold per school. There are only three schools for which a statistically significant (weaker) jump is identified at a 'second' threshold, in line with what we would expect just by chance. Excluding these three schools has no effect on our main estimates.[24]

The left-hand side of Figure 2 plots the discontinuity around the estimated thresholds, as such representing our first stage. There is a precisely estimated jump at the threshold: the probability of being assigned to the high track approximately doubles, from 30% to 60%. The still relatively large share of non-compliance directly around the threshold likely reflects that some schools (also) consider teacher recommendations, which are spread across multiple test scores. Moreover, students may also forego the opportunity of attending the high track. Nonetheless, there is a clear jump at the threshold, and the RDD approach is appropriate as long as other determinants of outcomes are

[23] While schools are lawfully obliged to consider either the test or the teacher recommendation, they may, for example, consider a very broad band of test scores for T^H. In this case, discontinuities may not be identified.

[24] Including the full set of schools in the analysis also leads to highly similar results as in our main analysis. 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.

smoothly continuous around the threshold.[25]

Figure 2: Discontinuity in high track attendance around the school-specific threshold

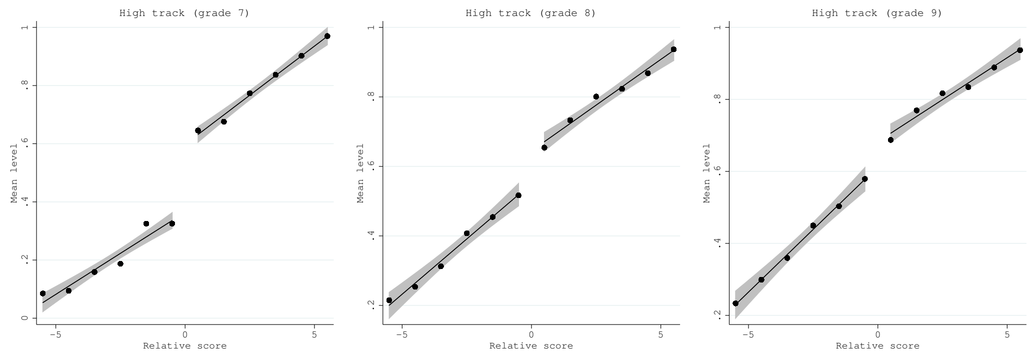


Figure 2 also shows assignment to the high track in subsequent grades. Discontinuities gradually reduce across grades, which reflects that students can be re-tracked in lower secondary education. Still, there appears to be a positive jump in grades 8 and 9. The treatment effect thus not only represents being in a higher track in grade 7, but also a higher probability of being in that high track in the two subsequent grades.

4.2 Fuzzy RD design

Given these established thresholds, we use a fuzzy RD design that instruments attendance of the high track in grade 7 with a dummy variable that takes value 1 if the student scores above the school threshold. The first stage is given by:

$$HT_{ic} = \lambda_0 + \lambda_1 I_{ic}(S_{ic} \geq \overline{S_s}) + \lambda_2 RA_{ic} + \lambda_3 S_{ic} + X'_{ic}\lambda_4 + \tau_c + \sigma_s + \mu_{ic} \tag{1}$$

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

[25] Thresholds are estimated per school and per cohort. This increases first stage power, and reflects that thresholds can be updated from year to year (the score bands reported by the testing bureau also change every few years). School-cohort thresholds are potentially more vulnerable to schools endogenously adjusting the threshold to the student population each year, which is why we estimate results with constant school-specific thresholds for robustness (see Appendix A7).

the high track. We include a linear control function for the test score S_{ic}, as this provides the best fit. RA_{ic} is the relative age of each student, measured by the date of birth of the student. 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:

$$Y_{ic} = \kappa_0 + \kappa_1 HT_{ic} + \kappa_2 RA_{ic} + \kappa_4 S_{ic} + X'_{ic}\kappa_4 + \tau_c + \sigma_s + \varepsilon_{ic} \tag{2}$$

Y_{ic} represents the outcome variable three years after track placement. These outcome variables can be classified in two categories: (i) cognitive and (ii) non-cognitive. Cognitive outcomes are track level, math scores and reading scores. Non-cognitive outcomes are conscientiousness, neuroticism, need for achievement, persistence, school motivation, self-confidence, and educational expectations. Standard errors are heteroskedasticity-robust, and correct for clustering at the class level. The optimal bandwidths are calculated using the Calonico et al. (2014) approach.[26]

To analyze treatment heterogeneity by relative age, the model is expanded with an interaction between T^H attendance and relative age, which is instrumented with an interaction between treatment eligibility and relative age. This adds a second instrument and a second first stage regression:

$$HT_{ic} = \lambda_0 + \lambda_1 I_{ic}(S_{ic} \geq \overline{S_s}) + \lambda_2 I_{ic}(S_{ic} \geq \overline{S_s}) * RA + \lambda_3 RA_{ic} + \lambda_4 S_{ic} + X'_{ic}\lambda_5 + \tau_c + \sigma_s + \mu_{ic}$$

$$HT_{ic} * RA_{ic} = \rho_0 + \rho_1 I_{ic}(S_{ic} \geq \overline{S_s}) * RA + \rho_2 I_{ic}(S_{ic} \geq \overline{S_s}) + \rho_3 RA_{ic} + \rho_4 S_{ic} + X'_{ic}\rho_5 + \tau_c + \sigma_s + \nu_{ic}$$

$$O_{ic} = \eta_0 + \eta_1 HT_{ic} + \eta_2 HT_{ic} * RA_{ic} + \eta_3 RA_{ic} + \eta_4 S_{ic} + X'_{ic}\eta_5 + \tau_c + \sigma_s + \varepsilon_{ic} \tag{3}$$

The regression models rely on interactions as this exploits all variation in relative age and provides

[26] As the running variable is discrete, we set the threshold in the regression models at -0.5 to ensure that bandwidths are symmetrical.

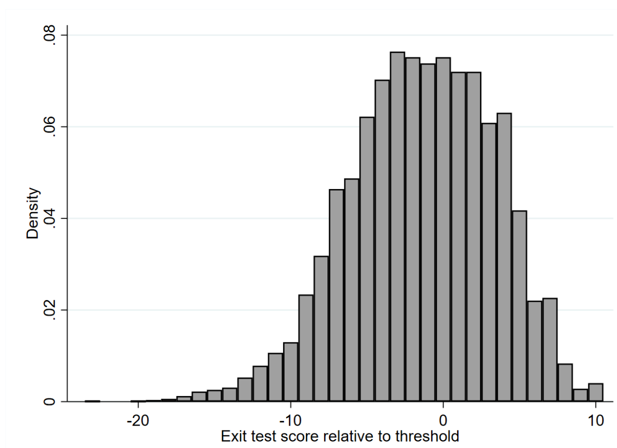
more precise results of treatment heterogeneity. We complement this with a graphical analysis which splits results between the older and younger half of the relative age distribution, as this is more visually intuitive. This follows recent work on 'difference in discontinuity designs' for heterogenous treatment effects in an RDD framework; see Grembi et al. (2016); Gilraine (2020).

4.3 Assumptions and tests

One main assumption behind our approach is that students are not able to self-select on any side of the threshold. As thresholds are not strictly enforced, (specific) students may 'shop around' by applying to different schools until they are placed in the high track. This type of selection is a threat to the validity of our results and, if present, prevents us from interpreting the results causally.

To address this concern, we first 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 3 shows that the distribution is very smooth. A McCrary density test confirms this (McCrary, 2008). This also implies that schools do not adjust the threshold to maximize the inflow of students. There is little incentive for schools to do so, as there is no school in our sample that offers only T^H.

Figure 3: Distribution of test score around school threshold



We further conduct tests that allow us to relax the no-sorting assumption in favor of two conditions which we believe to be less restrictive. First, we argue that *everything else equal, students prefer to attend the school which is closest to their home.*

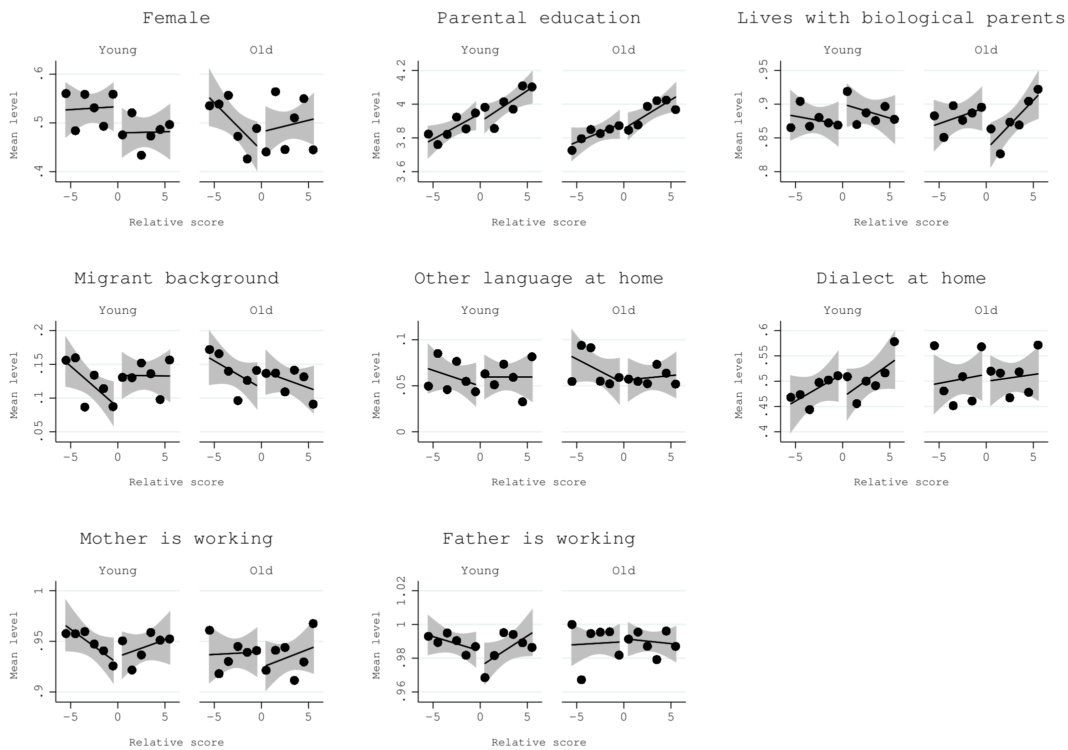
It appears unlikely that students/parents move residence based on the threshold score of the nearest secondary school, especially since this is not public information. Hence, the main concern lies in endogenous commuting of students to schools. Figure A4.1 in the Appendix plots the population of students, 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. The remainder largely live in an urbanized area with a high concentration of schools. Around 40% of students that do not attend their closest school travel less than one kilometer extra. It appears less likely that these students select the attended school based the sorting policy, as they are nearly indifferent in terms of commuting distance. We further find that there is no statistically significant jump in commuting distance at the threshold (Figure A4.2), and that students not commuting to the nearest school appear very similar to those that are (Table A4.1). While potential sorting issues thus appear limited in this setting, 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 (see section 6).

These figures lead to our second underlying assumption, namely that *student mobility is problematic if students are not treatment eligible in the closest school, and travel further away to be in a higher track.*

Endogenous self-selection of students to schools based on the school threshold occurs only when wanting to attend T^H, since going to T^M is always possible. Only 5% of students in our sample attends T^H in a school that is further away than the closest school and that has a lower score threshold than the closest school. These students originate disproportionately (60%) from the South-East region of Limburg. Again, this is likely due to the higher availability of schools in this area. In additional robustness checks, we show that excluding this region yields very similar estimates (Tables A4.4 and A4.5).

As a final validity test, we analyze whether observable characteristics are balanced around the threshold. Figure 4 plots average values for all controls against the exit test scores relative to the

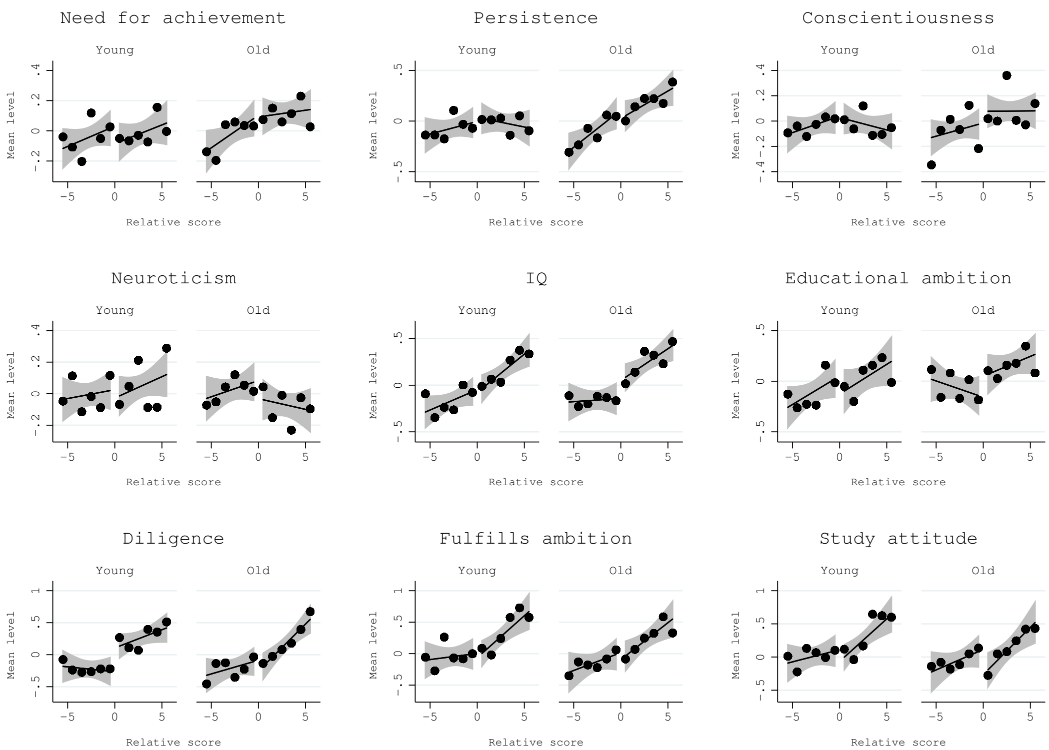
Figure 4: Balancing tests around the threshold: controls



threshold. We provide these separately for younger and older students; results for the full sample can be found in Figure A2.6 in the Appendix. No jumps are observed at the threshold, confirming that the instrument is exogenous to these characteristics. This is also confirmed by the estimates from a regression of the instrument on all controls, in Appendix Table A2.2. Joint significant tests fail to be rejected in all specifications.

We also run balancing tests across grade 6 measures of our outcome variables and other pre-treatment measures of cognitive and non-cognitive skills. Figure 5 shows that track assignment at the margin is also orthogonal to lagged measures of need for achievement, persistence, conscientiousness and neuroticism, several teacher-reported measures of non-cognitive skills, and IQ. Only one out of 18 discontinuities provided in Figure 5 is statistically significant at the 5% level (higher diligence for the younger half). Figures for the full sample can be found in Appendix Figure A2.7.

Figure 5: Balancing tests around the threshold: 6th grade outcomes



5 Results

We now provide results of the effect of high track attendance at the achievement margin and its interaction with relative age. We first provide graphical results, separately for old and young students in a difference-in-discontinuity design (5.1), and then provide results from the model that interacts with relative age (5.2 and 5.3).

5.1 Difference in discontinuity design

Figures 6 to 8 provide graphical results for the difference in discontinuity design. Outcomes are corrected for covariates, to increase precision (Appendix 2 provides raw results, as well as figures for the full sample). Figure 6 plots the reduced form for the two cognitive tests, while Figures 7 and 8 provide non-cognitive outcomes.

Figure 6: Reduced form graphs: test scores (9th grade)

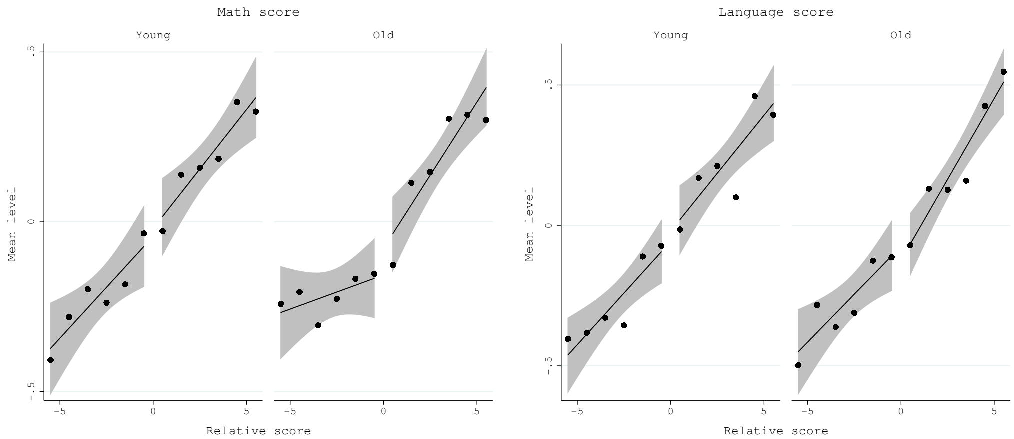


Figure 7: Reduced form graphs: personality traits (9th grade)

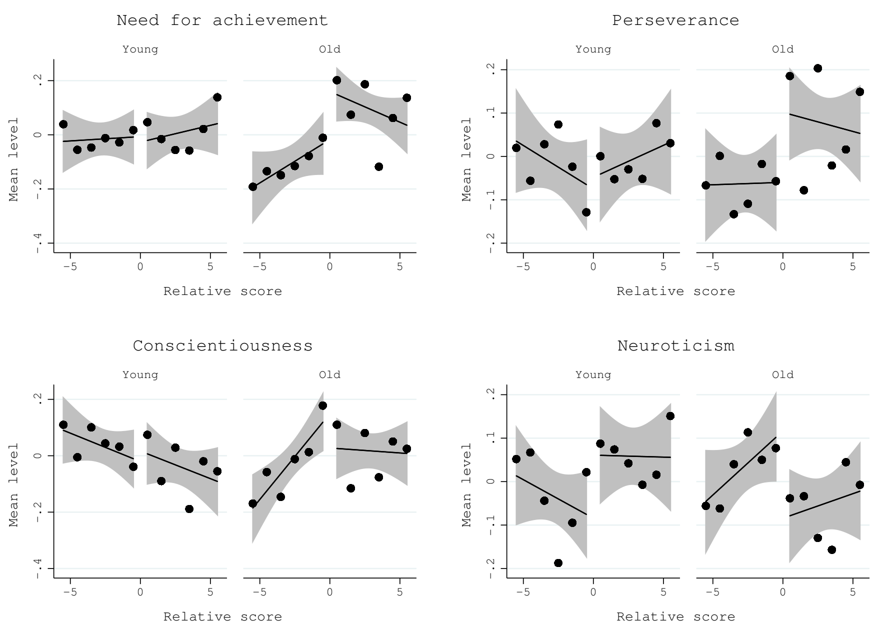
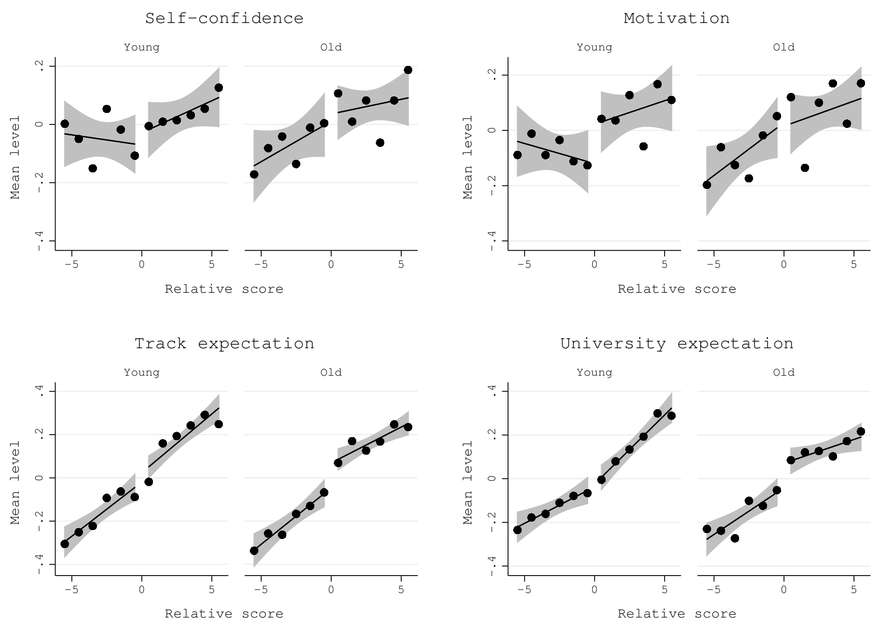


Figure 8: Reduced form graphs: other non-cognitive outcomes (9th grade)



Three years after track placement, math and reading achievement increase smoothly across the relative exit test score, with no jump at the threshold for either group of students. Figure 7 shows a smooth pattern for the personality traits for young students as well, but suggests positive gains for older students in terms of higher need for achievement and persistence, and lower neuroticism. Figure 8 suggests positive increases in educational expectations for older students, while the pattern is rather smoothly continuous for other outcomes. While suggesting positive non-cognitive gains for older students, the imprecision in this visual analysis is rather high. Regression analysis is needed for conclusive results. We present these results for both cognitive and non-cognitive outcomes in the following two subsections.

5.2 Cognitive outcomes

Table 3 presents the second stage results using equation (3), where the dependent variables are (i) being in T^H in grade 9 (ii) math score in grade 9 (iii) reading score in grade 9. Appendix Table

A3.1 shows first stage estimates.[27] All results reported below include the full set of controls and school and year fixed effects. Appendix Table A7.3 provides highly similar estimates for a model without controls, confirming that the instrument is orthogonal to observed characteristics.

Table 3: Cognitive Skills

	Track 9th grade	Math Score	Reading Score
High Track	0.346**	-0.164	-0.093
	(0.174)	(0.378)	(0.332)
High Track * RA	-0.007	0.016	0.016
	(0.011)	(0.027)	(0.024)
RA	0.004	-0.023	-0.013
	(0.006)	(0.016)	(0.013)
N	2,641	1,650	1,693
KP stat	17.60	13.13	10.03
Optimal BW	±3	±4	±4

Note: The regressions include a linear control function for the exit test score on each side of the threshold, and controls for gender, socio-economic status, 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). Math and reading estimates are in standard deviations. Standard errors are robust and corrected for clustering at the class level. The Kleibergen-Paap (KP) statistic assesses instrument relevance. $*p < 0.10, **p < 0.05, ***p < 0.01$

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 (34.6 percentage point). Naturally, the difference in high track assignment equals 1 in grade 7, so the difference has narrowed. This is to be expected as students in T^M/T^H classes 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 cognitive skills at the achievement margin, although the results should be interpreted with caution as the precision of these estimates is low. Nonetheless, the point es-

[27] 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 Appendix Tables A7.1 and A7.2) and 7.03 for the case with two instruments (which applies to the main results portrayed in Table 3).

timates suggest that the higher track environment does not translate into more efficient learning in these areas in lower secondary education. This also holds when we combine all math and language items into one overlapping IRT score to increase sample size and precision (Appendix Table A8.1). This suggests that, for cognitive achievement, track allocation is efficient such that no marginal students would benefit from switching track. It may be surprising that higher peer quality in T^H does not translate into achievement gains, but these findings are in line with other research for the Netherlands that identifies weak or no short-run achievement effects of track assignment or tracking age (Borghans et al., 2019, 2020).

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 despite the expected lower cognitive ability of older marginal students. The point estimate of the interaction is in the expected direction (i.e. the treatment effect is lower for older students) but statistically insignificant and low (suggesting a difference in treatment effects of around 0.09 between the very oldest and the very youngest student in class). For math and reading achievement, treatment effects do not differ significantly between the older and the younger students.

For comparison, table A7.1 in the Appendix estimates the model without interactions with relative age. These coefficients are highly similar to the estimates in Table 3, in line with the fact that the effect of high track attendance on cognitive skills is not heterogeneous across relative age.

5.3 Non-cognitive outcomes

Table 4 presents the results for the 9th grade non-cognitive outcomes. The top panel presents estimates for personality traits, and the bottom panel for other non-cognitive variables. Appendix Table A3.2 provides first stage estimates.

The top panel shows that attending the high track benefits older students (more) in terms of higher need for achievement and perseverance, and lower neuroticism. The estimates are statistically significant and substantial. No statistically significant treatment effects are identified for conscientiousness, neither for the baseline effect nor across relative age.

The bottom panel shows that marginal students obtain higher expectations from attending T^H. This suggests that the lower rank within class is 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 (2020) for the United Kingdom). Results on relative age heterogeneity for the educational expectations are somewhat inconclusive. Estimates are very low (with negative sign) in the regression model in Table 4 while effects are stronger for older students when splitting the sample (Figure 8 and Appendix Table A7.7). Baseline effects for confidence and motivation also have positive (and non-negligible) point estimates, but these are statistically insignificant. Interactions with relative age are small and insignificant, both in the interaction model and the split analysis.

Average effects, for the model without age interactions, are presented in Table A7.2 in the Appendix. They show that the favourable effect of the high track is also present for the (local) average student for perseverance, but not for 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 only improves need for achievement and neuroticism for the relatively older students in class.

The point estimates for the statistically significant interactions are large. They imply that the positive treatment effect of attending the high track is around 0.7 standard deviations larger for the very oldest student in class, compared to the very youngest student in class, with respect to need for achievement and perseverance, and around 1.0 standard deviation larger with respect to neuroticism. Alternatively, splitting the sample by young and old gives effects for the latter group of 0.66 to 0.83 (Appendix Table A7.7). It should be emphasized that the imprecision of these results is also high. Confidence intervals from Table A7.7 show that the lower bound of these effects do not exceed 0.10. As such, we refrain from drawing any strong conclusions about the

Table 4: Non-Cognitive Skills

	Need Achiev.	Perseverance	Conscient.	Neuroticism
High Track	0.271	0.305	-0.198	0.443
	(0.346)	(0.400)	(0.366)	(0.426)
High Track * RA	0.057**	0.059**	0.015	-0.084***
	(0.025)	(0.028)	(0.029)	(0.028)
RA	-0.037***	-0.025	-0.004	0.038**
	(0.013)	(0.015)	(0.015)	(0.015)
N	2,733	2,122	2,131	2,128
KP stat	25.02	17.69	18.33	17.36
Optimal BW	±4	±3	±3	±3
	Confidence	**Motivation**	**Exp. Track**	**Exp. University**
High Track	0.308	0.381	0.373**	0.309*
	(0.276)	(0.309)	(0.189)	(0.175)
High Track * RA	0.007	-0.001	-0.006	-0.002
	(0.021)	(0.018)	(0.009)	(0.010)
RA	0.007	0.004	0.006	0.001
	(0.013)	(0.010)	(0.005)	(0.006)
N	2007	2636	2241	2233
KP stat	19.01	23.86	23.43	22.90
Optimal BW	±4	±5	±5	±5

Note: 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 model includes a linear control function for the exit test score on each side of the threshold, and the control variables discussed in Table 3. Standard errors are robust and corrected for clustering at the class level. The Kleibergen-Paap (KP) statistic assesses instrument relevance. $*p < 0.10, **p < 0.05, ***p < 0.01$

effect sizes, beyond the fact that these are statistically significant and appear to be non-negligible, for the specific group of older students at the achievement margin. At the same time, a substantial effect is not implausible in this setting as the treatment is strong and directly impacts the daily learning environment of students for a period of multiple years. Moreover, it moves students between either the very bottom of the class or the very top, which likely relates to the formation of the non-cognitive skills that we study. This mechanism may be amplified by the fact that students at the bottom of the class face the incentive of working hard to prevent downgrading or retention while those at the top do not face those same incentives. Strong effects are also in line with the comparatively higher malleability of these skills in early adolescence (Cunha et al., 2010).

Taken together, we identify no negative interaction between high track attendance and relative age for cognitive outcomes, while we identify positive interactions for non-cognitive 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 can be considered 'above' one's cognitive ability induces a need for higher effort, in order to keep up with that more challenging environment. As such, older students compensate by working harder, reflected by increases in need for achievement and perseverance. These positive spill-overs on non-cognitive skills may help older students to remain in the highest track and perform as well as younger students on achievement tests, despite the latter being more cognitively able at the margin when attending T^H.

Psychological literature can provide insights why this not only leads to enhanced 'grit', but also to decreases in neuroticism. Our measure of neuroticism considers two of its facets: stress/anxiety and depression. Analysis shows that the interaction with relative age is predominantly present for the former. Psychological studies find that there are two main responses to stress: problem-solving (producing relief and a better capability to handle future stress) and avoidance (further enhancing future stress); see, e.g., Penley et al. (2002); Ursin and Eriksen (2004). Simplifying, our results may indicate that the pressures of being an academically marginal student in the high track are handled by older students through working harder, and are handled by younger students through procrastination and increased stress. These different responses may in turn relate to differences in personality at baseline: Mühlenweg et al. (2012) find that at age 11 younger students are more irritable and oversensitive, and less adaptive than their older peers. Analysis on 6th grade data in our sample also finds evidence of relative age effects for some personality traits at age 12 (see Appendix Table A1.4). These baseline differences may subsequently produce different gains from attending the high track.

The evidence for older students compensating lower cognitive ability with higher effort remains indirect. We therefore assess more direct measures of effort in the data. Two are available: time spent on homework and 'school attitude'. The latter asks students whether they work hard on assignments, learn for tests, etc. (see Appendix Table A9.7). It thereby relates to grit, but is more directly aimed at actual behavior in school. Table A8.2 presents results for these measures. We identify positive interaction effects for homework done at school and for parent-reported school attitude, but statistically insignificant estimates for homework done at home and student-reported school attitude.[28] Hence, there is some direct evidence of enhanced effort for older students from attending the higher track.

6 Robustness

We conduct a range of robustness tests, which are extensively documented in the Appendix. The traditional bandwidth sensitivity test, provided in Appendix A6, provides very consistent results. We further discuss robustness analysis centered around four issues: alternative instruments, other discontinuities around the threshold, different counterfactuals, and validity of the relative age measure.

Appendix A4 provides several analysis on the potential issue of selective mobility. Primarily, an alternative instrumental variable that measures treatment eligibility for the closest school, rather

[28] Homework time at school refers to time during the scheduled school day and not after-school hours (which are asked in a separate item). This can take the form of individual working time during regular classes, or scheduled unsupervised study hours.

than eligibility with respect to the attended school, provides findings consistent with the main approach. The same applies to an approach where thresholds are calculated across regions or school boards (exploiting the fact that there appears to be some coordination in the setting of school-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. While naturally less precise, the pattern of results is highly similar.

The more general issue at hand is whether there are no discontinuities in other variables at the threshold. Appendix Table A7.4 shows that results are similar when we include lagged measures of non-cognitive outcomes, which are available for a sub-sample of students. This confirms the results for the balance tests in Appendix Figure A2.6. We also provide results for a model that includes the teacher recommendation as an additional control (see Appendix Tables A7.10 and A7.11). Teacher recommendations jump slightly at the threshold, which is expected because 'feeding' primary schools may coordinate with secondary schools when mapping exit test scores to recommendations (or they may both follow the suggested recommendation from the testing bureau). Including it as a control thus weakens the first stage and leads to higher standard errors in the second stage. At the same time, the model may correct for potential discontinuities in characteristics that are not observed in our data but are incorporated by the teacher when giving the recommendation (e.g. other non-cognitive skills). The tables show that point estimates are highly similar when controlling for the teacher recommendation, indicating that such characteristics do not jump at the threshold.

A third issue is the existence of two counterfactual situations (the T^M track and the T^M/T^H track). We separately analyze treatment effects with respect to each counterfactual. Results are presented in Appendix A5. They should be interpreted with care as the samples are selected, and results are imprecise. We identify a substantially stronger effect on T^H attendance in grade 9 for the T^M counterfactual. This is not surprising, since T^M/T^H students still have a direct path to T^H. The non-cognitive gains for the older students are present with respect to both counterfactual situations. The appendix elaborates on potential mechanisms, but this analysis shows that the identified gains in non-cognitive skills for older students are not dependent on the counterfactual that we employ.

Two final tests pertain to the validity of the interaction estimate. First, our main model forces covariates to have a homogeneous effect on the outcome across relative age. The relative age interaction may therefore pick up on heterogeneity between relative age and other covariates (including the running variable). We therefore report separate estimates for old and young students in Appendix Table A7.7 (without age interactions). These estimates also formalize the visual analysis from the difference-in-discontinuity approach provided in Section 5.1. While point estimates are highly imprecise, they confirm the pattern of results identified in the main analysis. Finally, some studies have questioned the exogeneity of birth dates; see, e.g., Buckles and Hungerman (2013); Barua and Lang (2016). 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. It does correlate with having repeated a grade, but controlling for retention or excluding the youngest birth quarter (for which retention rates are especially high) 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.

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 affects non-cognitive skills in a heterogeneous manner across relative age. Older students benefit from attending the higher track in terms of higher perseverance, higher need for achievement and higher emotional stability. We show that results are not driven by selective mobility of students to schools, and are robust to alternative bandwidths, alternative approaches for threshold estimation, and the inclusion of lagged outcomes. Apart from shedding new light on the interaction between tracking and relative age, this study is also the first to estimate a causal effect of track assignment on non-cognitive skills in general, and shows that attending a higher track can produce benefits in non-cognitive development for certain groups of students.

Earlier research on relative age has documented that relatively younger students are tracked below their academic potential. This automatically implies that relatively older students are at-risk of being tracked into too demanding tracks. In light of the evidence on the decreasing 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. Hence, while these students might fall short of the cognitive ability level that is believed to be necessary for the highest track (i.e. they would fall short of achievement thresholds if tests would be corrected for relative age), they appear to better handle the stress that comes with being a marginal student, and apply more effort in response. 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 previous studies, 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, we show that this also holds when comparing younger students in a low track with younger students in a high track, at the achievement margin. Moreover, we have shown that the 'overtracking' of older students 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 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 attendance of the higher track, then there may be a group of younger students with equivalent non-cognitive treatment effects further below the threshold (i.e. those with the same age-corrected test scores as the marginal older students). The results can therefore also be interpreted more generally in that demanding learning environments can benefit non-cognitive skills.

We have analyzed the effects of track selection specifically for the margin of attending the highest 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. Our study identifies favorable effects on non-cognitive skills. An important open question is the extent to which these effects, which predominantly seem to reflect higher effort levels, are sustainable in the long run.

Funding

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.

Ethics declarations

The authors declare that they have no conflict of interest.

References

Arrow, K. J. (1971). A utilitarian approach to the concept of equality in public expenditures. *The Quarterly Journal of Economics 85*(3), 409–415.

Barua, R. and K. Lang (2016). School entry, educational attainment, and quarter of birth: A cautionary tale of a local average treatment effect. *Journal of Human Capital 10*(3), 347–376.

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

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

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.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica 82*(6), 2295–2326.

Cascio, E. U. and D. W. Schanzenbach (2016). First in the class? Age and the education production function. *Education Finance and Policy 11*(3), 225–250.

Cunha, F., J. J. Heckman, and S. M. Schennach (2010). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica 78*(3), 883–931.

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.

Gilraine, M. (2020). A method for disentangling multiple treatments from a regression discontinuity design. *Journal of Labor Economics 38*(4), 1267–1311.

Grembi, V., T. Nannicini, and U. Troiano (2016). Do fiscal rules matter? *American Economic Journal: Applied Economics 8*(3), 1–30.

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., 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.

Jackson, C. K. (2018). What do test scores miss? The importance of teacher effects on non–test score outcomes. *Journal of Political Economy 126*(5), 2072–2107.

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, National Bureau of Economic Research, nr. 19656.

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 (2020). Top of the class: The importance of ordinal rank. *The Review of Economic Studies 87*(6), 2777–2826.

OECD (2011). *PISA take the test: Sample questions from OECD's PISA assessments*. Paris, France: OECD Publishing.

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

Penley, J. A., J. Tomaka, and J. S. Wiebe (2002). The association of coping to physical and psychological health outcomes: A meta-analytic review. *Journal of behavioral medicine 25*(6), 551–603.

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.

Ursin, H. and H. R. Eriksen (2004). The cognitive activation theory of stress. *Psychoneuroendocrinology 29*(5), 567–592.

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.

Yamaguchi, S., H. Ito, and M. Nakamuro (2020). Month-of-birth effects on skills and skill formation. *Available at SSRN: https://ssrn.com/abstract=3632546*.

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.