The long-term effects of early tracking

Abstract: The Dutch education system provides variation in the timing of tracking of students. Students can enroll in a tracked class right away, or postpone tracking one or two years by enrolling in a comprehensive class. Using the VOCL cohort study of '93, the effect of early tracking on several labor market outcomes and educational outcomes are estimated. An instrumental variable approach using regional variation in the supply of categorical classes is implemented to overcome selection bias. The results show that MAVO advice students that are tracked early are less likely to complete higher education, whereas VWO advice students that are tracked early are more likely to complete university.

Master thesis Policy Economics

Koen van der Ven

453506

Supervisor:

Prof. dr. Dinand Webbink

Second assessor:

Matthijs Oosterveen MSc

July 2017

ERASMUS UNIVERSITEIT ROTTERDAM

ERASMUS SCHOOL OF ECONOMICS

Acknowledgements

I want to thank prof. dr. Dinand Webbink for his supervision and useful comments I received while writing this thesis. He also encouraged me to pursue this topic, and provided me with the opportunity to do the research at the CPB Netherlands Bureau for Economic Policy Analysis. I also want to thank drs. Sonny Kuijpers for advising me with some of the day-to-day problems I encountered. Finally, I want to thank dr. Karen van der Wiel and everyone else at the CPB I didn't mention for their comments and discussion.

Table of Contents

Introduction

On January 25th, 2017, the chairman of the Dutch Council for Secondary Education (VO-raad) presented a set of policy measures to increase the equality of opportunities in secondary school to the Dutch parliament (Rosenmöller, 2017). He proposed to review the school advice students receive after primary school after two years of secondary education, and to improve the availability of comprehensive classes. The second measure aims to delay tracking for students. Another recent report on the Dutch education system by the OECD (2016) recommends delaying the age of tracking together with reducing the number of tracks available in order to improve the mobility of students within the system. Both documents comment on the Dutch educational tracking system being imperfect.

The practice of educational tracking is separating a cohort of students into classes based on their academic ability. Students within the same class are subject to the same curriculum, whilst students from different classes are subject to differing curricula. The aim of tracking is to increase learning among all types of ability groups. Proponents of tracking argue that it is more efficient in raising average test scores of students than not tracking at all. Homogenous classes allow teachers to better cater their lessons toward the specific ability level of the students in his class. The lower ability groups benefit from properly paced instructions to keep up with classmates; the higher ability groups are able to reach their full potential because of focused curricula. Another, less-mentioned argument in favor of tracking is that it enables students preferring an academic or a vocational career to differentiate early on (OECD, 2016). Opponents, however, argue that tracking increases inequities, as the lower tracks tend to consist of students from disadvantaged backgrounds such as minorities (OECD, 2011). Besides the inequality argument, opponents also argue that the so-called peer effects are nonlinear. This means that if students are not separated, the lower-ability students benefit from interaction with higher-ability students, while the higher-ability students do not receive a penalty from the lower-ability students (Woessmann, 2009).

The timing of tracking is also important in this debate, as is apparent from the letter of the VO-raad to the Dutch parliament and the recent OECD report. In order to improve equality in education, the OECD (2011) argues to delay tracking to limit the negative effect of early tracking, as weaker students are systematically disadvantaged earlier on in the case of early tracking. Brunello, Giannini and Ariga (2007) also mention early tracking is noisy, since young students' capabilities are imperfectly identified. Whether (early) tracking is indeed more efficient, and whether it affects equality of opportunity, is a question to be answered by empirics.[1]

One way to investigate the efficiency/equality debate is to compare cross-country differences. There are clear distinctions at which age governments of different nations start tracking their students. Countries such as Germany, Austria, Hungary and the Slovak Republic start tracking as early as the age of 10. Other countries such as the UK, the US, Canada, Japan and Sweden do not track students at all during secondary education. Clearly, there is no consensus among policymakers how to address tracking. Hanushek and Woessmann (2006) investigate whether early tracking affects inequality in educational performance. Exploiting the distinctions in tracking starting age across countries mentioned above, they find that early tracking increases inequality in performance. They find very little evidence for efficiency gains, rather tracking early seems to lower average performance for reading and math tests. In their review of the literature Meier and Schütz (2007) confirm the findings of Hanushek and Woessmann (2006). On the one hand the effect on average test scores is unclear; some studies find positive effects, while others find negative effects. On the other hand they mention a clear trend that tracking increases the variance of test scores, i.e. tracking increases inequality. Meier and Schütz (2007) also conclude that *early* tracking negatively affects inequality, because delayed tracking seems to benefit disadvantaged students.

A country's tracking system is not only characterized by whether it tracks students and at what age, other features are also of interest. In a cross-country comparison, Ammermüller (2005) shows that inequality increases if a country's education system offers more tracks. Besides the number of tracks, the mobility between tracks is also an important aspect to take into account when evaluating a country's tracking system, as it is closely linked to equality of opportunity. Epple, Newlon and Romano (2002) add another dimension, namely that tracking in public-sector schools results in retaining high-ability students, but also entail losing wealthy, low-ability students to private schools. The positive result is more high-ability students being enrolled in public schools due to tracking. Of course, for this argument to hold one needs private sector education, which is limited in continental Europe compared to the English-speaking countries.

[1] A tracking system can also be more efficient if it has the same educational results but is less costly than a comprehensive system. However, there is little research on this type of efficiency; therefore it is not mentioned any further in this thesis.

This thesis investigates one particular aspect of educational tracking, namely whether a delay in tracking results in positive effects. It does so by analyzing the long-term effects of early tracking in the Dutch education system on individual outcomes. Specifically, it investigates the effect on educational attainment and labor market outcomes such as income and employment, leading to the following research question: *Does early tracking lead to long-term differences in educational outcomes and labor market outcomes of students?* The Dutch system starts tracking students at the beginning of secondary school, which is at the age of 12. However, the existence of comprehensive first- and (sometimes) second-year classes, the so-called *dakpan brugklassen* or *brede brugklassen*, creates opportunities to delay tracking for an additional two years. This thesis compares long-term outcomes of students enrolled in regular or categorical classes to students enrolled in such comprehensive classes using regression analysis. Ideally, a researcher wants to randomly assign students to the comprehensive and categorical class in order to overcome the selection bias, as it is probable that students in the comprehensive class are for example more motivated. Therefore this paper adopts an instrumental variable approach to control for this bias. The instrument used exploits regional variation in the supply of schools, as proposed by Van Elk, Van Der Steeg and Webbink (2011).

This paper adds to the literature by performing a long-term analysis of early tracking. Also, the nation-wide Dutch education system makes it possible to investigate tracking without having to control for regional effects. Another advantage is that the empirical strategy is not based on a change in a country's tracking system, and therefore there is no time dimension. Students from the same cohort are compared to each other. Finally, it refines the instrument used by Van Elk et al. (2011) by looking at a less aggregate level, thus improving the estimation results. The results show that MAVO advice students are hurt by early tracking in terms of completion of higher education as the propensity to finish higher education is reduced by 5.3 percentage points. This effect is insignificant when controlling for selection bias using an IV regression. In contrast, VWO advice students seem to benefit from early tracking in terms of completion of university. The OLS estimate indicates an increase of 7.3 percentage points. The corresponding IV estimate indicates that tracked students finish university more by 16.3 percentage points.

The structure is as follows. Section I provides a review of the literature on the effects of early tracking. Section II discusses the Dutch education system, followed by section III explaining the empirical strategy. Section IV gives a description of the data. Section V displays the main results and subsequently Section VI presents further analyses such as the robustness of the results. Section VII contains the discussion and conclusion.

Section I. Review literature

This section reviews the empirical findings of previous studies on the subject of early tracking. First, a set of papers is reviewed that use cross-country variation in early tracking to identify the effect of early tracking on different outcome variables, depending on whether the paper investigates equity or efficiency of tracking. Next, the papers that investigate the effect of early tracking using within-country variation are reviewed, followed by the set of papers concerning studies that focus on within-school variation. Finally, the subject of optimal tracking age is discussed.

In the literature there is confusion about the exact definition of tracking. In this thesis *tracking* is defined as grouping students based on their ability and providing different curricula per ability group. This is seen mostly throughout Europe in education systems. *Ability grouping* is an informal version of tracking, and is defined as grouping students based on their ability but providing a uniform curriculum for all students. This type of tracking is seen in countries such as the US and Canada.

Cross-country variation

The paper of Hanushek and Woessmann (2006) discussed above gives evidence from a cross-country perspective. They estimate the causal effect of early tracking on educational inequality measured by the variance in test scores using a DD approach to deal with unobserved country heterogeneity. The 26 countries in their sample do not track students in primary school (4th grade), but some of these countries do track students four years into secondary school (8th grade). They compare the change in the variance of test scores in these countries with the change in the variance of the countries that did not track early. The data comes from various international studies on student performance, namely TIMSS, PIRLS and PISA.[2] The results show a systematic

[2] TIMSS stands for Third International Mathematics and Science Study and is math and science oriented. It is conducted in 1995, 1999, 2003 and so on. PIRLS stands for Progress in International Reading Literacy Study and is oriented on reading literacy. It is conducted in 2001,

increase in variance in test scores, and thus inequality. They also do not find evidence for a positive effect of early tracking on average test scores, and thus efficiency. Two papers dispute the results of Hanushek and Woessmann (2006). First, Eisenkopf (2007) argues that the results from Hanushek and Woessmann (2006) are biased because they ignore incentive effects already present in a tracked schooling system. That is, tracking incentivizes students in primary education to perform better in order to enroll in a higher track. The approach of Hanushek and Woessmann (2006) does not take this effect into account. Second, Jakubowski (2010) argues that the results are affected by differences in the design of the international tests. In particular, the TIMMS and PIRLS surveys are by grade, whereas the PISA survey is defined by age. When Jakubowski controls for the differences in the samples of these tests, the results turn out insignificant.

Schütz, Ursprung and Woessmann (2008) use another method of evaluating the effect of early tracking in a cross-sectional setting. Using data from the TIMSS conducted in 1995 and 1999, and using a larger sample consisting of 54 countries, they investigate the effect of family background on student's test scores. They add a dummy for each country to deal with unobserved country heterogeneity and find that tracking students earlier exacerbates the family background effect on test scores, whereas early tracking itself does not significantly affect test scores. In other words, early tracking increases the impact of family influence, which is an indication of an increase in inequality of opportunity, as socio-economic factors influence test scores more in this case. The results of Schütz et al. (2008) are in line with Hanushek and Woessmann, supporting the increase in inequality due to early tracking, whilst providing no evidence for efficiency.

Woessmann, Luedemann, Schütz and West (2009) and Ammermüller (2005) confirm the finding of Schütz et al. (2008). Woessmann et al. (2009) find evidence that early tracking decreases test scores, but this evidence is not statistically significant. Furthermore they present statistically significant evidence for increasing influence of family background due to early tracking. Their econometric analysis is similar to Schütz et al. (2008) and uses PISA 2003 data, in which an index called Economic, Social and Cultural Status proxies for family background. Ammermüller (2005) uses a DD

2006, 2011 and so on. PISA stands for Programme for International Student Assessment and covers all subjects. In 2000 reading proficiency was tested, in 2003 math, in 2006 science, in 2009 again reading proficiency and so on. They are carried out by the International Association for the Evaluation of Educational Achievement (IEA).

approach similar to Hanushek and Woessmann (2006), comparing test scores over time and across countries with differences in institutional features, one of which is differences in the number of school types mentioned earlier. This tracking measure (number of tracks) is different from the tracking measures from the other papers (first age of tracking). It does not account for how many years a student is exposed to tracking, and thus may underestimate the effect of tracking (Waldinger, 2007). Data comes from the PIRLS study of 2001, and the PISA study of 2000. He reports that the impact of students' background characteristics increases in the presence of several school types, thus confirming that tracking increases inequality. Note that the paper of Ammermüller (2005) does not discuss the effect of *early* tracking on inequality.

The following papers come to a different conclusion. Waldinger (2007) uses the same country-dummy approach as Schütz et al. (2008) in combination with a DD approach. He uses PISA and PIRLS data for reading skills, and TIMMS data for mathematics. The results show that early tracking does not exacerbate the influence of family background because of insignificant results, both for reading and mathematics. In replicating the results of Hanushek and Woessmann (2006), but with a different measure of early tracking, he also ends up with insignificant results. Hanushek and Woessmann (2006) define early tracking as tracking before the age of 15, whereas Waldinger (2007) defines it as tracking at the end of grade 5 or before.

Brunello and Checchi (2007) take a longer-term perspective in the sense that they go beyond test score outcomes and also look at educational attainment, earnings, and literacy. Using the European Community Household Panel combined with the International Social Survey Programme, they obtain the necessary data, covering 21 countries. Employing a strategy similar to Schütz et al. (2008) and Waldinger (2007), their results show that the length of tracking tends to reinforce family background effects when it comes to educational attainment and earnings, but reduces it in the case of literacy. The latter means that early tracking improves the effect of family background on the ability to read or process documents, contradicting the findings of the family background effect on the reading proficiency in Hanushek and Woessmann (2006) for instance.

Finally, Ariga and Brunello (2007) use an instrumental variable (IV) approach to investigate the effect of tracking length on the International Adult Literacy Survey. When using cross-sectional data, even students within a country experience differences in exposure to tracking due to dropout and age. The instruments used come from the survey itself, in which dropouts are asked to indicate the reason they left school. The replies are divided into three dummies: one for financial constraints, one for family reasons, and one for personal constraints. The results from the IV regressions show that one additional year of tracking raises test scores by 3.3 percent compared to one additional year of comprehensive school. The authors are cautious with generalizing these results, as it is a local average treatment effect rather than the average treatment effect, a point which will be discussed further in Section III. Nevertheless, they state that the view that de-tracking is positive in terms of equity gains may be too optimistic. However, the used instruments are in my opinion not valid. The exclusion restriction is likely to be violated. Students who drop out of school due to financial constraints or family reasons are likely to be students from disadvantaged families, which is correlated with for instance motivation and ability, which in turn affects students' test scores. The authors try to control for this by adding variables to control for family background among others. Also the exogeneity assumption is likely to be violated, as it is not exogenously determined whether your family has financial constraints for instance.

Within-country variation

The advantage of using within-country data is that researchers do not have to deal with unobserved country characteristics. Instead, researchers rely on variation of a particular education system over time, or take advantage of a specific institutional feature of a particular education system, like this thesis, to identify the causal effect of early tracking on equality or efficiency.

In Sweden the age of tracking was delayed from the age of 13 to 16 nationwide in the 1960s, but this reform was implemented gradually. The first municipalities were selected (nonrandom) as an experiment; each year new municipalities were added to delay the age of tracking. Meghir and Palme (2005) use this policy reform to investigate the effect of early tracking with a DD approach, as it enables them to compare individuals from the same cohort but that are tracked at a different age. The results show that the reform increases educational attainment for low-ability students from a disadvantaged background, but it also increases schooling of high-ability students from a disadvantaged background thereby increasing intergenerational mobility. Individuals from a disadvantaged background also earn more, whereas individuals from an advantaged background seem to earn less. Detracking appears to have an equalizing effect in Sweden.

However, as it turns out, ability grouping remained part of the Swedish education system for math and English, until this was abolished in 1995. Sund (2006) investigates if this ability grouping affects the grade of the first math course in upper-secondary school and also graduating upper-secondary school. He finds no effect of ability grouping neither on the grade nor on the propensity to graduate. One takeaway point is that the results from Meghir and Palme (2005) are possibly biased. As tracking was delayed in the 60s, ability grouping was introduced for math and English (Sund, 2006), thus possibly interfering with the true causal effect of delaying tracking in the study of Meghir and Palme (2005).

Galindo-Rueda and Vignoles (2007) and Pischke and Manning (2006) investigate a similar reform in England and Wales. In the 60s and 70s the age of tracking in British education was delayed from 11 to 16. This transition was not done uniformly across the country, resulting in some areas already offering comprehensive secondary education but other areas still offering tracked education in this time period. Galindo-Rueda and Vignoles (2004) use a matching approach in which they compare test scores of students with similar characteristics, differing only in going to either a tracked or comprehensive school system. They find that high-ability students in the tracked system fared better in terms of test scores than the high-ability students in the comprehensive system. The effect for middle and low-ability students turns out insignificant. However, Pischke and Manning (2006) show that the differences between the students in the tracked and the comprehensive system already exist at the age of 11, indicating the presence of selection bias. The identification strategy of Galindo-Rueda and Vignoles (2007) does not control for this bias, thus there conclusions are invalid.

In Finland, the two-track school system was replaced by a comprehensive system. This reform was implemented gradually from 1972 to 1977, introducing the possibility to investigate the effect of this reform using a DD approach similar to Meghir and Plame (2005). The study of Pekkarinen, Uusitalo and Kerr (2009a) estimates the effect of the reform on Finnish Army Basic Skills test and find that a small positive effect on mean verbal test scores, but find no effect on mean arithmetic and logic tests. They do mention that the test scores of students with low-educated parents increase for all three tests. In another study, Pekkarinen, Uusitalo and Kerr (2009b) use the same approach to investigate the effect of the Finnish reform on intergenerational income elasticity. The reform reduces the intergenerational income elasticity with 23%. Comprehensive schooling thus has a positive effect on equality.

In the German state of Bavaria, a reform resulted in students in the basic track (*Hauptschule*) and middle track (*Realschule*) being tracked after grade four instead of grade six. Piopiunik (2014) employs a triple difference approach to investigate the causal effect of this reform on student achievement. In particular, he compares the change in performance of students as a result of the reform in Bavaria to other German states, but also compares the affected tracks (*Haupt- and Realschule*) to the unaffected track (*Gymnasium*) to additionally account for state-specific or school-type-specific effects respectively. The results show lower performance in both *Hauptschule* as well as in *Realschule*, and that this decrease persists several years after the reform. Piopiunik (2014) argues that the decline in the lower track is due to peer effects, whereas the decline in the middle track might be due to hiring of additional, inexperienced teachers and other implementation problems.

Other evidence from Germany comes from the fact that the sixteen German states differ in their age of first tracking. Woessmann (2010) exploits this fact in a regression analysis and finds that states that track later perform better at equality of opportunity for disadvantaged children. The author warns, however, that because the number of observations, equal to the number of German states, is small, the result should not be interpreted causally. Rather, the relationship between early tracking and equality of opportunity within Germany is a "controlled descriptive association" (Woessmann, 2010, p. 237).

Dustmann, Puhani and Schonberg (2016) take a long-term perspective by looking at whether early track allocation affects educational attainment and labor market outcomes in Germany. They do so by using a regression discontinuity design based on date of birth. The idea is that students born just after the school entry cut-off of July 1 are older than the students born just before this cut-off. The oldest children in class, those born in July, are more likely to attend the higher track. The effect of attending this higher track for students close to the cut-off on long-term outcomes, such as wages and days worked, is insignificant. So despite the early track allocation student earn the same wages and are equally employed. The authors attribute this finding to the flexibility of the German education system. That is students being able to easily attend higher tracks after completing a different track, or students that completed the higher track but fail to enroll for university. However, a causal interpretation of using date of birth as instrument has received much criticism since the seminal paper of Angrist and Krueger (1990), which first applied this strategy. Dustmann et al. (2016) reviewed 30 articles to verify if date of birth has a direct impact on long-term outcomes other than through track assignment, and conclude that if such a relation exists, it is positive. As the results are insignificant, the possible positive bias does not result in a positive relationship between high track attendance and wages (among others). Rather, the relationship remains insignificant, or even turns out negative, when controlling for the bias. This calls into question the efficiency of the German tracking system.

As early track choice resulting from differences in birthdays does not affect wages (Dustmann et al., 2016), early track choice influenced by parental background does affect wages, according to Dustmann (2004). He presents evidence that differences in parental background, i.e. parental education and occupation, propagate into substantial wage differences through early track choice. In other words, children with low-educated parents are more likely to attend lower tracks, which results in lower wages later on. Dustmann notes that this conclusion is not to be interpreted causally; rather he argues that the influence of parents on early track choice is an important factor in explaining intergenerational mobility. Bauer and Riphahn (2006) mention similar results for intergenerational educational mobility using Swiss data. They report that the correlation between parent education and child education is affected by the timing of tracking.

According to Van Elk et al. (2011), early tracking has detrimental effects on completion of higher education. In the Netherlands students can choose to postpone tracking one or two years in secondary education if they enroll for comprehensive first and second year's classes. Comparing the high school completion rates of low-ability students enrolled in these comprehensive classes with low-ability students directly enrolled in tracked classes, they find that the students that went to tracked classes are less inclined to finish higher education using an IV approach. They also mention that there is no effect on the educational attainment of the high-ability students enrolled in the same comprehensive classes, thus providing evidence for equalizing peer effects in a comprehensive school system.

Figlio and Page (2002) estimate the effect of tracking per ability level by comparing test score gains of students in tracked and untracked schools in the US. Note that in the US tracking takes place in the form of ability grouping. The results show no evidence of low-ability students being harmed by tracking. An IV regression even indicates test score improvements for low-ability students. The authors argue that tracking attracts high-income students, resulting in additional funding by parents, better teachers and positive school-level spillovers, all improving low-ability students' test scores.

Within-school variation

Besides studies focusing on between-country variation and studies focusing on within-country variation, there is a third group of studies, namely those focusing on within-school variation. Most of these studies investigate the effect of ability grouping on equity or efficiency, rather than the effect of tracking itself. One example is the well-known paper by Duflo, Dupas and Kremer (2008). It employs a randomized experiment in Kenya to investigate the causal effect of tracking. In this experiment primary schools that are selected to the treatment group assign students to certain classes according to their ability, whereas the control group schools do not assign according to ability. The results suggest that on average test scores in the treated schools are 0.14 SD higher, but also that the top half as well as the bottom half of the students increased in test scores (by 0.19 SD and 0.16 SD respectively). These results indicate that there are efficiency gains as argued by proponents of tracking, in addition to equality not being harmed.

Garlick (2016) investigates a reform in group assignment policy at the University of Cape Town in South Africa. Before 2006 first year students were tracked into dormitories, from 2006 onward they are placed randomly. Using a DD approach, he compares the GPA of the students in dormitories before and after the reform with the GPA of non-dormitory students as control group. The results show that tracking decreases the low-scoring students' GPA, whilst not affecting the GPA of high-scoring students, thus tracking increases inequality and decreases efficiency. He also investigates the effect of living among high-ability students or low-ability students using the randomization process of placing students into dormitories. He concludes that living with high-ability students raises students' GPA. This effect is larger for low-ability students. This study uses dormitory-level variation, rather than school-level variation, and therefore does not answer whether educational tracking (or more accurately ability grouping) affects test scores. Nevertheless, it provides evidence for nonlinear peer effects, in the sense that low-ability students benefit, while high-ability students are unaffected.

Other within-school evidence comes from Card and Giuliano (2016). They focus on the effect of being placed in a gifted/high-achieving class. These classes were designed to place all gifted students in one classroom. The remaining seats in the classroom are allocated to non-gifted students. Comparing the non-gifted students in a gifted class with the students in a regular class, the authors find positive effects for non-gifted students. These benefits mainly concern black and Hispanic students. They also report no spillovers to other students not participating in the gifted class.

Lastly, Betts and Shkolnik (2000) find no evidence that ability grouping increases average math achievement growth, using US data. Previous studies compare the average achievement of, for instance, a low-ability group to the average achievement of a non-tracked group, which might structurally be different. Once the authors address this problem and adequately control for mean group ability, they do not find differential effects for low-ability, middle-ability or high-ability groups.

Optimal age of tracking

On the subject of the optimal age of tracking, Brunello et al. (2007) introduce a model in which two opposing effects determine the optimal age of tracking. The "specialization" effect positively influences tracking age. In order to reap the full benefits of specialization, students should be tracked as early as possible. But there is also a countervailing effect, namely the "noise" effect, already mentioned in the introduction. This effect entails the misallocation of students if tracked early due to imperfect identification of abilities. The resulting prediction of this model is that an increase in total factor productivity growth results in an older optimal age of tracking. They argue that if a country accelerates in growth, there is more depreciation of skills from lower tracks. Delaying tracking age, and thus delaying stratification, reduces the depreciation as the acquired skills might already be outdated on entering the job market. Ariga, Brunello, Iwahashi and Rocco (2005) use a similar model to predict the first age of selection and the relative size of general to vocational jobs for European countries, as well as the US, Australia, Canada, Korea and Japan. The optimal tracking age for the Netherlands turns out to be higher than the actual age. This result is based on calibrated parameters used in the model, and assumption on for instance welfare maximization of the government. Ariga et al. (2005) argue that the differences between actual and calibrated values might result differences in preferences in each country, such as giving priority to less privileged households.

As is clear from this literature review, the available empirical evidence is far from uniform in its conclusions. In the cross-country analyses most studies conclude a negative effect of early tracking on inequality, but some of these results prove not to be robust as argued by other studies. Some studies even report opposite results. The same goes for within-country evidence. Most studies fail to provide evidence for overall gains in test scores due to early tracking, whereas a few indicate reduced test scores; most studies do provide evidence for an equalizing effect in the case of delaying tracking. A recurring theme is the difficulty to interpret the conclusions as causal due to statistical difficulties in the research design. As for the within-school variation evidence, they are in line with the other two, in the sense that there is no consensus among the results on the effects of tracking (or ability grouping) on equity or efficiency.

Section II: the Dutch education system

In this section an overview of the Dutch education system is given. The system changed in august 1997, but as the school-level data (see section 4) is from before this date, the "old" system is described.

Education in the Netherlands is divided into primary, secondary and tertiary education, with secondary education being characterized by a tracked system. Students in primary education are aged 4 to 12, with the first year being optional. All students face the same curriculum, and at the end of primary education they take a centralized exam, the so-called CITO test. The result of the test, together with the student's preference and teacher recommendation, determines the level of secondary education in which students enroll.

Secondary education consists of four tracks. VWO is the highest track, and gives direct access to university. It takes six years to complete, thus students presumably leave secondary education at the age of 18. HAVO is the upper-middle track and takes five years. Students graduating from HAVO pass on to higher professional education (HBO). The lower-middle track is called MAVO. It takes four years and prepares students for vocational education (MBO). Finally, the lowest track is called VBO. It is similar in design to MAVO, as it also takes four years and prepares for MBO, although for lower levels. There are possibilities to move to a higher track after the student has finished her current track. MAVO students have the possibility to enroll for the final two years of HAVO; HAVO students have the possibility to enroll for the final two years of VWO. These promotions are designed to improve the mobility of students and diminish the noise associated with tracking at an early age. Moving to a lower track is always possible.

Tertiary or higher education consists of university or HBO. Universities offer research based education (WO), which generally takes up to four years. HBO institutions offer education in applied sciences, and are geared towards specific professions. The vocational track MBO provides vocational training at different levels taking up to four years, depending on the level. Similar to secondary education, there are possibilities to move to a higher level of education. Students that completed vocational education can enroll for HBO; students that completed HBO can enroll for WO.

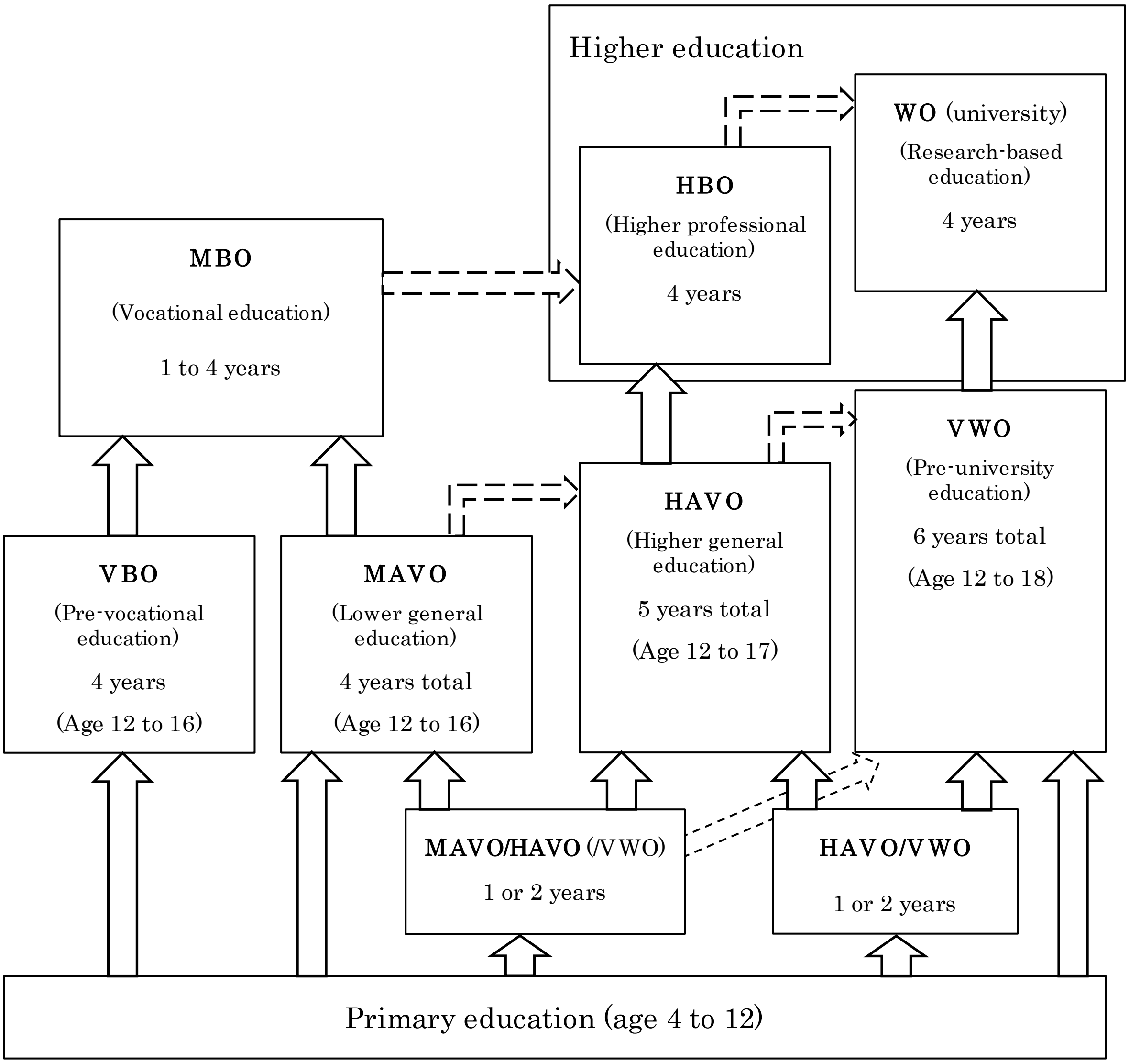
In order to further reduce tracking noise, there is the opportunity for students entering secondary education to spend one or potentially two years in a comprehensive class. This way, students can determine the track most suited to their ability and/or motivation. Specifically, students can enroll in a class combining MAVO and HAVO, HAVO and VWO, or MAVO, HAVO and VWO, besides the categorical classes for MAVO and VWO. There are no categorical first-year HAVO classes. Attending a comprehensive class effectively delays the age of tracking. This aspect gives rise to the empirical strategy of this thesis, a point which will be discussed in depth in the next section. Figure 1 provides a depiction of the education system.

The *current* Dutch education system differs from the old system in one aspect, namely the VBO and MAVO tracks are combined but divided into 4 sub tracks.[3] The previously mentioned report on the current Dutch education system by the OECD (2016) points out that the Dutch system is regarded as one of the best of all OECD countries, but there is still room for improvement. These improvements highlight the weak spots of the old system as well, as the old system and the current system differ only slightly. They recommend reducing the extent of early tracking as track promotion has become more difficult according to the report. Delaying the age of tracking and reducing the number of tracks are options to reform. The fact that secondary education enrollment is based on preferences, teacher recommendation and test results, rather than on test results only, is another potential point of improvement. The discretion of primary teachers and secondary schools for which the students apply invoke a potential bias in the decision making, while the test results are objective. Relying only on these test results reduces the noise of early tracking. A final recommendation is the encouragement of upward transitions between tracks and the merger of some tracks. These recommendations are in line with the results from the literature from the

[3] The 4 tracks range from the higher theoretical track to the lower pre-vocational track

previous section. Reducing tracking in the form of delaying tracking could improve the system, a question this thesis tries to answer.

Figure 1. The Dutch education system before August 1997



Section III: empirical strategy

This section is dedicated to the empirical strategy of this thesis. It makes use of the specific structure of the Dutch education system addressed in the previous section. Next, the identification strategy is explained, along with a discussion of the interpretation of the results.

The existence of comprehensive first- and second-year classes creates variation in the timing of tracking within a cohort of Dutch students. This thesis investigates the effect of this variation by comparing several long-term outcomes of similar students. The effect of early tracking can be ascertained for students with a MAVO advice and a VWO advice but not for students with a HAVO advice. MAVO and VWO students have the possibility to enroll for either a categorical school or a comprehensive school. HAVO students can only enroll in a comprehensive school, thus tracking is delayed for all HAVO advice students. For MAVO students the outcomes of categorical MAVO students are compared to the outcomes of students enrolled in a MAVO/HAVO (MH) class or a MAVO/HAVO/VWO (MHV) class; for VWO students the outcomes of categorical VWO students are compared to the outcomes of students enrolled in a HAVO/VWO (HV) class or a MHV class. The effect of early tracking is analyzed using the following equation:

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 X_i' + \varepsilon_i \qquad (1)$$

Here Y_i is the outcome variable of interest.[4] T_i is a dummy variable indicating early tracking, so the dummy is equal to 1 if student i starts secondary education in a categorical class, and the dummy is equal to 0 if she starts in a comprehensive class. Consequently, β_1 estimates the effect of early tracking on the outcome variable. X_i' denotes a vector of controls. Finally, ε_i is the error term.

The outcome variables of interest are separated into education outcomes and labor market outcomes. Completion of higher education, completion of university education and dropout rates are considered as education outcomes. Van Elk et al. (2011) answer the question whether tracking early affects the completion of higher education in the case of the MH class. Comparing the effect on both completion of higher education (HBO and WO) and completion of WO provides additional insight in the final level of educational attainment. The effect on dropout rates is also interesting as dropouts are a vulnerable group in terms of for instance labor market prospects and crime. The labor market outcomes of interest are wages and employment, but also whether someone receives welfare assistance. Control variables include personal characteristics, socioeconomic background, and measures for ability. They are added to reduce any confounding effect of these covariates.

In the introduction several channels are introduced through which tracking might affect outcomes. The foremost channels are the effect of a different curriculum,

[4] In case of a binary outcome variable, the effect is estimated using the linear probability model.

the effect of spillovers of the other students present in the class (this is referred to as peer effects), and the effect of the appropriate teaching style the teacher adopts in different classes. In this study there is no attempt to disentangle these effects. Consequently, only the net effect of these channels is estimated.

Identification strategy

An important lesson from the literature discussed in section I is the difficulty of obtaining a causal relationship. Some studies specifically mention their results as correlations, while other studies' identification strategy turns out invalid. The advantage of using the Dutch system is that it is a nationwide system, so one does not have to control for regional differences (cf. Woessmann, 2010), and there is no time dimension of a policy change (cf. Meghir and Palme, 2005). For a causal interpretation of the effect of early tracking a researcher wants to compare long-term outcomes of a particular student enrolled in a comprehensive class to the counterfactual scenario of that *same student* being enrolled in a categorical class. Estimating equation (1) compares students in a comprehensive class to students in a categorical class, but this does not provide a causal inference, as students enrolled in a comprehensive class might fundamentally differ from those in a categorical class. The estimated effect is then confounded by these differences, the so-called selection bias. For instance, in case of the HV class, motivated parents of students with HAVO advice want to send their children to the HV class if they believe this comprehensive class provides better opportunities for their children. For students with VWO advice the opposite is conceivable. Motivated parents send their children to the categorical VWO class believing it provides better opportunities. Early tracking is thus confounded by motivation, resulting in the early tracking variable being correlated with the error term. For the MV class a similar argument can be made.

In order to overcome this selection bias an IV approach is used. Van Elk et al. (2011) propose an instrument exploiting regional variation in the supply of schools. Other studies also employ instruments based on geographic variation for estimating educational attainment (Card, 1993; Currie and Moretti, 2003; Park and Kang, 2008). The idea behind such an instrument is that some students are forced to go to a particular school because it is the only school available in the region. This availability creates exogenous variation in education choices. Specifically, Van Elk et al. (2011) use the relative supply ratio of categorical schools per municipality type. These municipality types are characterized by the degree of urbanization among others. This instrument is valid as long as the exclusion restriction holds. That is, the instrument only affects the outcome variable of interest through early tracking. The supply ratio of schools is considered endogenous if it is linked to the demand for comprehensive or categorical schools, which would invalidate the instrument. They argue that the change in the supply ratio might reveal changes in demand for school types. Also, the fraction of highly educated parents might indicate demand for school types, as these parents value the availability of comprehensive schools in order to advance the opportunities of their children. Including these variables do no significantly alter the results, so Van Elk et al. conclude that there is no bias caused by the demand for education. A second potential problem concerns the quality of schools. If the supply ratio is related to the quality of schools across regions the exclusion restriction does not hold, because then the supply ratio affects educational attainment through quality. The inclusion of an urbanization indicator variable results in insignificant estimates. Additionally, there is no relationship between school mean test scores and the supply ratio, suggesting that school quality is uncorrelated to the instrument.

Borghans, Diris, Smits and De Vries (2012) use a similar instrument. It is based on the fraction of students in a municipality that is enrolled in a school offering a specific track. However, the validity of this instrument is debatable. Van Elk et al. (2011) base their instrument on the availability of certain schools in a municipality type. The exogenous variation comes from some students being forced to go to a particular school type, as there is no other school type available nearby. Borghans et al. (2012) rely on variation in the number of students enrolled in a particular school type, but this might already reflect the selection it tries to solve.

This thesis employs a similar instrument as Van Elk et al. (2011). One disadvantage of their instrument is that it is based on aggregate values at municipality type level. The supply ratio of a student in her municipality might differ from the supply ratio of the municipality type. As such, looking at the relative supply ratio per *municipality* instead of *municipality type* could improve the accuracy of the estimation, as it provides a better account of the actual exposure of a student to certain school types. This would improve the first stage predictions, resulting in higher F-statistics of the F test of excluding the instrument. This point is discussed further in section V.

An additional improvement is to look at regions instead of municipalities. The exposure of students to certain school types goes across the borders of the municipality of residence, as it is common for students to go to school in a nearby municipality, but less common to travel to another region (the numbers substantiating this claim are discussed in the next section). The regions used in this thesis are the so-called labor market regions (*arbeidsmarktregio's*). These 35 regions consist of several municipalities and are used by these municipalities and the Dutch Employee Insurance Agency (UWV) to employ their services. One advantage of these regions is that the municipalities themselves decided on the division, thus respecting close ties among certain municipalities. However, students near the border of a labor market region face a different relative supply ratio than the ratio of their labor market, as they can go to school in the nearby labor market region. The best way to construct the instrument would be to use a spatial approach, in which a circle is drawn on the map around the living address of the student and subsequently determine the relative supply ratio within that circle. However, time constraints do not allow me to pursue this approach. The first stage of the 2SLS procedure is as follows:

$$T_i = \alpha_0 + \alpha_1 Z_i + \alpha_2 X' + \eta_i \tag{2}$$

Here T_i is the early tracking dummy, Z_i is the relative supply ratio of categorical schools in labor market region of student i, X is a vector of controls, and finally η_i is the error term. The resulting second stage is the following:

$$Y_i = \beta_0 + \beta_1 \hat{T}_i + \beta_2 X' + \varepsilon_i \tag{3}$$

$\hat{T}_i$ is the predicted value of T_i coming from the first stage. If the exclusion restriction of the instrument holds, β_1 gives the unconfounded effect of early tracking on outcome variable Y_i. However, this is not the average treatment effect, but the *local* average treatment effect as argued by Imbens and Angrist (1994). This is the treatment effect for the so-called compliers. These are the individuals that respond to a change in the instrument status. That is, the individuals that would change to a tracking school resulting from an increase in the relative supply ratio in their municipality. As a result, the local average treatment effect is not generalizable to every individual in the population as it only affects the subset of compliers. The estimated effect of β_1 does not apply to the never-takers, those that always choose a comprehensive school, and the always-takers, those that always choose a tracking school.

The empirical strategy of this thesis tries to uncover the causal effect of early tracking on several labor market outcomes and educational outcomes. To do this, it uses control variables to limit the confounding effect of these covariates. On top of that it tries to control for hidden confounding covariates, such as motivation, by employing an IV approach. The instrument uses regional variation in the supply of schools at the level of labor market regions to overcome the selection bias.

Section IV: data

This section provides a description of the data used in the analyses of this thesis. It also gives the sample statistics of the two student groups under consideration, followed by information about the location of schools and students.

The VOCL data set, which stands for Secondary Education Cohort Students, is a random sample of students starting secondary education in the Netherlands. The data set contains three cohorts, namely 1989, 1993 and 1999. The size of each cohort is about 20.000 students, which accounts for 10% of the student population. This study investigates long-term outcomes; therefore the '89 cohort is naturally preferred. However, the instrument makes use of the municipality of residence of the students. Living addresses for the sample students are available from '95 and onward. The '93 cohort is therefore the most preferred option, as the instrument is much less noisy for this cohort. The validity of the instrument is discussed in depth in Section VI. The VOCL keeps track of the educational career of each student in the sample, up to 2012. Each year for all students it is known what level of education and what grade the student follows, but also if they left the education system. There are students still enrolled in tertiary education in 2012. These students are matched with more recent graduation data up to 2015. Completing higher education is defined as graduating from HBO or WO at bachelor level; completion of university is defined as completion of WO at least at bachelor level. Dropouts are defined as individuals with primary education as their highest degree. The advice of the students at the beginning of secondary education is also given in the VOCL. Therefore the VOCL offers the opportunity to easily isolate the two student groups under consideration, namely MAVO advice and VWO advice. Furthermore, the VOCL provides information on personal characteristics, socioeconomic background and test scores of a test taken in the first year of secondary education. The test consists of arithmetic, information processing and language.

For the long-term comparison of students in the two advice groups, the VOCL data is supplemented with labor market outcomes available from Statistics Netherlands. Income is defined as hourly wage in October 2015, coming from data collected by the UWV. It only concerns employee wages; profits resulting from self-employment are not taken into account. Profits reflect the remuneration of selling a certain good or service, while wages reflect the remuneration of selling ones labor. The employment and welfare assistance variables are defined by the Social Economic Category (SEC) data set of Statistics Netherlands using October 1st, 2015, as date of inquiry. This dataset characterizes every Dutch citizen to 1 of 13 categories based on a number of sources. The categories consist of the type of employment (4 categories), and whether someone receives a particular form of social assistance (4 categories). The other categories are not of interest for this thesis. The 4 employment categories specify whether someone is employed.[5] 2 of the social assistance categories determine if someone receives welfare assistance.[6] If someone receives more welfare assistance than she receives wage in a particular month, she is characterized as social assistance receiving rather than being employed according to the SEC. Although these individuals are not considered employed, their hourly wage is still used in the analysis of the effect on wages, as the hourly wage remains indicative of their productivity. In other words, their wage is still informative of how effective someone's labor is.

Sample statistics

Table 1 and 2 show the sample statistics for the MAVO advice students and the VWO advice students respectively. Column (1) displays the mean of the corresponding variable for tracked students; column (2) displays the mean for students enrolled in a comprehensive class. Column (3) reports the p-value of the difference of the means of column (1) and (2). A p-value below 0.05 means there is a statistically significant difference at a 5% significance level. That means that tracked students are systematically different from comprehensive students in regard of that variable. The age variable is given in years. The test score variables indicate how many of 20 questions were answered correctly. Finally, wage is given as hourly earnings. All other variables are in percentages.

As is clear from table 1, comprehensive MAVO students score higher than their tracked counterparts. There are no significant differences for the labor market outcomes, but the completion of higher education and university also favor the

[5] These categories are employee, director/major shareholder, self-employed, other self-employed.
[6] These are unemployment benefits and welfare benefit. Sickness benefits and other social benefits are not accounted to the welfare assistance variable.

comprehensive students. Dropouts also seem more prevalent in the comprehensive classes. These students are also from more urbanized areas and have parents with a higher professional level.[7] There seems to be no structural difference in the parents' educational background. Finally, comprehensive students are more often from a different nationality than the Dutch nationality.

Table 2 shows that tracked VWO students have higher wages, but do not differ significantly in terms of employment and welfare assistance. There is also no distinction between dropout rates and higher education completion, but completion of university is more frequent among tracked students. Their parents are more highly educated, have higher professional levels, and they live in more urbanized areas. This is reasonable as their higher education results in more professional jobs, and these jobs most often are located in big cities. Tracked students score higher in the entry test at the beginning of secondary education. In contrast to the MAVO students, the VWO students do not differ in the composition of nationality. There is a statistically significant difference in the age, but it is not economically significant.

These descriptive statistics present the first evidence of the effect of early tracking on the outcomes of interest. Clearly there is a selection bias, as tracked students and comprehensive students differ in multiple ways besides some of the outcome variables. The next section will control for these differences to investigate true effect of early tracking. It is hypothesized that early tracking for the MAVO advice group *negatively* affects the outcome variables, whereas for the VWO advice group it is hypothesized that early tracking *positively* affects the outcome variables. The outcome variables most likely affected display a significant difference in table 1 and 2.

[7] Note that the difference of the professional level of parent and also the completion of university is statistically significant at a 10% level, rather than at a 5% level.

Table 1. Sample statistics of MAVO advice students

Variable	Tracked classes: MAVO (1)	Comprehensive classes: MAVO-HAVO (-VWO) (2)	p-value of difference (3)
Age	12.11	12.14	0.1594
Female	54.5	52.7	0.2955
Nationality			**0.0000**
Dutch	96.1	92.8	
Other (4 categories)	3.9	7.2	
Parent education			0.9373
Primary education	9.5	11.3	
Secondary education low	19.4	18.0	
Secondary education high	52.9	50.4	
Higher education first phase	14.5	16.5	
Higher education second phase	3.4	3.0	
Higher education third phase	0.4	0.9	
Professional level parent			**0.0819**
Self-employed without personnel	9.4	8.9	
Self-employed with personnel	12.6	11.6	
Worker	31.7	29.9	
Lower employee	13.3	12.5	
Intermediate employee	20.5	23.5	
Higher profession	12.5	13.4	
Urbanization level			**0.0000**
Very high	16.7	18.7	
High	17.7	33.9	
Moderate	28.7	29.1	
Low	21.8	10.9	
Very low	16.2	7.5	
Test score arithmetic	11.2	11.6	**0.0013**
Test score information processing	12.1	12.3	**0.0322**
Test score language	12.0	12.2	**0.0047**
Hourly wage	17.9	18.3	0.1688
Employment	85.8	85.9	0.9006
Welfare assistance	4.4	4.9	0.4222
Dropout	7.5	9.5	**0.0328**
Completion higher education	25.9	31.3	**0.0004**
Completion university	2.9	3.9	**0.0903**
Number of students	2223	1408	

All means are in percentages, except for age (years), test scores (correct answers out of 20), and wage (hourly earnings)

Table 2. Sample statistics of VWO advice students

Variable	Tracked classes: VWO (1)	Comprehensive classes: VWO-HAVO (-MAVO) (2)	p-value of difference (3)
Age	11.9	11.9	**0.0369**
Female	48.6	50.4	0.5754
Nationality			0.3921
Dutch	97.4	99.0	
Other (4 categories)	2.6	1.0	
Parent education			**0.0000**
Primary education	2.1	2.5	
Secondary education low	3.2	5.4	
Secondary education high	23.6	34.5	
Higher education first phase	27.7	33.7	
Higher education second phase	29.2	18.3	
Higher education third phase	14.2	5.7	
Professional level parent			**0.0039**
Self-employed without personnel	2.4	2.3	
Self-employed with personnel	9.4	6.7	
Worker	7.1	12.3	
Lower employee	9.4	13.8	
Intermediate employee	25.1	30.8	
Higher profession	46.3	33.9	
Urbanization level			**0.0000**
Very high	56.3	20.3	
High	24.4	28.1	
Moderate	15.2	35.2	
Low	4.1	15.2	
Very low	0	1.3	
Test score arithmetic	17.5	16.9	**0.0005**
Test score information processing	17.2	16.8	**0.0204**
Test score language	17.1	16.5	**0.0000**
Hourly wage	25.7	23.7	**0.0033**
Employment	91.1	91.0	0.9585
Welfare assistance	3.3	2.2	0.2855
Dropout	5.1	4.2	0.4653
Completion higher education	84.6	82.1	0.3044
Completion university	66.6	52.2	**0.0000**
Number of students	389	671	

All means are in percentages, except for age (years), test scores (correct answers out of 20), and wage (hourly earnings)

Location of schools

Table 3 provides more insight in the location of schools that are present in the VOCL sample. Column (1) gives the number of categorical MAVO schools per labor market region, together with the number of MAVO advice students attending such schools in parentheses. The location of the students is based on administrative data available from Statistics Netherlands. Column (2) gives the number of comprehensive MAVO schools per labor market region, that is MH and MHV schools, with the number of MAVO advice students attending in parentheses; column (3) gives the (in sample) relative supply ratio of categorical schools, which represents the instrument for the MAVO advice students. For example, in the region *Groningen* (1) there is 1 categorical MAVO school and 4 comprehensive MAVO schools which are present in the VOCL data set. 22 students attend a categorical school, and 147 students attend a comprehensive school. Columns (4) to (6) do the same, but for VWO schools and (VWO advice students).

It is known if students enter secondary education in a categorical class or a comprehensive class. Therefore, a school is characterized as a categorical MAVO (VWO) school if it offers at least one class of categorical MAVO (VWO). Similarly, a school is characterized as a comprehensive MAVO (VWO) school if it offers at least one MH (HV) or MHV class. Some schools offer both categorical classes and comprehensive classes. These schools are counted as both a categorical school and a comprehensive school, and are designated in table 3 with an asterisk. Not all schools in the VOCL sample are allocated to the regions because of missing data. For instance, the location of three of the categorical VWO schools is still unknown, but it is likely that one of these schools is located in *West-Brabant* (25), as there are 25 students from that region attending a categorical VWO. If these missing schools would be allocated, this would increase the predictive power of the instrument in the first stage regression. It is assumed that the allocation error is random, as the sample is randomized over students and not schools; therefore the error is distributed proportionately over regions.

The instrument is based on the exposure of students to certain schools within their labor market region. Students living at the border of a labor market region could go to school in a neighboring region. They face a different supply ratio than the implied ratio for their labor market region of residence, which could pose a problem to the quality of the first stage estimations. It turns out that only 7.1% of the MAVO advice students travel to a different region for school, and 9.6% of the VWO advice students. Note that these figures also suffer from the fact that not all schools are allocated to the regions. Another option considered for the instrument is looking at the relative supply ratio per municipality, but 34.1% of the MAVO advice students go to school in a different municipality. For the VWO advice students this is 48.7%. Therefore, taking labor market regions instead of municipalities is indeed an additional improvement, as the supply ratio per municipality is not a correct reflection of the supply ratio the students face.

Table 3. Supply ratio of schools present in VOCL

Labor market region	Supply ratio MAVO schools					Supply ratio VWO schools				
	(1)		(2)		(3)	(4)		(5)		(6)
Groningen (1)	1	(22)	4	(147)	0.2	0		5	(67)	0
Friesland (2)	1	(62)	2	(105)	0.33	1	(19)	2*	(18)	0.33
Drenthe (3)	1	(17)	1*	(18)	0.5	0		0	(3)	0
Ijsselvechtstreek (4)	0	(165)	2	(56)	0	1	(19)	2	(23)	0.33
Twente (5)	0	(71)	1	(76)	0	0	(16)	1	(4)	0
Stedendriehoek en Noordwest Veluwe (6)	0	(51)	1	(15)	0	0	(4)	1	(17)	0
Midden-Gelderland (7)	0	(6)	2	(49)	0	0		3	(23)	0
Rijk van Nijmegen (8)	0	(2)	0	(30)	0	0		0		0
Achterhoek (9)	0	(77)	1	(4)	0	0	(2)	0	(4)	0
Rivierenland (10)	1	(127)	1*	(14)	0.5	0		1	(8)	0
Flevoland (11)	1	(31)	0	(1)	1	0		0	(3)	0
Gooi en Vechtstreek (12)	1	(96)	1	(31)	0.5	0		2	(37)	0
Midden-Utrecht (13)	1	(125)	1*	(5)	0.5	0		2	(33)	0
Amersfoort (14)	3	(85)	3*	(42)	0.5	0		4	(24)	0
Noord-Holland Noord (15)	0	(17)	3	(92)	0	0		1	(25)	0
Zuid-Kennemerland (16)	1	(58)	0	(1)	1	2	(35)	0		1
Zaanstreek/Waterland (17)	0	(42)	1	(49)	0	0	(1)	2	(31)	0
Groot Amsterdam (18)	0	(74)	0	(37)	0	0	(5)	1	(10)	0
Holland Rijnland (19)	0	(49)	0	(22)	0	1	(77)	1	(25)	0.5
Midden-Holland (20)	0	(69)	0	(4)	0	0	(2)	0	(1)	0
Haaglanden (21)	0	(16)	2	(38)	0	0	(1)	3	(13)	0
Rijnmond (22)	3	(120)	3	(85)	0.5	3	(97)	5	(59)	0.375
Drechtsteden (23)	1	(83)	1	(20)	0.5	0	(1)	1	(5)	0
Zeeland (24)	0	(21)	3	(88)	0	0		2	(5)	0
West-Brabant (25)	0	(65)	1	(20)	0	0	(25)	1	(10)	0
Midden-Brabant (26)	2	(99)	2	(57)	0.5	0		2	(11)	0
Noordoost-Brabant (27)	0	(107)	1	(50)	0	1	(16)	2	(20)	0.33
Zuidoost-Brabant (28)	0	(66)	1	(32)	0	1	(15)	2*	(26)	0.33
Noord-Limburg (29)	1	(75)	1*	(4)	0.5	0		0	(14)	0
Zuid-Limburg (30)	3	(136)	2	(53)	0.6	1	(33)	6	(81)	0.143

Food Valley (31)	1	(53)	3*	(54)	0.25	0		2	(14)	0
Helmond-De Peel (32)	0	(82)	0	(1)	0	0		1	(22)	0
Midden-Limburg (33)	0	(16)	2	(51)	0	0		1	(3)	0
Zuid-Holland Centraal (34)	0	(8)	1	(50)	0	1	(46)	2*	(26)	0.33
Gorinchem (35)	1	(26)	1*	(3)	0.5	0		1	(5)	0
Not allocated to region	69	(4)	22	(4)		3		18	(1)	
Total	92		70			15		77		

* designates schools offering both categorical and comprehensive classes.

Number of students attending the schools are given in parentheses

Section V: main results

This section displays and discusses the main results, both the OLS and the IV estimates. First, the results for the MAVO students are presented, followed by the VWO students. Afterwards, the effect of early tracking on specific subgroups is analyzed.

MAVO advice group

Starting with the MAVO advice students, the descriptive correlations in table 1 indicate that there are no differences in labor market outcomes, but for educational outcomes the comprehensive students score better. Table 4 displays the estimated effect of early tracking on labor market outcomes and educational outcomes using OLS. The estimated coefficients in column (1) are the effect of early tracking on a particular outcome variable with no additional controls. The effect on hourly wage is given as the effect on the natural logarithm of hourly wage, thus the estimated effect can be interpreted as the *percentage change* in hourly wage. Column (2) presents the estimated coefficients after controlling for socioeconomic variables. These include age, age squared, gender, nationality, parent's education, parent's professional level and the urbanization indicator. Column (3) presents the estimated coefficients after controlling for the socioeconomic variables together with the test score variables of the arithmetic, information processing and language tests taking at the beginning of secondary education to control for ability. The standard errors are clustered at school level to control for within-school correlation of observations. The outcomes of students within the same school are possibly correlated due to school specific characteristics. In this case clustered standard errors are more appropriate than conventional standard errors for statistical inference. Furthermore, the mean of each outcome variable is given below the outcome variable name for comparison.[8]

Table 4. The effect of early tracking for MAVO advice students using OLS

Outcome variable	(1)	(2)	(3)
ln(Hourly wage)	-0.017	-0.005	-0.005
18.06	(0.014)	(0.014)	(0.013)
Observations	2751	2356	2288
R-squared	0.0007	0.0368	0.0507
Employment	-0.002	-0.016	-0.010
0.858	(0.014)	(0.014)	(0.014)
Observations	3515	2978	2885
R-squared	0.0000	0.0187	0.0220
Welfare assistance	-0.006	-0.001	-0.004
0.046	(0.008)	(0.008)	(0.008)
Observations	3515	2978	2885
R-squared	0.0002	0.0036	0.0072
Dropout	-0.020	-0.002	-0.006
0.083	(0.013)	(0.011)	(0.011)
Observations	3631	3066	2966
R-squared	0.0013	0.0181	0.0272
Completion of higher education	**-0.055*****	**-0.063*****	**-0.053****
0.280	(0.021)	(0.021)	(0.022)
Observations	3631	3066	2966
R-squared	0.0035	0.0466	0.0672
Completion of university	-0.010	-0.006	-0.004
0.033	(0.007)	(0.007)	(0.007)
Observations	3631	3066	2966
R-squared	0.0008	0.0060	0.0135
Socioeconomic controls	No	Yes	Yes
Test score controls	No	No	Yes

***, **, * indicates significance at a 1%, 5% or 10% significance level respectively

Standard errors clustered at school level are given in parentheses

[8] The mean of hourly wage is given instead of the mean of the natural log hourly wage, as the mean of the natural log hourly wage is uninformative.

Tracking early has a negative effect on all outcome variables, but is only statistically significant for completion of higher education, as indicated by table 4. Enrolling in a tracked class reduces the propensity of completing higher education by about 5 percentage points. This result is similar to the result of Van Elk et al. (2011). As the effect on completion of university education is statistically and economically insignificant, one can conclude that for MAVO advice students tracking early only affects finishing higher education at the HBO level. Moreover, the negative effect on log hourly wage, employment, welfare assistance, and dropout in the third specification is nowhere near statistical significance, besides the fact that they are also not economically significant. As all outcome variables are given in percentages, early tracking decreases the probability of dropping out by less than 1 percentage point for instance. The effect on hourly wage is denoted as the percentage change in hourly earnings, thus early tracking decreases hourly earnings by 0.5% (taking the point estimate at face value).

As discussed in Section III, the estimated effects of table 4 are biased due to selection bias. To overcome this bias an IV approach is implemented, using the relative supply ratio of categorical schools in the labor market region of residence of students as instrument. The strength of an instrument is determined by its relationship with the endogenous variable, in this case the early tracking variable. As Bound, Jaeger and Baker (1995) discuss, weak instruments can cause inconsistencies in the IV estimations. Therefore, Staiger and Stock (1997) propose that the F-statistic of the excluded instrument in the first stage regression should be at least 10 for instruments to be strong enough. In table 5 all F-statistics of the first stage are well above the cut-off value, indicating that the instrument is strong enough for all outcome variables and across all specifications. The F-values are also much larger than the F-values found by Van Elk et al. (2011), which do not exceed 12.0. This indicates that using the relative supply ratio per labor market region instead of per municipality type is indeed a substantial improvement of the first stage predictions.

The outline of table 5 is the same as before. The first specification does not control for covariates, the second specification controls for socioeconomic covariates, the third specification controls additionally for the test scores.

Table 5. The effect of early tracking for MAVO advice students using IV

Outcome variable	(1)	(2)	(3)
ln(Hourly wage)	-0.096	-0.090	-0.088
18.06	(0.065)	(0.062)	(0.060)
F-value excl. instrument	148.95	141.37	129.28
Observations	2471	2120	2060
R-squared	0.0000	0.0135	0.0323
Employment	-0.065	-0.113	-0.097
0.858	(0.077)	(0.085)	(0.084)
F-value excl. instrument	203.36	187.69	166.15
Observations	3152	2677	2595
R-squared	0.0000	0.0021	0.0084
Welfare assistance	-0.024	-0.014	-0.025
0.046	(0.035)	(0.036)	(0.038)
F-value excl. instrument	203.36	187.69	166.15
Observations	3152	2677	2595
R-squared	0.0000	0.0012	0.0040
Dropout	0.091	0.070	0.050
0.083	(0.069)	(0.046)	(0.045)
F-value excl. instrument	213.45	197.12	173.45
Observations	3253	2753	2664
R-squared	0.0000	0.0029	0.0176
Completion of higher education	-0.214*	-0.151	-0.130
0.280	(0.109)	(0.092)	(0.093)
F-value excl. instrument	213.45	197.12	173.45
Observations	3253	2753	2664
R-squared	0.0000	0.0432	0.0644
Completion of university	-0.018	-0.027	-0.021
0.033	(0.031)	(0.031)	(0.033)
F-value excl. instrument	213.45	197.12	173.45
Observations	3253	2753	2664
R-squared	0.0001	0.0021	0.0110
Socioeconomic controls	No	Yes	Yes
Test score controls	No	No	Yes

***, **, * indicates significance at a 1%, 5% or 10% significance level respectively

Standard errors clustered at school level are given in parentheses

The estimated coefficients for the IV regressions turn out to be larger (in absolute value). Oreopoulos (2006) emphasizes that this is to be expected, as the effect measured by IV is different from the effect measured by OLS. OLS measures the average treatment effect of the whole sample, whereas IV measures the local average treatment effect of the complier subsample. The effect on this particular subgroup, in case it is relatively small, can exceed that of the general population (Oreopoulos, 2006). The differences between the OLS results and the IV results indicate that there is indeed a selection bias.

The estimated coefficients in the third specification are all statistically insignificant. The IV result for the completion of higher education is not statistically significant, but is similar in magnitude to the finding of Van Elk et al. (2011). Although for the other outcome variables there are indications of a negative effect of early tracking, no definitive conclusions can be drawn as the estimates coefficients are not statistically significant.[9]

VWO advice group

Table 6 display the OLS estimation results for the VWO advice group, and has the same setup as the previous tables in this section. The first column reports the estimated coefficient without additional controls, whereas column 2 and 3 add the socioeconomic variables and the test score variables respectively.

Table 6. The effect of early tracking for VWO advice students using OLS

Outcome variable	(1)	(2)	(3)
ln(Hourly wage)	**0.069****	0.045	0.028
24.44	(0.032)	(0.029)	(0.029)
Observations	818	736	717
R-squared	0.0105	0.0590	0.0698
Employment	0.001	-0.010	-0.011
0.911	(0.018)	(0.018)	(0.018)
Observations	997	891	868
R-squared	0.0000	0.0144	0.0200
Welfare assistance	0.011	0.014	0.014
0.026	(0.010)	(0.012)	(0.011)

[9] The effect of early tracking is positive for dropout, but dropping out of school is considered undesirable and therefore designated as a “negative” effect.

Observations	997	891	868
R-squared	0.0011	0.0039	0.0076
Dropout	0.010	0.006	0.009
0.045	(0.013)	(0.013)	(0.014)
Observations	1060	948	923
R-squared	0.0005	0.0373	0.0503
Completion of higher education	0.025	0.019	0.000
0.830	(0.021)	(0.024)	(0.026)
Observations	1060	948	923
R-squared	0.0010	0.0311	0.0459
Completion of university	**0.144*** **	**0.108*** **	**0.073***
0.575	(0.033)	(0.039)	(0.040)
Observations	1060	948	923
R-squared	0.0198	0.0796	0.1221
Socioeconomic controls	No	Yes	Yes
Test score controls	No	No	Yes

***, **, * indicates significance at a 1%, 5% or 10% significance level respectively

Standard errors clustered at school level are given in parentheses

Tracking early significantly affects the completion of university. Attending a tracked VWO class increases the propensity of completing university with 7.3 percentage points in the third specification. The effect on dropout and completion of higher education is positive but statistically and economically insignificant in the third specification. As there is no effect of early tracking on completion of higher education, but there is an effect on completion of university, tracked students benefit by graduating at university level. Comprehensive students do finish tertiary education but more at the HBO level. The effect of early tracking on the labor market outcomes is insignificant for all three outcome variables in the third specification. In column 1 the effect of early tracking on the log hourly wage is positive and significant, but is confounded by the control variables, as the effect is reduced after the addition of these controls, thereby losing its significance. Table 7 presents the IV estimations for the VWO advice students in the familiar fashion.

Table 7. The effect of early tracking for VWO advice students using IV

Outcome variable	(1)	(2)	(3)
ln(Hourly wage)	**0.118***	0.080	0.068
24.44	(0.065)	(0.073)	(0.071)
F-value excl. instrument	349.26	231.34	208.22
Observations	795	713	694
R-squared	0.0016	0.0562	0.0656
Employment	-0.013	-0.041	-0.041
0.910	(0.033)	(0.039)	(0.041)
F-value excl. instrument	430.85	270.93	252.17
Observations	970	865	843
R-squared	0.0000	0.0125	0.0196
Welfare assistance	**0.028***	**0.036***	**0.040***
0.026	(0.015)	(0.021)	(0.022)
F-value excl. instrument	430.85	270.93	252.17
Observations	970	865	843
R-squared	0.0000	0.0000	0.0023
Dropout	-0.014	-0.033	-0.026
0.045	(0.023)	(0.022)	(0.022)
F-value excl. instrument	460.11	289.00	268.63
Observations	1030	920	896
R-squared	0.0000	0.0325	0.0454
Completion of higher education	0.026	0.039	0.014
0.830	(0.042)	(0.055)	(0.058)
F-value excl. instrument	460.11	289.00	268.63
Observations	1030	920	896
R-squared	0.0010	0.0297	0.0426
Completion of university	**0.186*****	**0.193*****	**0.163****
0.575	(0.058)	(0.075)	(0.072)
F-value excl. instrument	460.11	289.00	268.63
Observations	1030	920	896
R-squared	0.0192	0.0759	0.1163
Socioeconomic controls	No	Yes	Yes
Test score controls	No	No	Yes

***, **, * indicates significance at a 1%, 5% or 10% significance level respectively

Standard errors clustered at school level are given in parentheses

Just as the IV estimates are larger in absolute size for the MAVO advice students relative to OLS, so are the IV estimates for the VWO advice students, suggesting the presence of selection bias. The positive effect on log hourly wage is consistent but not significant in all specifications. Tracked VWO students earn about 7% more per hour than their comprehensive counterparts, but this effect is not significant in the third specification. The effect on employment is small and insignificant. There is a significant, positive effect on welfare assistance. This is an unexpected finding, as the descriptive statistics suggest a negative correlation. After further investigation, it turns out that this effect is driven by education level of the parents. Specifically, the effect of early tracking is only significant if the education level of the parent is equal to higher education first phase, which is equivalent to HBO bachelor or WO bachelor. The take-up of welfare assistance is affected by a number of factors like the expected amount and duration of the benefit, information costs regarding existence and application procedures, and the social stigma (Hernanz, Malherbet, & Pellizzari, 2004). However, for this particular group (tracked VWO advice students with highly educated parents) these factors are unlikely causal to their welfare uptake, as on average tracked VWO advice students earn more and are better educated. The effect remains significant if October 2014 (a year earlier) or November 2015 (a month later) is used as reference date. A possible explanation of this finding could lie in the labor market careers of the students. Of the tracked VWO advice students only 12 of the 361 (3.32%) receive welfare, compared to 14 of the 636 (2.20%) comprehensive VWO advice students. The careers of these 12 students cannot be directly identified, but information about their educational career (especially tertiary) is available. Three students did not finish secondary education, and are thus regarded as dropouts. Of the other students 4 finished university (one in language studies, and three technical studies), 4 finished HBO education (in teacher education, arts, business administration, and a technical study), and 1 did not receive a diploma for any study conducted at WO and HBO level. Their educational careers do not provide a clear explanation of why they receive welfare assistance. The fact that such a small number of students in the sample drive the result is remarkable. The significant result regarding welfare assistance is theoretically unjustifiable, but according to the data it is not temporary. The true cause of this finding is still unknown.

The results for completion of higher education and university confirm the conclusion from table 6. There is a significant, positive effect on completion of university, whilst the completion of higher education is unaffected. This means that comprehensive students finish higher education relatively more at HBO level, and the tracked students finish it relatively more at university level. Tracking early thus benefits students with VWO advice. There is again no effect on dropout. Furthermore, the first stage F-statistics of the excluded instrument are well above the cut-off value of 10 across all specifications and outcomes, indicating that the instrument is strong enough.

The results from table 4 to 7 indicate that tracking early benefits students with a higher school advice, but hurts students with a lower school advice. For both groups under consideration the IV estimations are larger (in absolute size) than the OLS results, but are in line with the pattern of the OLS results. Notably, for MAVO advice students, completion of higher education is affected negatively by early tracking. But this effect seems to be confounded, as the significance disappears in the preferred specification. Wages and employment are lower for tracked students, and dropout rates are higher, but these results are not significant, and therefore no definite conclusion can be drawn. For VWO advice students the results point in a different direction. Completion of university is significantly, positively affected across all specifications. Wages are higher for tracked students, but this effect is not significant in the preferred specification, indicating confoundedness by the control variables. Furthermore, the IV results suggest a positive effect on receiving welfare assistance.

Effect on subgroups

Table 8 and 9 display the estimated effects of early tracking for particular subgroups. Specifically, the subgroups are boys and girls, children from lower educated parents and higher educated parents, where higher educated is defined as at least completion of first phase higher education, and below and above the median test score. This test score is the combination of the three tests (arithmetic, language, and information processing) taken at the beginning of secondary education. The estimates come from OLS, indicated by the upper estimate with corresponding standard error, and IV, indicated by the lower estimate with corresponding standard error, using the third specification, that is, controlling for both the socioeconomic variables as well as the test score variables.

Table 8. The effect of early tracking on subgroups for MAVO advice students

Outcome variable	Boys	Girls	Lower education	Higher education	Below median	Above median
ln(Hourly wage)	0.005	-0.015	-0.014	0.028	-0.021	0.008
18.06	(0.023)	(0.016)	(0.014)	(0.031)	(0.022)	(0.018)
	-0.041	-0.121	**-0.120***	0.024	-0.076	-0.088
	(0.094)	(0.079)	(0.067)	(0.151)	(0.076)	(0.093)
Employment	-0.013	-0.010	0.007	**-0.081*****	-0.033	0.006
0.858	(0.017)	(0.020)	(0.017)	(0.026)	(0.021)	(0.019)
	-0.152	-0.063	-0.111	-0.020	-0.119	-0.041
	(0.099)	(0.098)	(0.094)	(0.157)	(0.109)	(0.107)
Welfare assistance	0.007	-0.013	-0.008	0.010	0.004	-0.011
0.046	(0.011)	(0.011)	(0.009)	(0.018)	(0.14)	(0.010)
	0.009	-0.050	-0.015	-0.082	-0.009	-0.059
	(0.053)	(0.045)	(0.040)	(0.093)	(0.047)	(0.062)
Dropout	-0.023	0.007	-0.015	0.027	-0.002	-0.010
0.083	(0.015)	(0.013)	(0.012)	(0.017)	(0.017)	(0.011)
	-0.025	**0.113***	0.037	0.104	0.057	0.034
	(0.057)	(0.060)	(0.046)	(0.100)	(0.052)	(0.067)
Completion of higher education	-0.025	**-0.077*****	**-0.052****	-0.053	**-0.087****	-0.023
	(0.030)	(0.026)	(0.024)	(0.046)	(0.033)	(0.027)
0.280	0.062	**-0.295****	-0.155	0.026	-0.076	-0.210
	(0.126)	(0.120)	(0.101)	(0.206)	(0.096)	(0.195)
Completion of university	0.016	**-0.022****	-0.009	0.014	-0.008	-0.002
	(0.010)	(0.010)	(0.007)	(0.021)	(0.008)	(0.010)
0.033	0.010	-0.050	-0.038	0.072	-0.003	-0.059
	(0.058)	(0.034)	(0.029)	(0.117)	(0.027)	(0.069)
Socioeconomic controls	Yes	Yes	Yes	Yes	Yes	Yes
Test score controls	Yes	Yes	Yes	Yes	Yes	Yes

***, **, * indicates significance at a 1%, 5% or 10% significance level respectively

Standard errors clustered at school level are given in parentheses

The results in table 8 indicate that girls with MAVO advice are hurt more by early tracking than boys with MAVO advice. Girls significantly complete less higher education, and even less university in case of the OLS result. This effect might be confounded by selection bias. Girls also drop out more according to the IV results. For the other outcomes the results are not significant, therefore no conclusions can be drawn.

Children of lower educated parents that enroll in a categorical class earn a lower hourly wage. The negative effect of early tracking on completion of higher education is the same for children of lower educated parents and children of higher educated parents, but the effect for lower educated parents is significant. Once controlling for selection, the significant effect is lost. Children of higher educated parents are significantly less employed, but this result is again only significant for the OLS specification. As it turns out, the effect is strongest for children with parents having finished higher education first phase. There are also no notable distinctions between above median students and below median students, except for completion of higher education in the OLS specification. The below median students are hurt more by early tracking than the above median students, which is logically justifiable. Table 9 gives the results for the VWO advice group.

Table 9. The effect of early tracking on subgroups for VWO advice students

Outcome variable	Boys	Girls	Lower education	Higher education	Below median	Above median
ln(Hourly wage)	0.006	**0.060****	0.034	0.025	-0.006	**0.071***
24.44	(0.049)	(0.029)	(0.038)	(0.044)	(0.037)	(0.042)
	0.110	0.017	0.099	0.051	-0.001	0.188
	(0.096)	(0.077)	(0.088)	(0.101)	(0.075)	(0.124)
Employment	-0.026	0.014	**0.059****	**-0.053****	0.014	-0.029
0.911	(0.026)	(0.027)	(0.026)	(0.026)	(0.029)	(0.027)
	-0.047	-0.021	-0.009	-0.068	-0.034	-0.040
	(0.047)	(0.083)	(0.070)	(0.054)	(0.052)	(0.083)
Welfare assistance	0.025	-0.004	-0.001	0.021	0.002	0.020
0.026	(0.016)	(0.017)	(0.017)	(0.015)	(0.014)	(0.017
	0.026	0.048	-0.012	**0.070****	0.035	0.032
	(0.027)	(0.046)	(0.030)	(0.031)	(0.022)	(0.037)
Dropout	0.010	0.007	0.009	0.011	0.014	-0.004
0.045	(0.017)	(0.019)	(0.026)	(0.011)	(0.022)	(0.014)
	-0.025	-0.029	-0.009	-0.029	-0.027	-0.028
	(0.027)	(0.029)	(0.038)	(0.025)	(0.035)	(0.027)
Completion of higher education	-0.063	**0.064***	-0.012	0.004	0.011	0.011
	(0.040)	(0.033)	(0.046)	(0.028)	(0.037)	(0.034)
0.830	-0.102	**0.157****	-0.083	0.058	0.046	-0.005
	(0.076)	(0.075)	(0.130)	(0.059)	(0.057)	(0.106)
Completion of university	**0.101***	0.048	0.071	0.064	**0.109***	0.053
	(0.054)	(0.050)	(0.060)	(0.044)	(0.057)	(0.052)
0.575						

	0.182**	0.142	**0.167***	0.146	**0.240*****	0.077
	(0.087)	(0.105)	(0.099)	(0.097)	(0.076)	(0.134)
Socioeconomic controls	Yes	Yes	Yes	Yes	Yes	Yes
Test score controls	Yes	Yes	Yes	Yes	Yes	Yes

***, **, * indicates significance at a 1%, 5% or 10% significance level respectively

Standard errors clustered at school level are given in parentheses

In the OLS regression girls earn significantly more than boys, but in the IV regression boys earn more than girls, though not significantly. For employment, welfare assistance, and dropout there are no differences between boys and girls. Furthermore, girls that go to a tracked VWO class complete significantly more higher education, but boys that go to a tracked class complete significantly more university. This result for boys is in line with the main regression results, as tracked students benefit from completion of university rather than HBO. But the result for girls is striking. First, there are positive but insignificant effects on completion of university, that is, it cannot be concluded that tracked girls differ from comprehensive girls in terms of completing university. Second, tracked girls complete more higher education than comprehensive girls, and these results are significant, despite the fact that there is no difference in university completion. So tracked girls finish more higher education at HBO level, which is the exact opposite finding of tracked boys.

The children from lower educated parents benefit more from early tracking as it significantly increases completion of university. Furthermore, there is again evidence from the IV regression that early tracking positively affects welfare assistance, but only for children from highly educated parents. As discussed above, this result is driven by a specific education level of parents. Another observation, this time from the OLS results, is that early tracking affects employment differently for children from higher educated parents and children from lower educated parents. Children from lower educated parents that enroll in a tracking school are employed more often, whereas children from higher educated parents that enroll in a tracking school are employed less often. Once controlling for selection bias, these effects disappear. In the main OLS regression results, there is little to no effect on employment as they are seemingly balanced out. For log hourly wage, dropout, and completion of higher education there are no differences.

The below median students benefit more from early tracking in terms of completion of higher education, but do not see their educational effort back in their hourly wage. The above median tracked students do not distinguish themselves in terms of educational attainment from their comprehensive counterparts, but there is some difference in hourly wages. However, this observation is not significant once controlling for selection. The other variables do not show clear distinctions.

This section provided the main estimation results. From OLS it follows that early tracking negatively affects the MAVO advice students, but this effect is only significant for completion of higher education. In contrast, early tracking affects VWO advice students positively, with a significant effect on completion of university and hourly wage, although the latter is only significant in the first specification. The IV results confirm the OLS results. For the other outcome variables there seems to be no effect, neither with OLS, nor with IV. Furthermore, there are some differences between subgroups, notably girls are more harmed by early tracking than boys for both the MAVO advice group and the VWO advice group in terms of completion of higher education and completion of university respectively.

Section VI: further analyses

This section presents further analyses related to the research question. First, the main estimation results of the '93 cohort are compared to the '89 cohort to see if there are differences between cohorts. Second, the effect on the outcome variables for the HAVO advice students attending MH and the HV are compared to investigate the net effect on these students. Next, the results of Section V are subjected to robustness analysis to see if the results are sensitive to minor changes in the data. After that follows a discussion of the validity of the instrument. Finally, the complier subpopulation is addressed.

'89 cohort

Comparing the effect of early tracking for different cohorts reveals if early tracking yields similar results at different points in time. The students of the '89 cohort are aged around 38 in 2015, the year from which the labor market outcomes are taken. As this study is interested in the long-term outcomes of early tracking, using the '99 cohort would be less insightful, as they are aged only 28 in 2015. Table 10 presents the OLS (upper estimate) and IV (lower estimate) results for the three specifications as in table 4 to 7. The first three columns are for the MAVO advice group; the columns (4) to (6) are for the VWO advice group. The numbers below the outcome variable names are the mean for the MAVO advice group and the mean for the VWO advice group respectively for that particular variable. The instrument used is the relative supply ratio of categorical schools in '89, constructed exactly the same as the '93 relative supply ratio.

Table 10. The effect of early tracking for MAVO advice and VWO advice for '89 cohort

Outcome variable	MAVO (1)	(2)	(3)	VWO (4)	(5)	(6)
ln(Hourly wage)	**-0.028****	**-0.033****	**-0.035****	**0.100*****	**0.059***	0.053
19.32	(0.014)	(0.014)	(0.014)	(0.034)	(0.034)	(0.034)
29.04	**-0.343***	**-0.291***	**-0.247****	**0.123***	0.096	0.082
	(0.192)	(0.148)	(0.116)	(0.068)	(0.086)	(0.085)
Employment	0.016	0.019	**0.022***	0.020	0.014	0.009
0.860	(0.013)	(0.013)	(0.013)	(0.019)	(0.023)	(0.021)
0.912	0.042	0.124	0.098	0.013	0.034	0.021
	(0.119)	(0.108)	(0.092)	(0.034)	(0.048)	(0.048)
Welfare assistance	**-0.016***	**-0.014***	**-0.013***	-0.007	-0.001	0.000
0.047	(0.009)	(0.007)	(0.008)	(0.014)	(0.015)	(0.013
0.031	-0.000	-0.028	-0.019	-0.031	-0.039	-0.031
	(0.054)	(0.049)	(0.045)	(0.021)	(0.027)	(0.024)
Dropout	-0.022	-0.017	**-0.022***	0.007	0.016	0.021
0.093	(0.016)	(0.013)	(0.012)	(0.016)	(0.018)	(0.018)
0.052	0.039	0.018	0.055	-0.010	-0.012	-0.002
	(0.092)	(0.070)	(0.068)	(0.024)	(0.032)	(0.031)
Completion of higher education	**-0.041****	**-0.042****	**-0.042****	**0.075****	0.046	0.031
0.235	(0.019)	(0.018)	(0.017)	(0.029)	(0.030)	(0.028)
0.801	**-0.419***	**-0.346****	**-0.406****	**0.104***	0.094	0.073
	(0.212)	(0.161)	(0.163)	(0.056)	(0.067)	(0.066)
Completion of university	**-0.012****	**-0.010***	-0.008	**0.213*****	**0.168*****	**0.147*****
0.025	(0.005)	(0.005)	(0.005)	(0.032)	(0.038)	(0.038)
0.541	-0.042	-0.031	-0.034	**0.157****	0.122	0.104
	(0.047)	(0.038)	(0.033)	(0.074)	(0.080)	(0.086)
Socioeconomic controls	No	Yes	Yes	No	Yes	Yes
Test score controls	No	No	Yes	No	No	Yes

***, **, * indicates significance at a 1%, 5% or 10% significance level respectively

Standard errors clustered at school level are given in parentheses

There are some differences between the '89 cohort and the '93 cohort. For the MAVO advice group, tracked students are employed slightly more often and receive slightly less welfare assistance. These effects are only significant for OLS, and therefore suffer from selection bias. Completion of higher education and hourly wages are negatively affected, just as for the '93 cohort, though some of these estimates are larger in magnitude. The estimated OLS effect on completion of higher education is in line with the results from Van Elk et al. (2011), but the IV coefficient is much larger (Van Elk. et al. found the effect to be -0.129). The estimated IV coefficients for hourly wage are also much larger. The validity of the instrument is not guaranteed, there is a possibility that the exclusion restriction is violated, invalidating the IV results in table 10. Regressing the outcome variables on the '89 ratio (and the controls) results in (highly) significant effects only for the log hourly wage variable and the completion of higher education variable in case of the MAVO advice group. This indicates that the instrument is directly related to these variables, which is in violation with the exclusion restriction (this validity test corresponds to the first validity test applied to the '93 instrument later in this section). The '89 ratio in its current form is not a valid instrument and thus the estimated effects for hourly wage and completion of higher education for the MAVO advice group are incorrect. A possible reason why the instrument is less valid for the '89 cohort is that, just as for the '93 cohort, not all schools are allocated to their respective labor market region. As a matter of fact, 126 of the 135 are not allocated. Once this problem in the data is solved, more constructive conclusions can be made about the effect of early tracking for the '89 cohort.

For the VWO advice group, the OLS and the IV estimates are in line with the results of the '93 cohort in both direction and magnitude. The only difference is the disappearance of the positive effect on welfare assistance, adding to the belief that it is indeed an uncommon finding. All in all, the effects of early tracking are similar to those found for the '93 cohort, but the quality of these estimates is not guaranteed due to missing data.

HAVO advice group

From the main estimation results it follows that the MAVO students attending the MH class benefit in terms of completion of higher education. They possibly gain from the peer effects of HAVO advice students attending the same class, but also from a different teaching style and different curriculum. Likewise, the VWO students in the HV class are negatively affected by these effects, as they are less likely to finish university education. The net result of these effects can be distinguished by comparing the HAVO advice students. As these students either start in the MH class or the HV class, there isn't a difference in the age of tracking. Subsequently this investigation is not about the effects of early tracking, rather it is about the net effect of attending the MH relative to the HV class. The HAVO advice students in the MH class are accompanied by MAVO advice students, whereas the HAVO advice students in the HV class are accompanied by VWO advice students. For instance, the peer effect HAVO advice students receive in these two classes might have opposite effects. Therefore, the outcomes of HAVO advice students attending the MH class are compared to the outcomes of HAVO advice students attending the HV class to investigate if the net effect is indeed different in these two classes. The MHV class is not taken into account because the HAVO advice students are accompanied by MAVO and VWO advice students, so the net effect they receive is confounded. The equation used becomes:

$$Y_i = \beta_0 + \beta_1 MH_i + \beta_2 X_i' + \varepsilon_i \tag{4}$$

Now, MH_i denotes if student i attends the MH class or not. Table 11 presents the OLS estimates for all outcome variables.

Table 11. The effect of enrolling in the MH class for HAVO advice students using OLS

Outcome variable	(1)	(2)	(3)
ln(Hourly wage)	**-0.121*** **	**-0.098*** **	**-0.071*** **
21.11	(0.034)	(0.029)	(0.026)
Observations	1082	953	933
R-squared	0.0097	0.0384	0.0526
Employment	-0.028	-0.010	0.000
0.892	(0.025)	(0.028)	(0.031)
Observations	1335	1172	1151
R-squared	0.0007	0.0171	0.0215
Welfare assistance	0.022	0.023	0.019
0.032	(0.018)	(0.022)	(0.021)
Observations	1335	1172	1151
R-squared	0.0016	0.0245	0.0272
Dropout	0.002	-0.013	-0.018
0.084	(0.019)	(0.026	(0.024)
Observations	1391	1218	1195

R-squared	0.0000	0.0231	0.0267
Completion of higher education 0.585	**-0.205******** (0.065)	**-0.126*** (0.068)	-0.082 (0.067)
Observations	1391	1218	1195
R-squared	0.0153	0.0620	0.0783
Completion of university 0.195	**-0.133***** (0.024)	**-0.105***** (0.027)	**-0.062**** (0.031)
Observations	1391	1218	1195
R-squared	0.0093	0.0401	0.0646
Socioeconomic controls	No	Yes	Yes
Test score controls	No	No	Yes

***, **, * indicates significance at a 1%, 5% or 10% significance level respectively

Standard errors clustered at school level are given in parentheses

Attending the MH class has a negative effect on hourly wage, and completion of university. Hourly wages are 7.1% lower and there is a 6.2 percentage point lower probability of finishing university. Note that these effects are likely to be biased by selection. Unfortunately, the instrument is not strong enough to also present the IV results in order to further investigate the net effect on HAVO advice students. Consequently, no final conclusions about the effect on HAVO advice students can be drawn.

Robustness analysis

Table 12 presents the robustness analysis of the main regression results. The regression results should be insensitive to sample restrictions. The first restriction looks only at students living in even-numbered labor market regions. The second restriction looks only at even-numbered observations. That is, the second student in the sample, the fourth student, the sixth, and so on. Column (1) presents the main OLS regression estimates, which is the upper estimate, and the main IV regression estimates, which is the lower estimate, for MAVO advice students; column (2) presents the OLS and the IV estimates for students living in even-numbered labor market regions; column (3) presents the OLS and IV estimates for the even-numbered observations. Columns (4), (5) and (6) do respectively the same, but for the VWO advice students. For the robustness analysis to be successful, the magnitude of the estimates is important.

Though some estimates turn out more significant, the results do not change appreciably in magnitude.

Table 12. Robustness analysis

Outcome variable	MAVO (1)	(2)	(3)	VWO (4)	(5)	(6)
ln(Hourly wage)	-0.005	-0.012	-0.011	0.028	0.002	0.031
	(0.013)	(0.018)	(0.018)	(0.029)	(0.033)	(0.041)
	-0.088	-0.277	-0.105	0.068	0.019	0.079
	(0.060)	(0.212)	(0.079)	(0.071)	(0.099)	(0.107)
Employment	-0.010	0.007	-0.008	-0.011	-0.001	-0.018
	(0.014)	(0.022)	(0.017)	(0.018)	(0.024)	(0.029)
	-0.097	-0.140	-0.043	-0.041	-0.065	-0.064
	(0.084)	(0.213)	(0.090)	(0.041)	(0.060)	(0.048)
Welfare assistance	-0.004	-0.008	0.001	0.014	0.016	0.020
	(0.008)	(0.011)	(0.010)	(0.011)	(0.011)	(0.020)
	-0.025	-0.086	-0.000	**0.040***	**0.075*****	-0.003
	(0.038)	(0.084)	(0.048)	(0.022)	(0.024)	(0.035)
Dropout	-0.006	-0.016	-0.003	0.009	0.016	0.006
	(0.011)	(0.016)	(0.013)	(0.014)	(0.018)	(0.022)
	0.050	0.071	-0.037	-0.026	**-0.056****	-0.026
	(0.045)	(0.109)	(0.047)	(0.022)	(0.027)	(0.030)
Completion of higher education	**-0.053****	**-0.065****	**-0.048***	-0.000	-0.019	0.016
	(0.022)	(0.027)	(0.028)	(0.026)	(0.034)	(0.042)
	-0.130	-0.223	0.072	0.014	-0.059	0.053
	(0.093)	(0.217)	(0.104)	(0.058)	(0.063)	(0.082)
Completion of university	-0.004	-0.005	0.002	**0.073***	0.031	**0.111***
	(0.007)	(0.009)	(0.008)	(0.040)	(0.049)	(0.058)
	-0.021	-0.020	-0.011	**0.163****	0.104	**0.282****
	(0.033)	(0.055)	(0.035)	(0.072)	(0.086)	(0.109)
Socioeconomic controls	Yes	Yes	Yes	Yes	Yes	Yes
Test score controls	Yes	Yes	Yes	Yes	Yes	Yes

***, **, * indicates significance at a 1%, 5% or 10% significance level respectively

Standard errors clustered at school level are given in parentheses

Validity instrument

In order for the instrument to be valid, the exclusion restriction and the first-stage assumption must hold. The first-stage F-values given in the tables in Section V confirm that the first stage is strong enough for legitimate estimation. The exclusion restriction states that the instrument can only have an effect on the outcome variables through the endogenous variable, in this case early tracking. This is investigated via four different tests.

The first test for the validity is adding the instrument to the OLS regression for all outcome variables and for both groups. The ratio should not have a significant effect on the outcome variable; otherwise the exclusion restriction is violated for that outcome variable. The instrument only affects the dropout rate significantly for VWO advice students. Accordingly, the IV results for VWO advice dropout are invalid. The following validity tests investigate further if this poses a problem for the validity of instrument in general.

If the supply ratio of schools is affected by demand factors, it could be that the exclusion restriction is violated. For instance, if in a certain area there are relatively more highly-educated people, this can be reflected in the school types that are available in that area. If this is the case, the supply ratio is affected by education and/or motivation, and therefore the supply ratio itself is influenced by selection. The second test examines if this poses a problem by adding variables to the regressions as controls and see if the main results are affected. Specifically, the fraction of highly educated parents per labor market region, those that have finished at least higher education, and the average level of education are added. Table 13 presents the results. Column (1) displays the main regression results of the MAVO advice group for comparison, column (2) adds the fraction of highly educated parent per labor market region, and column (3) adds average level of education per labor market region. Columns (4) to (6) do the same, but for the VWO advice students. As there are no significant changes to the results, the supply ratio is not affected by parents' demand for certain school types.

Table 13. The influence of demand factors on the main regression results

Outcome variable	MAVO			VWO		
	(1)	(2)	(3)	(4)	(5)	(6)
ln(Hourly wage)	-0.005	-0.004	-0.005	0.028	0.028	0.024
	(0.013)	(0.118)	(0.031)	(0.029)	(0.029)	(0.029)
	-0.088	-0.075	-0.090	0.068	0.064	0.042
	(0.060)	(0.069)	(0.063)	(0.071)	(0.070)	(0.072)
Employment	-0.010	-0.005	-0.006	-0.011	-0.010	-0.013
	(0.014)	(0.014)	(0.014)	(0.018)	(0.019)	(0.019)
	-0.097	-0.086	-0.088	-0.022	-0.051	-0.060

	(0.084)	(0.095)	(0.084)	(0.045)	(0.041)	(0.040)
Welfare assistance	-0.004	-0.006	-0.006	0.014	0.014	0.015
	(0.008)	(0.008)	(0.008)	(0.011)	(0.011)	(0.011)
	-0.025	-0.035	-0.030	**0.038***	**0.041***	**0.044****
	(0.038)	(0.044)	(0.040)	(0.022)	(0.021)	(0.022)
Dropout	-0.006	-0.008	-0.007	0.009	0.009	0.010
	(0.011)	(0.011)	(0.011)	(0.014)	(0.014)	(0.014)
	0.050	0.057	0.053	-0.033	-0.027	-0.020
	(0.045)	(0.049)	(0.045)	(0.023)	(0.023)	(0.024)
Completion of higher education	**-0.053****	**-0.049****	**-0.052****	-0.000	0.000	-0.003
	(0.022)	(0.022)	(0.023)	(0.026)	(0.026)	(0.026)
	-0.130	-0.118	-0.128	0.035	0.006	-0.008
	(0.093)	(0.106)	(0.096)	(0.059)	(0.057)	(0.056)
Completion of university	-0.004	-0.005	-0.021	**0.073***	**0.074***	**0.069***
	(0.007)	(0.007)	(0.019)	(0.040)	(0.040)	(0.074)
	-0.021	-0.022	-0.023	**0.186****	**0.153****	**0.138***
	(0.033)	(0.037)	(0.034)	(0.075)	(0.073)	(0.072)
Socioeconomic controls	Yes	Yes	Yes	Yes	Yes	Yes
Test score controls	Yes	Yes	Yes	Yes	Yes	Yes
Fraction high educated	No	Yes	No	No	Yes	No
Average education	No	No	Yes	No	No	Yes

***, **, * indicates significance at a 1%, 5% or 10% significance level respectively

Standard errors clustered at school level are given in parentheses

The third test concerns another channel through which the supply ratio might be linked to outcomes, namely via the quality of schools. If there is a correlation between the school quality and the supply ratio, the exclusion restriction would be violated as school quality affects the outcome variables. In order to test this, the average test score per school in ’89 is taken as a proxy for school quality. The advantage of taking the ’89 average test scores rather than the ’93 average test scores is that the quality of the school is determined before the data in the ’93 sample. The proxy is regressed on the ratio and the control variables to see if there is a correlation. Neither for the MAVO advice group, nor for the VWO advice group there is a significant effect of the supply ratio on school quality.[10]

A final test on the validity of the instrument is checking whether the instrument is correlated with the residuals of the main OLS regression results. If there is such a

[10] These results are not shown, and are available on request.

correlation, this indicates that the supply ratio is linked to unobservable factors present in the error term. As it turns out the supply ratio is not significantly correlated to the residuals of the outcome variables for both MAVO advice and VWO advice, except for the residual of dropout for the VWO advice group.[11] The results of this final test are consistent with the results of the first validity test.

Furthermore, the instrument is based on students' living addresses, which are available from the first of January, '95. As the VOCL data is from '93, it is possible that students have relocated from between September '93, the start of the school year, and until January '95. The VOCL denotes the reason why a student left the school they enrolled in in '93. As it turns out, 13 students relocated elsewhere in the period between September '93 and August '95. For 39 students in that same period the reason they left is labeled unknown. At its maximum, 1.1% of the sample under consideration could have moved, though it is likely only 0.27% has moved. The mismatch between living addresses does not pose a significant problem for the validity of the instrument.

Complier subpopulation

The effect measured by IV is not the average treatment effect but the local average treatment effect. This is the treatment effect on the complier subpopulation. The compliers are the individuals who go to a tracking school when exposed to a higher relative supply ratio. In order to characterize individuals in the sample as a complier, one can compare the first-stage coefficient of the instrument for certain subgroups. Table 8 and 9 displays the effect of early tracking for different subgroups, but not the first-stage coefficients. These first-stage coefficients of the instrument for the subgroups are displayed in table 14. For each subgroup three estimates are given, as there are three different first-stage estimations. The data on hourly wages comes from a different data set (UWV) than the data on employment and welfare assistance (both SEC), resulting in minor differences.[12] The data on dropout, completion of higher education, and completion of university come from the VOCL itself. So for each data set (with a different origin), there is a unique first-stage.

[11] See footnote 10.
[12] The differences in the number of observations given in the main regression results for each outcome variable demonstrate this.

Table14. First-stage estimated coefficients of the instrument

Subgroup	MAVO advice			VWO advice		
	(1)	(2)	(3)	(4)	(5)	(6)
Boys	0,424	0,441	0,450	0,872	0,892	0,889
Girls	0,480	0,475	0,478	1,005	0,960	0,982
Low educated parents	0,464	0,470	0,470	1,000	0,954	0,952
High educated parents	0,394	0,388	0,420	0,857	0,888	0,901
Below median	0,544	0,576	0,585	1,034	1,054	1,070
Above median	0,377	0,348	0,349	0,783	0,770	0,777

Three estimates are given per subgroup, as there are three different first-stage regressions.

Column (1) and (4) is for the wage sample (UWV), column (2) and (5) for the employment and welfare assistance sample (SEC), column (3) and (6) for the educational outcomes sample (VOCL).

As it turns out, there are no noteworthy differences in the instrument coefficients when comparing boys and girls in the MAVO advice group. This indicates that the complier subpopulation is not overrepresented by either boys or girls. For the VWO advice group the coefficients are higher for girls, but these differences are not obvious. For the MAVO advice group, the first-stage instrument coefficient is higher for children with lower educated parents and for below median students, but the differences are more profound for the below median group. These individuals are therefore more likely to be compliers, as the instrument is more indicative of enrolling in a tracked school. For the VWO advice group, there are clear differences in the first-stage instrument coefficients between the above and below-median students. The below-median students have higher coefficients and are thus overrepresented in the complier subpopulation.

Another method of characterizing compliers is comparing the compliance probability for certain subgroups (Angrist and Pischke, 2008). This compliance probability for a binary instrument is constructed of three parts. The first part is the size of the complier group, which is given by the Wald first-stage estimate: $E[T_i|Z_i = 1] - E[T_i|Z_i = 0]$. It is the increase in the probability of going to a tracking school when the instrument is "switched on". This probability is multiplied with $P[Z_i = 1]$, or the probability that the instrument is switched on. Finally, divide by $P[T_i = 1]$, or the probability of treatment (going to a tracking school) to get to the compliance probability. The formula becomes:

$$\frac{P[Z_i = 1](E[T_i|Z_i = 1] - E[T_i|Z_i = 0])}{P[T_i = 1]} \tag{5}$$

For a compliance probability of a subgroup the Wald first-stage estimate becomes:

$$\frac{E[T_i|Z_i = 1, x_{1i} = 1] - E[T_i|Z_i = 0, x_{1i} = 1]}{E[T_i|Z_i = 1] - E[T_i|Z_i = 0]} \tag{6}$$

Where $x_{1i} = 1$ indicates that individual i is part of subgroup 1, for instance female. In other words, the Wald first-stage is determined by the first stage for females relative to the overall first stage. However, the instrument used in this thesis is not binary, as it takes on different values between 0 and 1 indicating the relative supply ratio. This entails that the interpretation of this method is less insightful. Therefore, only the more intuitive and simpler method involving the first-stage estimates is used to characterize the compliers.

This section presented several analyses. The overall result for the '89 cohort is in line with the result for the '93 cohort, though there are some differences in the magnitudes of the estimated effects resulting from the instrument being invalid. HAVO advice students benefit from going to a HV class through higher wages and finishing more university, but these results might be driven by selection. Furthermore, the main regression results are robust to restrictions on the sample. The exclusion restriction is likely to hold, resulting in a valid instrument. Finally, there is evidence that the complier subpopulation consists mainly of below median students for both the MAVO advice group and the VWO advice group.

Section VII: discussion and conclusion

This thesis tries to answer the question whether early tracking in the Dutch education system leads to long-term differences. Students that enter secondary education in a comprehensive class are effectively tracked at a later age than students who enroll in a categorical class right away. Using the Secondary Education Cohort Student data from 1993, several labor market outcomes and educational outcomes of these students are compared. Potential selection bias is overcome by applying an instrumental variable approach, using an instrument based the regional variation in the availability of certain school types.

The effect of early tracking turns out differently for the two student groups under consideration. The MAVO advice students that enroll in a categorical MAVO class in the first year of secondary education are 5.3 percentage points less likely to finish higher education. After controlling for selection bias, this effect increases in absolute size, but it is not statistically significant. The other outcome variables are not significantly affected in the preferred specification, but most of the estimates are unfavorable for the tracked students. The VWO advice students tend to benefit from early tracking. They significantly graduate more often at university level (7.3 percentage points). This effect increases in size after controlling for the selection bias to 16.3 percentage points. The effect on other variables is insignificant, except for an increase in welfare assistance of 4.0 percentage points, which seems, after deeper investigation, a peculiarity in the data.

To interpret the IV results, it should be noted that they are not easily generalizable. Using an instrumental variable approach estimates not the average treatment effect on the whole population, but the local average treatment effect on the subgroup of the affected individuals, the so-called compliers. The effects estimated by OLS are generalizable, but they suffer from selection bias, and might therefore also differ from the average treatment effect.

There are two important limitations to this study affecting the quality of the conclusions. First, the instrument suffers from missing data, as not all schools in the data set are allocated to their respective labor market region. Improving the allocation would result in an even stronger instrument, making way for more precise estimations. The instrument could also be improved by using a spatial approach to determine the relative supply ratio in the direct vicinity of a student's living address. Additionally, the effect for the '89 cohort, currently obscured by imprecise estimation, can be ascertained to provide a comparison at different points in time. Second, the effects measured in this paper are a *net effect* of going to categorical class rather than a comprehensive class. This net effect is comprised of different elements, like peer effects and teacher effort. Isolating the effect of these elements requires follow-up research.

The net effect of going to a lower class for HAVO advice students suggests non-linear net effects, where the MAVO advice students benefit, but the HAVO advice students are hurt. Due to these non-linear effects it is hard to draw conclusions on the overall efficiency of the Dutch tracking system, as these non-linear effects highlight both the advantage and the disadvantage of the system. As an economist, I refrain from putting weights on the non-linear effects for each advice group, as this becomes increasingly subjective. With every policy decision, there are winners and losers. Abandoning the system affects students at the top of the ability distribution, but keeping the system as it currently is affects those at the bottom.

References

Ammermüller, A. (2005). *Educational opportunities and the role of institutions* (ZEW Discussion Paper 05-44). Mannheim: Centre for European Economic Research.

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

Angrist, J. D., & Krueger, A.D. (1991). Does Compulsory School Attendance Affect Schooling and Earnings?, *Quarterly Journal of Economics, 106*(4), 979-1014.

Angrist, J. D., & Pischke, J. S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Ariga, K., Brunello, G. (2007). *Does Secondary School Tracking Affect Performance? Evidence from IALS* (IZA Discussion Paper No. 2643).

Ariga, K., Brunello, G., Iwahashi, R., Rocco, L. (2005). *Why is the timing of school tracking so heterogenous?* (IZA discussion paper 1854).

Bauer, P., & Riphahn, R. T. (2006). Timing of school tracking as a determinant of intergenerational transmission of education. *Economics Letters*, *91*(1), 90-97.

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

Borghans, L., Diris, R., Smits, W., & De Vries, J. (2012). *The Impact of Early Tracking on Later-life Outcomes: An Instrumental Variable Approach*. (Discussion Paper, 7 (143)).

Bound, J., Jaeger, D. A., & Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American statistical association*, *90*(430), 443-450.

Brunello, G., & Checchi, D. (2007). Does school tracking affect equality of opportunity? New international evidence. *Economic policy*, *22*(52), 782-861.

Brunello, G., Giannini, M., & Ariga, K. (2007). The Optimal Timing of School Tracking, forthcoming in Peterson, P. and L. Woessmann (eds.), *Schools and the Equal Opportunity Problem*, MIT Press, Cambridge MA. (also IZA Discussion Paper 955)

Card, D. (1993). *Using Geographic Variation in College Proximity to Estimate the Returns to Schooling*. (NBER Working Paper No. 4483)

Card, D., & Giuliano, L. (2016). Can Tracking Raise the Test Scores of High-Ability Minority Students?. *The American Economic Review*, *106*(10), 2783-2816.

Currie, J., & Moretti, E. (2003). Mother's education and the intergenerational transmission of human capital: Evidence from college openings. *The Quarterly Journal of Economics*, *118*(4), 1495-1532.

Duflo, E., Dupas, P., & Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. *The American Economic Review*, *101*(5), 1739-1774.

Dustmann, C. (2004). Parental background, secondary school track choice, and wages. *Oxford Economic Papers*, *56*(2), 209-230.

Dustmann, C., Puhani, P. A., & Schönberg, U. (2016). The Long-Term Effects of Early Track Choice. *The Economic Journal*, forthcoming.

Eisenkopf, G. (2007). Tracking and incentives. A comment on Hanushek and Woessman, Thurgau Institute of Economics, Research Paper Series, (22).

Epple, D., Newlon, E., & Romano, R. (2002). Ability tracking, school competition, and the distribution of educational benefits. *Journal of Public Economics*, *83*(1), 1-48.

Figlio, D. N., & Page, M. E. (2002). School choice and the distributional effects of ability tracking: does separation increase inequality?. *Journal of Urban Economics*, *51*(3), 497-514.

Galinda-Rueda, F. and A. Vignoles (2007). The heterogeneous effect of selection in UK secondary schools. In: L. Woessmann and P. E. Peterson (eds.), *Schools and the Equal Opportunity Problem*, Cambridge, MA: MIT Press, 103-128.

Garlick, R. (2016). *Academic peer effects with different group assignment policies: residential tracking versus random assignment* (Economic Research Initiatives at Duke (ERID) Working Paper 220).

Hanushek, E. A., & Woessmann, L. (2006). Does early tracking affect educational inequality and performance? Differences-in-differences evidence across countries. *Economic Journal, 116*(510), C63-C76.

Hernanz, V., Malherbet, F., & Pellizzari, M. (2004). *Take-up of welfare benefits in OECD countries: A review of the evidence, vol. 17*. Paris: OECD.

Jakubowski, M. (2010). Institutional tracking and achievement growth: Exploring difference-in-differences approach to PIRLS, TIMSS, and PISA data. *Quality and inequality of education*, 41-81.

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

Meier, V., & Schütz, G. (2007). *The economics of tracking and non-tracking* (Ifo Working Paper, No. 50).

Oreopoulos, P. (2006). Estimating average and local average treatment effects of education when compulsory schooling laws really matter. *The American Economic Review*, *96*(1), 152-175.

Organisation for Economic Co-operation and Development (OECD). (2011). Equity and Quality in Education-Supporting Disadvantaged Students and Schools. Paris: OECD Publishing.

Organisation for Economic Co-operation and Development (OECD). (2016). *Netherlands 2016: Foundations for the Future*, Paris: OECD Publishing.

Park, C., & Kang, C. (2008). Does education induce healthy lifestyle?. *Journal of Health Economics*, *27*(6), 1516-1531.

Pekkarinen, T., Uusitalo, R., & Kerr, S. (2009a). *School tracking and development of cognitive skills* (IFAU– Institute for Labour Market Policy Evaluation Working paper No. 2009:6).

Pekkarinen, T., Uusitalo, R., & Kerr, S. (2009b). School tracking and intergenerational income mobility: Evidence from the Finnish comprehensive school reform. *Journal of Public Economics*, *93*(7), 965-973.

Piopiunik, M. (2014). The effects of early tracking on student performance: Evidence from a school reform in Bavaria. *Economics of Education Review*, *42*, 12-33.

Pischke, J. S., & Manning, A. (2006). *Comprehensive versus Selective Schooling in England in Wales: What Do We Know?* (NBER Working Paper No. 12176).

Rosenmöller, P. (2017, January 25). Gelijke kansen in het onderwijs [Letter to Dutch Lower Parliament]. VO-raad.

Schütz, G., Ursprung, H. W., & Wößmann, L. (2008). Education policy and equality of opportunity. *Kyklos*, *61*(2), 279-308.

Staiger, D., & Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica 65*, 557-586.

Sund, K. (2006). *Detracking Swedish Secondary Schools – any Losers any Winners?* (Working Paper No. 2). Stockholm: Swedish Institute for Social Research (SOFI)

Van Elk, R., Van Der Steeg, M., & Webbink, D. (2011). Does the timing of tracking affect higher education completion?. *Economics of Education Review*, *30*(5), 1009-1021.

Waldinger, F. (2007). Does ability tracking exacerbate the role of family background for students' test scores. *Unpublished manuscript, London School of Economics, London, England.*

Woessmann, L. (2009). International evidence on school tracking: A review. *CESifo DICE report – Journal for Institutional Comparisons 7*(1), 26-34.

Woessmann, L. (2010). Institutional determinants of school efficiency and equity: German states as a microcosm for OECD countries. *Jahrbücher für Nationalökonomie und Statistik*, *230*(2), 234-270.

Woessmann, L., Luedemann, E., Schuetz, G., West, M. R. (2009). *School Accountability, Autonomy and Choice around the World*. Cheltenham: Edward Elgar.