# ESSAYS IN APPLIED MICROECONOMICS

## NON-MONETARY INCENTIVES, SKILL FORMATION, AND WORK PREFERENCES

ISBN: 978 90 3610 605 4

Cover design: Crasborn Graphic Designers bno, Valkenburg a.d. Geul

This book is no. **761** of the Tinbergen Institute Research Series, established through cooperation between Rozenberg Publishers and the Tinbergen Institute. A list of books which already appeared in the series can be found in the back.

# Essays in Applied Microeconomics

## Non-Monetary Incentives, Skill Formation, and Work Preferences

# Essays over Toegepaste Micro-economie

## Niet-Monetaire Prikkels, Vorming van Vaardigheden en Werkvoorkeuren

**Thesis**

to obtain the degree of Doctor from the

Erasmus Universiteit Rotterdam

by command of the rector magnificus

Prof.dr. R.C.M.E. Engels

and in accordance with the decision of the Doctorate Board

The public defense shall be held on

Friday 18 September 2020

by

Maria-Alexandra Coţofan

born in Iaşi, Romania

Erasmus University Rotterdam

**Doctoral Committee:**

**Promotor:**
Prof. dr. A.J. Dur

**Other members:**
Prof. dr. A.C. Gielen
Prof. dr.ir. J.C. van Ours
Prof. dr. E.J.S. Plug

**Copromotor:**
Dr. J. Delfgaauw

# Acknowledgements

As I reach the end of my doctoral studies and the beginning of a new and exciting chapter, inevitably a moment of reflection settles in. In this pivotal moment I find myself eager (and admittedly a little nervous) to reflect on the many wonderful experiences and think about the amazing people without whom I would not be here today. In the following lines I would like to take a moment and thank all those who have inspired me, challenged me, and supported me throughout this transformative journey.

I would first like to thank the two people who have been by my side since the beginning of the road and without whom I am convinced I would not be standing here today: Mom and Dad. Thank you for loving me, for all my happy memories, and for working tirelessly to make sure my dreams could come true. For allowing me to grow into myself and for (discreetly) steering me into the right direction at times, I am forever grateful. To Buia and Bunu I am indebted for their unconditional love, for inspiring a passion for science and nature, and for planting within me the seed of critical thought. You will always be missed. To my grandparents who always take life lightly I am grateful for teaching me about kindness and good-humor, and for being by my side every step of the way. I also want to thank Wendy, Ian, Nan, and Uncle Ian for becoming like a second family to me over the past few years, and I look forward to many more exciting times together. I send my love to Alex, Anca, Victor, Delia, and Oana with whom I shared countless childhood stories that are not for the faint-hearted.

It feels like almost a lifetime ago when I embarked on the great adventure of moving to

The Netherlands. Dear Utrecht, there have been ups and downs, but it is safe to say that I quickly fell in love with you and that you became my new home. This would have never happened without good friends on the way. For that I thank Alpay, Ernst, Pim, Ursallah, Peter, and many others with whom I shared all the laughs and joys of student life. Things turned rather grim when I (full of enthusiasm for the future - oh how foolish!) joined the Tinbergen Institute. But through the arduous labors of becoming a 'true Economist', a light at the end of the tunnel appeared. Dearest acquaintances, it has been a relief to share these pains with you over a (guilty) beer at the infamous Blauwe Engel. To Pim, Laura, Magda, Robin, Timo, Benji, Sarah, David, Huaiping (and others that I might later realise I forgot to mention - the horror!), I say not farewell, but see you soon! I'm sure our countless trips, game nights, and Sinterklaas dinners will continue for years to come.

I am extremely happy to have been able to do my PhD at Erasmus University. If I dare to call it successful, that is in big part due to Robert. I am grateful for all the support that you have given me throughout and for your (tireless) effort in helping me find my research path. I couldn't have asked for a better supervisor. I will miss our frequent research talks, but I take solace in the confidence that many more are still to come. I also want to thank Josse who has an endless supply of excellent advice and his door is always open when in need of some. I am thankful to Anne Gielen, Anne Boring, Sacha, Olivier, Bauke, Otto, Jurjen, Benoit, Aart, Arjan, and Albert Jan who always had the time to discuss my work, provide advice, and who have been wonderful colleagues and a great help throughout the job market. I will definitely miss everyone at the Economics Department which has been extremely warm and welcoming. To all of you who provided me with so much guidance and support throughout I am indebted for making this journey a little easier.

I also want to thank Stephan Meier and Lea Cassar for hosting me, for bearing through endless Skype calls, and for always sharing my enthusiasm for research (a publication in Science is coming any day now, I can feel it!). Ron, Trudie, and Max, you have been a

pleasure to work with and I hope we will share other exciting projects in the future. Finally, I want to thank all the committee members who took the time to participate in my defence, and for all their valuable comments which markedly improved my dissertation.

The final words in this section I reserve for the person who has been the most important all along. None of this would have been possible without having you by my side. You have been my best friend, my partner in everything, my fiercest debater, and my biggest support. I look forward to whatever the future brings because everything is an adventure with you. So this book, like everything else, I dedicate to you.

*Maria Coţofan*

*Utrecht, February 2020*

*To Sam, who never stops believing in me.*

# Contents

# Chapter 1

# Introduction

Work is a central pillar in the organization of society and it plays an important part in the life of many individuals. It is for this reason that the well functioning of labor markets has long been a central theme in Economics. While much of the research in this field has focused on understanding how patterns in wages and employment arise and how these patterns impact workers, in recent times it has become increasingly clear that income is not the sole motivating factor for individuals. Life and work satisfaction are determined by a broad range of factors, out of which having a high income is only one aspect. For example, many people have a preference for jobs with a pro-social aspect or for flexible work arrangements, and are willing to earn lower wages in exchange for those features. An increasing number of firms and non-profit organizations take those preferences into account and offer a diverse set of benefits in order to retain the most productive workers, and increase the motivation of their employees.

Economists have traditionally studied the potential of such incentives in motivating workers to exert a higher level of effort. However, the use of incentives almost always involves a

trade-off and findings on their effectiveness are mixed and vary substantially across occupations. Monetary incentives, such as bonuses and promotions do lead to higher effort in some settings, but they are costly and can crowd-out the motivation of certain workers. Increased monitoring can reduce shirking, but it can also make workers feel distrusted or it can signal that the task at hand is unattractive, leading to a decrease in effort on the side of the employees. Gifts and rewards can prompt workers to be reciprocal towards managers, but they can also cause those who don't receive them to be more spiteful. The difficulty in assessing the effectiveness of incentive schemes partly stems from the large variation in the way firms are organized, making it difficult to compare different settings. But workers also have heterogeneous preferences and beliefs about the nature of their job and about their ability to perform it well, such that one incentive scheme might not work equally well for all employees.

However, little is know in the Economics literature about why workers have such diverse preferences for work, how they form, and how they influence performance. To better understand this process, one needs to study individuals, the environments they are placed in, and the decisions that they make, at all stages of their life. Early-life experiences and the environment growing-up can play a formative role in shaping preferences and beliefs about work, and determine the types of jobs individuals will end up performing during adulthood. Similarly, educational decissions in childhood play an important part in developing the cognitive and non-cognitive skills that children will later transfer to the labor market. Understanding this large variation in experiences can help explain why workers self-select into certain industries, why they demand different job attributes, and why they respond differently to incentives. This thesis aims to fill this gap in the literature through a collection of three essays meant to shed light on these important yet little researched questions, using a diverse set of empirical methods.

In Chapter 2 I present the results from a field experiment designed to study the effect of a non-monetary incentive, namely public praise for the best employees, on performance.

Non-monetary rewards such as praise are used widespread to increase the workplace performance of employees. However, the effects of praise on performance have so far been almost exclusively assessed in lab-like settings where workers perform simple and repetitive tasks. Such experiments often fail to capture the complexity of many work settings, where employees face complicated incentive schemes, find it difficult to increase performance when faced with complex tasks, and are not solely motivated by income. The experiment focuses on a group of employees embedded in such a complex work setting, namely school teachers. Despite a growing literature on the effects of teacher incentives on performance, little is known about the efficiency of non-monetary rewards in improving educational outcomes. I address this gap in the literature by measuring how repeated public praise for the best teachers impacts the performance of 900 teachers in 39 schools, over the course of a full academic year.

The students of teachers in the treatment group who receive unannounced public praise perform significantly better in subsequent months; the students of teachers who do not receive unannounced public praise perform significantly worse following the intervention. I investigate whether these changes in performance are the result of teachers in treated schools manipulating the grades they give their students as a response to the intervention. To do so, I analyze how well students perform on high-stake standardized and anonymously graded final exams. The positive effects of unannounced public praise are large and persistent, and reflect real learning gains. The negative effects of unannounced public praise disappear over time, and do not influence the exam performance of students. Repeated rounds of public praise do not impact teacher behavior significantly.

As the experiment makes use of a dynamic treatment design which allows for teachers in treated schools to be both recipients and non-recipients of praise at different points in time, I attempt to disentangle the competing mechanisms that drive teacher behavior in my setting. Results are best explained by a mechanism where praise sends a comparative message about

performance. Updating their beliefs, teachers become more motivated if they receive good news through praise, and become discouraged when the news is bad. However, as teachers become accustomed to the reward they stop responding to repeated interventions.

The results provide a cautionary tale on the use of non-monetary teacher incentives such as praise, showing that one needs to carefully consider the trade-off between boosting the motivation of the best performing teachers at the expense of demotivating the worst performing ones. However, the experiment also indicates that for this group of workers, this trade-off becomes less important in the long-run. This could be because teachers realize that decreasing effort when not praised leads to negative externalities on the educational outcomes of their students. In other words, an initial crowding-out of intrinsic motivation due to not being praised appears to be compensated by teachers exerting higher effort in the long run. This finding has important implications for workers in pro-social jobs and suggests that in the long-run it is easier to motivate than to demotivate such workers through public praise.

The results in Chapter 2 emphasize that teacher motivation plays an important part in the educational attainment of students. While teachers are known to have large and persistent effects on the life outcomes of their students, many other factors can impact childrens' school performance. One other crucial factor is the environment in which children are placed. Such environmental factors typically include a child's family situation, the quality of their school, or the type of peers they are surrounded by. Chapter 3 takes a closer look at how the learning environment in which a child is placed influences their individual achievement.

Being placed in a more challenging learning environment can put students on a different path than similar peers who are placed in a less challenging class, given their ability. This observation motivated a large literature on the effects of assigning students to different academic tracks, based on their ability. Critics of such assignment mechanisms have argued that, especially when tracking is performed at early ages, the ability of students will be measured in a noisy manner. That is because when track assignment is performed early on, those

children in class who are just a few months younger perform significantly worse due to differences in maturity at the time of tracking. It has been well documented in the literature that such differences in maturity at the time of tracking can be wrongly labelled as differences in ability, and that relatively younger students are less frequently assigned to academic tracks. Thus, even small differences in age at the time of track assignment can potentially lead to large miss-allocations if younger students are classed as less able due to simply being less mature. However, little is known about how such miss-allocations impact future educational outcomes and the development of non-cognitive skills.

In chapter 3 we investigate this issue, by estimating the effect of track assignment at the achievement margin on both cognitive and non-cognitive skills, as well as how this effect differs across relative age. Previous studies have commonly found that younger students in class have a lower probability of assignment to higher tracks. One might therefore expect that the effects of attending such tracks is heterogeneous across relative age. We use a regression discontinuity design that exploits school-specific admission thresholds to estimate the effect of top track attendance at the achievement margin, and also identify interactions with relative age. We find no effect on cognitive outcomes, across relative age. However, attending the higher track increases perseverance, need for achievement, and emotional stability for the older students. The results show that placing more mature students in a learning environment that is challenging given their cognitive potential can have positive spillovers on their non-cognitive skills and on the effort they put into learning.

These spillovers appear to mitigate the expected complementarity between ability and academic track attendance, and explain why older students do not perform worse on cognitive tests despite their higher susceptibility to being tracked above their ability level. This suggests that when evaluating educational decisions, both cognitive and non-cognitive skills should be taken into account. Results in chapter 3 also have important implications for the future labor market outcomes of children. A growing literature emphasizes that non-cognitive

skills are particularly important for later-life outcomes. Such skills appear to be especially malleable in early adolescence, but are thought to to stabilize during adulthood. With tracking taking place early in life, those relatively older students just above the track assignment threshold who were 'lucky' enough to be placed in a more challenging learning environment end up being more motivated and more emotionally stable. If such shocks to non-cognitive skills are permanent, the consequences of track assignment can persist well into adulthood.

The final chapter of this thesis studies why people find some aspects of work more important than others. While Chapter 2 reveals that non-monetary rewards can be effective for the employees who receive them, we know little about why workers value such job attributes in the first place. However, being able to explain what determines preferences for different features of work is crucial in understanding the motivation of employees and the organization of firms. Despite their clear importance, virtually nothing is known about how those job preferences are shaped and how they change over time.

In Chapter 4 we propose that such preferences are shaped by the shared experiences that different cohorts had in the past. In particular, we posit that experienced macroeconomic conditions at crucial times in one's life play a very important part in determining what types of job attributes people end up preferring later in life. Building on insights from both Economics and Psychology, we focus on shared experiences of macroeconomic conditions during the 'Impressionable years' (aged 18 to 25), as this period has been shown to be particularly important for the development of preferences, beliefs, and attitudes. Using a representative sample of 20,000 US survey respondents, we investigate how experienced income per-capita during one's impressionable years relates to how important they find having a high income or having a meaningful job at the time of the survey. We construct experiences during the 'Impressionable years' using variation in income per capita across US regions and over time since the 1920s.

We find that job preferences vary in systematic ways with macroeconomic conditions,

with job meaning gaining much more priority in good times and with income being ranked as more important in bad times. Our findings are particularly pronounced for young people, confirming that indeed they are the group most susceptible to being affected by macroeconomic shocks. Most importantly, we show that macroeconomic conditions during the 'Impressionable years' have permanent effects on job preferences. Deep recessions thus create cohorts of workers who give higher priority to income for the rest of their career, whereas booms make cohorts permanently care more about job meaning.

Even though the chapters in this thesis may come across as rather distinct, they actually share a number of similarities. Chapter 2 and Chapter 4 both look at the importance of non-monetary work attributes. On the one hand, Chapter 4 studies a representative sample of the US population in order to explain why certain individuals prefer non-monetary work aspects at the expense of monetary ones. On the other hand, Chapter 2 looks at a group of school teachers and asks how their performance is affected when introducing a non-monetary incentive scheme.

Chapter 3 and Chapter 4 study how experiences when young affect individuals at later stages of their life. Both chapters exploit the fact that individuals are exposed to different environments depending on their 'luck' early in life, to measure the extent to which this variation in experiences shapes their preferences and their skills in adulthood. In Chapter 3 we exploit 'luck' in the form of a discontinuity around school admission thresholds which randomly determines whether some students are placed into an academic track, and we measure the effects of being placed in a more challenging learning environment on the cognitive and non-cognitive skills of the marginal students. In Chapter 4, we exploit regional variation in experienced macro-economic conditions when young, to explain how work preferences are affected by the 'luck' of growing up in relatively good times.

Finally, Chapter 2 and Chapter 3 are similar to the extent that they both study various aspects of education, with each chapter focusing on a different determinant of student per-

formance. Chapter 2 asks whether simple non-monetary incentives such as public praise for the best teachers can be used as a tool to improve the educational outcomes of their students. Chapter 3 looks at how students' educational outcomes are shaped by the learning environment in which they are embedded in at a young age.

# Chapter 2

# Learning from Praise - Evidence from a Field Experiment with Teachers[1]

## 2.1 Introduction

Non-monetary incentives are playing an increasingly important role in many firms (Gallus and Frey, 2016). From best employee awards and verbal recognition to a sense of identifying with the company's mission, managers can use a broad set of tools to increase the performance of workers. Praise, in particular, now features extensively in popular publications and the business literature, as an effective way to motivate employees (see e.g. Nelson (1997)). However, the effect of praise on effort and performance remains largely unknown. A growing body of experimental research provides evidence for a positive effect of praise on performance (Stajkovic and Luthans, 2003; Grant and Gino, 2010; Kosfeld and Necker-

[1]This chapter is based on Coţofan (2019). It reports the results from a field experiment for which the design was pre-registered at https://www.socialscienceregistry.org/trials/2604/history/32360. I am grateful to all those who provided valuable comments and feedback on early drafts, especially to Robert Dur, Josse Deelfgauw, Bauke Visser, Anne Boring, Jan Stoop, and the participants in many conferences and seminar presentations. This research would not have been possible without the collaboration of 'Adservio Social Innovation SRL' who provided the data and the experimental platform. I am particularly indebted to Alexandru Holicov and Marian Andrei for their invaluable contributions.

mann, 2011; Anderson et al., 2013; Ashraf et al., 2014; Lourenço, 2015; Bradler et al., 2016; Gallus, 2016; Gubler et al., 2016; Hoogveld and Zubanov, 2017). However, the existing evidence is largely limited to short-run effects. Moreover, the evidence is silent when it comes to the effects of repeated praise, is speculative about mechanisms driving such effects, and is confined to jobs involving simple and repetitive tasks. In this paper, I contribute to this body of literature by designing a large-scale field experiment to investigate the long-run effects of praise on performance, and the interplay between announced, un-announced, and repeated praise. I study this question in a setting where employees - 900 teachers in 39 Romanian schools - perform cognitively complex tasks.

There is a growing literature on the effects of providing teacher incentives aimed at improving educational outcomes. However, empirical papers have focused almost exclusively on monetary incentives and have found mixed effects on student performance (Leigh, 2012). Studies in developing countries generally find positive effects of teacher incentives on student test scores and teacher attendance (Glewwe et al., 2010; Muralidharan and Sundararaman, 2011). However, Springer et al. (2011) and Fryer (2013) study large-scale and costly interventions in the US, and find no treatment effects. While it is the fact that providing monetary incentives can increase teacher effort and can lead to better student performance, it can also crowd out teacher intrinsic motivation (Firestone and Pennell, 1993), or lead to cheating or teaching to the test (Holmstrom and Milgrom, 1991; Jacob and Levitt, 2003). What's more, if the incentive scheme is too complex and teachers feel as if they have little control, interventions can have no impact on student achievement (Fryer, 2013).

Little is known, however, about how non-monetary incentives such as public praise impact teacher performance. A number of mechanisms have been put forward to explain why individuals respond to praise in the workplace. First, when praise is provided publicly and only to top performers, it sends a signal about the performance norm at work, such that information about relative performance induces higher effort levels from bottom performers and

lower effort levels from top-performers, as both strive to move closer to the apparent performance norm (Bernheim, 1994; Sliwka, 2007; Fischer and Huddart, 2008; Chen et al., 2010; Bradler et al., 2016). Second, when status awards such as praise or job titles are valued and anticipated, they motivate workers to increase effort. Praise activates reputation concerns on the side of the worker, and engages them in a status contest in anticipation of future praise (Moldovanu et al., 2007; Besley and Ghatak, 2008). Third, an agent uninformed of their own ability can get (de)motivated if the principal's actions signal their true ability (Benabou and Tirole, 2003). When effort and ability are complementary, sending a message about relative performance implies a trade-off for the principal between boosting the self-image of some employees, while hurting that of others (Crutzen et al., 2013). This mechanism is also in line with evidence from psychology, on how workers use appraisals as a source of information to gain more accurate self-knowledge (Felson, 1993; Baumeister, 1998).

In this paper, I exploit a dynamic treatment design to shed light on the potential mechanisms driving teacher responses to public praise. I set-up a randomized intervention in which teachers are praised based on the performance gains of their students. The intervention is repeated at regular time intervals for an entire academic year. In a sample of 900 teachers in 39 Romanian schools, I rank teachers based on improvements in the performance of their students. Teachers are ranked within their own discipline, across all schools. The 25% best teachers within each discipline are labeled as top performers and qualify for praise. I exploit the fact that all schools in the sample use an on-line platform environment to have platform managers publicly praise the top performing teachers in a random half of these schools. In the other half, no praise is provided. Within each school, the platform is regularly used by teachers, students, and parents. While the intervention clearly targets teacher performance, in section 2.7 I further discuss how the treatment could interact with parent and student behavior, and show that this is not likely to influence the results. In treated schools, the intervention gives teachers a very coarse partition of their rank, namely whether they are in the

top 25%, or not. In control schools, teachers do not receive any information.

The intervention is repeated twice more in the treated schools, at regular time intervals, throughout the remainder of the academic year. A teacher can be praised repeatedly during the academic year, but teachers can also become top performers for the first time in later rounds. The first intervention is deliberately not announced. During the first intervention it is announced that praise will be repeated, without disclosing an exact date for future rounds. The literature on providing praise distinguishes between an unannounced reward and an announced one. Empirical evidence suggests that announced praise increases the performance of all individuals (Kosfeld and Neckermann, 2011), while unannounced praise has a particularly positive ex-post effect on the performance of non-recipients (Bradler et al., 2016; Hoogveld and Zubanov, 2017). How a combination of the two impacts behavior remains unexplored. The design of my study does not allow for isolating the effect of repeated praise, from that of unannounced praise. However, the experiment I conduct is a significant improvement on the state of the art, because it sheds light on how individuals respond to being repeatedly praised, and it explores how effective the intervention is once they learn to expect it.

The purpose of my experiment is to study the effect of the intervention on student performance gains (based on grades given by the teacher), student attendance, and student performance on anonymously graded standardized exams. My main results are as follows. At the school level, unannounced praise does not have any effect on teacher performance on average. While the average treatment effect at the school level is not statistically significant, there are sizable heterogeneous treatment effects for recipients and non-recipients of praise. Non-praised teachers in the treatment group decrease performance, while praised teachers increase it. The performance of a non-praised teacher in the treatment group decreases by 0.30 standard deviations as compared to similar teachers in the control group. On the other hand, the performance of a praised teacher increases by 0.23 standard deviations as compared

to similar teachers in the control group. The effects are large and economically significant.

The treatment response does not vary substantially with the distance from the 25% "top-performer" threshold, confirming that indeed teachers do not know their rank. The results are best explained by a mechanism where praise sends teachers a signal about their performance. As such, updating their beliefs, teachers become more motivated if they receive good news through praise, and become demotivated if they receive bad news through not being praised, in line with the theoretical prediction in Benabou and Tirole (2003) and Crutzen et al. (2013). Repeated interventions do not seem to have any effect on teacher performance. This is true both for teachers who were praised in the past and those who are praised for the first time, suggesting that when teachers learn to expect the reward praise becomes less effective.

Some critics of providing rewards based on performance argue that once incentives or monitoring are conditioned on a performance measure, the said measure ceases to be effective, also known as "Goodhart's law" (Goodhart, 1984). For instance, problems arise when the performance measure can be manipulated by employees. Since teachers grade their own students, praise based on the performance gains of their students can incentivise gaming on the side of the teachers. This concern can be addressed when an objective performance measure is available. I use results on high-stake anonymously graded standardized exams, undertaken by final year students. Based on these exam grades, I test whether teachers respond to praise by increasing performance, or if they simply manipulate the performance measure by grading more leniently.

The results indicate that the subjective performance measure does not become a poorer predictor of standardized exam performance in the treated schools. Moreover, I find that positive changes in the performance of students are driven by real learning gains. Praising teachers in the first round raises the grades of their students by 0.17 standard deviations on the anonymously marked exams, undertaken six months after the intervention. The persistence and magnitude of the effect is remarkable, given that final exams cover a broad range of

topics. On the other hand, students whose teachers were not praised in the first round do not perform significantly worse on standardized exams as compared to similar students in the control group. Hence, the positive effect that is also observed in the subjective performance measure survives, while the negative effect on subjectively assessed performance disappears over time.

The remainder of this paper is organized as follows. Section 2.2 provides an overview of relevant theoretical mechanisms and formulates predictions. Section 2.3 introduces the setting, Section 2.4 describes the experimental design, Section 2.5 presents the results on unannounced public praise, and Section 2.6 describes the results on announced and repeated public praise. Finally, Section 2.7 addresses parent and student behavior and Section 2.8 discusses broader implications and concludes.

## 2.2 Theoretical background

Individuals are said to value praise. But why is praise desirable, and what are the underlying mechanisms that drive behavioral responses to praise? Do these channels predict different outcomes for recipients and non-recipients of public praise, and does it matter whether employees expect such an incentive?

There is little to no evidence on the long-run effects of praise,[2] and the existing theoretical mechanisms are limited in predicting behavioral responses to repeated interventions. In this section I review a number of mechanisms that can drive teacher behavior, and discuss the extent to which some of their features apply to my setting, while some dimensions are likely

[2]Somewhat related, a number of experimental papers have looked at the effects of public feedback on performance (Blader et al. (2016), Bandiera et al. (2013), Delfgaauw et al. (2013)), with mixed findings. While feedback and praise are closely related, the former is focused on conveying pure relative performance information, while the latter also sends a clear signal of appreciation to workers. Ashraf et al. (2014) disentangle the effects of recognition and feedback, and find that they have opposing effects on performance. While social recognition increases performance, both public and private disclosure of rank information reduce performance.

to differ. Specifically, I discuss (i) status contests, (ii) conformity to the norm and (iii) changes in motivation due to learning about performance.

Generally, these mechanisms are integrated in a principal-agent framework. However, for the purpose of this paper, I will focus on the choices of agents. While providing public praise might not always be an optimal strategy for the principal, such considerations are beyond the scope of this paper and I will abstract from them in the remainder of this section.

**Status contests**

A number of studies have shown that agents care about their reputation. This can be driven by the desire to signal a high ability due to career concerns (Bénabou and Tirole, 2006; Swank and Visser, 2006) or because agents wish to be respected by their peers (Grant and Gino, 2010). Praise can send a signal about the quality of a worker. In particular, when praise is given publicly and only to top performers, an element of social comparison is introduced. In my setting, praising top-performers in a way that is visible to colleagues, parents, and students sends a strong signal about the quality of a teacher.

Besley and Ghatak (2008) have postulated that status awards, such as a better job title or calling someone the "employee of the month", are incentive compatible and they increase effort on the side of the agent while reducing the optimal level of monetary incentives. Moldovanu et al. (2007) predict that agents will seek status awards, as they lead to a higher status within the group. Given the common expectation that praise will be repeated in the future, all agents should increase effort, following the introduction of the reward. Else, in a one-off unannounced intervention, changes in effort due to status concerns should be zero for all agents. To accommodate this mechanism in my setting, during the first intervention all teachers are told that praise will be provided again in the future.

As such, the first hypothesis is:

*H1: If teachers' behavior is driven by status concerns, teachers in a treated school will*

*increase performance after the first intervention, independent of whether they were praised or not.*

**Conformity to the norm**

Bernheim (1994) argues that when social status is important to individuals, they will conform to social norms. That is because social groups can penalize individuals who deviate from accepted norms, a penalty reflected in a loss of social reputation (Akerlof, 1980). When individuals fear that departures from the social norm will diminish their position within the group, they will conform to a homogeneous standard of behavior, in spite of having heterogeneous underlying preferences.

The provision of public praise sends a signal, to both recipients and non-recipients, about the performance norm in the workplace. Individuals who have a preference for conformity will want to adjust their effort such that they are in line with the performance norm (Sliwka, 2007; Fischer and Huddart, 2008; Chen et al., 2010; Bradler et al., 2016). In this case, praise will have opposite effects on the performance of recipients and non-recipients. Those that receive praise learn that they belong to the top 25% of workers, while those who do not receive it learn that they belong to the bottom 75%. If teachers are conformists and like to behave like their peers, then in treated schools top performers should decrease performance, and bottom performers should increase it, so as to get closer to the apparent work norm.

The second hypothesis is:

*H2: If teachers' behavior is driven by conformity to the norm, teachers in a treated school will decrease performance if they were praised, and will increase performance if they were not praised.*

**Effect of learning about performance on motivation**

Extrinsic and intrinsic motivation have been extensively discussed in the context of incen-

tives, be they monetary or symbolic. Economists generally argue that incentives are a useful tool in promoting effort and performance, and a significant number of empirical papers support this claim (Gibbons, 1998; Lazear, 2000). However, more recent literature focused on the potential negative spill-overs of such incentives on motivation. The seminal papers of Fehr and Falk (1999), Gneezy and Rustichini (2000b), and Gneezy and Rustichini (2000a) are some of the early examples to report such "hidden costs" of rewards.

Crowding out of intrinsic motivation, in the terminology of Frey (1997), can be a potential mechanism through which rewards reduce the effort of employees. Some lab and field evidence confirms that financial rewards crowd out the intrinsic motivation of agents (Deci, 1971; Kohn, 1999; Fehr and Falk, 1999; Gneezy and Rustichini, 2000b,a). Motivation crowding out appears particularly relevant for public sector employees, such as teachers: a number of studies in the public administration literature and in economics show that public servants tend to be more intrinsically motivated (Buelens and Van den Broeck, 2007; Crewson, 1997; Dohmen and Falk, 2010; Buurman et al., 2012). Georgellis et al. (2010) show that public workers' motivation can be crowded out by incentives, while Bellé (2015) finds that financial incentives can crowd out the image motivation of workers in jobs with pro-social impact.

To understand why and how rewards impact the motivation of employees, Benabou and Tirole (2003) use the concept of "looking-glass self", as coined by Cooley (1902). This mechanism postulates that in a principal-agent setting where the principal uses some form of reward or incentive, the agents learn about their own ability through the reward. In other words, such a reward impacts the agent directly through their payoff, and indirectly through their inference process. Benabou and Tirole (2003) argue that empowerment, encouragement, and praise are examples of confidence-enhancement strategies on the side of the principal.

Crutzen et al. (2013) expand on the "looking-glass self" in a setting where ability and

effort are complementary in the production function, such that effort levels increase in a workers' beliefs about their ability. Since agents do not know their ability, they gain self-knowledge and update their beliefs contingent on a comparison signal sent by the employer. The crucial trade-off that the principal faces is between boosting the self-confidence of the best workers, while harming that of the relatively worse ones (Crutzen et al., 2013; Kamphorst and Swank, 2016). As such, in this setting, public praise not only sends a message to the best-performing teachers, but also to those teachers who are not being praised. In other words, if teachers in treated schools learn about their performance through the intervention, then praise sends "good news" about their ability, and not being praised sends "bad news" about their ability.

The third hypothesis is:

*H3: If teachers' behavior is driven by learning about their performance, a teacher in a treated school will become more motivated and increase performance when praised, and will become demotivated and decrease performance when not praised.*

**Predictions**

In line with the three mechanisms discussed, the possible treatment effects can be summarized in Table 2.1. However, it is also possible that several mechanisms play a role at the same time.

Table 2.1: Treatment effects of public praise according to hypotheses H1, H2, and H3

| **Recipient** | **Non-Recipient** | **Mechanism** |
|---|---|---|
| **+** | **+** | Effect of status incentives (H1) |
| **-** | **+** | Effect of conformity to the norm (H2) |
| **+** | **-** | Effect of learning about performance on motivation (H3) |

First, a positive treatment effect for bottom performers in treated schools can be driven by

both status incentives and conformity to the norm. If status incentives drive teacher behavior, Moldovanu et al. (2007) predict that all teachers should increase effort proportional to their ability. If the worst performing teachers are also the least able ones, then bottom ranked teachers will increase performance less than those whose performance falls just below the threshold. On the other hand, if a teacher is a conformist, she will increase performance in line with her beliefs about how far below the thresholds her performance falls. As such, a teacher should increase performance more if she believes that she is at the bottom of the distribution than if she believes that she is ranked just below the performance threshold. However, as teachers do not know their rank, on average all teachers who are not praised should increase performance similarly if their response is driven by conformity.

Second, a positive treatment effect for praised teachers can be explained by both H1 and H3. Repeating the intervention can shed more light on this issue, by looking at those who are praised multiple times. If in the first period teachers learn about relative performance and become more motivated to exert higher effort, repeated praise provides less information. Such a teacher has already learned about their relative performance, and being a top performer again should result in more modest updating.[3]

The repeated provision of praise over a long period of time is an important innovation of this paper. Not only is this, to the best of my knowledge, the first experiment to explore how persistent the effects of praise are over longer periods of time, but it also investigates whether the intervention loses bite once agents get used to the award system. Rogers and Frey (2016) argue that individuals may become desensitized to repeated exposure to a given stimuli. However, in certain instances, repeated interventions can have an effect on behavior. This is the case if the proprieties of the stimulus are dynamic, or if it is presented at unpredictable

[3] A similar prediction can be derived if the utility from praise is concave in the frequency of praise. Looking at the response of those who are praised only once as compared to the response of teachers who are praised repeatedly can provide additional evidence on whether this is a likely mechanism. The findings in section 2.6 show that this mechanism is not likely to drive the results.

intervals. In my setting, repeated praise is announced, but the exact date of the intervention is not known to teachers. Furthermore, different teachers can be praised in repeated rounds, such that subsequent messages still contain a substantial amount of new information. While integrating the short and long run responses to repeated praise in a theoretical framework is beyond the scope of this paper, teacher responses to repeated praise can shed more light on the underlying mechanisms, and can provide useful guidelines for future theoretical and experimental work.

## 2.3 Setting

The experiment targets roughly 900 teachers in 39 Romanian schools, who in total have about 20,000 students aged 11 to 18. In Romania, the school year starts in September and continues until the end of June. The education system runs through three 4-year pre-university education cycles: primary school (aged 7-10), secondary school (aged 11-14), and high school (aged 15-18). This experiment will focus on teachers from secondary schools and high schools.

Romania has a centralized education system, and all schools follow the academic curriculum designed by the Ministry of Education. The curriculum provides a detailed guideline of the teaching material. Furthermore, schools use comparable textbooks which are typically approved by the Ministry of Education, ensuring that teachers use the same materials and proceed with the curriculum in a similar order. As such, schools are homogeneous with respect to the type of information that students learn, and the competencies and skills they are expected to acquire throughout the school year. My experiment focuses on teachers who teach one of the following nine academic subjects: Romanian language, English language, Mathematics, Physics, Chemistry, Biology, History, Geography, and Computer Science.

At the end of the second and the third cycle (at age 14 and age 18), students are required

to undertake standardized national-level examinations in order to graduate from the current cycle, and continue to the next. These standardized exams are high stake, as they help determine high-school and university admission. Undertaken in strictly invigilated exam centres, students work under the supervision of exam inspectors, and class teachers are not present. Exams are graded through a double-blind procedure, by randomly assigned teachers from a different school. As such, class teachers cannot influence their student's performance on these tests by either designing the test, helping students during the examination, or by deciding the grade.

Teachers' wage is independent of their students' performance. The performance of students does not impact teachers' probabilities of promotion either. Teachers typically are subjected to standardized examinations and procedures to earn the right to be hired (*examen de titularizare* for becoming a teacher) or promoted (*gradul didactic I/II*), which do not take student grades into account. As a consequence of that, there is no career incentive for teachers to artificially inflate the grades of students, since they cannot get fired and will not be promoted based on this performance measure. This unique setting allows for cleanly identifying the effect of non-monetary incentives, as teachers cannot leverage praise to gain future monetary benefits.

The format of the academic curriculum is such that each academic year covers new material. For example, while 5th grade students study plant biology, 6th grade students study animal biology, etc. The consequence of this design is that the first grade that the students receive at the beginning of the academic year should reflect the baseline ability of a student, and should be by and large independent from previous learning. As such, I use the first grade that students receive in the beginning of the new academic year, as a proxy for the baseline ability of the student, and in section 2.4 I provide additional evidence that this appears to be a reliable measure.

## 2.4 Experimental design

This experiment follows 39 schools, located in 15 different regions in Romania.[4] All the schools in this experiment are making use of an on-line education platform which tracks student progress. Schools can decide for themselves whether they want to implement the system, and the usage of the platform comes at a small monthly cost.

The platform allows parents to see their childrens' performance and attendance in real time and makes it easier to keep track of their school progress which is regularly updated by teachers. By working directly with the platform providers and not with individual schools I can ensure that schools, teachers, and parents are not aware of being part of an experiment, and avoid any selection effects into the sample. My experiment thus qualifies as a natural field experiment, following the terminology in Harrison and List (2004). Access to the anonymized data allows me to monitor the performance of all students and teachers in the school for an entire academic year.

Schools are randomly assigned to either treatment or control. Teachers at schools which are assigned to the treatment group will receive "public praise". More precisely, the "best performing teachers" will be publicly praised through a message posted on-line through the management platform. The best performing teachers are those who score among the top 25% across all schools, within their own subject. The first intervention is unannounced. The first intervention announces that public praise will be given again in the future. However, the exact date and frequency of future interventions is not disclosed. Subsequent rounds of praise take place at regular time-intervals until the end of the academic year. Appendix A.1

[4]The 39 schools in this experiment perform better than the national average, with average exam grades of 8.48 at the end of secondary school and 8.18 at the end of high school (on a scale from 1 to 10). According to the most recent statistics from the Ministry of Education in Romania (https://www.edu.ro/rapoarte-publice-periodice), the average grade for final exams at the country level are 7.44 at the end of secondary school, and 7.83 at the end of high school. However, students from schools in rural areas typically perform worse on the final exams, bringing the national average down. As a result, my sample is roughly representative of schools in the urban area. Since schools are randomly assigned to either treatment or control, quality differences between sampled schools and average Romanian schools are not a threat to the internal validity of this experiment.

discusses the experimental time-line in detail.

**Determining Top Performing Teachers**

Building on the literature on teacher productivity, top performers will be determined on the basis of the performance gains of their students. There is by now a fairly large literature on calculating such improvements in student performance due to teacher impact, by looking at teacher value added (Hanushek, 1971; Chetty et al., 2014). While it is common in the literature to extract teacher value added from the teacher fixed effect in a regression explaining test score changes (Chetty et al., 2014), this approach is not chosen here for two reasons. First, such an approach might be too hard to explain to the teachers. Fryer (2013) finds that a large scale monetary incentive scheme in New York public schools had no effect on student achievement, despite the intervention totalling a cost of $75 million. He argues that the most likely explanation for the zero treatment effects was the fact that the scheme was too complex, and provided teachers with too little control. Similarly, teachers might find it difficult to increase performance, when the ranking mechanism is too complex to understand.

Second, calculating teacher value added in the standard way requires that test scores are comparable in terms of content and level, and assumes that students would score similarly across years, in the absence of a teacher effect. In my setting, teachers have some freedom in designing and grading the tests on which the students' performance gains are calculated, so changes in student performance might not only capture a teacher effect, but also variation in test difficulty over time. To accommodate these two considerations, I calculate performance gains (henceforth PG) using an alternative method to the standard procedure in the literature. While this measure is not directly comparable to the standard teacher value added, in section 2.5 I investigate how noisy my measure of performance gains is, and show that it is accurate in predicting student performance on standardized exams, indicating that it is a good measure of student learning.

In my experiment, the school year is divided into four periods. Teacher performance is evaluated for each one of these time periods, namely before each of the three rounds of public praise, and once after the third and final round. Teachers are ranked according to an average of all the individual performance gains of their students, $pg_i$. Each student's performance gain $pg_i$ for a period is given by the difference between their baseline performance for the period (denoted by $\theta_{ib,t}$) and their subsequent performance that period $\theta_{i,t}$, where $t \in \{1,2,3,4\}$:

$$pg_{i,t} = \theta_{i,t} - \theta_{ib,t}.$$

$\theta_{i,t}$ is a weighted average of all the subsequent grades of a student within each period, where the final grade is given a weight of 50% and for all other intermediate grades, the remaining weight is equally distributed.[5] The final grade is given a higher weight because it measures the longest period of time to pass since the baseline performance grade is recorded.[6]

For the repeated rounds (t=2,3,4), the performance gain is calculated in a similar manner, where the new baseline performance is replaced by the performance in the previous period, such that:[7]

$$\theta_{ib,t+1} = \theta_{i,t} = pg_{i,t} + \theta_{ib,t}$$

The first grade at the beginning of the school year is used as the baseline performance for the first period, and represents a proxy for student ability. I argue that this is a reliable proxy,

[5]To better illustrate this, take a simple example where at the end of the first period, a student has four grades namely $g_1$, $g_2$, $g_3$ and $g_4$, in this exact order. Then the initial performance gain $pg_{i,t=1}$ for each student will be calculated using the following formula: $pg_{i,t=1} = \theta_{i,t=1} - \theta_{ib,t=1} = \frac{\frac{g_2+g_3}{2}+g_4}{2} - g_1$.

[6]This weighting method was agreed jointly with educational experts managing the online platform, and it has been pre-registered in the experimental design.

[7]When $\theta_{i,t}$ is missing, $\theta_{i,t-1}$ will be used, and so on. If no previous average exists, $\theta_{ib,t=1}$ is used.

as prior to the intervention teachers have no incentive to manipulate grades (teachers are not monitored or rewarded based on student grades). A potential threat to this approach is that a teacher could influence the performance of her students across academic years. For example, a good teacher could put her students on a higher learning path than an average teacher. In that case, students who had good teachers in the past, could have a higher starting baseline performance in the beginning of the new academic year. As a result, the PG of such teachers would be mechanically lower, implying that the best teachers (according to this definition) might not be the ones who are publicly praised.

To investigate whether this is an issue, I make use of a subset of 20 schools (7,742 students and 380 teachers) where data for the previous academic year is also available. Table A.2.1 in Appendix A.2 shows the relationship between current pre-intervention PG, the measure on which teachers are rewarded, and last year's PG. In other words, I test whether a teacher with a high PG in the past is less likely to have a high PG in the current academic year. Even after controlling for a number of student, teacher, and school characteristics, the relationship between performance gains in the previous year and performance gains in the pre-intervention period in the current academic year is weak. Furthermore, the coefficient of interest is positive and rather small (a one standard deviation increase in the previous year's PG translates into a 0.08 standard deviations larger pre-intervention PG in the current year). This suggests that, according to this measure, well performing teachers in the previous year are not less likely to be labelled as top performers in the current year.

In yet another robustness check, I exploit the fact that some students have just joined the school in the beginning of a new cycle, and they have not had the same teacher in the past.[8] Table A.2.2 in Appendix A.2 shows the relationship between baseline performance and performance gains (column 1) and how this relationship differs by whether a student

[8]These are students who just started secondary school, or students who just started high-school in schools that don't offer a secondary education program.

had the same teacher in the past or not (column 2). As expected, there is a negative relationship between the baseline performance and the subsequent performance gains: if the baseline performance of a student increases by one standard deviation, the pre-intervention student performance gains decrease by 0.63 standard deviations. This is a mechanical effect, such that if students have a very high starting level, they will naturally have less room for improvement. However, new students do not appear to learn more than recurring ones, nor does the relationship between baseline performance and learning differ across new and recurring students.[9] This indicates that having the same teacher for two years in a row does not impact learning differently than having a teacher for the first time.

Table 2.2 presents the average PG per academic subject, across all schools. On average PG are always positive within a subject, varying between 0.09 and 0.25 points. This variation underlines the importance of selecting top performing teachers within their own subject. Specifically, teachers of different subjects most likely require different skills and use different teaching methods, making the comparison between say a math teacher and a history teacher less relevant than between two math teachers or two history teachers.

Teachers' PG each period is defined as an average of all the individual performance gains of their students in that period. A teacher is a top performer if, based on their students' performance gains, they are ranked in the top 25% best performing teachers, within their own subject. Top performing teachers at schools assigned to the treatment group are publicly praised. There are no treated schools, at any point throughout the experiment, in which no teacher is publicly praised. The share of top performers within each school is fairly comparable across schools with a standard deviation of 12.9%.[10]

[9]These results are also robust to estimating the coefficients separately for top performers and for bottom performers.

[10]The perceived scarcity of public praise could also influence how teachers respond to the intervention. Controlling for the share of school-level top-performing teachers within each subject does not change the results either qualitatively or quantitatively, confirming that this variation does not drive the results.

Table 2.2: Average performance gains per academic subject, in the beginning of the school year

| Subject | Mean | Standard Deviation | No. teachers |
|---|---|---|---|
| Biology | 0.253 | 0.621 | 63 |
| Chemistry | 0.148 | 0.943 | 48 |
| Computer Science | 0.116 | 0.712 | 60 |
| English Language | 0.092 | 0.615 | 146 |
| Geography | 0.249 | 0.609 | 65 |
| History | 0.139 | 0.650 | 65 |
| Mathematics | 0.119 | 0.651 | 151 |
| Physics | 0.247 | 0.920 | 84 |
| Romanian Language | 0.120 | 0.614 | 173 |

*Notes: Columns show the mean and the standard deviation of PG across all subjects, prior to the intervention. PG is expressed in points, and can in principle take any value between -9 and 9.*

**Intervention**

After a period of collecting data on teacher and student baseline performance, the first intervention takes place on January 22nd 2018, following the Christmas break. The messages are unannounced and unanticipated. In the schools which were assigned to the treatment group, a message (for the full intervention text, see Appendix A.3) is posted on the front page of the platform, which is visible to all those with a user account (teachers, parents, and students) immediately as they log-in.

The message is addressed to teachers and it states that the platform is interested in how student performance has improved since the beginning of the school year, as it is one of the ways to measure academic progress. The message announces that for a number of academic subjects platform managers have assessed the improvement in student grades across all the schools that implement the electronic platform. Based on this assessment, teachers are informed that a number of teachers in their schools are among the top 25% performers within their subjects, across all the schools using the platform. The top performing teachers are listed by name, and thanked for their effort and contribution. Finally, the announcement mentions that such messages will be sent again in the future, to show the platform's gratitude

towards teachers' hard work.

The message is highly visible, and seen by all teachers who log-in to post grades or record attendance. To further ensure that all teachers read the message carefully, an additional private message is sent to their personal inbox. The e-mail informs them again about the intervention and provides them with a link to the original post.

The same procedure is repeated twice more throughout the remainder of the academic year, in March and May respectively. Following each intervention, teacher performance is measured on roughly equal intervals of two months.

**Data and Randomization**

Data spans 39 schools from 15 Romanian regions.[11] Data collection records the performance of all the students in the school, across the 9 academic subjects of interest. In total, there are 855 teachers[12] in the sample, and 19,748 students. Since each student takes on average about 7 of the 9 academic subjects,[13] there are in total 130,316 data entries.

Randomization is performed at the level of the treated unit, namely the school, and stratified across three important dimensions:[14]

(i) Student baseline performance : A school-level weighted average of the initial grade that students receive at the beginning of the school year across all subjects, and a proxy for the average student ability in the school.

[11] Some of the schools that use the platform had only recently purchased the rights and were still largely inactive at the time. I drop the schools in which less than 20% of teachers use the platform. In the remaining 39 schools in the sample, 87% of teachers regularly use the platform.

[12] Some teachers never record any grades in the mentioned period, more precisely 13% of the sample. This indicates some selection with respect to the "type of teacher" that uses the on-line platform. However, this is not a threat to the validity of the experiment: These teachers are similarly distributed between treated and control schools (p-value$=0.455$). For the 87% of teachers who use the platform, remaining active did not differ by the treatment status after the intervention, as can be seen in Appendix A.4.

[13] Some subjects are only introduced in later years, and some students only choose, for example, a subset of science subjects.

[14] Together, these three stratification variables capture the main sources of heterogeneity across the 39 schools which are otherwise very similar in the curriculum they use and their administrative structure.

(ii) Teacher baseline performance: A school-level weighted average of the pre-intervention (since the beginning of the school year) teacher PG, and a proxy for the average teacher quality in the school.

(iii) School size: The number of teachers in the school (who actively use the platform and teach academic subjects).

Due to the limited number of schools, stratification variables are re-coded as binary indicators, as opposed to continuous measures. For example, if the student baseline performance in a school is above the sample average, the binary indicator takes value one, and zero otherwise. Using the three binary indicators, eight strata are constructed. Within each strata, I randomly assign the 39 schools to either treatment or control. Due to a strata with just one school and by splitting the ties in favor of the treatment group, the randomization process assigned 21 schools (55% of teachers in the sample) to the treatment group and 18 schools (45% of teachers in the sample) to the control group.

Table 2.3 shows that the randomization process was successful. When comparing schools in the treatment group with schools in the control group, there appear to be no significant differences in terms of either the stratification variables[15] or a number of additional important controls.

Before the first intervention, roughly 70% of the students for whom a baseline performance measure exists have at least one additional grade. As such, for these students, the PG can be calculated. At the teacher level, this is calculated based on a weighted average of all the individual performance gains of their students.

[15]To capture potentially fine grained differences, the continuous stratification variables are used in Table 2.3, as opposed to the binary indicators.

Table 2.3: Balance tests for mean differences between treatment and control

| Variable | C | T | P-value |
|---|---|---|---|
| Student baseline performance | 7.741 | 7.873 | 0.681 |
| | (0.230) | (0.219) | |
| Teacher baseline performance | 0.142 | 0.187 | 0.472 |
| | (0.043) | (0.045) | |
| School size (no. teachers) | 21.611 | 22.238 | 0.890 |
| | (3.240) | (3.130) | |
| % Urban schools | 0.833 | 0.810 | 0.851 |
| | (0.090) | (0.088) | |
| % Publicly funded | 0.833 | 0.762 | 0.594 |
| | (0.090) | (0.095) | |
| % Female students | 0.524 | 0.542 | 0.540 |
| | (0.027) | (0.023) | |
| % Male teachers | 0.254 | 0.257 | 0.279 |
| | (0.060) | (0.049) | |
| No. skipped classes | 0.745 | 0.650 | 0.585 |
| | (0.139) | (0.105) | |
| N | 18 | 21 | |
| **Multivariate t-test statistics** | | | |
| F-value | 0.233 | | |
| P-value | 0.982 | | |

*Notes: The first two columns show variable means between the control group of schools, and the treated group of schools. In brackets, standard deviations are presented. The third column shows the p-values from two-sample t-tests on the null hypothesis that group means are equal. Significance levels: *** p<.01, ** p<.05, * p<.1.*

Since some students might not have their performance assessed between interventions, the composition of students who determine teacher PG can differ over time. However, the average teacher has 230 students across the multiple classes that they teach, and teachers' PG is calculated based on a substantial share of their students. On average, for each teacher, their PG is calculated based on 112 (54%) students pre-intervention, 160 (72%) students after the first round, 125 (55%) students after the second round, and 135 (61%) students in the last round.[16] There is no evidence that teachers in treated schools start recording more grades post-intervention.[17] As such, performance gains are determined by a large number of

[16]Not all grading takes place through a class-level written examination. Students within one class can be graded at different times, for example based on class participation.

[17]Calculated by looking at the difference in the number of recorded new grades per student after the inter-

students for each teacher.

From the 855 active teachers in the pre-intervention sample for whom PG is calculated, for 821 (96%) of them PG is also calculated in the second round of intervention, for 758 (89%) of them PG is also calculated in the third one, and for 729 (85%) of them PG is also calculated in the last round. This attrition is not due to teachers leaving the school, but because none of their students are graded between interventions. Appendix A.4 shows that this attrition does not depend on being assigned to the treatment, on whether a teacher was a top performer or not, nor on the interaction between the two. In total, 56% of teachers qualify for public praise at least once throughout the experiment.

## 2.5 Results of unannounced public praise

The effects of public praise on teacher performance are estimated by looking at three outcome variables: (i) PG calculated using class grades given by the teacher, (ii) student attendance, and (iii) standardized exam performance of their students. Data on PG and attendance are collected prior to the intervention, and following each of the three interventions. Standardized exams take place at the end of the school year, for a subset of students ending an academic cycle, aged 14 and 18.

**Student performance gains**

To assess the effects of unannounced praise on student performance gains and attendance, I estimate the following two period teacher fixed effects model:

$$Perf_{i,t+1} = \alpha_1 T_{i,t} + \alpha_2 Top_{i,t} + \alpha_3 T_{i,t} * Top_{i,t} + \mu_i + \tau_t + \varepsilon_{i,t} \tag{2.1}$$

vention. The p-value for the coefficient that regresses the number of new grades after the first round on the treatment dummy is 0.686.

where $Perf_{i,t+1}$ is teacher performance two months after the intervention, measured by either PG or attendance. $T_{i,t}$ is a treatment dummy, indicating whether a teacher was exposed to the treatment or not, such that the treatment dummy takes value 0 for all schools prior to the intervention, and values 0 or 1 after the first intervention, depending on whether the school was assigned to the treatment or the control group. $Top_{i,t}$ is an indicator for being a top performer, namely if a teacher qualifies for being praised at time t+1, by being ranked in the top 25% at time t. $T_{i,t} * Top_{i,t}$ is the interaction between being a top performer and being in a treated school, which takes value 1 for teachers who are publicly praised. $\mu_i$ is a teacher-specific fixed effect which captures all time-invariant teacher characteristics, and $\tau_t$ is a time fixed effect. The analysis is performed at the teacher level, and the standard errors are clustered at the school level.

Two months after the first intervention, PG is calculated again for 96% of the active teachers in the pre-intervention sample, having a mean of 0.33 points and a standard deviation of 0.62 points (where PG can in principle take any value between -9 and 9, but in sample it ranges between -5 and 5). Table 2.4 estimates equation (2.1), using PG as an outcome variable.[18] Appendix A.5 presents the results for attendance.

Column 1 shows that at the school level there is no statistically significant treatment effect of the intervention, although the point-estimate of the average treatment effect is negative. Coefficient $\alpha_2$ in column 2 reveals that in the control group, teacher performance is in line with mean reversion. If students experience a steep learning curve in the first period, PG will consequently be lower next period, as there is less room for improvement. Reversely, when PG is low in the pre-intervention period, student grades will subsequently increase, as

[18]The standard errors are clustered at the school level. While the number of clusters is larger than the minimally required number of 30, I perform additional robustness checks to exclude the possibility that the standard errors are biased by the fact that there are only 39 schools in the sample. Following Cameron et al. (2008), I implement the wild bootstrap procedure, designed to produce reliable standard errors even when the number of clusters is small. The bootstrapped p-values on the coefficients do not change the significance of the results in Table 2.4, indicating that the number of schools is not a concern for the reliability of the estimated standard errors.

improving is comparatively easy.[19]

In column 2, treatment effects are separated by whether a teacher was publicly praised or not. In treated schools, those teachers who do not qualify for public praise, but do observe their colleagues being praised decrease performance by 0.30 standard deviations ($\alpha_1$). On the other hand, those teachers who are top performers and are praised increase performance by 0.23 standard deviations ($\alpha_1 + \alpha_3$), as opposed to top-performing teachers in the control group.[20]

Table 2.4: The effect of unannounced public praise on PG

| | **New PG** | **New PG** |
|---|---|---|
| ($\alpha_1$) Treatment | -0.150 | -0.303** |
| | (0.115) | (0.121) |
| ($\alpha_2$) Top performer | | -1.980*** |
| | | (0.147) |
| ($\alpha_3$) Treatment * Top performer | | 0.528** |
| | | (0.233) |
| Teacher Fixed Effects | yes | yes |
| Time Fixed Effects | yes | yes |
| N | 821 | 821 |
| F-value | 11.75 | 169.61 |

*Notes: The dependent variable is the PG calculated two months after the first intervention, expressed in standard deviations. In parentheses, heteroskedasticity robust standard errors are presented, clustered at the school level. Significance levels: *** p<.01, ** p<.05, * p<.1.*

The behavior of top-performing teachers in treated schools is in line with hypothesis *H3*, as being publicly praised sends teachers a positive message about their performance. Updating their beliefs in response to this signal, praised teachers improve their performance even further. However, a positive treatment effect for praised teachers is also in line with status incentives (*H1*), or a combination between status incentives and learning about relative performance. I further explore this possibility in the end of this section and in section 2.6.

[19] Another competing explanation for the observed mean reversion is that PG is in fact a noisy measure of performance. In Table 2.5 I show that PG has substantial explanatory power for final exam performance.

[20] The p-value on the difference between $\alpha_1$ and $\alpha_1 + \alpha_3$ is $p - value = 0.195$

The behavior of the bottom 75% of teachers is also in line with hypothesis *H3*, as both *H1* and *H2* predict that following the first intervention, teachers ranked below the threshold in treated schools will increase performance. However, since the bottom category is comparatively large, the crowding out of motivation might not be the same for all the teachers who fall below the threshold in treated schools. For example, teachers who know that their students have made little progress since the beginning of the year could become strongly demotivated if they perceive it to be very difficult to catch up and become top performers in the future. On the other hand, the motivation of teachers who know that their students have improved might be crowded out to a lesser extent, or they could even increase performance, in the hope of being praised in the future.

To investigate this, I rank the bottom 75% of teachers in three different categories: those in the 1st quantile who qualified for praise (the top 25%), those in the 2nd quantile (between 25% and 50% of the performance distribution), those in the 3rd quantile (between 50% and 75% of the performance distribution), and those in the 4th quantile (the bottom 25% of the distribution) as compared to similar teachers in the control group. The results in Appendix A.6 show that all bottom performing teachers decrease performance following the intervention. Furthermore, teachers in the three bottom quantiles do not respond significantly different to the treatment, indicated by the associated p-values. This confirms that teachers do not have additional information on their relative performance, and only respond to the given binary partition of their rank. In other words, all teachers in treated schools appear to exclusively use the signal that the intervention sends to update their beliefs about their relative performance. In line with the mechanism proposed by Benabou and Tirole (2003) and Crutzen et al. (2013), all bottom-performers seem to become demotivated and perform worse.

**Standardized exams**

To assess the effect of unannounced praise on an objective performance measure which teachers cannot manipulate, I make use of grades on the standardized exams that final year students undertake at the end of the school year. I use a linear regression model which controls for a rich set of student, teacher, and school characteristics. The effect of unannounced praise on exam performance is cleanly identified, because the first round of the intervention is unanticipated and there is no selection into either treatment or into being a top-performer. I estimate the following equation:

$$Grade_{i,j} = \beta_1 T_{i,j} + \beta_2 Top_j + \beta_3 T_{i,j} * Top_j + \beta_4 X_{i,j} + \nu_{i,j} \tag{2.2}$$

where $Grade_{i,j}$ is the final exam performance of student i, who has teacher j for that specific exam subject.[21] $T_{i,j}$ is the treatment dummy and $Top_j$ is an indicator for whether that student's teacher was a top performer prior to the first intervention. The interaction term $T_{i,j} * Top_j$ captures the effect of having a teacher who was praised in the first round, and $X_{i,j}$ is a vector of controls at the student-level (gender, year of study, baseline student performance on the subject, track), teacher-level (subject), and school-level (region, in a rural area, publicly funded, size, baseline student and teacher performance at the school level, and past exam performance of the school). The standard errors $\nu_{i,j}$ are clustered at the school level.

At the end of the academic year, all year 8 (aged 14) and year 12 (aged 18) students undertake nationally standardized anonymously graded exams. The Ministry of Education makes the exam results publicly available in July. In total, 3,423 students are matched to the their exam grade, equivalent to 75% of the total number of final year students in the sample.[22] The students belong to 335 teachers from the original sample who teach final year

[21] Students take two exams at the age of 14, in Mathematics and Romanian Language. At the age of 18, students take three exams, in Romanian language, a compulsory track-specific subject, and a track-specific subject of their choice.

[22] The matching success rate is contingent on the name of the student being spelled identically in both the

classes.

To asses whether the results can be generalized to all students, I test whether final year students systematically differ from the rest of the sample in terms of any observed characteristics. The results in Appendix A.7 shows that final year students do not appear to be different from other students, neither are they more likely to be over-sampled from the treated group. As such, despite only having exam data for final year students, the results are likely to generalize to the whole sample.

I begin by investigating whether pre-intervention PG is a good measure of student learning, as measured by standardized exam performance. Table 2.5 shows the relationship between the pre-intervention performance gains of each student and their exam performance, controlling for a rich set of student-level, teacher-level, and school-level variables.

Table 2.5: Relationship between pre-intervention PG and exam performance

| | **Exam grade** |
|---|---|
| Pre-intervention PG | 0.317*** |
| | (0.030) |
| Student controls | yes |
| Teacher controls | yes |
| School controls | yes |
| N | 5,308 |
| F-value | 483.92 |
| R-squared | 0.557 |

*Notes: The dependent variable is the student's exam performance, expressed in standard deviations. The pre-intervention PG is also expressed in standard deviations. OLS regression controls for baseline student performance, student gender, subject fixed effects, class profile fixed effects, degree of urbanization, being a publicly-funded school, school size, baseline student and teacher quality at the school level, past year exam performance at the school level, and region fixed effects. In parentheses, heteroskedasticity robust standard errors are presented, clustered at the school level. Significance levels: *** p<.01, ** p<.05, * p<.1.*

A one standard deviation increase in pre-intervention performance gains translates into a

platform database, and the database containing the exam grades, provided by the Ministry of Education. For unmatched students, some form of name miss-match occurred. Miss-matches are caused by middle names being omitted in one database, parent initials missing, or varying diacritics. The fully random matching errors do not impact the composition of the sample.

0.32 standard deviations increase in performance on standardized exams. The coefficient is both statistically and economically significant, and suggests that PG is a good predictor of student learning: even if the pre-intervention performance gain is only based on a subset of the curriculum, it explains a substantial share of the variation in final exams which cover a broad range of topics.

While exam grades are an objective measure of performance, PG could be manipulated by teachers, after the intervention. An extensive cross-disciplinary literature argues that once a measure of performance is used to reward or to monitor, it becomes less effective over time. This has frequently been referred to as Goodhart's law (Goodhart, 1984) and has been linked to the idea of gaming on the side of the agent. In other words, since in treated schools public praise is conditioned on PG, the measure can become less informative about actual performance over time because teachers have the freedom to grade their own students. To address the concern that PG can become a noisy performance measure after the introduction of the reward, I look at the relationship between PG and exam grades, over time. More precisely, I perform a difference in difference analysis on the correlation between the objective performance measure (the standardized exam grade) and the subjective incentivized measure of performance (PG), by estimating:

$$Corr_{i,t} = \lambda_0 + \lambda_1 Post + \lambda_2 T + \lambda_3 Post * T + \lambda_4 X_{i,t} + \varepsilon_{i,t}$$

where the correlation is calculated at the teacher level, across all of their students who undertake the standardized exam. *Post* is an indicator for the post-intervention period, T is the treatment dummy, and $X_{i,t}$ is a vector of controls.

The time trend coefficient in Table 2.6 indicates that PG becomes somewhat less predictive of exam performance over the course of the school year, in all schools. There are no baseline differences between the treated and the control group prior to the intervention,

nor does PG become less predictive of exam performance in the treated group. These results clearly indicate that teachers do not attempt to manipulate the incentivized performance measure, once exposed to the treatment.[23]

Table 2.6: DiD analysis on the correlation between PG and standardized exam performance

| | **Corr(PG, exam)** |
|---|---|
| Post-intervention period | -0.068** |
| | (0.030) |
| Treatment | -0.006 |
| | (0.051) |
| Post-intervention period * Treatment | 0.014 |
| | (0.055) |
| Student controls | yes |
| Teacher controls | yes |
| School controls | yes |
| N | 497 |
| F-value | 17.60 |
| R-squared | 0.080 |

*Notes: The dependent variable is the correlation coefficient, at the teacher level, between the student's exam performance (expressed in standard deviations) and the student's pre-intervention performance gains (expressed in standard deviations). OLS regression controls for average baseline performance of a teacher's students, student gender composition, subject fixed effects, degree of urbanization, being a publicly-funded school, school size, baseline student and teacher quality at the school level, past year exam performance at the school level, and region fixed effects. In parentheses, heteroskedasticity robust standard errors are presented, clustered at the school level. Significance levels: *** p<.01, ** p<.05, * p<.1.*

For 305 teachers who are observed throughout the experiment (and 3,017 of their students), I assess the effect of public praise on their students' exam performance,[24] by estimating equation (2.2). The results in Table 2.7 show that the intervention has no statistically significant effect at the school level. The coefficients in the second column are consistent

[23]This finding is also robust to estimating the equation separately for the sample of top performers, and the sample of bottom performers. The results are available on request.

[24]The analysis is performed at the student level. Due to either matching errors, or due to the fact that not all subjects have mandatory exams (for example Science students can choose between Biology, Chemistry, Physics and Computer Science), a small share of teachers have a very small number of students who undertake final exams. As a result of that, their response to the intervention is measured imprecisely, such that giving them equal weight in a teacher-level regressions introduces considerable noise. Teacher level regressions show that all the results are qualitatively the same if one excludes teachers with very few students, or restricts the analysis to subjects which are mandatory.

with the effects of unannounced public praise on PG. Those students whose teacher was praised in the first round score 0.17 standard deviations ($\beta_1 + \beta_3$) higher on the final exam in the end of the school year, as compared to their counterparts in the control group whose teacher was a top performer in the first round.[25] By comparison, top performers in the control group increase the performance of their students on the final exam by about 0.05 standard deviations ($\beta_2$).[26]

Table 2.7: The effect of unannounced public praise on standardized exam performance

| | **Grade** | **Grade** |
|---|---|---|
| ($\beta_1$) Treatment | -0.055 | -0.089 |
| | (0.051) | (0.056) |
| ($\beta_2$) Top performer | | 0.054 |
| | | (0.052) |
| ($\beta_3$) Treatment * Top performer | | 0.258*** |
| | | (0.094) |
| Student Controls | yes | yes |
| Teacher controls | yes | yes |
| School Controls | yes | yes |
| N | 6,639 | 6,639 |
| F-value | 349.27 | 214.38 |
| R-squared | 0.486 | 0.492 |

*Notes: The dependent variable is the student's exam performance, expressed in standard deviations. OLS regressions control for baseline student performance, student gender, subject fixed effects, profile type fixed effects, degree of urbanization, being a publicly-funded school, school size, baseline student and teacher quality at the school level, past year exam performance at the school level, and region fixed effects. In parentheses, heteroskedasticity robust standard errors are presented, clustered at the school level. Significance levels: *** p<.01, ** p<.05, * p<.1.*

In a meta-analysis of studies on interventions in education, Sanders et al. (2015) find that effect sizes are usually no larger than 0.17 standard deviations. The magnitude of the effects on student achievement following a simple non-monetary incentive scheme for teachers are remarkable. By comparison, Duflo and Hanna (2005) find an increase of 0.17 standard deviations in student achievement when teacher's pay is conditioned on their attendance.

[25] The p-value on the difference between $\beta_1$ and $\beta_1 + \beta_3$ is $p - value = 0.051$

[26] While a high pre-intervention PG is a good predictor of better exam performance, the effect of being a top-performer on exam grades in the control group is a combination of high pre-intervention PG, and strong mean-reversion in the following period, explaining why the $\beta_2$ coefficient is comparatively small.

Muralidharan and Sundararaman (2011) find that incentive pay increased student achievement by 0.12 standard deviations in language and 0.16 standard deviations in math, during the first year.

This indicates that the effects of unannounced praise have persistent effects on student performance, six months after the intervention. The increase in PG due to public praise translates into better performance on final exams, confirming that teachers do not cheat when grading their own students. Rather, the results are in line with real learning on the side of the students, as a result of better teacher performance. Interestingly, the negative effect of not being praised on PG (0.30 standard deviations in Table 2.4) disappears over time and does not affect final exam performance (0.09 standard deviations in Table 2.7) in a significant manner.

Previously in this section I have argued that teacher behavior is best explained by hypothesis *H3*. Re-visiting these arguments, the results in Table 2.7 are also in line with this mechanism. While both status contests and conformity to the norm are incentive compatible with teachers "cheating" when grading their own students, hypothesis *H3* which works through the intrinsic motivation of teachers is not. As the negative effect of not being praised in the first round is phased-out over time, it suggests that in settings where intrinsic motivation is important, in the long run it is easier to motivate than to demotivate employees through public praise.

## 2.6 Results of announced and repeated public praise

### Student performance gains

It is not easy to cleanly identify the effects of repeated interventions on teacher performance. In treated schools, teachers who are praised in the second or in the third round have

already been exposed to and affected by unannounced praise. To nevertheless shed some light on the effects of repeated interventions, I undertake an exploratory analysis. I begin by estimating the per-period effect of the combination of unannounced and announced public praise on PG and attendance, using the following equation:

$$Perf_{i,t+1} = \sum_{t=1}^{t} \gamma_{1,t} * T_{i,t} + \sum_{i,t=1}^{t} \gamma_{2,t} * Top_{i,t} + \sum_{i,t=1}^{t} \gamma_{3,t} * T_{i,t} * Top_{i,t} + \mu_i + \tau_t + \omega_{i,t} \quad (2.3)$$

where $t = 1$ is the pre-intervention period, and at $t = \{2,3,4\}$ the three intervention rounds take place. As such, at $t = 2$ top performers in treated schools receive unannounced praise. At $t = \{3,4\}$ top performers in treated schools receive announced praise, and a subset of them receive repeated praise. Standard errors are clustered at the school level.

To explore the effects of repeated public praise, I estimate (i) the effect of not being praised in a given round, (ii) the effect of being praised for the first time in any given round, and (iii) the effect of being praised repeatedly in any given round, as compared to similar teachers in the control group:

$$Perf_{i,t+1} = \delta_1 T_{i,t} + \sum_{i,j=0}^{2} \delta_{2,j} Type_{i,j,t} + \sum_{i,j=0}^{2} \delta_{3,j} T_{i,t} * Type_{i,j,t} + \mu_i + \tau_t + \psi_i \quad (2.4)$$

where $Type_{i,j,t}$ is a categorical variable which records the type $j$ of a teacher $i$ within each period $t$. Specifically, $Type_{i,j,t}$ takes value 0 if teacher $i$ is not a top performer at time $t$, value 1 if teacher $i$ is a top performer for the first time at time $t$, and value 2 if teacher $i$ is a top performer for the second or third time at time $t$.

Announced (and for some teachers repeated) praise is given two times throughout the remainder of the school year, namely two months and four months after unannounced public

praise. After the second round PG is calculated again for 89% of the active teachers in the original sample,[27] and after the final round PG is calculated for 85% of the active teachers in the original sample.[28] As described in Appendix A.4, this attrition is random and does not relate to being in the treated group, to being a top performer, or to the interaction between the two.

The remainder of this section presents the effects of announced and repeated public praise on PG. Appendix A.8 presents the corresponding results for attendance. There are no differences between treatment and control in the number of new grades (per student) that teachers record, confirming that there is no gaming on the side of the teachers at the extensive margin.[29] To shed more light on the way learning is distributed throughout the academic year, Table 2.8 shows the average PG across treated and control schools, throughout the experiment.

Table 2.8: PG throughout the intervention

| | **Treatment** | **Control** | **N** |
|---|---|---|---|
| Pre-treatment | 0.152 | 0.147 | 855 |
| Post unannounced praise | 0.282 | 0.387 | 821 |
| Post announced praise 1 | -0.078 | -0.032 | 758 |
| Post announced praise 2 | 0.473 | 0.519 | 729 |

*Notes: Columns show the average PG in the treatment and the control group, throughout the intervention. PG is expressed in points, and can in principle take any value between -9 and 9.*

Table 2.9 provides an overview of the treatment effects across all periods, by estimating equation (2.3). Announced praise does not seem to have any significant effects on either top or bottom performers, a finding consistent in both repeated interventions. This is in line with the idea that once rewards are anticipated, they tend to lose their effectiveness in moving

[27] Based on roughly 144 students for each teacher, equivalent to 61% of all of their students on average.

[28] Based on roughly 103 students for each teacher, equivalent to 44% of all of their students on average.

[29] Calculated by looking at the difference in the number of recorded new grades per number of students that a teacher has, after each round. The p-value for the coefficient that regresses the number of new grades on the treatment dummy is 0.125 after the second round and 0.947 after the third one.

performance.

Table 2.9: The effects of unannounced and announced public praise on PG

| | **New PG** |
|---|---|
| ($\gamma_{1,1}$) Treatment Round 1 | -0.232** |
| | (0.113) |
| ($\gamma_{1,2}$) Treatment Round 2 | -0.043 |
| | (0.112) |
| ($\gamma_{1,3}$) Treatment Round 3 | -0.069 |
| | (0.123) |
| ($\gamma_{2,1}$) Top performer Round 1 | -0.883*** |
| | (0.098) |
| ($\gamma_{2,2}$) Top performer Round 2 | -0.760*** |
| | (0.133) |
| ($\gamma_{2,3}$) Top performer Round 3 | -0.621** |
| | (0.236) |
| ($\gamma_{3,1}$) Treatment * Top performer Round 1 | 0.291** |
| | (0.143) |
| ($\gamma_{3,2}$) Treatment * Top performer Round 2 | -0.327 |
| | (0.195) |
| ($\gamma_{3,3}$) Treatment * Top performer Round 3 | 0.127 |
| | (0.269) |
| Teacher Fixed Effects | yes |
| Time Fixed Effects | yes |
| N | 821 |
| F-value | 56.47 |

*Notes: The dependent variable is the PG calculated two months after the previous intervention, expressed in standard deviations. In parentheses, heteroskedasticity robust standard errors are presented, clustered at the school level. Significance levels: *** p<.01, ** p<.05, * p<.1.*

To shed light on the effects of repeated praise on teacher performance, I estimate equation (2.4) which compares the performance of teachers who were not praised in any given round, with the performance of teachers who were praised for the first time, and the performance of teachers who were praised for a repeated time within that round. Table 2.10 presents the results. Praising teachers for the first time (in any round) has no significant effect on their PG in the following period. Since public praise in the first round has a large and significant effect on the performance of both top and bottom performing teachers, the small and insignificant

coefficients $\delta_1$ and $\delta_{3,1}$ suggests that the positive effects of public praise observed in Table 2.4 disappear when teachers anticipate the intervention.

Table 2.10: The effects of repeated public praise on PG

| | **New PG** | **New PG** |
|---|---|---|
| ($\delta_1$) Treatment | -0.079 | -0.115 |
| | (0.086) | (0.089) |
| ($\delta_{2,1}$) Top performer first time | | -0.658*** |
| | | (0.081) |
| ($\delta_{2,2}$) Top performer repeated time | | -1.240*** |
| | | (0.291) |
| ($\delta_{3,1}$) Treatment * Top performer first time | | 0.037 |
| | | (0.109) |
| ($\delta_{3,2}$) Treatment * Top performer repeated time | | -0.229 |
| | | (0.353) |
| Teacher Fixed Effects | yes | yes |
| Time Fixed Effects | yes | yes |
| N | 821 | 821 |
| F-value | 55.72 | 43.32 |

*Notes: The dependent variable is the PG calculated two months after the previous intervention, expressed in standard deviations. In parentheses, heteroskedasticity robust standard errors are presented, clustered at the school level. Significance levels: *** p<.01, ** p<.05, * p<.1.*

Being praised repeatedly has no significant effect on teacher performance, and the point-estimate is negative. This is in line with hypothesis *H3*, as teachers who were already praised once in the past update their beliefs to a much lesser extent when receiving an additional positive signal. An alternative explanation for coefficient $\delta_{3,2}$ is that the utility from praise is concave in the frequency of praise. However, concavity implies that those teachers who are praised only once should increase performance in all rounds. The much smaller size of coefficient $\delta_{3,1}$ in Table 2.10, as compared to coefficient $\alpha_3$ in Table 2.4, suggests that this is not a plausible mechanism.

Equation (2.4) imposes the restrictive assumption that in treated schools the response of any teacher $i$ at time $t$ is independent of their experiences in previous rounds. However, repeating the intervention in treated schools and exposing the same group of teachers to

multiple treatments gives rise to increasingly complex combinations of effects with each additional round. To relax this assumption, in Table 2.11 I estimate a flexible specification controlling for each type of experience that a teacher could have had throughout the year, such that at each point in time the previous performance of a teacher is taken into account. The reference category is made up of teachers who were never praised, up to that period.

While the effects of unannounced praise remain sizable and precisely estimated, repeated interventions do not appear to significantly affect teacher performance. However, some interesting patterns arise. First, teachers who are praised for the first time in subsequent rounds ($\psi_4$ and $\psi_9$) do not improve performance. Second, those teachers who were only praised in the first round and became more motivated as a result of that, appear to exert additional effort to maintain a high performance throughout ($\psi_3$ and $\psi_6$). This is particularly visible following the final intervention, with a marginally significant increase in PG, as compared to similar teachers in the control group. Third, there are no benefits from praising a teacher in two consecutive interventions ($\psi_5$, $\psi_{11}$ and $\psi_{12}$).

These results suggest a number of additional takeaways. When rewards are given repeatedly, being first to the prize seems to matter more. Teachers who are praised in the first round increase performance ($\psi_2$) and appear to remain more intrinsically motivated throughout the remainder of the experiment ($\psi_3$ and $\psi_6$). On the other hand, being second or third to the prize does not translate into better performance. Finally, repeated rewards over short periods of time do not achieve the desired results, as coefficients on being praised two or three periods consecutively are always negative ($\psi_5$, $\psi_{11}$ and $\psi_{12}$). However, being praised in the first round and in the third one returns a large and positive coefficient ($\psi_{10}$) following the final intervention. This indicates that while praise looses bite as it becomes less scarce, it remains a powerful tool for those who receive it sparingly.

Table 2.11: Treatment effects throughout all periods

| | **PG 1 round** | **PG 2 rounds** | **PG 3 rounds** |
|---|---|---|---|
| ($\psi_1$) **Treatment** | -0.303** | -0.249** | -0.265** |
| | (0.121) | (0.099) | (0.104) |
| **Treatment * Type** | | | |
| ($\psi_2$)T*Top1 | 0.528** | 0.473** | 0.490** |
| | (0.233) | (0.215) | (0.221) |
| ($\psi_3$)T*(NTop2 & Top1) | | 0.298 | 0.314 |
| | | (0.279) | (0.278) |
| ($\psi_4$)T*(Top2 & NTop1) | | -0.124 | -0.116 |
| | | (0.524) | (0.254) |
| ($\psi_5$)T*(Top2 & Top1) | | -0.068 | -0.051 |
| | | (0.445) | (0.449) |
| ($\psi_6$)T*(NTop3 & NTop2 & Top1) | | | 0.525* |
| | | | (0.301) |
| ($\psi_7$)T*(NTop3 & Top2 & NTop1) | | | 0.179 |
| | | | (0.261) |
| ($\psi_8$)T*(NTop3 & Top2 & Top1) | | | 0.401 |
| | | | (0.483) |
| ($\psi_9$)T*(Top3 & NTop2 & NTop1) | | | 0.106 |
| | | | (0.196) |
| ($\psi_{10}$)T*(Top3 & NTop2 & Top1) | | | 0.606 |
| | | | (0.536) |
| ($\psi_{11}$)T*(Top3 & Top2 & NTop1) | | | -0.625 |
| | | | (0.490) |
| ($\psi_{12}$)T*(Top3 & Top2 & Top1) | | | -0.455 |
| | | | (1.028) |
| Type Fixed Effects | yes | yes | yes |
| Teacher Fixed Effects | yes | yes | yes |
| Time Fixed Effects | yes | yes | yes |
| N | 821 | 821 | 821 |
| F-value | 169.61 | 88.53 | 122.81 |

*Notes: The dependent variable is the PG calculated two months after the previous intervention, expressed in standard deviations. In parentheses, heteroskedasticity robust standard errors are presented, clustered at the school level. Significance levels: *** p<.01, ** p<.05, * p<.1.*

**Standardized exams**

To explore the relationship between repeated public praise and final exam performance, I estimate:

$$Grade_{i,j} = \phi_1 T_{i,j} + \psi_2 Freq_j + \psi_3 T_{i,j} * Freq_j + \psi_4 X_{i,j} + \nu_{i,j} \tag{2.5}$$

where $Grade_{i,j}$ is the final exam performance of student $i$, under teacher $j$. $Freq_j$ is a categorical dummy which indicates whether, and how many times, a teacher was a top performer throughout the academic year. In other words, $Freq_j$ takes value 0 if a teacher was always a bottom performer, value 1 if a teacher was a top performer only once, and value 2 if a teacher was a top performer repeatedly throughout the year. The vector of controls $X_{i,j}$ is defined as in equation (2.2), and standard errors $\nu_{i,j}$ are clustered at the school level.

In Table 2.12 I undertake an exploratory analysis on the relationship between repeated public praise and final exam performance. Specifically, I estimate equation (2.5) which classifies teachers as always bottom performers (135 teachers and 2,192 students), top performers only once throughout the experiment (138 teachers and 2,281 students) and top performers more than once (45 teachers and 1,027 students).

Dis-aggregating the results by the frequency with which a teacher was a top performer reveals some heterogeneity. The students of those teachers who were never praised do not perform any different than their peers in the control group, with a point estimate of zero. The students of teachers who were only praised once throughout the academic year do not perform significantly different either. On the other hand, the students of repeatedly praised teachers perform 0.33 standard deviations better on final exams, as compared to their counterparts in the control group.[30] This effect appears to be predominantly driven by teachers who were praised in the first and in the third round (roughly 60% of the teachers who were

[30]The p-value on the difference between $\psi_1$ and $\psi_1 + \psi_{3,2}$ is $p - value = 0.000$

praised repeatedly). In line with coefficient $\psi_{10}$ in Table 2.11, these findings suggest that in the long-run, repeated public praise can be an effective tool if given sparingly. However, those teachers who are top performers multiple times in the treated group are a select type and have predominantly also been praised in the first round. Thus, the results should be interpreted accordingly and with caution.

Table 2.12: The effect of repeated public praise on standardized exam performance

| | **Grade** | **Grade** |
|---|---|---|
| ($\phi_1$) Treatment | -0.055 | 0.006 |
| | (0.051) | (0.079) |
| ($\phi_{2,1}$) Top performer only once | | 0.132 |
| | | (0.132) |
| ($\phi_{2,2}$) Top performer more than once | | 0.038 |
| | | (0.052) |
| ($\phi_{3,1}$) Treatment * Top performer only once | | -0.176 |
| | | (0.144) |
| ($\phi_{3,2}$) Treatment * Top performer more than once | | 0.338*** |
| | | (0.082) |
| Student Controls | yes | yes |
| School Controls | yes | yes |
| N | 6,639 | 6,639 |
| F-value | 349.27 | 587.47 |
| R-squared | 0.486 | 0.493 |

*Notes: The dependent variable is student exam performance, expressed in standard deviations. OLS regressions control for baseline student performance, student gender, subject fixed effects, class profile fixed effects, degree of urbanization, being a publicly-funded school, school size, baseline student and teacher quality at the school level, past year exam performance at the school level, and region fixed effects. In parentheses, heteroskedasticity robust standard errors are presented, clustered at the school level. Significance levels: *** p<.01, ** p<.05, * p<.1.*

## 2.7 Teacher and parent response

Publicly praising top performers in this experiment implies that parents and students also learn about the quality of a teacher. Thus, it is possible that they adjust their behavior as a response to the intervention. Parents learning about the quality of teachers could partly

drive the treatment effect if the amount of time spent home with children changes due to the message (Pop-Eleches and Urquiola, 2013). In other words, parents whose children's teacher is revealed to be a bottom performer could increase the amount of time they spend helping them with homework, and reduce effort if they find out that the teacher is particularly good. Student behavior could also be affected by the intervention through both a direct channel, and indirectly through teacher effort. In other words, the student's optimal choice of effort could also be affected by the signal their teacher receives.

However, if praising a teacher sends a clear positive signal about their performance, not praising a teacher sends a much more ambiguous message to teachers, students, and parents; and while the intervention provides an explicit incentive for teachers, for students and parents it is at most an information treatment. For example, while teachers internalize the reward and stop responding over time, there is less reason for students to stop responding when their teacher is praised for the first time, in a later round.

To nevertheless address these concerns, I perform an additional robustness check to investigate whether parents and students adjust their behavior as a response to the intervention. Specifically, I focus on one decision that students or parents can make as a response to the treatment, by looking at the subset of students finishing high school and undertaking the exit standardized exam at the end of the school year (*examenul de bacalaureat*). Depending on the class track, these students undertake two compulsory exams on track-specific subjects, and additionally select one subject of their choice for their third exam.[31] Students must officially choose the elective discipline and register for the exam in the end of May, roughly one month before the exam and five months after the first intervention.

Using the Ministry of Education track-specific guidelines, I restrict the analysis to exam grades for elective subjects. I then ask whether a student is more likely to choose to undertake

[31] For example, students in the science track can choose one discipline between Biology, Chemistry, Physics and Computer science, but exams in Mathematics and Romanian Language are compulsory.

an exam on a subject where the class teacher received unannounced praise, or less likely to do so if they learn that their teacher is a bottom performer. This would indicate that students or parents directly respond to the intervention, by adjusting educational decisions based on the signal sent by public praise.

Results in Appendix A.9 show that students are not more likely to select an exam subject once they learn that the teacher is particularly good, neither are they more likely to avoid a subject when they learn that the teacher is a bottom performer. This appears to indicate that parents and students do not adjust educational decisions after learning about the performance of teachers. While the results in Appendix A.9 only look at a subset of the total sample (137 teachers teaching elective subjects in the final year to 1,990 students), they provide some evidence that indeed the conclusions of this paper are driven by teacher behavior.

## 2.8 Conclusion

This paper has shown that introducing public praise as an incentive can have large and persistent effects on the performance of teachers. Unlike experiments studying the short-run effects of praise in simple jobs, this is the first study to assess how persistent the effects of public praise are and to measure the effects of praise on employees who perform cognitively complex tasks. By analyzing the interplay between unannounced and announced praise, this experiment exploits a dynamic treatment design to shed light on the different theoretical mechanisms that drive teacher behavior in this setting.

Unannounced praise has large effects on teacher performance, consistent with a theoretical mechanism where teachers do not know their performance, but learn about it through public praise. In treated schools, being praised sends a positive signal to top performing teachers, while not being praised sends a negative signal to bottom-performing teachers. Receiving good news boosts confidence and increases teacher performance, while bad news

demotivates and reduces teacher performance. The findings are in line with the comparative cheap talk literature, where principals face a trade-off between boosting the confidence of the best performing agents, while harming that of the worst performing ones. A public message where praise is scarce appears to be credible, as teachers seem to update their beliefs accordingly.

Conformity to the norm does not seem to have a bite in this setting. Since teachers are compared to both peers from within their own school and peers form different schools, the pressure to conform might be reduced. Top performing teachers can more easily increase performance without fearing social punishment, while those at the bottom feel less pressure as the message does not reveal how poor their performance is within their own school. Future experimental work could vary the salience of workplace social norms to further explore this channel.

The behavior of bottom performing teachers does not seem to be driven by status contests either. This could be partly explained by the fact that teachers are intrinsically motivated public servants, such that the utility from winning a status contest is likely outweighed by crowding out of intrinsic motivation when not rewarded. The fact that teachers do not respond to praise in repeated rounds, even when experiencing it for the first time, is in line with the idea that once rewards for a certain performance measure are internalized, expected, and perceived as less scarce, they tend to lose bite. This suggests that the marginal benefits to continued treatment are small even when the composition of top-performing teachers changes over time.

The findings of this paper do not support the idea that teachers attempt to manipulate performance gains. This is to be expected if teacher behavior is driven by an intrinsic channel (*H3*), as opposed to extrinsic motivators (*H1* and *H2*). Unannounced praise leads to real learning gains and better performance on anonymously graded standardized exams for the students of top performing teachers. By the end of the school year, the negative effect of

unannounced praise due to motivation crowding out does not reflect the exam performance of students whose teachers were not praised. The fact that the negative responses are short lived suggests that when employees are sufficiently intrinsically motivated, in the long-run it is easier to motivate, than to demotivate through public praise.

An extensive literature has discussed the complexity of rewarding teachers: incentive based subjective assessments can lead to gaming, or teaching to the test. Monetary incentives in such settings are costly and often inefficient (Goodman and Turner, 2013; Eberts et al., 2000). Reasons usually invoked are the complexity of the incentive scheme, or the fact that incentives are too small (Fryer, 2013). In particular, small monetary incentives can backfire in pro-social jobs as the motivation of workers is crowded out. A simple system of non-monetary rewards could be a cheap and efficient alternative, as the positive effect of unannounced praise is persistent, while the negative effect fades away as teachers who are demotivated at first appear to overcompensate in the long-run.

The fact that repeated public praise is ineffective suggests that managers should not provide the same non-monetary reward too often- but rather be creative in varying the incentive, as well as credibly commit to a public and comparative message. To limit the negative effect on bottom performers, praise should have a positive focus, and implicit shaming should be avoided by keeping the bottom category broad. Finally, much more can be said about the different settings in which these incentives are effective, contingent on the type of task and the preferences of the worker. Focus has recently shifted on the reduced importance of income, as many modern workers are increasingly in search of non-monetary work aspects such as meaningful work, flexibility, and appreciation (Cassar and Meier, 2018a). As the nature of work changes, better understanding such rewards and re-thinking optimal job design becomes increasingly important.

# Appendix A

## A.1 Data Collection and Time Line

This experiment follows roughly 900 teachers in 39 Romanian schools over the course of an entire academic year, from September 2017 to August 2018. All the 39 schools in the experiment use an online management platform designed to monitor student performance. Between the 1st of September 2017 and the 21st of January 2018, data on the baseline performance of teachers is collected, and used to compute student performance gains.

Based on their performance, teachers are ranked across all schools, within their own subject. Those teachers who are in the top 25% best performers within their subject are labelled as 'Top performers' and qualify for receiving public praise. The 39 schools are then randomly assigned between a 'Treatment' group and a 'Control' group, stratified on the baseline performance of students (first grade in the beginning of school year) and teachers (pre-intervention performance gains), and on the size of the school. The treatment assignment of schools remains unchanged throughout the entire academic year. There is no selection into the treatment, and no schools opt out of receiving the messages following the first intervention.

The first intervention takes place on the 22nd of January 2018, following the end of the Christmas break. The platform managers post the messages on the platform page of each of the treated schools. The messages posted in each school are identical in terms of content. The only source of variation in the messages is the names of the top performing teachers within each school. The message is only visible to teachers, parents, and students within that school. The message is posted on the main page of the platform, visible immediately after logging-in. An additional email is sent by the platform managers to all teachers in the school, reminding them to read the public message and providing them with a link to the

original post.

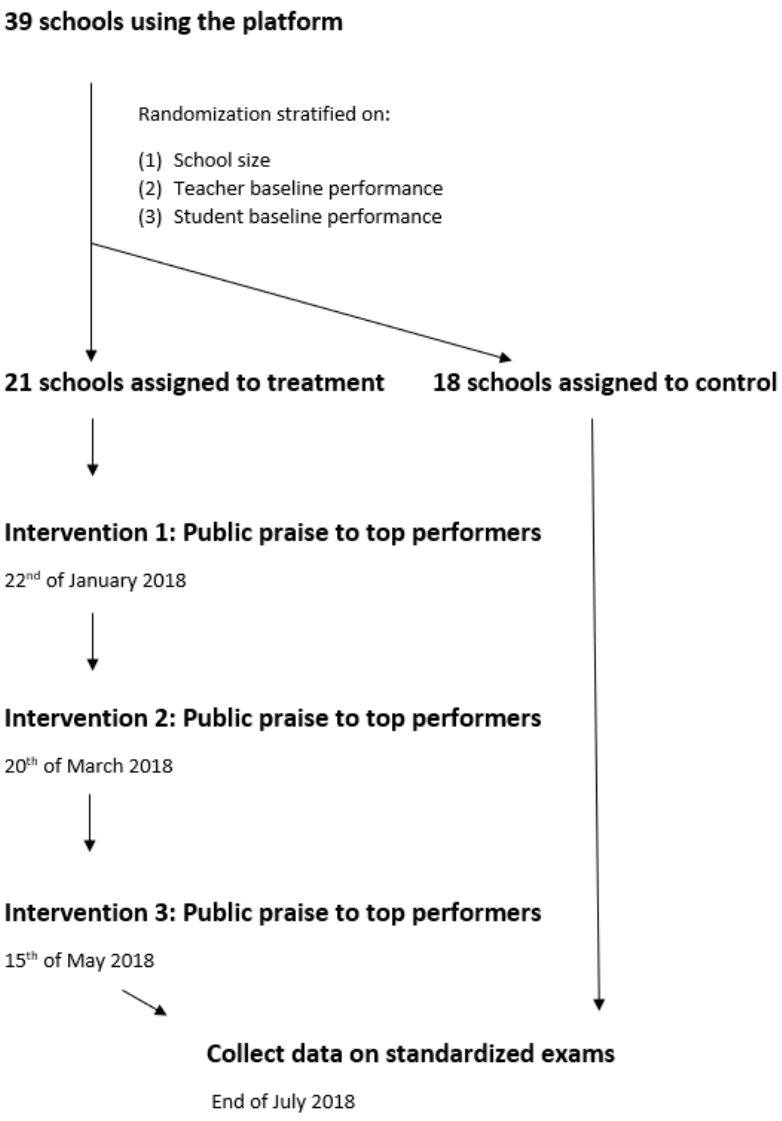
The second intervention takes place on the 20th of March 2018, two months after the first round of messages is sent. Student grades between the 22nd of Januray 2018 and 20th of March 2018 aree used to calculate the new performance gains for all teachers across all schools. Teachers are ranked again based on this new performance gain, and labelled as 'Top performers' if they are among the 25% best performing teachers (within their own subject). Top performing teachers in treated schools are publicly praised in a new round of messages posted on the 20th of March 2018.

The third and final intervention takes place on the 15th of May, two months after the second round of messages. Analogous, a final round of public messages is posted in all the treated schools. Finally, teacher performance gains are calculated again between the 15th of May 2018 and the end of the academic year.

In the week of 11th of June 2018, students finishing secondary school (aged 14) undertake high stake standardized exams, which are anonymously graded ('Examen de capacitate'). In the week of 25th of June 2018, students finishing high school (aged 18) undertake high-stake standardized exams which are anonymously graded ('Examen de Bacalaureat').

Figure A.1 presents a schematic overview of the experiment design and timeline.

Figure A.1: Overview of the experiment design and timeline



## A.2 PG robustness checks

Table A.2.1: The relationship between current learning and previous learning

| | **Pre-treatment PG** |
|---|---|
| Last year's PG | 0.079 |
| | (0.060) |
| | [0.096*] |
| N | 371 |
| F-value | 4.39 |
| R-squared | 0.19 |

*Notes: The dependent variable is pre-intervention PG expressed in standard deviations. Last year's PG is also expressed in standard deviations. OLS regression controls for student gender composition, average baseline performance of a teacher's students, subject fixed effects, degree of urbanization, being a publicly-funded school, school size, baseline student and teacher quality at the school level, and region fixed effects. In parentheses, heteroskedasticity robust standard errors are presented. Since the analysis is performed on only 20 schools, in brackets the p-value from applying the wild bootstrap procedure on standard errors clustered at the school level are presented (Cameron et al, 2008). Significance levels: *** p<.01, ** p<.05, * p<.1.*

Table A.2.2: The relationship between pre-treatment PG and learning for new and old students

| | **Pre-treatment PG** | **Pre-treatment PG** |
|---|---|---|
| Base perf. | -0.626*** | -0.630*** |
| | (0.027) | (0.027) |
| New students | | 0.047 |
| | | (0.033) |
| Base perf. * New students | | 0.033 |
| | | (0.010) |
| N | 48,547 | 48,547 |
| F-value | 84.64 | 86.01 |
| R-squared | 0.32 | 0.32 |

*Notes: The dependent variable is pre-intervention PG at the student level, expressed in standard deviations. The baseline performance of students is also expressed in standard deviations. 'New students' is a dummy variable which takes value one for students who just joined the school and 0 otherwise. OLS regressions control for student gender, average baseline student performance, subject fixed effects, degree of urbanization, being publicly-funded school, school size, baseline student and teacher quality at the school level, and region fixed effects. In parentheses, heteroskedasticity robust standard errors are presented, clustered at the school level. Since the same teacher can teach across different classes both new and recurring students, the analysis is performed at the student level. Significance levels: *** p<.01, ** p<.05, * p<.1.*

## A.3 Full text of intervention

The intervention text, original in Romanian language, is posted by the platform managers on the front page of the website, visible to all teachers, parents, and students immediately after logging-in.

**Unannounced public praise message:**

”Dear Teachers,

We are interested in how the performance of this school’s students is improving over time, since we want to encourage progress in education.

One way to measure the progress of students, is to see how much their grades improved since the beginning of the year. For a number of subjects (Mathematics, Romanian, English, Biology, Chemistry, Physics, History, Geography and Computer Science) we have looked at the improvement in student grades across all the schools that implement the *[platform name]* school management solution.

We are happy to announce that a number of teachers in your school are among the top 25% performers for their subject, across all the schools in our database. For these subjects, their student’s grades have improved the most since the beginning of the semester, as compared to the grades of students from other schools! These teachers are:

*[Teacher 1 name]*

.....

*[Teacher n name]*

We would like to thank these teachers in particular for their contribution!

In the future we plan to send such messages more often, to show our gratitude towards your hard work!

Best,"

**Announced and repeated public praise message:**

"Dear Teachers,

As you know, in the past we have analyzed, for a number of subjects (Mathematics, Romanian, English, Biology, Chemistry, Physics, History, Geography, English and Computer Science), the improvement in student grades across all the schools that implement the *[platform name]* school management solution.

We have now repeated this analysis. We are happy to announce that a number of teachers in your school are among the top 25% performers for their subject, across all the schools in our database. For these subjects, their student's grades have improved the most over the last 2 months, as compared to the grades of students from other schools! These teachers are:

*[Teacher 1 name]*

.....

*[Teacher n name]*

We would like to thank these teachers in particular for their contribution!

In the future we plan to send such messages more often, to show our gratitude towards your hard work!

Regards,"

Additionally, each teacher in a treated school is sent a reminder about the public message, through a personal e-mail from the platform managers. This measure is implemented to ensure that the treatment is as visible as possible.

Following the first intervention, the private message sent to all teachers is:

"Hello,

We are pleased to announce that for a number of subjects we have reviewed the increase in student performance in schools which implement the *[platform name]* school management solution.

Based on this analysis, *[number teachers]* teachers in your school are among the top 25% teachers in existing schools in our database!

If you want to see who these teachers are (or if you are one of them) you can see the list here: *[link to public message]*

Regards,"

Following the second and the third intervention, the private message sent to all teachers is:

"Hello,

We are pleased to announce that we have reviewed again the increase in student performance over the past 2 months. This analysis included all the schools which implement the *[platform name]* school management solution.

Based on this analysis, *[number teachers]* teachers in your school are among the top 25% teachers in existing schools in our database!

If you want to see who these teachers are (or if you are one of them) you can see the list here: *[link to public message]*

Regards,"

## A.4 Attrition

Attrition occurs when a teacher did not record any grades, for any of her students, during one experimental period. As such, for this teacher, no performance gain can be calculated. This type of attrition happens if a teacher did not test her students within a given period, or if a teacher stopped using the system all together. From the 855 teachers in the original sample for 4% of teachers ($n = 34$) performance gains cannot be calculated in period 2, for 7% of teachers ($n = 63$) in period 3, and for 3% of teachers ($n = 29$) in the last period. For roughly 86% of teachers performance gains can be calculated throughout the intervention, making the so-defined attrition rate reasonably low.

Table A.4 below shows the results from the Hotelling's T-squared test for multivariate data. The test verifies whether two sets of means are equal to each other across two groups, namely between a group of teachers who opt-out by recording no grades ('Attrition') and a group of teachers who do not opt-out ('No attrition'). The test has the advantage of jointly testing multiple variables at the same time, in this case the treatment status, being a top performer, and the interaction between the two. In the case of only one variable, the test reduces to a standard t-test. According to the results in Table A.4, attrition each round does not depend on either treatment status, on being a top performer, or on the interaction between the two.

Table A.4: Balance test for joint mean differences between the 'Attrition' and 'No attrition' group, each round

| **Round 1** | No attrition | Attrition |
|---|---|---|
| Treatment | 0.55 | 0.44 |
| Top performer | 0.25 | 0.32 |
| Treatment * Top performer | 0.14 | 0.15 |
| N | 821 | 34 |
| F-value joint difference | 1.15 | |
| P-value joint difference | 0.33 | |
| **Round 2** | No attrition | Attrition |
| Treatment | 0.55 | 0.54 |
| Top performer | 0.26 | 0.21 |
| Treatment * Top performer | 0.12 | 0.14 |
| N | 758 | 63 |
| F-value joint difference | 1.53 | |
| P-value joint difference | 0.21 | |
| **Round 3** | No attrition | Attrition |
| Treatment | 0.54 | 0.76 |
| Top performer | 0.24 | 0.17 |
| Treatment * Top performer | 0.12 | 0.14 |
| N | 729 | 29 |
| F-value joint difference | 1.93 | |
| P-value joint difference | 0.12 | |

*Notes: Columns show mean differences between the 'No attrition' and the 'Attrition' groups. Significance levels: *** p<.01, ** p<.05, * p<.1.*

## A.5 Unannounced public praise and attendance

I investigate the effects of unannounced praise on school attendance. Teachers are not praised on the basis of class attendance. However, as teacher performance changes as a result of unannounced praise, students can also become more or less likely to attend class. Furthermore, when the platform managers send a public praise message, they also signal that they are paying attention. This might induce teachers to feel monitored and take active steps to increase the attendance rate. Table A.5.1 presents the cumulative number of skipped classes among all the students of a teacher, up to that period in time.

Table A.5.1: Cumulative number of skipped classes among all the students of a teacher

| | **Treatment** | **Control** | **N** |
|---|---|---|---|
| Pre-treatment | 0.754 | 0.788 | 855 |
| Post unannounced praise | 1.229 | 1.064 | 821 |
| Post announced praise 1 | 3.472 | 3.577 | 756 |
| Post announced praise 2 | 4.136 | 4.556 | 729 |

*Notes: Columns show mean differences in cumulative number of skipped classes between the 'Treatment' and the 'Control' groups. The cumulative number of skipped classes is calculated at the teacher level, across all of their students.*

Pre-intervention attendance does not differ across treatment and control groups. After two rounds of the intervention, the students of teachers in the treatment group appear to have fewer classes skipped on average. However, both the level and the variation in skipped classes is very low, suggesting that teachers have very limited control over the measure, and little space for improvement. On average, by the end of the year, a teacher will record less than five skipped classes across all of her students. Given the limited variation in attendance, the minimally detectable effects are also large.

I estimate equation (2.1) where the main outcome variable is average skipped classes per teacher, as accumulated up to time t. Table A.5.2 shows the average treatment effects on attendance, following unannounced praise. The point estimates indicate that the treatment does not have a significant impact on class attendance, and the coefficients are small and

imprecisely estimated.

Table A.5.2: The effect of unannounced public praise on attendance

| | **Skipped classes** | **Skipped classes** |
|---|---|---|
| ($\alpha_1$) Treatment | 0.185 | 0.247 |
| | (0.290) | (0.259) |
| ($\alpha_2$) Top performer | | 0.213 |
| | | (0.213) |
| ($\alpha_3$) Treatment * Top performer | | -0.243 |
| | | (0.255) |
| Teacher Fixed Effects | yes | yes |
| Time Fixed Effects | yes | yes |
| N | 821 | 821 |
| F-value | 3.45 | 1.96 |

*Notes: The dependent variable is the cumulative number of skipped classes at the teacher level, two months after the first intervention. In parentheses, heteroskedasticity robust standard errors are presented, clustered at the school level. Significance levels: *** p<.01, ** p<.05, * p<.1.*

## A.6 Heterogeneous treatment effects for bottom performers

I estimate the following equation:

$$Perf_{i,t+1} = \alpha_0 + \alpha_1 T_t + \alpha_2 Quant_{i,t} + \alpha_3 T_t * Quant_{i,t} + \mu_i + \tau_t + \varepsilon_{i,t}$$

where $Perf_{i,t+1}$ is the new performance gain for teachers following the first intervention, and $Quant_i, t$ is a set of dummies for each of the four quantiles of the teachers's performance distribution. Table A.6 below shows the results from estimating the equation above for teachers in the 1st quantile who qualified for praise (the top 25%), teachers in the 2nd quantile (between 25% and 50% of the performance distribution), teachers in the 3rd quantile (between 50% and 75% of the performance distribution), and teachers in the 4th quantile (the bottom 75% of the distribution) as compared to similar teachers in the control group.

Table A.6 shows that teachers in treated schools at different quantiles of the bottom 75% performance distribution do not respond differently following the intervention. A t-test of joint equality of the coefficients on the three bottom quantiles returns an F-value of 1.90 and a p-value of 0.16.

Table A.6: The effect of unannounced public praise on PG, for different quantiles of the performance distribution

| | **New PG** |
|---|---|
| Treatment | -0.374** |
| | (0.160) |
| Top performers (Q1) | -2.765*** |
| | (0.187) |
| Quantile 2 | -1.505*** |
| | (0.134) |
| Quantile 3 | -0.928*** |
| | (0.176) |
| Treatment * Top performers (Q1) | 0.599** |
| | (0.251) |
| Treatment * Quantile 2 | 0.199 |
| | (0.161) |
| Treatment * Quantile 3 | 0.154 |
| | (0.220) |
| Teacher Fixed Effects | yes |
| Time Fixed Effects | yes |
| N | 821 |
| F-value | 152.71 |

*Notes: The dependent variable is the PG calculated two months after the first intervention, expressed in standard deviations. In parentheses, heteroskedasticity robust standard errors are presented, clustered at the school level. Significance levels: *** p<.01, ** p<.05, * p<.1.*

## A.7 Balance tests for final years students

Table A.7 reports the coefficients, standard errors and p-values from regressing a number of controls on a dummy variable which takes value one if a student is in the final year, and zero otherwise. With one exception, final year students do not appear to be different across any dimension, neither are they more likely to be over-sampled from the treated group. Final year students are slightly less likely to be sampled from schools that have some private funding. However, this is mechanically determined by the fact that these schools are mostly focused on secondary education, and typically do not offer classes for final year high school students. The main specifications control for school funding.

Table A.7: Differences between final year students and non-final year students

| | | **Is a final year student** | |
|---|---|---|---|
| **Variable** | **Coefficient** | **Standard error** | **P-value** |
| Baseline performance | 0.011 | 0.008 | 0.172 |
| Pre-treatment PG | 0.007 | 0.006 | 0.264 |
| Female student | -0.011 | 0.006 | 0.352 |
| In treated group | 0.009 | 0.026 | 0.745 |
| Urban school | 0.030 | 0.031 | 0.336 |
| Private funding | -0.089*** | 0.030 | 0.005 |
| Randomization variables | | | |
| School size | -0.001 | 0.001 | 0.493 |
| Baseline teacher performance | -0.091 | 0.075 | 0.234 |
| Baseline student performance | 0.013 | 0.016 | 0.392 |
| F- value | 2.49 | | |
| P-value | 0.024 | | |
| R-squared | 0.004 | | |
| N | 48,101 | | |

*Notes: The dependent variable is an indicator taking value 1 if a student is in their final year (aged 14 and 18) and 0 otherwise. Student baseline performance and pre-treatment PG are expressed in standard deviations. In the first column, coefficients from an OLS regression of the dependent variable on controls are presented. In the second column, heteroskedasticity robust standard errors are presented, clustered at the school level. In the final column, the associated p-value on each coefficient is displayed. Significance levels: *** p<.01, ** p<.05, * p<.1.*

## A.8 Announced public praise and attendance

Table A.8 shows the per-period treatment effect on student attendance throughout the remainder of the school year. The intervention has no significant effect on attendance.

Table A.8: The effect of unannounced and announced public praise on attendance

| | **Skipped Classes** | **SE** |
|---|---|---|
| **Treatment** | | |
| ($\gamma_{1,1}$) Round 1 | 0.203 | (0.300) |
| ($\gamma_{1,2}$) Round 2 | -0.130 | (0.547) |
| ($\gamma_{1,3}$) Round 3 | -0.162 | (0.657) |
| **Praise** | | |
| ($\gamma_{2,1}$) Round 1 | 0.287 | (0.125) |
| ($\gamma_{2,2}$) Round 2 | -0.112 | (0.125) |
| ($\gamma_{2,3}$) Round 3 | 0.380 | (0.387) |
| **Treatment * Praise** | | |
| ($\gamma_{3,1}$) Round 1 | -0.058 | (0.246) |
| ($\gamma_{3,2}$) Round 2 | 0.316 | (0.259) |
| ($\gamma_{3,3}$) Round 3 | -0.783 | (0.503) |
| Teacher Fixed Effects | yes | |
| Time Fixed Effects | yes | |
| N | 821 | |
| F-value | 20.93 | |

*Notes: The dependent variable is the cumulative number of skipped classes at the teacher level, two months after each intervention. In parentheses, heteroskedasticity robust standard errors are presented, clustered at the school level. Significance levels: *** p<.01, ** p<.05, * p<.1.*

However, by the third round, praised teachers appear to record less skipped classes for their students. While the coefficient is not significant, it is quite sizable. On the other hand, no consistent pattern is observed in the control group, in line with the fact that there is little variation in skipped classes, and with the fact that most students appear to attend regularly regardless of teacher quality.

## A.9 Unannounced public praise and exam choice

To test whether students are more likely to select an elective exam if their teacher was praised in the first (unannounced) intervention, I define subject choice as a dummy variable which takes value 0 if the subject is part of the track specific electives, but the student does not choose to undertake the exam. Alternatively, the variable takes value 1 if the subject is part of the track specific electives and the student chooses to undertake the exam.

As the regression only includes schools who offer a high-school program, the number of clusters drops to 22. To ensure that the standard errors are not biased by the small number of clusters I bootstrap standard errors using the wild bootstrap procedure developed by Cameron, Gelbach, and Miller (2008), with 5000 replications. Table A.9 below also reports the coefficients and associated the p-values from this exercise, confirming that the treatment does not influence elective choice.

Table A.9: Unannounced public praise and exam choice

| | **Subject chosen** |
|---|---|
| Treatment | 0.043 |
| | (0.045) |
| | [0.395] |
| Top performer | 0.053 |
| | (0.109) |
| | [0.647] |
| Treatment * Top performer | -0.096 |
| | (0.126) |
| | [0.496] |
| N | 4,563 |
| F-value | 9.16 |
| R-squared | 0.027 |

*Notes: The dependent variable is an indicator which takes value 1 if the student chose the respective elective course as a final exam, and 0 otherwise. OLS regression controls for baseline student performance, student gender, class type fixed effects, degree of urbanization, being a publicly-funded school, school size, baseline student and teacher quality at the school level, past year exam performance at the school level, and region fixed effects. In parentheses, heteroskedasticity robust standard errors are presented, clustered at the school level. Significance levels: *** p<.01, ** p<.05, * p<.1.*

# Chapter 3

# The Heterogeneous Effects of Early Track Assignment on Cognitive and Non-Cognitive Skills[1]

## 3.1 Introduction

The use of educational tracking to sort students into different learning environments is common practice worldwide. Tracking has traditionally been criticized because it is argued to enhance inequality by concentrating peer and school quality at the top of the achievement distribution; see, e.g. Hanushek and Woessmann (2006). Another point of critique towards (early) tracking is that students can be misallocated to tracks that do not fit with their abilities, as the ability signals that tracking is based on are noisy (Brunello et al., 2007). Several

[1]This chapter is based on joint work with Ron Diris and Trudie Schils. We would like to thank Robert Dur, Marjolein Muskens, Bart Golsteyn, Roxanne Korthals, Sergio Parra Cely, and seminar participants at Maastricht university for their helpful comments and feedback. Data collection has been funded by the government of the province of Limburg (Provincie Limburg), Maastricht University, school boards in primary, secondary and vocational education, and institutes for higher education in Limburg, within the program Educatieve Agenda Limburg.

studies have particularly focused on relatively young students in class as a group that is especially harmed by (early) tracking. It has been well-documented that there are differences in student achievement by relative age, and that these decrease as students grow older; see, e.g., Bedard and Dhuey (2006); Elder and Lubotsky (2009); Mühlenweg and Puhani (2010). As tracking is based on achievement, younger students are sorted more often to tracks that are below their academic potential, especially when they are tracked at early ages. It has been recognized to a much lesser extent that there is an equivalent issue on the other end of the relative age distribution; relatively older students are at risk of being placed above their potential, which can be harmful for their learning development as well. Hence, the effects of tracking can be highly heterogeneous by relative age, which could thereby be an important source of misallocation of students to optimal tracks. Moreover, such potential misallocation can be expected to not only affect cognitive learning development but also non-cognitive skills.

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

We find that attending the higher track has no effects on math and reading achievement across relative age, but do identify heterogeneous treatment effects for non-cognitive skills. Attending the higher track benefits older students especially in terms of perseverance, need

for achievement, and emotional stability. These gains in non-cognitive skills are absent for the relatively young in class. We further find that older students who attend the higher track are not more likely to fall back to lower tracks in subsequent grades despite the fact that their cognitive abilities are lower on average (contingent on being in the top track), which could be explained by these compensating spillovers on non-cognitive skills.

Previous studies have examined the interplay between relative age and tracking. Korthals et al. (2016) find that the relation between relative age and academic track attendance is stronger in early tracking countries, but simultaneously identify that the relation between relative age and achievement at the end of compulsory education is smaller in such countries. Moreover, they find that younger students have *higher* wages than their older peers when tracking is done early. Dustmann et al. (2017) provide a key contribution to the field by directly estimating the causal effect of track attendance at the achievement margin, and doing so for both educational and labour market outcomes. The authors use the variation in academic track attendance by month of birth to identify causal effects. They find that attending a higher track does not lead to more favorable long-run outcomes.[2]

These studies suggest that underestimating the potential of the relatively young in early tracking systems is not necessarily detrimental for their future attainment. However, these studies do not elicit the direct effect of track attendance for either relatively young or relatively old students, because they rely on a comparison between each group. What is typically neglected in discussions around tracking and relative age is that the relatively older students can also be harmed from being sent to a track that is above their potential. For example, not identifying a tracking effect in Dustmann et al. (2017) could be the result of the effect on the younger students sorted into a 'too low' track being canceled out by the effect on the

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

older students being sorted into a 'too high' track. If this would be the case, the near-zero effect sizes would hide that there can be considerable welfare gains by reallocating students on each side of the relative age distribution. To assess whether allocation that is partly based on fleeting relative age effects is harmful or not, it is required to directly identify the effects of track allocation across relative age.

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

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

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

This remainder of this study is organized as follows. Section 3.2 provides an overview of relevant aspects of the Dutch educational system. A description of data sources is given in Section 3.3, while Section 3.4 discusses the methodological approach. Section 3.5 presents and discusses the main results. Robustness analysis is provided in Section 3.6. Section 3.7 concludes.

## 3.2 Setting

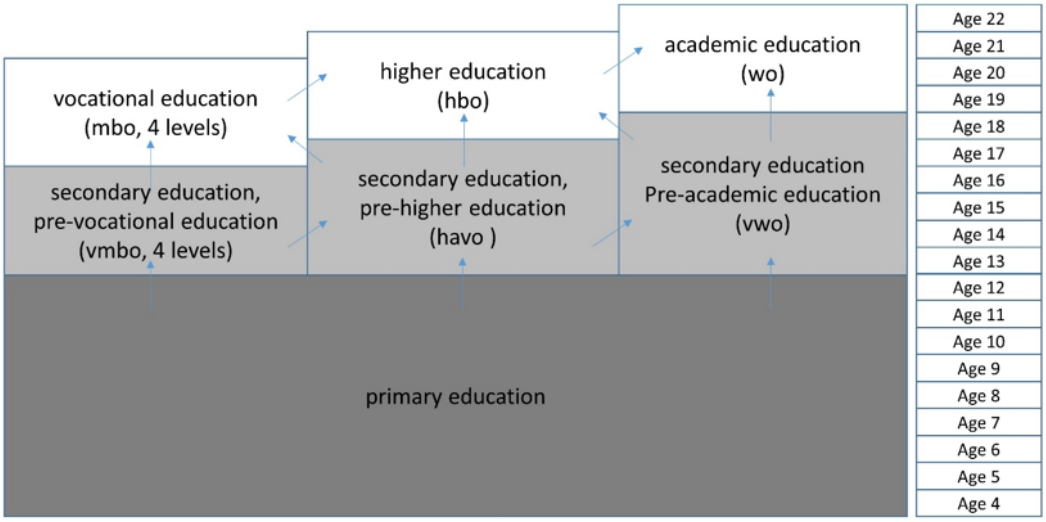
Figure 3.1 shows an overview of the main stages of the Dutch educational system. Primary education consists of eight years out of which the first two are spent in kindergarten. As of the third year of primary school (1st grade) children formally learn how to read and write. Most children start kindergarten at the age of 4, enter 1st grade at the age of 6, and finish primary school at the age of 12. When entering secondary school (grade 7), students are sorted into different school tracks. There are two underlying indicators for the sorting of students to tracks:

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

adjacent tracks.

2. A teacher recommendation: In 6th grade, teachers provide a subjective assessment of the child's ability and the associated track in secondary school. In the recommendation, the teacher is supposed to summarize the history of achievement of the student, but also an assessment of the broader (cognitive and non-cognitive) development, based on the teachers' classroom observations throughout the year. As with the 'test recommendation', the teacher recommendations can be towards a single track, or a mixed recommendation for two adjacent tracks. The teacher recommendation is granted after the results on the exit test and may thus be strongly affected by the test. The correlation between teacher recommendation and test recommendation equals 0.82.

Figure 3.1: Dutch education system




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

The Dutch secondary education system is hierarchically structured by ability and consists of three main tracks that differ in duration and qualification. The four-year track (*vmbo*) qualifies children for vocational education, the five-year track (*havo*) qualifies children for

higher (professional) education and the six-year track (*vwo*) qualifies children for academic education/university. On average, 55 percent of the children end up in the pre-vocational track, 25 percent in the pre-higher education track, and 20 percent in the pre-academic track. The pre-vocational track is further divided into four sub-tracks (*vmbo-b*, *vmbo-k*, *vmbo-g*, *vmbo-t*) that differ in their focus on practical versus theoretical content in the curriculum. Time spent on more theoretically oriented courses increases from *vmbo* b (25% of time) to *vmbo* t (100% of time). The *vmbo-b* and *vmbo-g* tracks have reduced considerably in recent years and in many schools have effectively been integrated within the *vmbo-k* and *vmbo-t* tracks respectively. Track recommendations also do not distinguish between the *vmbo-g* and *vmbo-t* track, which are jointly put under the label *vmbo-gt*. Recommendations do distinguish between *vmbo-b* and *vmbo-k* but the former track is relatively small (around 8% of the total student population). Hence, in practice the distinction is between one (largely) practical and one theoretical subtrack within *vmbo*, and the Dutch system as a whole can be argued to have four main tracks.

Table 3.1 shows for each track the average score on the primary school exit test and the percentage of students attending it (for all students in the OML data).

Table 3.1: Share of students per track recommendation given by 6th grade teacher

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

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

Roughly half of the students in the sample receive mixed recommendations, which are given when the teacher believes a student to be at the margin of two tracks. As shown, a

mixed recommendation for *vmbo-k* and *vmbo-gt* is not given (by rule).

Final tracking decisions can still be postponed for one year. Schools have the discretion to keep two adjacent tracks together in grade 7. Around 50% of students are tracked directly in grade 7 and the remainder only in grade 8.[3] These temporary comprehensive classes are most common for the combination of the practical and theoretical subtracks in *vmbo* and for the combination of *havo* and *vwo*.

## 3.3 Data

The data we use stem from the Dutch Onderwijsmonitor Limburg (OML). This is a cooperative project between Maastricht University and schools, school boards, and government bodies in the province of Limburg which is situated in the South of the Netherlands.[4] The province has about 1.1 million inhabitants, and a population density of 520 inhabitants per square kilometer, which is slightly above the average population density in the Netherlands of 502 inhabitants per $km^2$ (Statistics Netherlands, 2017). The average disposable household income in the region is about 34000 Euro per year, which is somewhat below the national average of 36200 Euro (Statistics Netherlands, 2017). Despite this, average scores on the standardized 6th grade exit test are above the national average.

The OML aims to collect and analyze information about the educational development of students in this region in order to provide feedback to schools and policy-makers about students' performance. The OML supplements administrative data with surveys among students, parents, and teachers at different stages in the education career, i.e. in kindergarten, grade 6 (final year of primary school), and grade 9 (secondary school). The administrative data include the exit test score and teacher recommendation from grade 6, and the track

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

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

placement from 7th to 9th grade. Questionnaires collect additional information on demographic indicators, socio-economic status, and non-cognitive skills of students. Additionally, the OML administers an IQ test in grade 6 and in 9th grade tests in math and reading. Parents can decide to withdraw from the survey based on the passive consent principle. A legal cooperation agreement states that the data collection is allowed to reach the overall goal of the program and is signed by all partners. The data collection is approved by the local ethical committee (ERCIC-092-12-07-2018). Researchers get fully anonymous data files.

The OML collects data for schools across the province in secondary school and for schools in the South of Limburg in secondary school. The coverage of schools is around 90 percent of the full population in that area in each case. For students that attended primary school in the North of the province, data on teacher recommendations and exit test scores are still collected, as they are also registered in the administrative data of secondary schools. Non-participating schools include special education schools, schools with a philosophy not to test children, and schools unable to plan the survey activities (participation takes ones hour per class in primary school and two hours per class in secondary school). Since most schools in the region participate in the project, sample selection problems are expected to be small.

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

Below we describe the data and variables in more detail.

**Demographic variables**

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

**Cognitive measures**

The OML contains data on the high-stake exit test that students take at the end of grade 6 (see Section 3.2). The test is standardized for all students and externally graded. It contains 200 multiple-choice questions testing the students on three main domains: Dutch language, mathematics, and study skills.

Data are also available for low-stakes tests on math and language in 9th grade. These tests were developed for the research project and administered digitally.[5] As the time for testing students was limited, not all students were administered both tests. Using a random algorithm, one third of the students were assigned to take only the language test, one third to take only the math test, and one third to take (shorter versions of) both tests. While

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

this reduces the sample size for estimating effects on cognitive outcomes, the randomization should ensure that this does not lead to any bias in our estimates. In addition, there are some missing test data on the school-cohort level, indicating that schools decided not to administer the test in class.[6] As we include school and year fixed effects, this is not a major concern for the internal validity of the point estimates. Exit test scores and background characteristics are not correlated with having missed the test, suggesting that external validity concerns are minor as well. Students from different tracks received items with different difficulty levels, but overlap in the test items ensures that we can construct a comparable scale using item response theory (IRT).

The data additionally contain information of the attended track in secondary education in grades 7, 8, and 9 which are all derived from the school administrative records. This includes any indication of mixed tracks that can still occur in grade 7 (and in some rare exceptions in grade 8 as well). Data on 7th and 8th grade placement are retrieved retrospectively from the school administration systems. Not all schools register this in an identifiable manner, and therefore this information is missing for around 20% of the students, who are excluded from the sample. As these missing observations are predominantly at the school-level rather than the individual level, this type of attrition is unlikely to be a major concern for our identification.[7] In addition, school switchers are likely to be overrepresented among these missing values. Since all of the schools in the sample offer both of the top tracks, there is no direct concern that school switching is linked to being retracked to the lower of the two tracks. A comparison further shows that the sample with missing data is slightly worse off in terms of exit test score, teacher recommendation, and parental background, but this disappears when we control for school fixed effects. In other words, schools with more low-

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

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

ability students are more likely not to report this information. As we control for school fixed effects, this is corrected for in the analysis.

**Non-cognitive measures**

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

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

Motivation is measured through 20 student-reported items (1-5 Likert scale) and self-confidence is measured through 25 student-reported items (1-4 Likert scale). We again create factor variables for each measure. There are two measures of educational aspirations, either on the secondary school track students expect to complete or on the highest level of post-secondary education they expect to complete. As the empirical analysis is focused on the top tracks in Dutch education, we create two dummy indicators taking value 1 if students expect to obtain a diploma from the top track (Exp. Track) or from university (Exp. Univ.). Expectations are not measured in the 2012 cohort of the OML. Appendix B.11 provides an overview of the items that the non-cognitive outcomes are based upon.

**Secondary school tracks**

Our analysis is motivated by the common finding in the literature that younger students in class obtain lower scores on achievement tests, especially early in life. When tracking occurs relatively early, such differences in maturity are expected to be more prominent in the allocation of students to tracks. In Appendix B1 we use primary school data on student achievement on the high-stakes exit test and on the tracking recommendation and estimate how these relate to age of testing (predicted by the month and day of birth). We show that our results are in line with the literature on relative age and performance, such that older students perform better on cognitive tests. Additionally, older students are more likely to be recommended to higher tracks by their teacher even after controlling for test performance. Hence, teachers appear to weigh maturity as an additional factor in the recommendation. This in turn also induces a positive relation between attendance of the high track in grade 7 and relative age: the predicted probability of high track attendance is around 8.1 percentage points higher for the very oldest student compared to the very youngest student in class, within the sample used in the empirical analysis.[8] We also find evidence that these relative

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

age effects substantially reduce as students grow older, in line with previous studies. As such, we expect that students who are older at the time of assignment are also more likely to struggle academically in higher tracks when at the achievement margin.

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

We define a student as being assigned to a high track if they were placed in *vwo* in 7th grade, as opposed to those placed in the next available track, namely *havo* or *havo-vwo*, depending on the school. In other words, those students who are placed at the *vwo* level are in the 'high track', while those who are placed in the second highest track that the school offers, are considered to be placed in the 'low track'. If the school has (next to a *vwo* track) a *havo-vwo* track, that will be the lower track. If the school only has *vwo* and *havo* tracks, then *vwo* will be the high track and *havo* will be the low track.[9] This implies that our treatment effect is essentially compared to two different counterfactuals jointly. We choose this approach because the alternative would imply endogenously excluding schools based on their sorting policies. Those in the mixed *havo-vwo* track in grade 7 comprise 60% of the group that is in the 'lower track'. Hence, this counterfactual carries more weight towards our results. The group in the mixed track still has a substantial chance of getting into the top

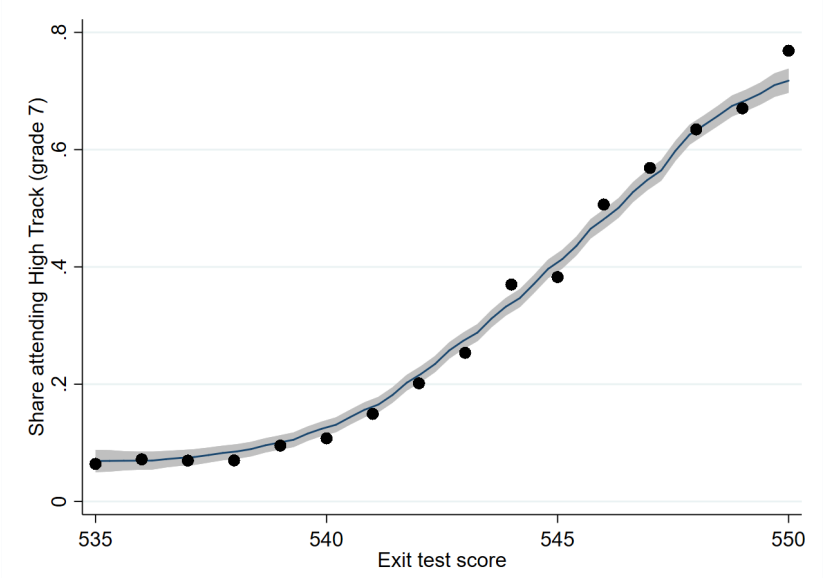
[9]When a school only has a *havo-vwo* class in 7th grade, there is no school threshold to estimate and the observations are excluded. This applies to 12 out of the 54 school-cohort observations.

track one year later, in which case their treatment would only consist of the one year being in a mixed grade. Section 3.6 will assess to what extent this may affect the results of the analysis.

## 3.4 Methodology

To assess the treatment effect of being in the highest track on cognitive and non-cognitive outcomes, we employ a regression discontinuity design. While the testing bureau links each of their final test scores to a recommended track, these score thresholds are not formally enforced for secondary schools in the Netherlands. Figure 3.2 below plots the the primary school exit test score against the probability of being assigned to the highest track. As shown, the probability of being assigned to the highest track increases smoothly with the exit test score, between 0 and 1. There is no clear discontinuity for any particular score, when all schools in our data are pooled together. The same applies when we drop one of the two counterfactual 'low track' situations.

Figure 3.2: Share of students attending high track across exit test score



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

**Fuzzy RD design**

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

$$HT_{ic} = \lambda_0 + \lambda_1 I_{ic}(S_{ic} \geq \overline{S_s}) + \lambda_2 RelAge_{ic} + f^k(S_{ic}) + X'_{ic}\lambda_4 + \tau_c + \sigma_s + \mu_{ic} \tag{3.1}$$

where $HT_{ic}$ is an indicator for being in the high track, for student i in cohort c, $S_{ic}$ represents the exit test score of each individual and $\overline{S_s}$ is the (school-specific) threshold score for eligibility for the high track. $f^k(S_{ic})$ is a polynomial control function of order $k$ for the test score $S_{ic}$. In all of the main analyses, we use separate linear specifications on each side of the threshold, as this provides the best fit. $RelAge_{ic}$ is the relative age of each student, measured by the month and day of birth of the student.[10] $X_{ic}$ is a vector of individual-level controls including gender, parental education, ethnicity, living with both parents, language spoken at home and the employment status of each parent. We further include fixed effects for each cohort c ($\tau$) and each secondary school s ($\sigma$). The second stage becomes:

$$Y_{ic} = \kappa_0 + \kappa_1 \widehat{HT_{ic}} + \kappa_2 RelAge_{ic} + f^k(S_{ic}) + X'_{ic}\kappa_5 + \tau_c + \sigma_s + \varepsilon_{ic} \tag{3.2}$$

[10] See Appendix B1 for a detailed account on constructing the relative age variable.

$Y_{ic}$ represents the outcome variable 3 years after track placement in the beginning of secondary school. These outcome variables can be classified in two categories: (i) cognitive and (ii) non-cognitive. Cognitive outcomes include the track level, math, and reading scores. Non-cognitive outcome variables include the personality traits of conscientiousness, neuroticism, need for achievement and persistence, measures of school motivation and self-confidence, as well as expectations students have with respect to future diplomas.

To analyze whether there are any heterogeneous treatment effects with respect to age, the model is expanded such that the interaction between being in the high track and relative age is instrumented with the interaction between treatment eligibility and relative age, adding a second first stage regression. The estimated equation becomes:[11]

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

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

$$O_{ic} = \eta_0 + \eta_1 \widehat{HT_{ic}} + \eta_2 HT_{ic} \widehat{* RelAge_{ic}} + \eta_3 RelAge_{ic} + f^k(S_{ic}) + X'_{ic}\eta_5 + \tau_c + \sigma_s + \varepsilon_{ic} \quad (3.3)$$

**Estimating school-specific thresholds**

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

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

determining their admission or sorting policy, although they are free to decide *how* to use that information.[12] The latter reflects the very high autonomy that schools have in decision-making within the Dutch educational system; see, e.g., Hanushek et al. (2013) who show that school autonomy across the OECD is highest in the Netherlands.

Little is known about the exact decision process of schools in setting thresholds, but evidence shows that it is common to have agreements about admission and sorting policies between different schools in the region. Reports indicate that around 25% of schools have an agreement with at least one other secondary school, 25% have agreements with all the schools in the region, and 10-15% indicate that there are formal agreements at the municipal level (Inspectie van het Onderwijs, 2014). Agreements at the municipal level are especially common in bigger cities. For example, all schools in Amsterdam have been subject to a formal agreement which states that all schools have to allow students to *vwo* above a certain exit score threshold, and indicating another range of scores just below for 'further consideration'.[13] To the best of our knowledge, such written agreements are lacking for the region of Limburg. Nonetheless, the school-specific thresholds that we will discuss in this section show substantial within-region correlation, which suggests that schools in this area operate in a similar way, though in a less formal manner.

We use secondary-school-level data on students admitted to the highest track and their exit test scores, to assess the prevalence of school-specific discontinuities. Porter and Yu (2015) show that program effects in an RD design can also be identified using a two-stage

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

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

procedure where the cutoff is estimated from the data. Furthermore, they show that estimating the cutoff in the first stage in this way does not affect the efficiency of the estimate in the second stage. In large samples, the fact that the threshold is unknown therefore does not affect the treatment estimates.

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

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

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

Table 3.2: Treatment eligibility and track assignment

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

As using a threshold to assign students to the highest track is not a strictly enforced method, it could be the case that some schools simply do not employ any threshold.[14] The potential problem of our approach is that we would then estimate the threshold where the noise in the data is strongest. This would elicit uncommonly good draws on the right hand side of the 'fake' threshold, or uncommonly bad draws on the left hand side of the 'fake' threshold, which could lead to biased results.

Figure B4 in the Appendix separately plots the jumps around empirically estimated thresholds for each secondary school. For a number of schools, the discontinuity is very fuzzy and most likely driven by random variation. As such, for our main estimates, we restrict the sample to the schools for which the jump at the threshold is statistically significant. After imposing this additional restriction, we capture 70% of the relevant schools and students in the region. This indicates that the majority of schools in our data use a threshold to assign students to the highest track.[15]

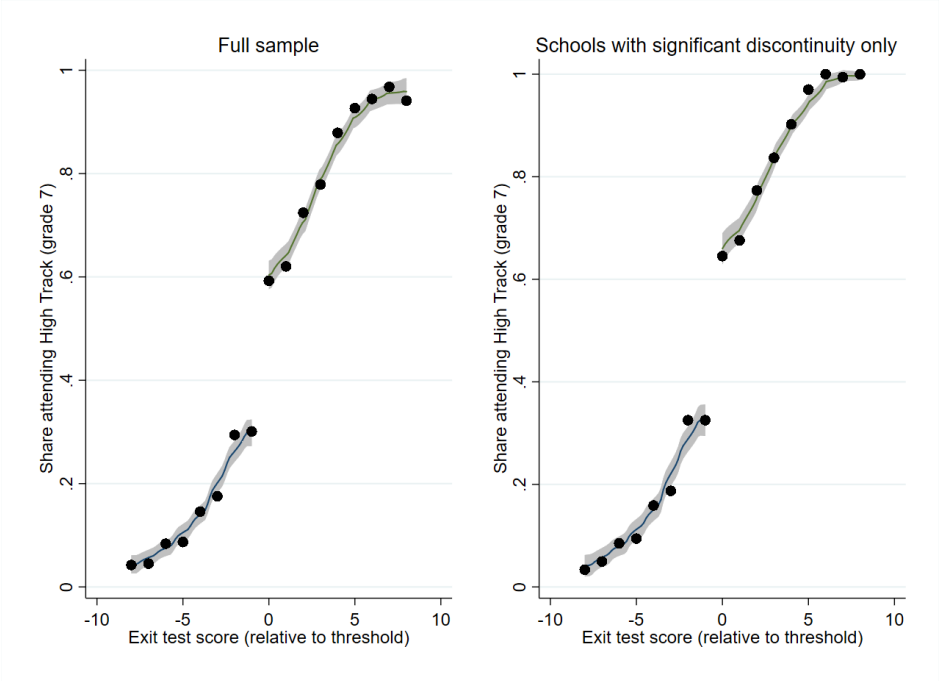
Figure 3.3 below plots the discontinuity around the estimated thresholds, for both the complete and the 'strict' sample. There is a precisely estimated jump in the probability of being assigned to the high track at the threshold: the probability of being assigned to the

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

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

high track approximately doubles, from 30% to 60%.

Figure 3.3: Discontinuity around the school-specific threshold



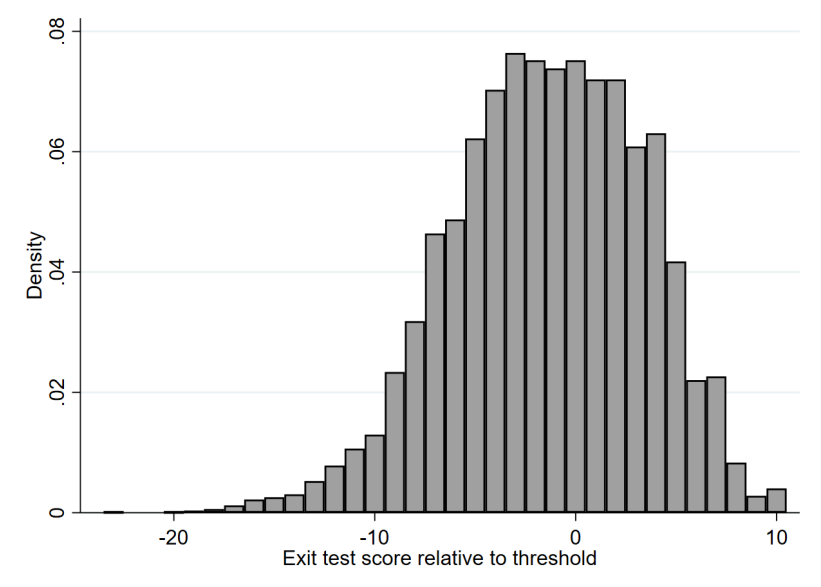

Note that the thresholds are estimated here per school and per cohort. We conduct a robustness analysis that uses thresholds that are estimated per school across the three cohorts as well (see Appendix B5), but choose the former as the main approach. For one, this approach increases first stage power. More importantly, we believe that this approach can accommodate the fact that thresholds can be updated from year to year. In fact, the score bands that the testing bureau reports also change every few years, and the formal agreements that exist in bigger cities in the Netherlands also show change over time. At the same time, we should recognize that a school-cohort threshold is potentially more vulnerable to schools endogenously adjusting the threshold to the composition of the student population in that year. The robustness test using thresholds estimated across years investigate whether this is a potential concern.

### Assumptions

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

To address this concern, we first examine the distribution of the exit test score, centered around the threshold. If sorting would occur, one would expect bunching in the distribution at 0. Figure 3.4 shows that the distribution is very smooth. A McCrary density test confirms this (McCrary, 2008). This also implies that schools do not adjust the threshold to maximize the inflow of students. There is little incentive for schools to do so in this setting, as there is no school in our sample that offers only the top track.

Figure 3.4: Distribution of test score around school threshold




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

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

It appears unlikely that students/parents move residence based on the threshold score of the nearest secondary school, especially since this is not public information. Hence, the main concern lies in endogenous commuting of students to schools. Figure B6.1 in the Appendix plots the population of students, location of schools, and the effective school catchment areas, in the region of Limburg.

In our sample, 74% of students attend the school closest to their home. Table B6.1 in the Appendix checks whether students who attend the closest school differ form students who don't. The students who attend the closest school are less likely to speak Dutch at home, as opposed to the regional dialect. This is likely a consequence of the fact that there is more school choice in the southeast, where the speaking of a dialect at home is more common. They do not differ on other characteristics.

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

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

In other words, endogenous self-selection of students to schools based on the school threshold occurs only when wanting to attend the higher track, since going to a lower track is always possible. Only 5% of students in our sample fall in that category. These students originate disproportionately from the region in the South-East of Limburg (3% of the total

5%). This is likely due to the higher availability of schools in this area. In additional robustness checks (see tables B6.2 and B6.3 in the Appendix), we show that excluding this region yields very similar estimates, confirming that this is not driving our results.

Finally, Figure B6.2 in the Appendix plots the student scores relative to the school threshold against the distance students travel to school. There is only a minor increase in travel distance at the threshold, which is statistically insignificant. This is further indicative evidence that students do not consistently choose to travel further away from home in order to select into the higher track.

While potential sorting issues appear limited in this setting, in Section 3.6 we show that our results are robust to an alternative specification where we instrument track assignment with treatment eligibility as determined by the threshold of the closest school.

**Balance tests**

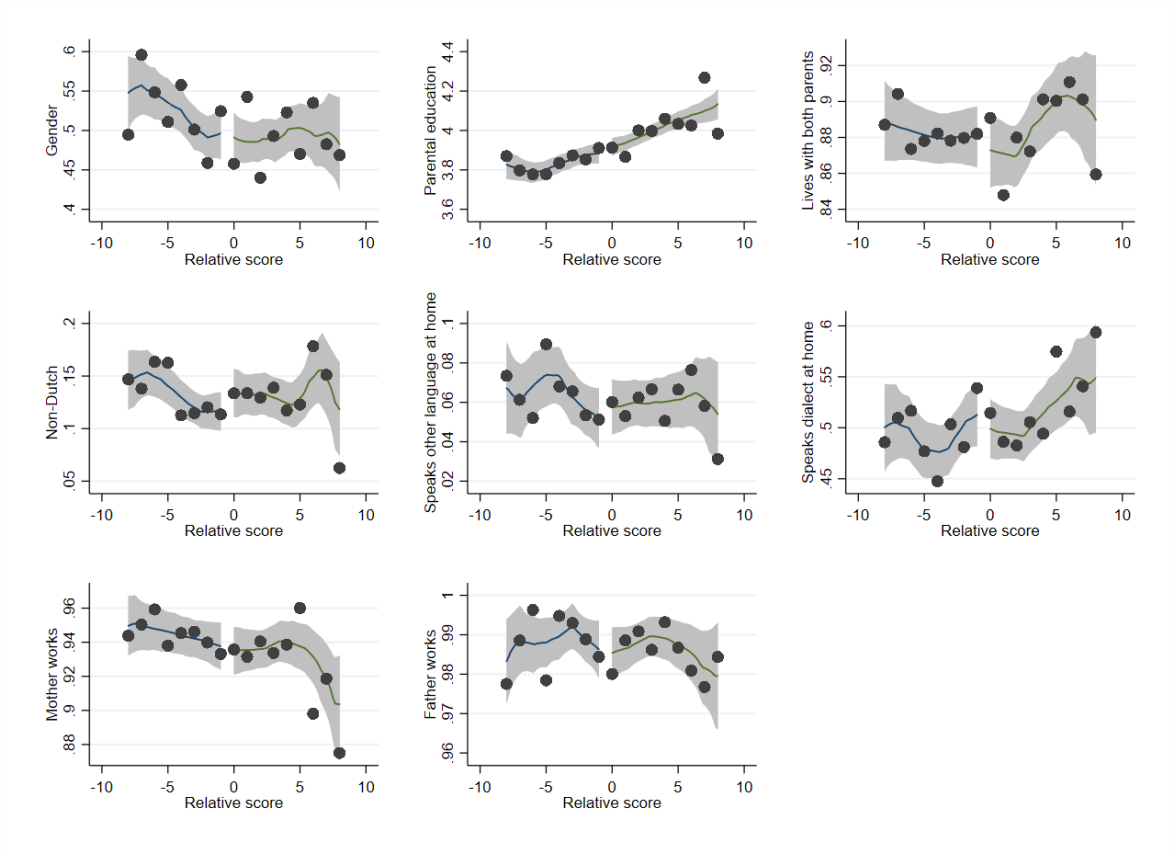
We now test whether important student characteristics and potential confounders are balanced around the threshold. As our dataset is very comprehensive as well as longitudinal, we can test this across a wide set of dimensions. We first look at our set of control variables. These are parent socio-economic status, student gender, ethnicity (Dutch or other), whether student is living with both parents, language spoken at home, and whether the parents are employed. In Figure 3.5, we plot the exit test scores around the threshold for each of these control variables.

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

Because our data are longitudinal, we can also run similar balancing tests across lagged measures of our outcome variables or other pre-treatment measures of cognitive and non-

cognitive skills. In Appendix B.8 we show that the track assignment is also orthogonal to 6th grade measures of need for achievement, persistence, conscientiousness and neuroticism, several teacher-reported measures of non-cognitive skills, and the score on a non-verbal IQ test.

Figure 3.5: Balancing tests around the threshold



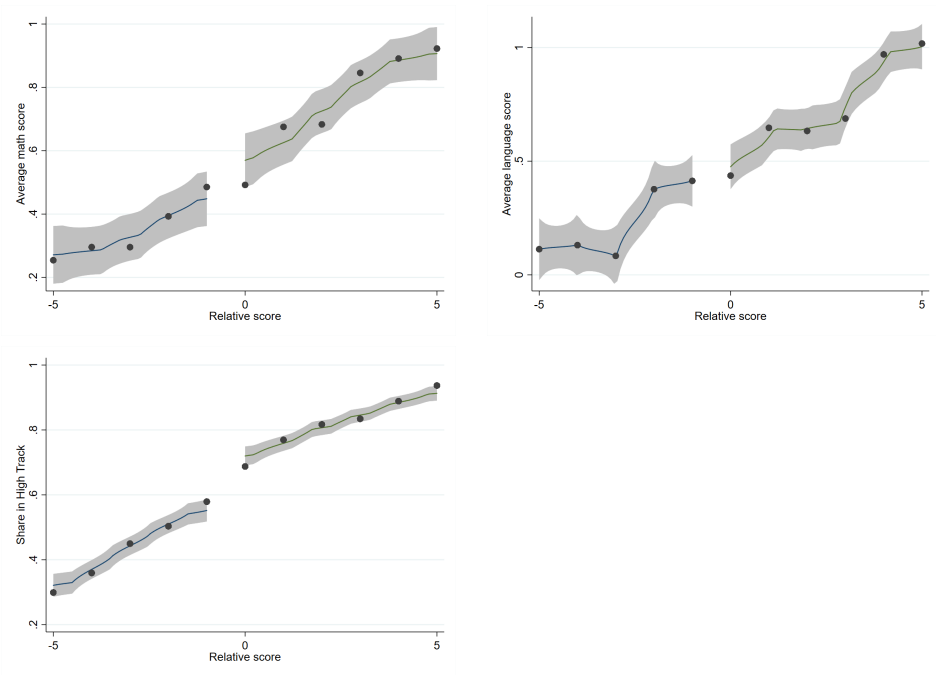

## 3.5 Results

We estimate the fuzzy regression discontinuity model described in equations (3.1) to (3.3). To do so we make use of the empirically estimated school-specific thresholds, restricting the sample to the schools for which the discontinuity is statistically significant.

Before exploring the regression results and the heterogeneous treatment effects with re-

spect to relative age, we first plot the second-stage for each individual outcome variable. Figure 3.6 below plots the second-stage for each of the three cognitive outcomes in 9th grade, while Figure 3.7 plots the reduced forms for the non-cognitive outcomes measured in 9th grade.

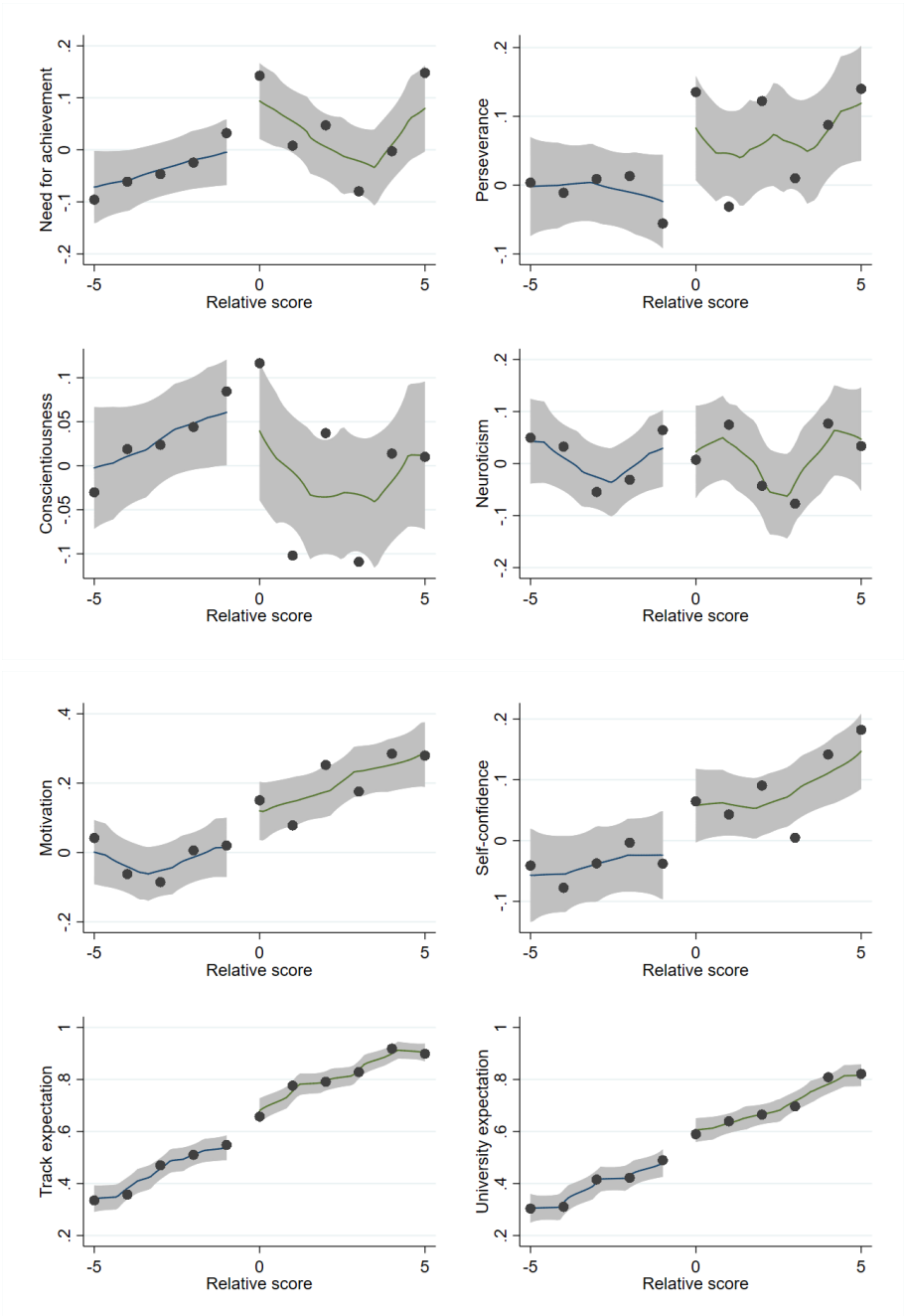
Figure 3.6: Reduced form graphs cognitive outcomes (9th grade)




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

Likely as a result of being locked into specific tracks, students' expectations for finished track and post-secondary education also jump at the assignment threshold. For the remaining non-cognitive outcomes, there is no clear and precisely estimated jump at the threshold, although the patterns for need for achievement and perseverance point somewhat in that direction. Effects for the latter two seem especially concentrated for those who just obtained

Figure 3.7: Reduced form graphs non-cognitive outcomes (9th grade)



the threshold score. This suggests that these academically marginal students may have a need to compensate lower cognitive ability by better non-cognitive development to stay on track. Regression analysis is needed for conclusive results, as well as for any indications of heterogeneity across relative age.

We present and discuss these results in turn for both cognitive and non-cognitive outcomes. The optimal bandwidths are calculated using the approach developed by Imbens and Kalyanaraman (2011), and their sensitivity is discussed in section 3.6.

**Cognitive outcomes**

Table 3.3 below presents the second stage results using equation (3.3), where the dependent variables are (i) track position 3 years after track placement[16], (ii) math scores 3 years after track placement, and (iii) reading scores 3 years after track placement. Appendix B10 shows the first stage results of these estimations.[17] All results reported below include the full set of controls as well as school and time fixed effects. A comparison to results without controls and school fixed effects is provided in Appendix Table B3.3. These estimates are highly similar, confirming that our instrument is orthogonal to observed characteristics.

As expected, students that are placed in the high track at the age of 12, are significantly more likely to still be in the high track by the age of 15. The coefficient size equals 0.386. Naturally, the difference in high track assignment in grade 7 equals 1, so the difference has narrowed. However, this is to be expected because students in the mixed *havo*/*vwo* track in grade 7, which make up 60% of those not in the high track, still need to be allocated. Part of this also reflects the downgrading of students that are in the high track in grade 7; descriptive statistics show that this occurs for around 13% of these students, and for around 18% of those

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

[17]We report Kleibergen-Paap test statistics on the relevance of the instrument. The Stock-Yogo critical values for weak instruments equal 16.38 for the case with one instrument (which applies to the results reported in Appendix B3) and 7.03 for the case with two instruments (which applies to the main results portrayed below).

at the school-specific threshold.

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

Table 3.3: Cognitive Skills

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

*Notes: The regressions include a control function for the exit test score on each side of the threshold, and controls for gender, socio-economic status, ethnicity, family structure, language spoken at home, parent employment status, cohort and school fixed effects. Optimal bandwidths are calculated (throughout the analysis) using the method by Imbens and Kalyanaraman (2011). Standard errors are robust. The Kleibergen-Paap (KP) statistic is provided to asses the strength of the identification. Significance levels: *** p<.01, ** p<.05, * p<.1.*

The results for the age interaction show that older students are not more likely to fall back to a lower track in the 3 years following track assignment. This is somewhat surprising, given that the older students in the top track are expected to be of lower ability and given that retracking to lower tracks happens rather frequently in Dutch education. The point estimate of the interaction is in the expected direction (i.e. the treatment effect is lower for older students) but statistically insignificant. It predicts a difference in treatment effects of around

0.09 between the very oldest and the very youngest student in class, suggesting a very minor treatment heterogeneity at best. In terms of math and reading achievement, treatment effects do not differ significantly between the older and the younger students.

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

For comparison, table B3.1 in the Appendix estimates the main effects from equation (3.3) without controlling for interactions with relative age. These coefficients are highly similar to the estimates in Table 3.3, in line with the fact that the effect of high track attendance on cognitive skills is not heterogeneous across relative age.

**Non-cognitive outcomes**

Table 3.4 presents the results for the 9th grade non-cognitive outcomes.[18] Appendix B10 shows the first stage results of these estimations.

The columns in Table 3.4 estimate equation (3.3) for: (1) need for achievement, (2) perseverance, (3) neuroticism, (4) conscientiousness, (5) confidence, (6) school motivation, (7) track expected to finish, and (8) post-high school education expected to finish. Note that the baseline effects for being in the high track reflect treatment effects for the youngest student in class. For comparison, table B3.2 in the Appendix presents the coefficients from estimating equation (3.3) without controlling for interactions with relative age, reflecting treatment effects for the (local) average student.

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

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

This means that the positive treatment effect of attending the high track is 0.48 and 0.54 standard deviations larger for the oldest student in class, compared to the youngest student in class. Attending the high track also reduces neuroticism by 0.084 standard deviations for every month that the student is older. This leads to a sizable treatment heterogeneity of 1 standard deviations between the oldest and youngest students in the sample.

Results for the model without age interactions presented in Table B3.2 in the Appendix show that the favourable effect of the high track on perseverance is also present for the (local) average student. However, the positive effects of attending the high track on need for achievement and neuroticism are statistically insignificant for the average student. In other words, high track assignment generally improves perseverance for the average student but more so for older students. On the other hand, attending the higher track only improves need for achievement and neuroticism for the relatively older students in class. We do not find any statistically significant treatment effect on conscientiousness, neither for the average student nor across relative age.[19]

The magnitude of these non-cognitive treatment effects is sizable. When we center the baseline effect on the oldest (rather than the youngest) student, the estimates for need for achievement, perseverance, and neuroticism are 0.44, 1.06, and 0.75 standard deviations respectively. To provide some comparison, the non-cognitive effects of the Perry Preschool Program center around 0.5 standard deviations (Heckman et al., 2010). Heckman et al.

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

(2006) identify a causal effect of completing high school on Locus of Control of 0.4 standard deviations and a causal effect of (some) college on self-esteem of 0.6 standard deviations. The estimates in this study for the oldest students are slightly above this (at least for the non-cognitive outcomes with a statistically significant effect). This could reflect the high malleability of these skills in early adolescence. At the same time, one has to keep in mind that this elicits the specific group that benefits the most. Moreover, these effects may be particularly large at the achievement margin, as high track assignment involves an allocation from the very bottom to the very top of the achievement distribution for these students. In any case, the effects we identify are of substantial size and as such represent important improvements in non-cognitive skills, which in turn have been shown to have strong links to multiple later-life outcomes.

The non-cognitive results could explain why older students are not more likely to be retracked to a lower track: being assigned to a track that may be 'too high' for their cognitive ability may challenge the older students more. In order to keep up with a more challenging environment, older students compensate by working harder, as reflected by the coefficients relating to need for achievement and perseverance. By being more challenged, older students assigned to the highest track appear to compensate a relatively lower cognitive ability with higher effort and a better ability to deal with psychological stress. Thus, these positive spill-overs on non-cognitive skills appear to help the older students to remain in the highest track and perform as well as the younger on achievement tests, despite the latter being more cognitively able on average when in the top track.

These (heterogeneous) estimates of the effect of high track attendance could be driven by a wide range of mechanisms. Being in a different track environment can represent a different curriculum, different teachers, different peers, a different class rank, and different expectations about future educational attainment. While we cannot disentangle all these potential forces, the setting and set of outcomes allow us to provide some insights. First of

all, the formal curricula in these two tracks are virtually identical in terms of the followed set of school subjects, but there can be differences in how advanced the level of instruction is.[20] All schools in our sample offer both tracks and thus students in each track are subject to the same teacher staff, making it highly unlikely that teacher (or school) quality effects partially drive our results.

It does not appear that rank effects have operated through self-confidence, given the insignificant estimate in column 5 of Table 3.4. This may be a result of the negative effect of a lower rank within class being offset by the positive effect of being in a high track in itself. As such, the setting differs from studies that show negative impact of class rank on self-image, expectations, and achievement in comprehensive schooling settings (Elsner and Isphording (2017) for the United States and Murphy and Weinhardt (2018) for the United Kingdom). In fact, students have higher expectations in the top track, although this effect does not differ by relative age and therefore does not appear to explain the favourable non-cognitive impacts of high track attendance for older students. We identify no treatment effects of high track attendance on motivation, either for the average student or across relative age. In summary, it appears that having different peers and the (informal) effects that this has on the level of instruction in class is the main driver of the results, as this seems to induce higher levels of effort from especially those students that are at the lower end of the ability distribution within the higher track (i.e. academically marginal students with a high relative age).

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

Table 3.4: Non-Cognitive Skills

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

*Notes: Outcome variables are, in order, need for achievement, perseverance, conscientiousness, neuroticism, self-confidence, school motivation, expected to finish high track and expected to finish university. The latter two are dichotomous, all other variables are standardized. The second stage regressions include a control function for the exit test score on each side of the threshold, and the control variables discussed in Table 3.3. Standard errors are robust. The Kleibergen-Paap (KP) statistic is provided to asses the strength of the identification.* $*p < 0.10, **p < 0.05, ***p < 0.01$

## 3.6 Robustness

### Student mobility

We further address the potential selection issue of students sorting to schools in our sample. In section 3.4, we have argued that, under some mild assumptions, student mobility is only an issue if students travel further away from home to potentially select into the track of their choice. We have shown that 74% of our sample attends the closest school, and that the ones that do not are not particularly different with respect to background characteristics. Additionally, only 5% of the sample attends a school further away that also has a more lenient threshold in order to be in the high track.

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

Tables 3.5 and 3.6 below present the results for cognitive and non-cognitive outcomes, based on this alternative approach. The findings are largely consistent with the main estimation results, presented in section 3.5. For the cognitive outcomes, we find a very similar effect for the 9th grade track. The coefficients for math and language are difficult to interpret as the KP-statistics are markedly below acceptable thresholds.

In Table 3.6, we find the same pattern of baseline effects as well as treatment heterogeneity as in the main analysis in terms of magnitude, sign and significance. Namely, we capture a positive interaction between high track attendance and relative age for need for achievement and perseverance, and a negative interaction for neuroticism. Again, expectations regarding track completion and university degree are markedly higher in the upper track, although the estimate for university expectations is statistically insignificant because

of higher imprecision. This confirms that our conclusions are not driven by a small sample of potentially self-selecting students.

Table 3.5: Closest school instrument: Cognitive Skills

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

*Notes: The second stage regressions include a control function for the exit test score on each side of the threshold, and the control variables discussed in Table 3.3. Standard errors are robust. The Kleibergen-Paap (KP) statistic is provided to asses the strength of the identification. Optimal bandwidths are re-estimated for this alternative approach. Significance levels: *** p<.01, ** p<.05, * p<.1.*

**Bandwidth sensitivity**

The bandwidths for Tables 3.3 and 3.4 have been optimally selected for each outcome variable, using the Imbens-Kalyanaraman method. To check the sensitivity of our results to variations in the bandwidth, we provide additional results where the bandwidth used is larger or smaller than the optimal one.

In the tables 3.7 and 3.8 we extend or restrict each bandwidth by 1 point on the exit test. As expected, wider bandwidths add more precision to the estimates, while smaller bandwidths decrease power. However, all our results remain very similar. The only true difference is that the point estimate on track in the 9th grade is statistically insignificant for the smallest bandwidth of 2 points. The estimate is still clearly positive, and the results for any other bandwidths are statistically significant. Moreover, the estimate for the age interaction is low and statistically insignificant across all bandwidths.

In Appendix B9 we show results of stronger changes in bandwidths, for a subset of

Table 3.6: Closest school instrument: Non-Cognitive Skills

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

*Notes: The second stage regressions include a control function for the exit test score on each side of the threshold, and the control variables discussed in Table 3.3. Standard errors are robust. The Kleibergen-Paap (KP) statistic is provided to asses the strength of the identification. Optimal bandwidths are re-estimated for this alternative approach. Significance levels:* $*p < 0.10, **p < 0.05, ***p < 0.01$

relevant outcome variables. The main results remain consistent.[21] In particular, Table B9 shows that, while the optimal bandwidth with respect to Need for Achievement is broad in the main estimation, the favourable track effect for older students is also present for more narrow bandwidths. Hence, our findings are not dependent on the employed bandwidth. Results are also robust to the inclusion of higher-order polynomials in the control function (available on request).

Table 3.7: Bandwidth sensitivity: Cognitive Skills

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

*Notes: The second stage regressions include a control function for the exit test score on each side of the threshold, and the control variables discussed in Table 3.3. Standard errors are robust. The Kleibergen-Paap (KP) statistic is provided to asses the strength of the identification. Significance levels: *** p<.01, ** p<.05, * p<.1.*

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

Table 3.8: Bandwidth sensitivity: Non-Cognitive Skills

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

*Notes: The second stage regressions include a control function for the exit test score on each side of the threshold, and the control variables discussed in Table 3.3. Standard errors are robust. The Kleibergen-Paap (KP) statistic is provided to asses the strength of the identification.* $*p < 0.10, **p < 0.05, ***p < 0.01$

**Controlling for lagged outcomes**

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

Table 3.9 below shows results when we control for these lagged outcomes (III), which are compared to the main results (I) and the results for the main specification run on the (smaller) sample for which the 6th grade measures are available (II). The table shows that results are highly similar when lagged outcomes are controlled for. The age interaction for perseverance even slightly increases and is now statistically significant at the 5% level, although this is primarily due to the sample change. These results solidify our main conclusion. They show that the favorable non-cognitive impacts for older students truly arise during the tracking period and that our estimates are not affected by baseline differences in these non-cognitive skills between those on each side of the threshold.

This further confirms the the balance tests for 6th grade measures discussed in Section 3.4, and available in Appendix Figure B8. Importantly, Figure B8 also shows that there is no particularly high value for the first data point above the threshold. The reduced form graphs for especially perseverance and need for achievement in grade 9 (Figure 3.7) suggested that the treatment effects are mainly concentrated at this truly marginal student. The different analyses have shown that the same marginal student is not different from those at the other side of the threshold in background characteristics or a wide set of pre-treatment non-cognitive skills.

Table 3.9: Non-Cognitive Skills: lagged outcomes

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

*Notes: Column I provides the main results, column II for the main model on the sample with available lagged outcomes and column III for the model including the lagged outcome as additional control. The second stage regressions include a control function for the exit test score on each side of the threshold, and the control variables discussed in Table 3.3. Standard errors are robust. The Kleibergen-Paap (KP) statistic is provided to asses the strength of the identification. Optimal bandwidths are re-estimated for the reduced.*

**Different counterfactuals**

As previously mentioned, the 'low track' condition jointly comprises the *havo*-track and the *havo-vwo* mixed track in grade 7. These two counterfactual settings differ in several dimensions, most prominently on the fact that *havo-vwo* students still have a considerable probability of assignment to the top track in case of good performance in grade 7. In this section, we separately analyze treatment effects with respect to each counterfactual situation, by dropping students in the other counterfactual from the analysis. These results, presented in Appendix Tables B7.1 through B7.4, should be interpreted with care as the estimates are naturally less precise and there may be issues of sample selection bias as well. Nonetheless, these results can indicate to what extent the existence of a mixed counterfactual can influence the main estimates and shed further light on some of the mechanisms.

Tables B7.1 and B7.3 show one particular difference for the cognitive results, namely a substantially stronger effect on *vwo*-attendance in grade 9 for the *havo* counterfactual. This is not surprising, since these students are effectively locked in the lower track, while the *havo-vwo* students still have a direct path to the top track.[22] The effect on 9th grade track is statistically insignificant for the *havo-vwo* counterfactual and does not preclude a meaningful effect, although the point estimate is positive. Interestingly, we observe in Tables B7.2 and B7.4 that the non-cognitive gains for the older students are present with respect to both counterfactual situations.

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

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

younger and the older half). Hence, the non-cognitive gains for this group of older students may still operate through a higher probability of being in the more demanding track, inducing compensating gains in non-cognitive skills. Still, the results could suggest that other mechanisms may be important as well.

The different peer environment in grade 7 may create a different reference point that has lasting effects on non-cognitive skills.[23] It could also be that, because *havo-vwo* students only get into the top track when they show high achievement in grade 7, these are students that are not likely to struggle academically in *vwo*, thereby reducing the need for compensation through effort or non-cognitive skills. We can only speculate on the relative importance of each of these potential mechanisms, and also need to be careful in interpreting these estimates that are rather imprecise and potentially subject to sample selection bias. In any case, this analysis shows that the identified gains in non-cognitive skills for especially older students are not dependent on the counterfactual that we employ.

**Additional robustness tests**

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

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

show the same pattern of results and identify a similar gain in non-cognitive skills for older students from attending the high track (these results are available on request).

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

## 3.7 Conclusion

This study has analyzed the effect of academic track attendance on cognitive and non-cognitive outcomes and its interaction with relative age. Our results show that assignment to the high track has no effect on cognitive outcomes for students at the achievement margin, irrespective of relative age. Track assignment does affect non-cognitive skills, and heterogeneously across relative age. Relatively older students benefit from attending the higher track in terms of higher perseverance, higher need for achievement, and higher emotional stability.

These effects are identified using a regression discontinuity design that empirically estimates school-specific thresholds from the data. We show that the results are not driven by selective mobility of students to schools, and they are also robust to alternative bandwidths, alternative approaches for threshold estimation, and the inclusion of lagged outcomes.

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

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

that the 'overtracking' of older students actually has benefits in terms of non-cognitive development, when we compare older students in the high track with older students in the low track.

In this light, one could conclude that high-stakes tests for track selection should not be corrected for relative age effects, but rather that a higher relative age is a positive signal for achieving a better non-cognitive development from attending the higher track, conditional on test performance. We emphasize, however, that these treatment effects by relative age may be highly dependent on the location of the threshold. If the lack of non-cognitive gains for younger students indeed occurs because they are of higher ability given top track attendance, then there may be a group of younger students with equivalent treatment effects on non-cognitive skills further below the threshold (i.e. those with similar *age-corrected* test scores as the older students with high non-cognitive treatment effects). The results can therefore also be interpreted more generally, namely that demanding learning environments can benefit non-cognitive skills.

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

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

question arises to what extent these effects, which predominantly seem to reflect higher effort levels, are sustainable in the long run.

# Appendix B

## B.1 Relative age effects

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

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

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

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

$$A_{ic} = \beta_1 + \beta_2 RelA_{ic} + \beta_3 X'_{ic} + \beta_4 S_{ic} + \varepsilon_{ic}$$

where $A_{ic}$ is the age of each student i in each cohort c, $RelA_{ic}$ is the relative age of the student as defined above, $X_{ic}$ is a vector of controls and $\varepsilon_{ic}$ is an individual specific error term clustered at the school level.

The second stage equation then equals:

$$Y_{ic} = \alpha_1 + \alpha_2 \widehat{A_{ic}} + \alpha_3 X'_{ic} + + \alpha_4 S_{ic} + \nu_{ic}$$

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

Table B.1.1 presents the results from both OLS estimates and IV coefficients where the dependent variable is the exit test score. Following the IV results, we identify that being one month older leads to an increase in the test score of 0.033 of a standard deviation higher. This means that the oldest in class score 0.4 standard deviations higher that the youngest students. This is a sizable difference and equal to around 40% of the difference in average scores between the *havo* and *vwo* tracks.

Table B.1.2 shows results when we use the teacher recommendation as an outcome. Older students receive higher recommendations, which is not surprising as the recommendation follows the test, which is characterized by large gaps by relative age. More importantly, we still identify a positive relation between age and teacher recommendation when the exit test score is controlled for. This indicates that teachers do not compensate the youngest based on their cognitive immaturity at the time of the test. In fact, they penalize the younger students

additionally, possibly by associating their immaturity with a lower ability that is not captured by the high stake test, or valuing relative maturity in itself as an important quality for being successful in future education.

Finally, Table B.1.3 reports results for 9th grade achievement. These estimates need to be interpreted with care, because the relation between relative age and attendance of the high track may impact these estimates (through the established gaps in exit test scores and teacher recommendations). Nonetheless, they can provide indicative results for the development of relative age effects over time. Table B.1.3 shows that relative age effects have virtually disappeared by grade 9. Moreover, estimates become negative once we control for the 6th grade ability indicators. The latter suggests that younger students indeed experience faster achievement growth between grades 6 and 9. While differences in high track attendance by age may endogenously contribute to these estimates, they are unlikely to be behind the narrowing of the relative age gap. For one, we identify no effects of attending the higher track on achievement in the main analysis. Second, a number of related studies suggest that if there is any effect of tracking on relative age gaps, it is in favour of the older students (Bedard and Dhuey, 2006; Korthals et al., 2016), implying that we might even underestimate the catching up of younger students. Therefore, we conclude that the expected narrowing of relative age gaps as students grow older appears to also be present within our sample, motivating the main analysis conducted in this study.

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

Table B.1.1: The effect of relative age on 6th grade achievement

| | **OLS** | **IV** | **N** |
|---|---|---|---|
| Exit test | -0.013*** | 0.033*** | 10.122 |
| | (0.0013) | (0.0038) | |
| Math subscore | -0.0094*** | 0.018*** | 5.677 |
| | (0.0025) | (0.0059) | |
| Language subscore | -0.013*** | 0.030*** | 5.677 |
| | (0.0025) | (0.0067) | |
| N | 10,122 | 10,122 | |

*Notes: Coefficients reflect the effect of an increase in relative age by one month. All regressions control for gender, parental education, ethnicity, living with both parents, language spoken at home and the employment status of each parent, captured by the vector $X'_{ic}$. Subscores for the exit test are only available for a subset of students. Standard errors are robust and clustered at the primary school level.* $*p < 0.10, **p < 0.05, ***p < 0.01$

Table B.1.2: The effect of relative age on teacher recommendation

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

*Notes: Coefficients reflect the effect of an increase in relative age by one month. All regressions control for gender, parental education, ethnicity, living with both parents, language spoken at home and the employment status of each parent, captured by the vector $X'_{ic}$. Subscores for the exit test are only available for a subset of students. Standard errors are robust and clustered at the primary school level.* $*p < 0.10, **p < 0.05, ***p < 0.01$

Table B.1.3: The effect of relative age on 9th grade achievement

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

*Notes: Coefficients reflect the effect of an increase in relative age by one month. All regressions control for gender, parental education, ethnicity, living with both parents, language spoken at home and the employment status of each parent, captured by the vector $X'_{ic}$. Sample size equals 5,450 for math scores and 5,082 for language scores. Standard errors are robust and clustered at the secondary school level. Significance levels: *** p<.01, ** p<.05, * p<.1.*

## B.2 Descriptives

Table B.2: Descriptive Statistics

| | **Share (%)** | **N** | |
|---|---|---|---|
| High track 9th grade | 52.6 | 6,080 | |
| Female | 50.8 | 6,056 | |
| Parental education | | 6,056 | |
| Primary | 1.6 | 95 | |
| Lower Secondary | 2.6 | 157 | |
| Upper Secondary | 21.0 | 1,269 | |
| Higher Professional | 52.2 | 3,160 | |
| University | 22.7 | 1,375 | |
| Lives with both parents | 88.4 | 6,056 | |
| Ethnicity | | 6,056 | |
| Limburg | 66.0 | 3,999 | |
| Other Dutch | 20.7 | 1,255 | |
| Non-Dutch | 13.2 | 802 | |
| Language at home | | 6,056 | |
| Dialect | 43.4 | 2,626 | |
| Dutch | 50.5 | 3,058 | |
| Other | 6.1 | 372 | |
| Mother employed | 93.9 | 6,080 | |
| Father employed | 98.8 | 6,080 | |
| Region | | 5,938 | |
| Central/North Limburg | 13.7 | 811 | |
| Western Limburg | 30.2 | 1,791 | |
| South-West Limburg | 35.5 | 2,105 | |
| South-East Limburg | 20.7 | 1,231 | |
| | **Mean** | **Standard deviation** | **N** |
| CITO score | 543.3 | 4.60 | 6,080 |
| Math score | 0.54 | 0.92 | 2,995 |
| Reading score | 0.49 | 0.98 | 3,053 |
| Need for achievement | 0.01 | 1.01 | 4,911 |
| Perseverance | 0.04 | 0.99 | 4,910 |
| Conscientiousness | 0.01 | 1.00 | 4,917 |
| Neuroticism | 0.02 | 1.03 | 4,920 |
| Confidence | 0.04 | 0.800 | 3,663 |
| School motivation | 0.10 | 1.03 | 3,789 |
| Expectation Track | 0.59 | 0.49 | 3,216 |
| Expectation University | 0.51 | 0.50 | 3,217 |

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

## B.3 Alternative model specifications

Below, we report results for two small changes to the main model specification. Tables B3.1 and B3.2 show estimates for the RD specification that only include the high track indicator, without the age interaction. As such, estimates represent the (local) effect for a student of average relative age, while the baseline effects in the main tables reflect the effect for the very youngest students in class. Table B3.3 shows a comparison between results with and without the control vector and school fixed effects.

Table B.3.1: Without age interaction: Cognitive Skills

| | **Track 9th grade** | **Math Score** | **Reading** Score |
|---|---|---|---|
| High track | 0.536*** | 0.054 | -0.225 |
| | (0.184) | (0.365) | (0.426) |
| Age | -0.001 | -0.014** | -0.004 |
| | (0.003) | (0.006) | (0.007) |
| N | 3,001 | 1,908 | 1,518 |
| KP stat | 35.36 | 26.28 | 20.71 |
| Optimal BW | $\pm 3$ | $\pm 4$ | $\pm 3$ |

*Notes: The second stage regressions include a control function for the exit test score on each side of the threshold, and the control variables discussed in Table 3.3. Standard errors are robust. The Kleibergen-Paap (KP) statistic is provided to asses the strength of the identification. Significance levels: *** p<.01, ** p<.05, * p<.1.*

Table B.3.2: Without age interaction: Non-Cognitive Skills

| | **Need Ach.** | **Persev.** | **Conscient.** | **Neurot.** | **Confid.** | **Motiv.** | **Exp. Track** | **Exp. Univ.** |
|---|---|---|---|---|---|---|---|---|
| High track | 0.200 | 0.774** | -0.255 | -0.218 | 0.478 | 0.217 | 0.384** | 0.273 |
| | (0.166) | (0.315) | (0.309) | (0.389) | (0.407) | (0.161) | (0.170) | (0.204) |
| Age | -0.009 | 0.003 | 0.005 | -0.007 | 0.009 | 0.003 | 0.002 | -0.000 |
| | (0.006) | (0.005) | (0.005) | (0.006) | (0.007) | (0.006) | (0.002) | (0.003) |
| N | 4366 | 2433 | 3114 | 2440 | 1781 | 3378 | 2409 | 2042 |
| KP stat | 144.75 | 40.23 | 46.96 | 31.66 | 26.79 | 148.05 | 24.79 | 22.13 |
| Optimal BW | ±7 | ±3 | ±4 | ±3 | ±3 | ±7 | ±5 | ±4 |

*Notes: The second stage regressions include a control function for the exit test score on each side of the threshold, and the control variables discussed in Table 3.3. Standard errors are robust. The Kleibergen-Paap (KP) statistic is provided to asses the strength of the identification. Significance levels: *** p<.01, ** p<.05, * p<.1.*

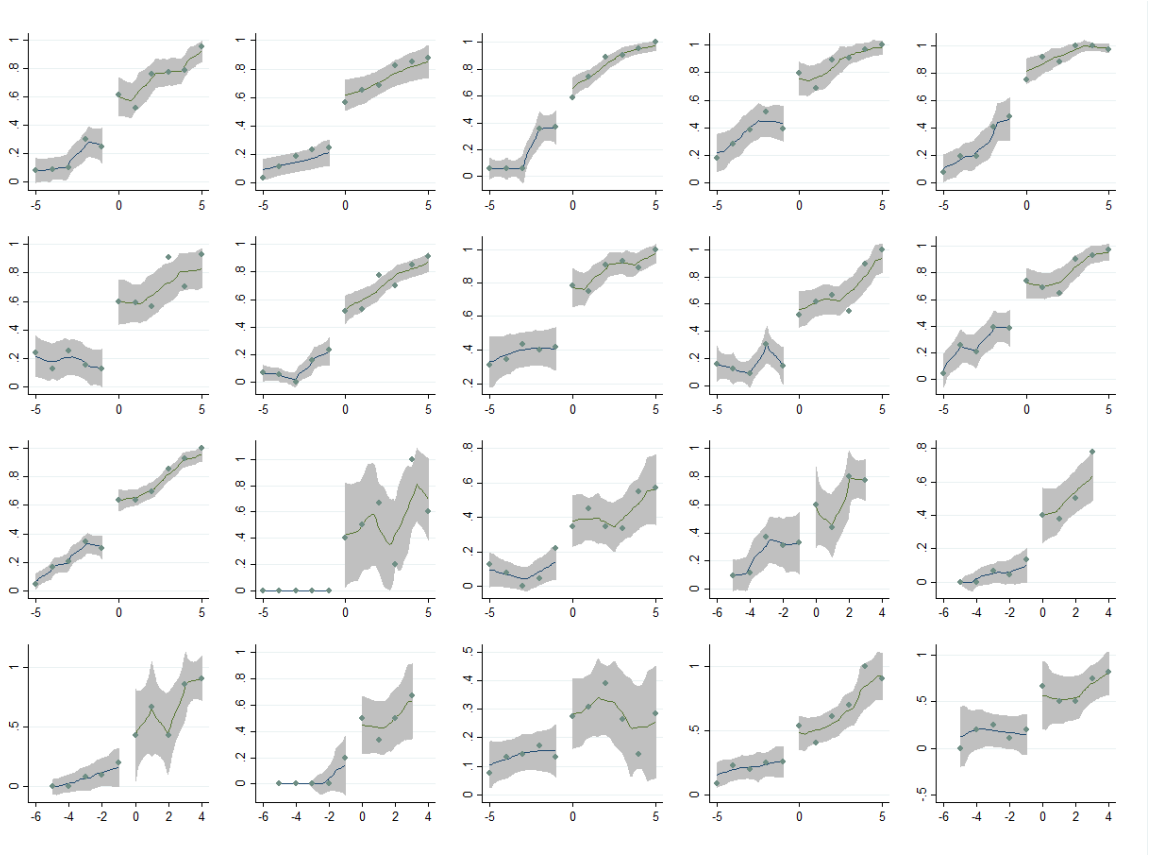
Table B.3.3: Estimates with and without control variables

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

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

## B.4 School-specific thresholds

Figure B.4: Discontinuity at threshold: school-specific figures



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

## B.5 Estimates when thresholds are estimated across cohorts

Table B.5.1: Cognitive Skills

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

*Notes: The second stage regressions include a control function for the exit test score on each side of the threshold, and the control variables discussed in Table 3.3. Standard errors are robust. The Kleibergen-Paap (KP) statistic is provided to asses the strength of the identification. Optimal bandwidths are re-estimated for this alternative approach.* $*p < 0.10, **p < 0.05, ***p < 0.01$

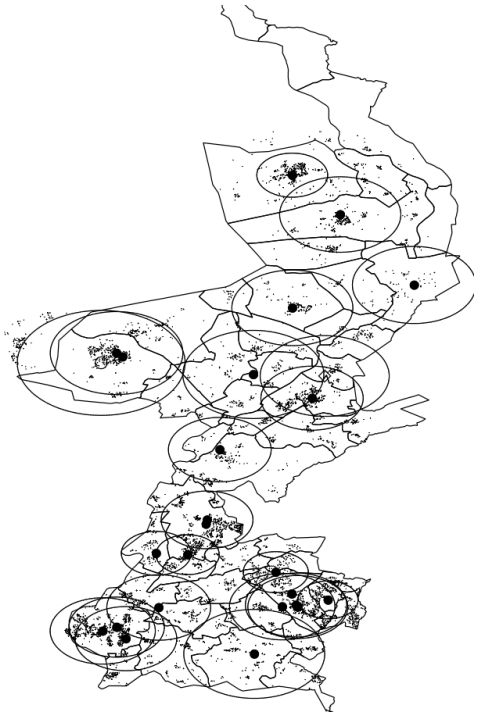
Table B.5.2: Non-Cognitive Skills

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

*Notes: The second stage regressions include a control function for the exit test score on each side of the threshold, and the control variables discussed in Table 3.3. Standard errors are robust. The Kleibergen-Paap (KP) statistic is provided to asses the strength of the identification. Optimal bandwidths are re-estimated for this alternative approach.* $*p < 0.10, **p < 0.05, ***p < 0.01$

## B.6 Sample selection

Figure B6.1: Distribution of schools and students across the province of Limburg



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

Figure B6.2: Average distance traveled to school relative to school threshold

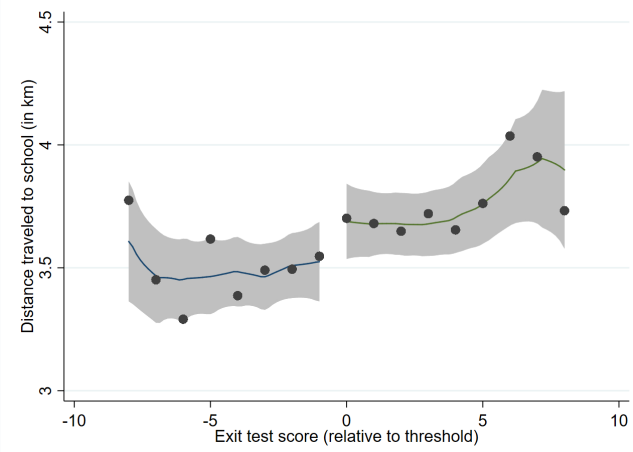


Table B6.1: Balancing tests for students not attending closest school

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

*Notes: Estimates are from a logit regression. The dependent variable is a dummy indicating whether a student attends the closest school or not. The regression controls for gender, parental education, ethnicity, language spoken at home, living with both parents, employment status of both the mother and the father, and exit test score.* $*p < 0.10, **p < 0.05, ***p < 0.01$

Table B6.2: Cognitive Skills: Excluding South-East

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

*Notes: The second stage regressions include a control function for the exit test score on each side of the threshold, and the control variables discussed in Table 3.3. Standard errors are robust. The Kleibergen-Paap (KP) statistic is provided to asses the strength of the identification. Optimal bandwidths are re-estimated for this alternative sample.* $*p < 0.10, **p < 0.05, ***p < 0.01$

Table B6.3: Non-Cognitive Skills: Excluding South-East

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

*Notes: The second stage regressions include a control function for the exit test score on each side of the threshold, and the control variables discussed in Table 3.3. Standard errors are robust. The Kleibergen-Paap (KP) statistic is provided to asses the strength of the identification. Optimal bandwidths are re-estimated for this alternative sample.* $*p < 0.10, **p < 0.05, ***p < 0.01$

## B.7 Separating the counterfactuals

Table B.7.1: Cognitive Skills: vwo vs havo-vwo

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

*Notes: The second stage regressions include a control function for the exit test score on each side of the threshold, and the control variables discussed in Table 3.3. Standard errors are robust. The Kleibergen-Paap (KP) statistic is provided to asses the strength of the identification. Optimal bandwidths are re-estimated for this alternative sample.* $*p < 0.10, **p < 0.05, ***p < 0.01$

Table B.7.3: Cognitive Skills: vwo vs havo

| | **Track 9th grade** | **Math Score** | **Reading Score** |
|---|---|---|---|
| High Track | 0.616*** | 0.806* | 0.096 |
| | (0.136) | (0.486) | (0.386) |
| High Track * Age | -0.010 | -0.009 | 0.004 |
| | (0.008) | (0.027) | (0.027) |
| Age | 0.005 | -0.004 | -0.003 |
| | (0.005) | (0.019) | (0.019) |
| N | 2046 | 1095 | 1125 |
| KP stat | 17.29 | 7.47 | 9.69 |
| Optimal BW | $\pm 4$ | $\pm 4$ | $\pm 4$ |

*Notes: The second stage regressions include a control function for the exit test score on each side of the threshold, and the control variables discussed in Table 3.3. Standard errors are robust. The Kleibergen-Paap (KP) statistic is provided to asses the strength of the identification. Optimal bandwidths are re-estimated for this alternative sample.* $*p < 0.10, **p < 0.05, ***p < 0.01$

Table B.7.2: Non-Cognitive Skills: vwo vs havo-vwo

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

*Notes: The second stage regressions include a control function for the exit test score on each side of the threshold, and the control variables discussed in Table 3.3. Standard errors are robust. The Kleibergen-Paap (KP) statistic is provided to asses the strength of the identification. Optimal bandwidths are re-estimated for this alternative sample.* $*p < 0.10, **p < 0.05, ***p < 0.01$

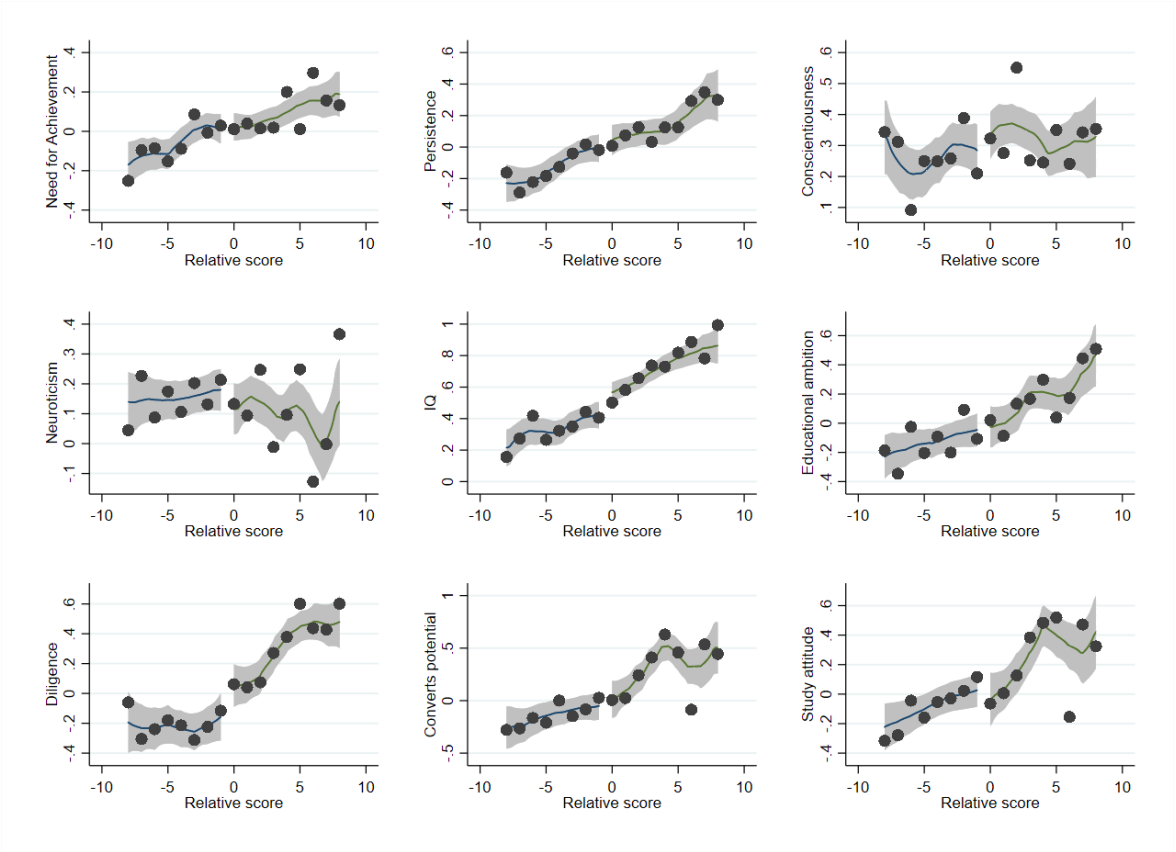
Table B.7.4: Non-Cognitive Skills: vwo vs havo

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

*Notes: The second stage regressions include a control function for the exit test score on each side of the threshold, and the control variables discussed in Table 3.3. Standard errors are robust. The Kleibergen-Paap (KP) statistic is provided to asses the strength of the identification. Optimal bandwidths are re-estimated for this alternative sample.* $*p < 0.10, **p < 0.05, ***p < 0.01$

## B.8 Balance tests for 6th grade non-cognitive skills and IQ

Figure B.8: Balance tests 6th grade outcomes




***Notes:*** *The figure shows balance tests for a number of outcomes measured in the 6th grade. Need for achievement, persistence, conscientiousness, neuroticism and educational ambition are student-reported. IQ is based on a non-verbal test. The final three outcomes are reported by the 6th grade teacher.*

## B.9 Additional bandwidth sensitivity

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

*Notes: For Track, Perseverance, and Neuroticism, the optimal bandwidth equals 3 and a reduction of more than 1 is not feasible. The second stage regressions include a control function for the exit test score on each side of the threshold, and the control variables discussed in Table 3.3. Standard errors are robust. The Kleibergen-Paap (KP) statistic is provided to asses the strength of the identification.* $*p < 0.10, **p < 0.05, ***p < 0.01$

## B.10 First stage coefficients

The tables below present the first stage coefficients for our two instruments, namely track eligibility and the interaction between track eligibility and age. The first stage coefficients corresponds to the our main second stage results, presented in Section 3.5.

Table B.10.1: Cognitive Skills: First stage

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

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

Table B.10.2: Non-Cognitive Skills: First stage

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

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

## B.11 Items non-cognitive skills

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

Table B.11.1: Items Need for Achievement

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

Table B.11.2: Items Perseverance

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

Table B.11.3: Items Conscientiousness

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

Table B.11.4: Items Neuroticism

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

Table B.11.5: Items self-confidence

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

Table B.11.6: Items Motivation

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

# Chapter 4

# Macroeconomic Conditions When Young Shape Job Preferences For Life[1]

Workers are not exclusively motivated by income. Many workers also deeply care about the meaning of their job and other non-monetary job attributes (Cassar and Meier, 2018b) and are willing to give up income for it (Maestas et al., 2018). This is not only of great relevance for mission-oriented organizations, such as not-for-profits and public sector organizations. Also for-profit firms typically do not just maximize profits, but take into account social factors as well (Hart and Zingales, 2017). While there is by now an extensive literature about how organizations attract, motivate, and retain a motivated workforce (Besley and Ghatak, 2018), little is known about how workers' preferences for different job attributes form, how the balance between income and meaning is shaped, and how and why this balance changes over time.

Since the nineteenth century, it has been claimed that different generations vary in their

[1]This chapter is based on joint work with Lea Cassar, Robert Dur, and Stephan Meier. We are grateful to Andrew Oswald, Antonio Spilimbergo, Paola Giuliano, Simon Lüchinger, Till von Wachter, Nicola Fuchs- Schündeln, Matthias Schündeln, Alois Stutzer, Michaela Slotwinski, Emily Bianchi, Otto Swank, Sacha Kapoor, Anne Boring, Josse Delfgaauw, and the participants in many conferences and seminars presentations for valuable comments and feedback.

preferences and beliefs based on their shared experience (Jaeger, 1985). Most recently, Millennials (the generation born between 1981 and 1996) are portrayed as having different work values (Twenge et al., 2010) and different preferences in general (Ertas, 2016; Rooney et al., 2018; Knittel and Murphy, 2019; Koczanski and Rosen, 2019) – resulting in a shift in how work is organized, in what types of firms get founded by Millennials, and in the mission of firms trying to compete for Millennials as workers and/or consumers. Classifying generations this way is to a large extent arbitrary, as it groups together individuals with widely different experiences. Thus, making claims based on this definition is problematic, even after controlling for time and life-cycle effects. In contrast, we follow an alternative approach, where individuals form preferences based on shared experiences. In particular, we empirically investigate how the shared experience of macroeconomic conditions affects job preferences for job meaning and income. Earlier studies have shown that macroeconomic conditions affect humans' well-being profoundly (Di Tella et al., 2001, 2003; Luechinger et al., 2010; Bianchi, 2013).

We combine insights from economics and psychology to develop our key hypothesis. Economists predict that if job meaning is a normal or luxury good, workers' demand for it should decrease in bad times: that is, when income is low (in absolute terms and/or compared with a relevant peer group). Psychologists have argued that the years between age 18 and 25 (the so-called impressionable years) are particularly important for the formation of people's preferences, beliefs, and attitudes. They are shaped during those years and change little after (Krosnick and Alwin, 1989). Research shows that macroeconomic conditions during the impressionable years affect preferences for redistribution (Giuliano and Spilimbergo, 2014) and personality traits (Bianchi, 2014, 2016). Other research shows that persons' experiences more generally shape their economic preferences and political views (Malmendier and Nagel, 2011, 2015; Fuchs-Schündeln and Schündeln, 2015; Slotwinski and Stutzer, 2018; Alesina and Fuchs-Schündeln, 2007; Laudenbach et al., 2019). Together these two insights

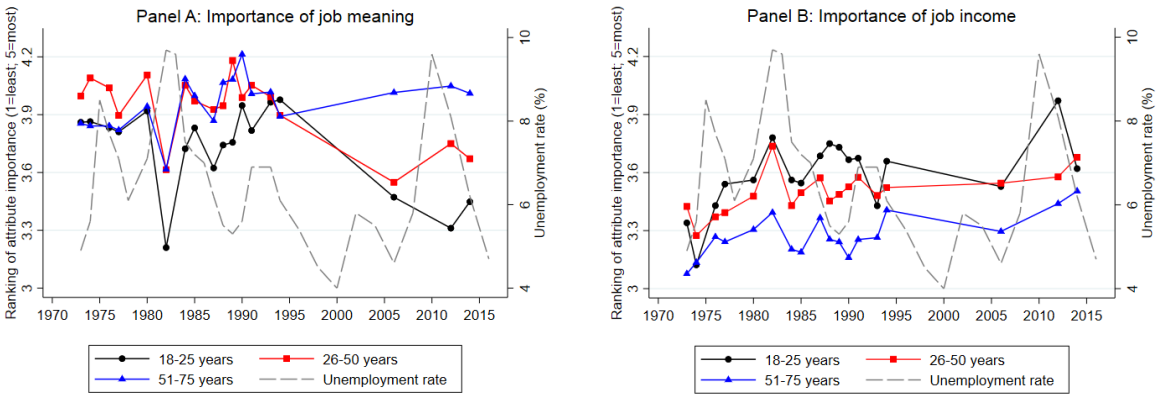
from economics and psychology suggest that lasting differences in job preferences between cohorts may be due to different macroeconomic experiences when young – beyond the impact of macroeconomic conditions at job market entry on lifetime earnings and education (Kahn, 2010; Oreopoulos et al., 2012).

We provide evidence on how preferences for job attributes are shaped and change over time based on data from the General Social Survey (GSS). From 1973 until 2014, a representative sample of the US population was asked in 18 of those 42 years to rank five job attributes: a high income, job security, opportunity for advancement, short working hours, and the meaning of work (see the Appendix C.1 for more details about the GSS and about the exact wording of the question and Table C5.1 for some descriptive statistics). Meaning of work and having a high income are the two most important attributes and also show a fair amount of variation over time. We therefore focus here on these two job aspects (and show results for the other three attributes in Table C5.6 in the Appendix).

In exploring differences in job preferences between cohorts we need to carefully control for time effects and life-cycle effects. In a first step, Figure 4.1 plots the average rank that people of different age groups give to high income and job meaning during the last four decades. Figure C4.1 in the Appendix shows the same for the remaining three job attributes. The charts also include the national unemployment rate as a key indicator of macroeconomic conditions. Three results are important. First, the ranking of job meaning and income varies substantially over time. Job preferences follow a cyclical pattern, with income (meaning) becoming more (less) important when unemployment increases. Job preferences of young people seem to be most affected by macroeconomic conditions. Second, there are substantial life-cycle effects. The young (18-25 years old) clearly rank income higher and meaning lower than do older respondents. This is not just the case in recent waves; it has been a consistent pattern throughout most of the sample period. Third, despite life-cycle effects, it seems that some cohort effects remain. For instance, the young in recent years value meaning

much less and income more than earlier cohorts did at the same age. We observe a similar pattern in the early eighties. In both these periods, the economy was in a deep recession (the Volcker Recession and the Great Recession, respectively).

Figure 4.1: Preferences for meaning (Panel A) and income (Panel B) across different age groups and over time.




*Notes: Based on a sample of 19,000 respondents who ranked preferences for job attributes in 18 waves, between 1973 and 2014. Preferences are ranked by respondents on a scale from 1 (least important) to 5 (most important). Right axis plots national unemployment rate.*

In a next step, we explore how much of the variation in Figure 4.1 can be explained by shared experiences of macroeconomic conditions. Identifying the effect of macroeconomic conditions on preferences is difficult with cross-time variation, as cohorts share many experiences such as technological progress and other national and global shocks. We therefore look at regional variation in macroeconomic conditions during the impressionable years. This allows us to identify the effect of macroeconomic conditions controlling for year fixed effects, age effects, birth cohort effects, and region effects, in addition to a rich set of socio-demographic variables. A closely related empirical approach is used in Giuliano and Spilimbergo (2014).

Using regional income data since 1929 from the US Bureau of Economic Analysis, we construct a measure of macroeconomic conditions during one's impressionable years by calculating the $IncomeLevel^{18–25}$. This measure is given by the logarithm of the average regional

income per capita experienced in each of the eight years between 18 and 25 years of age, in the region in which a respondent resided at the age of 16.[2] In Table C5.3 in the Appendix we additionally control for the standard deviation of experienced income during the impressionable years, to allow for the fact that some respondents have lived through much more volatile times. This measure does not appear to predict job preferences, nor does adding it to the regression specification change our key conclusions in any important way. The yearly regional income per capita is adjusted for inflation, using national-level CPI indexes, so that it is expressed in 2017 US dollars (see the Appendix C.3 for more details on how the measure is constructed).[3]

Our main regression specification is

$$JobPref_{i,r,t} = \beta_0 + \beta_1 IncomeLevel_{i,r,t}^{18-25} + \beta_2 X_{i,r,t} + \tau_t + \rho_r + \rho_r^{age16} + \varepsilon_{i,r,t} \tag{4.1}$$

where the dependent variable $JobPref_{i,r,t}$ is a ranked response, on a scale from 1 to 5, indicating how important respondent $i$ living in region $r$ in year $t$ finds a certain job attribute. $IncomeLevel_{i,r,t}^{18-25}$ is the logarithm of the real income per capita during the impressionable years in the region in which a respondent resided at the age of 16. $X_{i,r,t}$ is a vector of individual specific demographics, including gender, years of education, father's and mother's education, race, marital status, number of children, squared household size, age dummies, the logarithm of household income, the logarithm of household income at the age of 16,

[2]The region in which a respondent resided at the age of 16 is our best proxy for where the respondent resided between 18 and 25 years of age. The underlying assumption is that during the impressionable years, individuals lived in the same region as they did at age 16. Table C5.2 in the Appendix shows that restricting the sample to respondents who at the time of the survey live in the same region as they did at the age of 16 does not alter our results.

[3]We use regional income per capita and not regional unemployment because the latter is only available from 1976 onward. Regional unemployment is negatively correlated with regional income per capita (-0.31). Using experienced *national* unemployment at age 18-25 (which is available from the 1920s onward) yields consistent results but is identified by age differences at time of the survey (see Table C5.4 in the Appendix).

work status, and decade-of-birth dummies. The term $\tau_t$ represents year fixed effects. To avoid the well-known collinearity issue between age, year, and cohort fixed effects but still capture cohort differences, we assume that the effect of birth year on job preferences is the same for all the individuals born within the same decade. Table C5.5 in the Appendix shows that results are robust to alternative specifications which vary the definition of birth and age categories, and confirms that our conclusions do not hinge on this restriction.

To capture time-invariant region effects, we add region-of-interview fixed effects, $\rho_r$, and region-at-age-16 fixed effects, $\rho_r^{age16}$.[4] To control for the possibility that there are common shocks at the region level, we cluster our standard errors at the level of the region in which a respondent lived at the age of 16.[5] Since there are only nine regions in the GSS panel, we use the wild bootstrap procedure suggested by Cameron et al. (2008), which estimates reliable standard errors, even with a small number of clusters. In all our tables, we report the $p$ values from these wild bootstrap regressions, and we base statistical significance on the bootstrapped standard errors. For ease of interpretation, we use ordinary least squares (OLS) regressions. Our results are robust to using a rank-ordered Probit model instead.

Table 4.1 reports the results from estimating equation (4.1) for "Meaning" and "Income." As described above, the regressions control for many variables – importantly time and life-cycle effects. Figures C.4.2 and C.4.3 in the Appendix plot the age and year fixed effects, and show that both life-cycle effects and general time trends are important. Moving from age 20 to age 75, our estimates predict a strong increase in the importance of job meaning and a strong decrease in the importance of a high income, both of about a full point. Over

[4]To keep the specification as simple as possible, we diverge from (Giuliano and Spilimbergo, 2014), who also add the term $\rho_r^{age16} * age_{i,r,t}$, which interacts the region at 16 fixed effects with the age (linearly) of respondent $i$ at time $t$. Our results are also robust to, and become stronger when, including this interaction. Additionally, we investigated whether our conclusions change when adding interactions between the region at 16 and a linear time trend, to control for region-specific trends. Our coefficients are also robust to this alternative specification.

[5]Clustering the standard errors at the level of the region in which the respondent lives in at the time of the survey does not alter the significance of our estimates.

the last four decades, the average ranking of job meaning has decreased by about 0.7 points, while high income has increased by almost a full point. As the average ranking ranges from minimum 1 to maximum 5, these are sizeable changes. In contrast, the decade of birth of respondents, that is cohort effects, seems less relevant (see Figure C4.4 in the Appendix).

Table 4.1: Experienced income during the impressionable years and job preferences for meaning and income

| | **Meaning** | **Meaning** | **Income** | **Income** |
|---|---|---|---|---|
| Income level 18-25 | 0.340 | 0.474 | -0.292 | -0.382 |
| | (0.113) | (0.115) | (0.103) | (0.103) |
| | [0.002***] | [0.002***] | [0.004***] | [0.001***] |
| Household income | ✓ | X | ✓ | X |
| Years of education | ✓ | X | ✓ | X |
| Labor market status | ✓ | X | ✓ | X |
| Demographic variables | ✓ | ✓ | ✓ | ✓ |
| Age FE | ✓ | ✓ | ✓ | ✓ |
| Decade of birth FE | ✓ | ✓ | ✓ | ✓ |
| Year FE | ✓ | ✓ | ✓ | ✓ |
| Region FE | ✓ | ✓ | ✓ | ✓ |
| Region at 16 FE | ✓ | ✓ | ✓ | ✓ |
| N | 19,026 | 19,026 | 19,022 | 19,022 |
| F-value | 24.61 | 18.94 | 8.59 | 8.57 |
| R-squared | 0.161 | 0.118 | 0.068 | 0.057 |

*Notes: Regressions are estimated using OLS. The 'Income level 18-25' is log-linearized. Demographic variables include controls for gender, race, father and mother education, marital status, number of children, household size (squared), and household income at the age of 16. In parentheses, heteroskedasticity robust standard errors are reported. In brackets, p-values are reported estimated using the wild bootstrap procedure suggested by Cameron et al. (2008), by clustering standard errors at the level of the region at age 16. Since the number of clusters is small, the more conservative Webb weights are used (Webb, 2013), implemented using the boottest estimator developed by Roodman et al. (2019), with 5000 replications. Sample re-weighted using the wtssall population weights in the GSS. Significance levels: *** p<.01, ** p<.05, * p<.1.*

Regarding our key variable of interest, columns (1) and (3) in Table 4.1 show that macroeconomic conditions during young adulthood shape job preferences in important ways. Individuals who experience a higher level of income during their impressionable years rate meaningful work as significantly more important. This happens at the expense of finding a

high income important.[6] The extensive control variables include personal income, years of education, and labor market status at time of survey, which have been shown to be affected by macroeconomic conditions at the time of entering the labor market (Kahn, 2010; Oreopoulos et al., 2012). Hence, the effects on job preferences that we identify hold beyond any possible effect through current labor market experience.

In columns (2) and (4) in Table 4.1 we estimate equation (4.1) without these variables. Our results are robust to this alternative specification, and the coefficients of interest become larger, suggesting that experienced income at a young age affects job preferences at a later age partly through affecting current income, labor market status, and attained education.[7]

The sizes of the coefficients indicate that the effects are economically significant. A one-standard-deviation increase in experienced income during the impressionable years translates to a move of -0.14 (0.12 of a standard deviation) in the average ranking of income, and a move of 0.17 (0.13 of a standard deviation) in the average ranking of meaning, where the lowest possible rank is one and the highest possible rank is five. To put this into context, the magnitude of the effect of (a one-standard-deviation increase in) experienced income on preferences for income is over 1.8 times that of the effect of gender, and as large as the effect of unemployment.[8] Comparatively, the magnitude of the effect of (a one-standard-deviation increase in) experienced income on preferences for meaning is 0.65 times as large as the gender effect and 3.4 times that of the unemployment effect.[9]

[6]The importance of the other three job aspects (job security, chances for advancement, and short working hours) is hardly affected by macroeconomic conditions during the impressionable years, see Table C5.6 in the Appendix. The null finding for job security is to some extent surprising given the evidence in Malmendier and Nagel (2011) that macroeconomic experiences affect willingness to take risk. Note, however, that in contrast to our study, the macroeconomic experiences studied by Malmendier and Nagel (2011) consist of the experienced histories of stock and bond returns and that their outcome variables are about financial risk taking.

[7]In a robustness check, we added a set of industry dummies, and found that the coefficient for $IncomeLevel^{18-25}$ did not change substantially. We have also examined whether macroeconomic conditions have a stronger impact when household income at age 16 is lower, which turns out not to be the case.

[8]The coefficient for gender is -0.08 ($p$ value=0.000) and the coefficient for unemployment is -0.14 ($p$ value=0.008).

[9]The coefficient for gender is 0.26 ($p$ value=0.000) and the coefficient for unemployment is 0.05 ($p$ value=0.406).

To shed more light on the magnitude of the coefficients, we look at regional variation in income level (see Figure C4.5 in the Appendix). At the start of our sample period (in 1929), regional differences in income were as high as 105%. Our estimates predict that those residing in particularly wealthy regions (such as Middle Atlantic and the Pacific area) would rate the importance of meaning 0.34 points higher and the importance of income 0.29 points lower (on a 5-point scale) than similar individuals residing in the poorest region (the East South Central area).[10] In more recent years, percentage differences in regional income have decreased, but still amount to about 42%. There is also substantial variation in regional income over time (see Figure C4.5 in the Appendix).

In Table C5.7 in the Appendix we study whether it is really only macroeconomic conditions during the impressionable years (18-25 of age) that permanently affect job preferences, or whether macroeconomic conditions during other stages of one's life matter too. The regression results show a very consistent pattern: Macroeconomic conditions during the impressionable years matter most; those in other periods matter much less or not at all. Likewise, we investigated how our results change if we additionally control for the income level in each region at the time of the survey. While we find that current income matters too and in the same way as income during the impressionable years, our conclusions in Table 4.1 regarding the permanent effect of income during the impressionable years are not affected, either qualitatively, or quantitatively (see Table C5.8 in the Appendix of this chapter).

Finally, we ask whether the effects that we find in Table 4.1 persist into old age or decay over time. The regressions in Table 4.2 allow the effect of income during the impressionable years on job preferences to vary with current age. Results show that there is very little decay of the effect of macroeconomic conditions during the impressionable year over a person's lifetime. Our results thus suggest that there are long-run consequences of recessions and

[10]Note that a 0.34 increase in the average ranking of meaning is equivalent to 34% of the population putting meaning a full rank higher.

booms for job preferences: positive macroeconomic experiences during the impressionable years lead cohorts of workers to give higher priority to meaning and lower priority to income for the rest of their lives, while recessions have the opposite effect.

Table 4.2: Experienced income during the impressionable years and job preferences for meaning and income: lifetime decay

| | **Meaning** | **Income** |
|---|---|---|
| Income level 18-25 | 0.397 | -0.484 |
| | (0.291) | (0.289) |
| | [0.013**] | [0.116] |
| Income level 18-25 * 26-50 age group | -0.138 | 0.073 |
| | (0.206) | (0.209) |
| | [0.278] | [0.703] |
| Income level 18-25 * 51-75 age group | -0.057 | 0.170 |
| | (0.242) | (0.242) |
| | [0.709] | [0.432] |
| Household income | ✓ | ✓ |
| Years of education | ✓ | ✓ |
| Labor market status | ✓ | ✓ |
| Demographic variables | ✓ | ✓ |
| Age FE | ✓ | ✓ |
| Decade of birth FE | ✓ | ✓ |
| Year FE | ✓ | ✓ |
| Region FE | ✓ | ✓ |
| Region at 16 FE | ✓ | ✓ |
| N | 19,026 | 19,022 |
| F-value | 24.28 | 8.51 |
| R-squared | 0.161 | 0.069 |

*Notes: Regressions are estimated using OLS. The 'Income level 18-25' is log-linearized. Demographic variables include controls for gender, race, father and mother education, marital status, number of children, household size (squared), and household income at the age of 16. In parentheses, heteroskedasticity robust standard errors are reported. In brackets, p-values are reported estimated using the wild bootstrap procedure suggested by Cameron et al. (2008), by clustering standard errors at the level of the region at age 16. Since the number of clusters is small, the more conservative Webb weights are used (Webb, 2013), implemented using the boottest estimator developed by Roodman et al. (2019), with 5000 replications. Sample re-weighted using the wtssall population weights in the GSS. Significance levels: *** p<.01, ** p<.05, * p<.1.*

Our findings also contribute to the ongoing discussion about generational differences in preferences and beliefs. Contrary to the popular opinion that groups such as the Millennials share similar experiences and consequently develop similar preferences, we argue that

such conventional definitions of generations span too large a period, and overlook substantial variation in experienced macro-economic conditions among group members. As Figure C4.6 in the Appendix for this chapter helps illustrate, each generation has both 'lucky' and 'unlucky' individuals, who according to our findings will value different attributes in their job depending on their macroeconomic experiences during their impressionable years. Overlooking such differences in preferences for work among members of the same generation can have important consequences for the organization and efficiency of labor markets.

Our paper has shown that shared experiences during the impressionable years shape job preferences for life. Macroeconomic shocks (such as the IT boom or the Great Recession) can have long-lasting effects on what people value in work. This has possible repercussions for the dynamics of business cycles. When booms create cohorts of workers who care less about income and more about meaning, economic growth may slow down as a result of the revised priorities of the workforce in favor of non-monetary aspects. Conversely, economies may grow out of a recession faster as they produce cohorts of workers who put a high income first. Our results also suggest that mission-driven organizations such as public-sector organizations and not-for-profits may suffer less from labor market tightness during booms than typically thought, as over time young workers enter the labor market with a stronger desire for meaningful work as a result of the favorable macroeconomic circumstances. This may in turn affect the response of policymakers and voters to the business cycle. Last, and perhaps most importantly, our study points to an explanation for why some cohorts seem to be more focused on earning a high income, while other cohorts put priority on seeking meaning and purpose. Instead of some hard-to-explain, deeply ingrained, exogenous difference in preferences between cohorts, temporary macroeconomic conditions may be the key to understanding persistent generational differences.

# Appendix C

## C.1 General Social Survey

The General Social Survey has gathered data on attitudes and behaviors in contemporary American society in 30 waves, since 1973 and up to and including 2016. The study is a repeated cross-section on a representative sample of the adult US population, conducted through predominantly in-person interviews.

In this paper we focus on 18 waves, between 1973 and 2014, in which 21,000 respondents were asked to rank their preferences for five job attributes: job meaning, having a high income, opportunities for promotions, job security, and short working hours.

Respondents were given a card with the five job characteristics, labeled in this order as:

1. High income
2. No danger of being fired
3. Working hours are short, lots of free time
4. Chances for advancement
5. Work is important and gives a feeling of accomplishment

After reading the card, they were asked the following question: *"Would you please look at this card and tell me which one thing on this list you would most prefer in a job? Which comes next? Which is third-most important? Which is fourth-most important?"*

If a job attribute was not chosen, it was labelled as the fifth-most important attribute. In our analysis we re-code the ranks of preferences such that a higher rank corresponds to a higher importance of an attribute.

*Gender* is a dummy variable taking value 0 for males, and 1 for females. *Race* is a

categorical variable, divided into white, black, and other. *Marital status* is classified as married, widowed, divorced, separated, and never married. The *number of children* and the *household size* are numerical variables on a scale from 1 to 8 or more, and 1 to 16 respectively. *Labor market status* is a categorical variable divided into working full-time, working part-time, temporarily not working, unemployed, retired, in school, keeping house, or other. *Age* and *education* are continuous variables, where age runs from 18 to 75 in our selected sample, and years of education run from 0 to 20.

*Household income* represents the real family income in constant $US. When a respondent did not fill in an amount (7% of the relevant sample), we imputed their household income using responses on socio-demografic questions (respondent's education, labor market status, age, household size, gender, marital status), and dummies for survey year and region of residence at the time of the survey. In all our specifications we control for respondents whose income was imputed, using a binary indicator. Imputation is performed using the *impute* function in Stata.

*Birth decades* are defined using the birth year of each respondent, in intervals of 10 years between 1898 and 2000. According to this definition, 10 different generations exist in our sample, with the oldest generation including those born between 1904 and 1910, and the youngest generation being made up of respondents born between 1990 and 1998.

*Parent education* is captured by two numerical variables counting the years of education of the mother and of the father of each respondent, ranging from 0 to 20. When a respondent did not fill in a number (20% of the relevant sample for mother education and 30% for father education), we imputed their parents' education using the average mother and father education level in the sample. In all our specifications we control for respondents whose parents' education was imputed, using a binary indicator. Imputation is performed using the *impute* function in Stata.

*Household income at the age of 16* is defined as a categorical variable on a 5-point scale, ranging from "far below average" to "far above average". When a respondent did not fill in a category (7% of the relevant sample), we imputed their household income at the age of 16 using the average level in the sample. In all the relevant specifications we control for respondents whose income at the age of 16 was imputed, using a binary indicator. Imputation is performed using the *impute* function in Stata.

In the GSS, states are grouped into nine macro regions: 1. New England (Maine, Vermont, New Hampshire, Massachusetts, Connecticut, Rhode Island), 2. Middle Atlantic (New York, New Jersey and Pennsylvania), 3. East North Central (Wisconsin, Illinois, Indiana, Michigan and Ohio), 4. West North Central (Minnesota, Iowa, Missouri, North Dakota, South Dakota, Nebraska, Kansas), 5. South Atlantic (Delaware, Maryland, West Virginia, Virginia, North Carolina, South Carolina, Georgia, Florida, District of Columbia), 6. East South Central (Kentucky, Tennessee, Alabama, Mississippi), 7. West South Central (Arkansas, Oklahoma, Louisiana, Texas), 8. Mountain (Montana, Idaho, Wyoming, Nevada, Utah, Colorado, Arizona, New Mexico), 9. Pacific (Washington, Oregon, California, Alaska, Hawaii).

Those who only moved to the US after the age of 16 are coded as foreigners (5.4%). Since we do not know whether these respondents were in the US during their impressionable years, their experiences in that period are unknown and they are not included in the sample.

## C.2 Income and unemployment

The U.S. Bureau of Economic Analysis (BEA) provides yearly data on state-level personal income (SAINC1 Personal Income Summary: Personal Income, Population, Per Capita Personal Income) since 1929.

The Bureau of Labor Statistics provides yearly data on the unemployment rate at the state

level since 1976. Since using this measure would restrict our sample size significantly, in regressions with unemployment experience during the impressionable years we use national-level data on unemployment. National unemployment rates are available from the BLS since 1929.

## C.3 Constructing experienced income during the impressionable years

Income data spans from 1929 to 2016. As the BEA data is at the state level, we use state-level income per capita and state level-population to calculate the regional income per capita:

$$IncCapR_{j,t} = \frac{\sum^{i} IncCapS_{i,t} * PopS_{i,t}}{\sum^{i} PopS_{i,t}}$$

where income per capita in each state $i$ in region $j$ at time $t$ ($IncCapS_{i,t}$) is weighted by the population of each state $i$ at time $t$ $(PopS_{i,t})$ in region $j$ to obtain the regional income per capita $IncCapR_{j,t}$.

In the next step, the regional income per capita is adjusted to control for inflation. To do this, we re-weight regional income per capita using data on US national-level CPI factors since 1929. We choose US$2017 as the base, and adjust regional income per capita with the corresponding factor of 245.1, such that:

$$IncCapR^{adj}_{j,t} = \frac{IncCapR_{j,t} * 245.1}{cpi_t}$$

where $cpi_t$ is the consumer price index each year, between 1929 and 2014.

Next, using the date of birth of each respondent in the survey and the year of the survey, we identify the years in which individuals were between 18 and 25 years of age. Knowing $IncCapR^{adj}_{j,t}$ each year between 1929 and 2016, we create the average experienced income

during the impressionable years, such that:

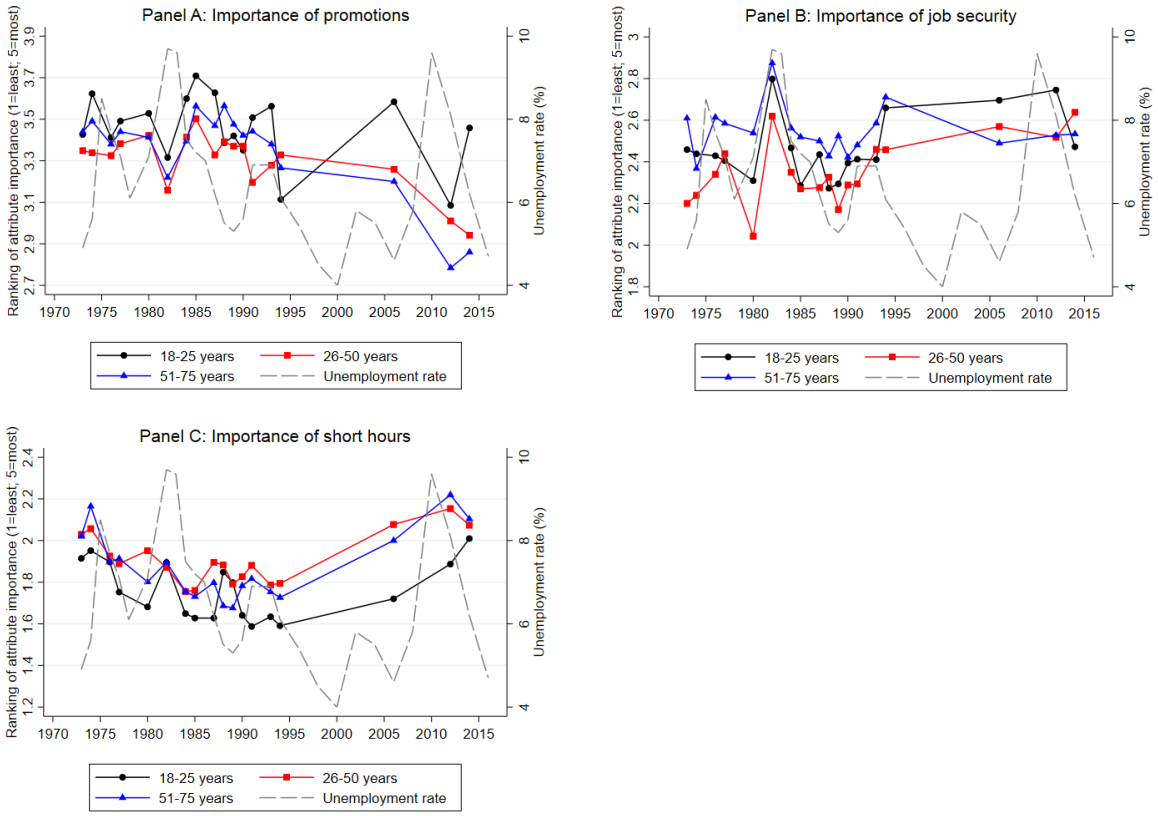
$$Exp_{i,r,t}^{18-25} = \frac{\sum_{t=1}^{T} IncCapR_{j,t}^{adj}}{T}$$

where $Exp_{i,r,t}^{18-25}$ is the average of the adjusted regional income per capita in each of the eight years when respondent $i$ was in their impressionable years (between 18 and 25 years old). When a respondent is below 25 at the time of the survey, the experience is a weighted average of income in the subset of years between 18 and up to the current age. The independent variable of interest used throughout the paper is the logarithm of $Exp_{i,r,t}^{18-25}$.

## C.4 Additional Figures

### C.4.1 Preferences for promotions, job security, and short hours across different age groups and over time

Figure C.4.1: Reported importance for promotions, job security, and short hours for three different age groups

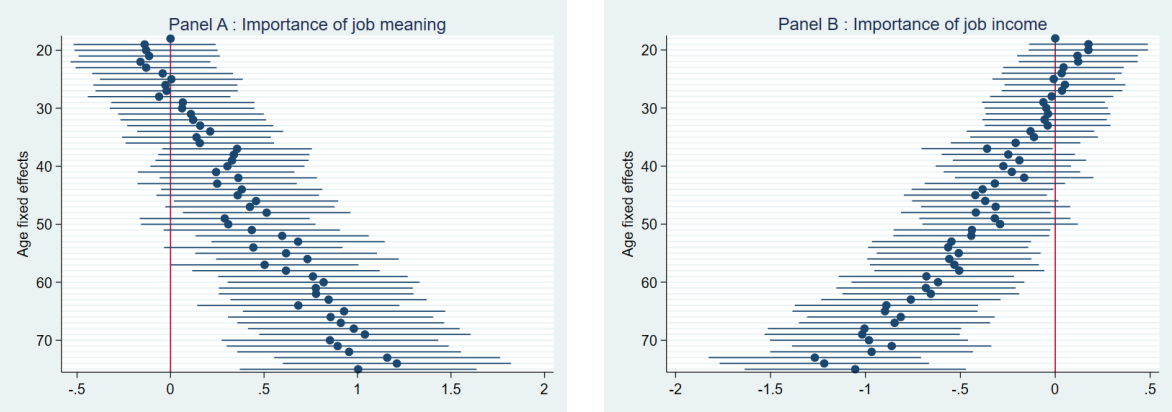



*Notes: Preferences for promotion (Panel A), job security (Panel B), and short hours (Panel C) across different age groups and over time. Note: Based on a sample of 19,000 respondents who ranked preferences for job attributes in 18 waves, between 1973 and 2014. Preferences are ranked by respondents on a scale from 1 (least important) to 5 (most important). Right axis plots national unemployment rate.*

### C.4.2-C.4.4: Age, year, and decade-of-birth fixed effects

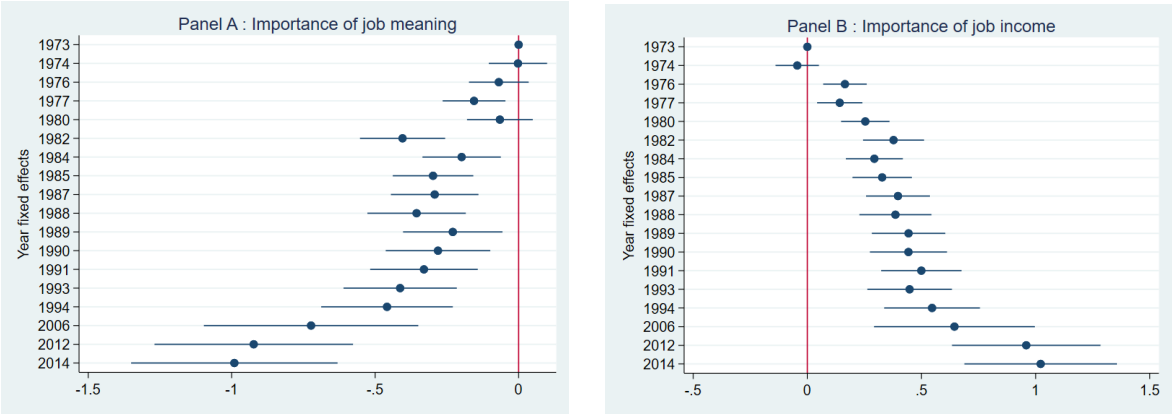
Figures plot coefficients and standard errors of age, time, and birth-decade fixed effects from estimating equation (4.1). Figure C4.2 shows strong life-cycle effects: As respondents become older, meaning starts playing a more important role, at the expense of how important having a high income is. Figure C4.3 show strong time trends: Over time, having a high income has become more important for all Americans at the expense of how important meaning is. Figure C4.4 shows that while income has become slightly more important and meaning has become slightly less important for more recent generations, the decade of birth plays a less substantial role in the ranking of preferences for meaning and income.

Figure C4.2: Age fixed effects



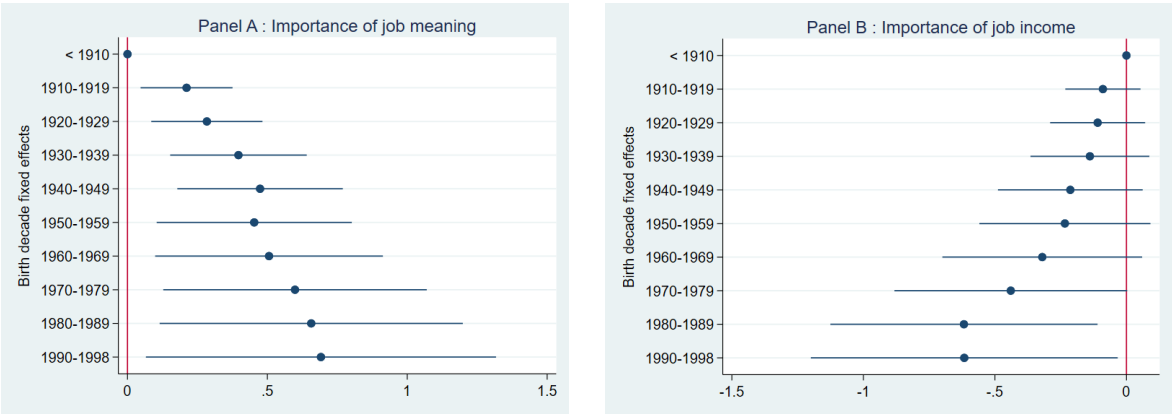

*Notes: Figure shows coefficients and standard errors of age fixed effects from estimating equation (4.1). Panel A shows results for importance of job meaning and Panel B for income. Reference group is age 18.*

Figure C4.3: Year fixed effects




*Notes: Figure shows coefficients and standard errors of year-of-survey fixed effects from estimating equation (4.1). Panel A shows results for importance of job meaning and Panel B for income. Reference group is year 1973.*

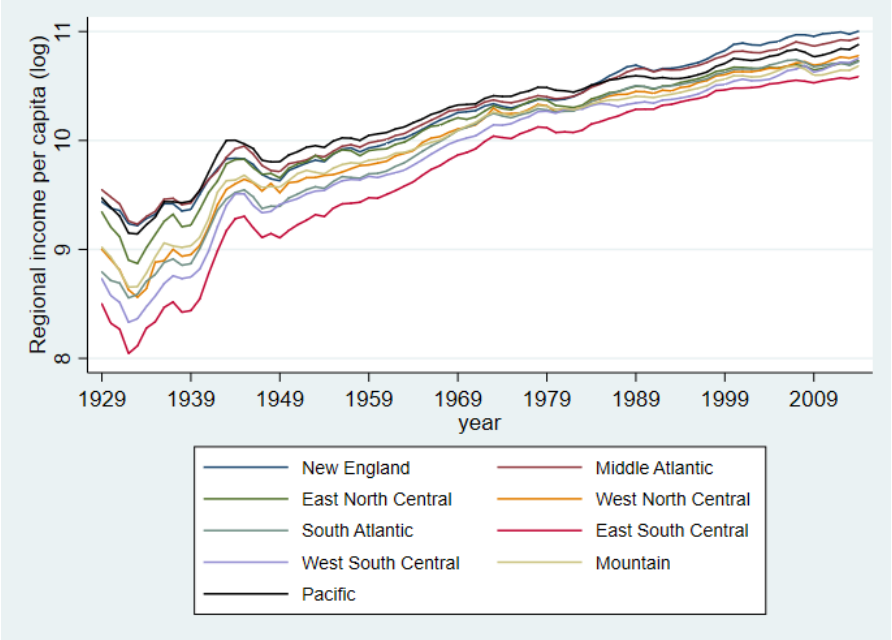
Figure C4.4: Birth-decade fixed effects




*Notes: Figure shows coefficients and standard errors of decade-of-birth fixed effects from estimating equation (4.1). Panel A shows results for importance of job meaning and Panel B for income. Reference group is those born before 1910.*

## C.4.5: Income per capita across the nine US regions between 1929 and 2014

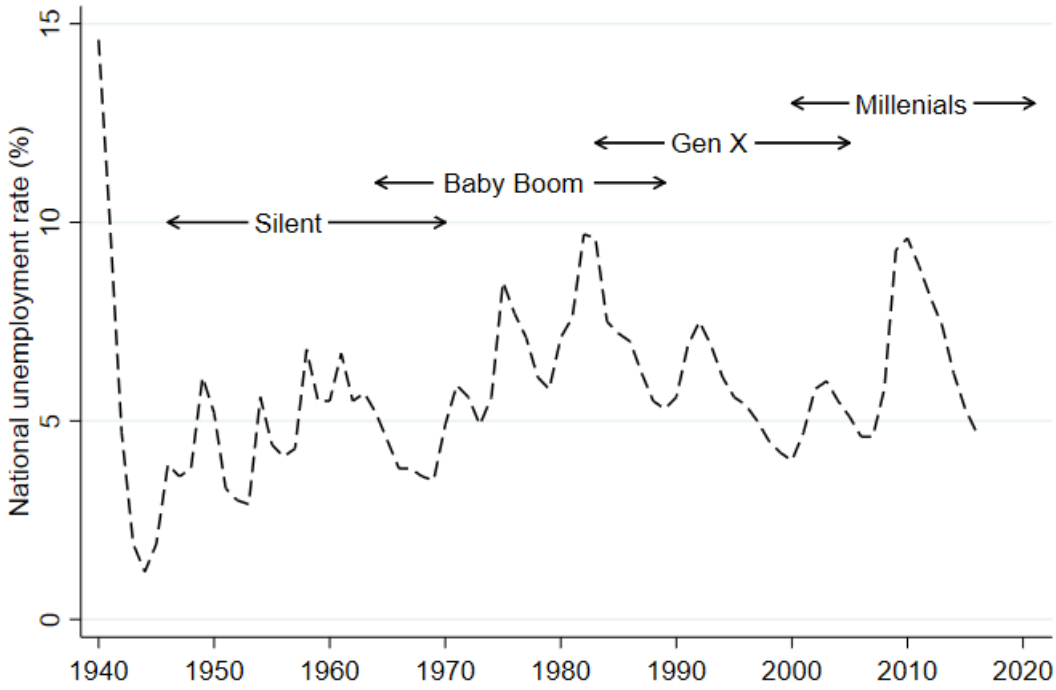
Figure C.4.5: The fluctuations in regional income per capita in the nine US regions between 1929 and 2014




*Notes: The figure plots the logarithm of regional income per capita each year, between 1929 and 2014. The regional income per capita is adjusted for inflation and expressed in US$2017. The nine US regions in the General Social Survey are defined in Appendix C.1.*

## C.4.6: The national unemployment rate and the "impressionable years" for different generations

Figure C.4.6: Different generations in their "impressionable years" (or aged 18-25), compared with the fluctuations in the national unemployment rate within that period.




*Notes: The figure shows the national unemployment rate (%) within the periods in which members of different generations were experiencing their "impressionable years", covering the period between 1940 and 2016. We define generations according to the Pew Research Center (2018) (see also Koczanski and Rosen, 2019) in the following way: the Silent Generation (1928-1945), Baby Boomers (1946-1964), Gen X (1965-1980), and Millenials (1981-1996).*

## C.5 Additional Tables

### C.5.1 Descriptive statistics

Table C5.1: Descriptive Statistics

| | Mean | Standard deviation | N |
|---|---|---|---|
| **Preferences** | | | |
| Meaning | 3.93 | 1.29 | 19,026 |
| Income | 3.43 | 1.52 | 19.020 |
| Promotions | 3.35 | 1.20 | 19,020 |
| Security | 2.41 | 1.22 | 19,020 |
| Hours | 1.89 | 1.15 | 19.018 |
| **Socio-Demographics** | | | |
| Male | 0.45 | 0.50 | 19.026 |
| Years education | 12.65 | 3.00 | 19.026 |
| Age | 42.80 | 15.38 | 19.026 |
| Birth year | 1944.53 | 18.71 | 19.026 |
| Annual income | 31,539.96 | 26,557.01 | 19,026 |
| Household size | 2.87 | 1.56 | 19,026 |
| No. children | 1.97 | 1.82 | 19,026 |
| % Married | 0.58 | 0.49 | 19,026 |
| % White | 0.83 | 0.37 | 19,026 |
| % Full-time employed | 0.52 | 0.50 | 19,026 |
| % Part-time employed | 0.11 | 0.31 | 19,026 |
| % Temporarily not working | 0.02 | 0.15 | 19,026 |
| % Unemployed | 0.03 | 0.18 | 19,026 |
| % Retired | 0.09 | 0.29 | 19,026 |
| % In school | 0.03 | 0.18 | 19,026 |
| % Keeping house | 0.18 | 0.39 | 19,026 |
| Mother years education | 10.54 | 3.60 | 15,264 |
| Father years education | 10.17 | 4.29 | 13,195 |
| Household income at 16 (1-5) | 2.78 | 0.85 | 17,665 |
| **Experiences 18-25** | | | |
| Unemployment | 6.78 | 3.54 | 19,026 |
| Income (2017$US) | 23,913.76 | 9,757.81 | 19,026 |

## C.5.2: Restrict analysis to non-movers

In the main text, we assume that people who live in a certain region at age 16 still live there during their impressionable years – but we only know where they live at age 16 and at the time of the survey. To test the robustness of our assumption, in Table C5.2, we restrict the sample to those respondents who at the time of the survey live in the same region that they did when they were 16 years old, assuming that these individuals never moved regions. With 79% of the original sample residing in the same place (as they did when they were 16) at the time of the survey, some power is lost. However, all coefficients are similar and remain statistically significant, indicating that this type of selection does not present a threat to the validity of our results.

Table C5.2: Experienced income during the impressionable years and preferences for income and meaning: non-movers

| | **Meaning** | **Income** |
|---|---|---|
| Income level 18-25 | 0.368*** | -0.298** |
| | (0.131) | (0.121) |
| | [0.011**] | [0.001***] |
| Household income | ✓ | ✓ |
| Years of education | ✓ | ✓ |
| Labor market status | ✓ | ✓ |
| Demographic variables | ✓ | ✓ |
| Age FE | ✓ | ✓ |
| Decade of birth FE | ✓ | ✓ |
| Year FE | ✓ | ✓ |
| Region at 16 FE | ✓ | ✓ |
| N | 14,982 | 14,980 |
| F-value | 19.17 | 6.97 |
| R-squared | 0.150 | 0.066 |

*Notes: Regressions are estimated using OLS. The 'Income level 18-25' is log-linearized. Demographic variables include controls for gender, race, father and mother education, marital status, number of children, household size (squared), and household income at the age of 16. In parentheses, heteroskedasticity robust standard errors are reported. In brackets, p-values are reported estimated using the wild bootstrap procedure suggested by Cameron et al. (2008), by clustering standard errors at the level of the region at age 16. Since the number of clusters is small, the more conservative Webb weights are used (Webb, 2013), implemented using the boottest estimator developed by Roodman et al. (2019), with 5000 replications. Sample re-weighted using the wtssall population weights in the GSS. Significance levels: *** p<.01, ** p<.05, * p<.1.*

## C.5.3: Average income vs. standard deviation of income

The standard deviation of income during the impressionable years is calculated for all respondents in our sample. To keep the interpretation comparable and simple, we use the logarithm of the measure. The small difference in observations between Table C5.3 and Table 4.1 is caused by the fact that for those respondents who are 18 at the time of the survey and have only had one impressionable year, there is no variance in experience. Hence, these subjects are left out of the specification.

Table C5.3: Experienced income during the impressionable years and preferences for income and meaning: average income vs. standard deviation of income

| | **Meaning** | **Income** |
|---|---|---|
| Income level 18-25 | 0.324 | -0.255 |
| | (0.120) | (0.109) |
| | [0.014**] | [0.002***] |
| Standard deviation of income 18-25 | 0.022 | -0.032 |
| | (0.027) | (0.025) |
| | [0.410] | [0.234] |
| Household income | ✓ | ✓ |
| Years of education | ✓ | ✓ |
| Labor market status | ✓ | ✓ |
| Demographic variables | ✓ | ✓ |
| Age FE | ✓ | ✓ |
| Decade of birth FE | ✓ | ✓ |
| Year FE | ✓ | ✓ |
| Region at 16 FE | ✓ | ✓ |
| N | 18,903 | 18,899 |
| F-value | 24.58 | 8.54 |
| R-squared | 0.163 | 0.068 |

*Notes: Regressions are estimated using OLS. The 'Income level 18-25' is log-linearized. Demographic variables include controls for gender, race, father and mother education, marital status, number of children, household size (squared), and household income at the age of 16. In parentheses, heteroskedasticity robust standard errors are reported. In brackets, p-values are reported estimated using the wild bootstrap procedure suggested by Cameron et al. (2008), by clustering standard errors at the level of the region at age 16. Since the number of clusters is small, the more conservative Webb weights are used (Webb, 2013), implemented using the boottest estimator developed by Roodman et al. (2019), with 5000 replications. Sample re-weighted using the wtssall population weights in the GSS. Significance levels: *** p<.01, ** p<.05, * p<.1.*

## C.5.4: Unemployment experience and preferences for job attributes

We analyze how unemployment experienced during the impressionable years relates to job preferences at the time of the survey. To do so, we use US-level unemployment rates since 1929 and up to 2016. Since there is no regional variation in these unemployment rates, the variation in experiences comes from age differences at the time of the survey. We estimate the same regression as in equation (4.1), but with national unemployment rate during the impressionable years.

Table C5.4 shows that in line with the findings in Table 4.1 in the main text using experienced income per capita, a similar substitution between income and meaning is observed when using experienced unemployment rate. Experiencing higher unemployment during the impressionable years leads to ranking the importance of income higher at the time of the survey, at the expense of meaning.

Table C5.4: Experienced unemployment during the impressionable years and preferences for income and meaning

| | **Meaning** | **Meaning** | **Income** | **Income** |
|---|---|---|---|---|
| Unemployment level 18-25 | -0.009 | -0.012* | 0.014** | 0.015** |
| | (0.007) | (0.007) | (0.006) | (0.006) |
| | [0.136] | [0.022**] | [0.018**] | [0.005***] |
| Household income | ✓ | X | ✓ | X |
| Years of education | ✓ | X | ✓ | X |
| Labor market status | ✓ | X | ✓ | X |
| Demographic variables | ✓ | ✓ | ✓ | ✓ |
| Age FE | ✓ | ✓ | ✓ | ✓ |
| Decade of birth FE | ✓ | ✓ | ✓ | ✓ |
| Year FE | ✓ | ✓ | ✓ | ✓ |
| Region FE | ✓ | ✓ | ✓ | ✓ |
| Region at 16 FE | ✓ | ✓ | ✓ | ✓ |
| N | 19,026 | 19,026 | 19,022 | 19,022 |
| F-value | 24.53 | 18.75 | 8.59 | 8.52 |
| R-squared | 0.161 | 0.117 | 0.068 | 0.057 |

*Notes: Regressions are estimated using OLS. Demographic variables include controls for gender, race, father and mother education, marital status, number of children, household size (non-linearly), and household income at the age of 16. In parentheses, heteroskedasticity robust standard errors are reported. In brackets, p-values are reported estimated using the wild bootstrap procedure suggested by Cameron et al. (2008), by clustering standard errors at the level of the region at age 16. Since the number of clusters is small, the more conservative Webb weights are used (Webb, 2013), implemented using the boottest estimator developed by Roodman et al. (2019), with 5000 replications. Sample re-weighted using the wtssall population weights in the GSS. Significance levels: *** p<.01, ** p<.05, * p<.1.*

## C.5.5: Alternative definition of generations and age categories

In our main results, we define a birth cohort as all those individuals born in the same decade, in order to avoid the well-know collinearity issue of age, birth year, and survey year. However, one could alternatively choose to control for birth-year fixed effects and replace the age fixed effects with broader categories. To ensure that our results are not dependent on the (somewhat arbitrary) delimitation of birth cohorts, we employ an additional robustness check. Columns (1) and (3) in Table C5.5 replace birth decades with categories of birth years in groups of 5 years. Columns (2) and (4) replace birth decades with fixed effects for each birth year, and groups the age fixed effects in categories of 5 years instead.

Our results are robust to both alternative specifications, indicating that our results hold regardless of the chosen specification for respondent age and birth year.

Table C5.5: Income experiences during the impressionable years and preferences for meaning and income: alternative specifications for birth and age fixed effects

| | **Meaning** | **Meaning** | **Income** | **Income** |
|---|---|---|---|---|
| Income level 18-25 | 0.325 | 0.343 | -0.241 | -0.226 |
| | (0.121) | (0.130) | (0.110) | (0.118) |
| | [0.034**] | [0.004***] | [0.012**] | [0.023**] |
| Household income | ✓ | ✓ | ✓ | ✓ |
| Years of education | ✓ | ✓ | ✓ | ✓ |
| Labor market status | ✓ | ✓ | ✓ | ✓ |
| Demographic variables | ✓ | ✓ | ✓ | ✓ |
| Age FE | ✓ | X | ✓ | X |
| Age groups (intervals of 5) | X | ✓ | X | ✓ |
| Birth year FE | X | ✓ | X | ✓ |
| Birth year groups (intervals of 5) | ✓ | X | ✓ | X |
| Year FE | ✓ | ✓ | ✓ | ✓ |
| Region FE | ✓ | ✓ | ✓ | ✓ |
| Region at 16 FE | ✓ | ✓ | ✓ | ✓ |
| N | 19,026 | 19,026 | 19,022 | 19,022 |
| F-value | 23.15 | 19.76 | 8.21 | 7.34 |
| R-squared | 0.162 | 0.163 | 0.069 | 0.071 |

*Notes: Regressions are estimated using OLS. The 'Income level 18-25' is log-linearized. Demographic variables include controls for gender, race, father and mother education, marital status, number of children, household size (squared), and household income at the age of 16. In parentheses, heteroskedasticity robust standard errors are reported. In brackets, p-values are reported estimated using the wild bootstrap procedure suggested by Cameron et al. (2008), by clustering standard errors at the level of the region at age 16. Since the number of clusters is small, the more conservative Webb weights are used (Webb, 2013), implemented using the boottest estimator developed by Roodman et al. (2019), with 5000 replications. Sample re-weighted using the wtssall population weights in the GSS. Significance levels: *** p<.01, ** p<.05, * p<.1.*

## C.5.6: Experienced income during the impressionable years and preferences for promotions, job security, and short working hours

We estimate equation (4.1), where the dependent variables are the ranked preferences (on a 5-point scale) for promotions, job security, and short working hours. There doesn't appear to be a meaningful relationship between experienced income during "Impressionable years" and the remaining three job preferences. The coefficients are substantially smaller than in Table 4.1, and they are not significant at any conventional level.

Table C5.6: Income experiences during the impressionable years and preferences for promotions, job security, and short working hours

| | **Promotions** | **Security** | **Short hours** |
|---|---|---|---|
| Income level 18-25 | 0.162 | -0.038 | -0.172 |
| | (0.108) | (0.107) | (0.104) |
| | [0.125] | [0.715] | [0.235] |
| Household income | ✓ | ✓ | ✓ |
| Years of education | ✓ | ✓ | ✓ |
| Labor market status | ✓ | ✓ | ✓ |
| Demographic variables | ✓ | ✓ | ✓ |
| Age FE | ✓ | ✓ | ✓ |
| Birth decade FE | ✓ | ✓ | ✓ |
| Year FE | ✓ | ✓ | ✓ |
| Region FE | ✓ | ✓ | ✓ |
| Region at 16 FE | ✓ | ✓ | ✓ |
| N | 19,021 | 19,021 | 19,019. |
| F-value | 6.44 | 13.01 | 4.89 |
| R-squared | 0.052 | 0.094 | 0.041 |

*Notes: Regressions are estimated using OLS. The 'Income level 18-25' is log-linearized. Demographic variables include controls for gender, race, father and mother education, marital status, number of children, household size (squared), and household income at the age of 16. In parentheses, heteroskedasticity robust standard errors are reported. In brackets, p-values are reported estimated using the wild bootstrap procedure suggested by Cameron et al. (2008), by clustering standard errors at the level of the region at age 16. Since the number of clusters is small, the more conservative Webb weights are used (Webb, 2013), implemented using the boottest estimator developed by Roodman et al. (2019), with 5000 replications. Sample re-weighted using the wtssall population weights in the GSS. Significance levels: *** p<.01, ** p<.05, * p<.1.*

## C.5.7: Impressionable years vs. other years

Following research in psychology, we focus on the impressionable years between 18 and 25. In the following section we investigate the possibility that the impressionable years are not the only important period in shaping job preferences. We separately investigate the effects of experienced income between different ages, in similarly constructed intervals of eight years. Specifically, we look at two additional intervals prior to the impressionable years (ages 0-9 and ages 10-17), and two equal-length intervals after them (ages 26-33 and ages 34-41). To assess the effect of experiences at each of these ages on preferences for income and meaning requires additional restrictions on the sample size. Specifically, controlling for experienced macro-economic conditions after the 'Impressionable years' (26-33 and 34-41) mechanically restricts the sample to those individuals who are at least as old as that. For this reason, we do not incorporate all five experience variables in one model but instead look at their effect on preferences for meaning and income separately.

As the alternative "impressionable years" are further away from the age of 16, the likelihood that the individual moved between the age of 16 and the time of the survey is higher. In line with Giuliano and Spilimbergo (2014), we address this issue by restricting the sample to those individuals who did not move between the age of 16 and the time of the survey. As results in Table C5.2 indicate, these non-movers appear to be representative of the whole sample. Columns (1) and (3) in Table C5.7 show that, in general, experiences during years other than at age 18 to 25 do not appear to explain preferences for job attributes. In line with Giuliano and Spilimbergo (2014), we add in columns (2) and (4) experienced income during the impressionable years. In these "horse races", the impressionable years are almost without exception the most important.

Table C5.7: Experienced income during other years and preferences for income and meaning

| | **Meaning** | **Meaning** | **Income** | **Income** |
|---|---|---|---|---|
| **Panel A: Ages 0-9** | | | | |
| Income level 0-9 | 0.062 | -0.115 | 0.032 | 0.215 |
| | (0.114) | (0.128) | (0.104) | (0.120) |
| | [0.563] | [0.250] | [0.821] | [0.215] |
| Income level 18-25 | | 0.627 | | -0.650 |
| | | (0.232) | | (0.219) |
| | | [0.028**] | | [0.005***] |
| N | 13,298 | 13,298 | 13,296 | 13,296 |
| **Panel B: Ages 10-17** | | | | |
| Income level 10-17 | 0.221 | 0.053 | -0.101 | 0.035 |
| | (0.109) | (0.126) | (0.103) | (0.119) |
| | [0.046**] | [0.489] ] | [0.184] | [0.741] |
| Income level 18-25 | | 0.428 | | -0.346 |
| | | (0.171) | | (0.160) |
| | | [0.008***] | | [0.012**] |
| N | 14,454 | 14,454 | 14,452 | 14,452 |
| **Panel C: Ages 26-33** | | | | |
| Income level 26-33 | 0.398 | 0.203 | -0.020 | 0.369 |
| | (0.180) | (0.253) | (0.163) | (0.233) |
| | [0.165] | [0.227] | [0.880] | [0.014**] |
| Income level 18-25 | | 0.186 | | -0.503 |
| | | (0.184) | | (0.176) |
| | | [0.203] | | [0.006***] |
| N | 12.690 | 12,690 | 12.688 | 12,688 |
| **Panel D: Ages 34-41** | | | | |
| Income level 34-41 | 0.648 | 0.356 | -0.055 | 0.130 |
| | (0.232) | (0.314) | (0.207) | (0.291) |
| | [0.090*] | [0.410] | [0.812] | [0.684] |
| Income level 18-25 | | 0.203 | | -0.278 |
| | | (0.181) | | (0.171) |
| | | [0.033**] | | [0.078*] |
| N | 9,672 | 9,672 | 9,670 | 9,670 |
| Household income | ✓ | ✓ | ✓ | ✓ |
| Years of education | ✓ | ✓ | ✓ | ✓ |
| Labor market status | ✓ | ✓ | ✓ | ✓ |
| Demographic variables | ✓ | ✓ | ✓ | ✓ |
| Age FE | ✓ | ✓ | ✓ | ✓ |
| Decade of birth FE | ✓ | ✓ | ✓ | ✓ |
| Year FE | ✓ | ✓ | ✓ | ✓ |
| Region at 16 FE | ✓ | ✓ | ✓ | ✓ |

*Notes: Regressions are estimated using OLS. The 'Income level 18-25' is log-linearized. Demographic variables include controls for gender, race, father and mother education, marital status, number of children, household size (squared), and household income at the age of 16. In parentheses, heteroskedasticity robust standard errors are reported. In brackets, p-values are reported estimated using the wild bootstrap procedure suggested by Cameron et al. (2008), by clustering standard errors at the level of the region at age 16. Since the number of clusters is small, the more conservative Webb weights are used (Webb, 2013), implemented using the boottest estimator developed by Roodman et al. (2019), with 5000 replications. Sample re-weighted using the wtssall population weights in the GSS. Significance levels: *** p<.01, ** p<.05, * p<.1.*

## C.5.8: Income level at ages 18-25 vs. income level at the time of the survey

Table C5.8: Experienced income during the impressionable years and preferences for income and meaning: Income level 18-25 vs. income level at the time of the survey

| | **Meaning** | **Income** |
|---|---|---|
| Income level 18-25 | 0.325 | -0.286 |
| | (0.113) | (0.104) |
| | [0.002***] | [0.010**] |
| Income level at survey | 0.503 | -0.182 |
| | (0.314) | (0.303) |
| | [0.047**] | [0.410] |
| Household income | ✓ | ✓ |
| Years of education | ✓ | ✓ |
| Labor market status | ✓ | ✓ |
| Demographic variables | ✓ | ✓ |
| Age FE | ✓ | ✓ |
| Decade of birth FE | ✓ | ✓ |
| Year FE | ✓ | ✓ |
| Region at 16 FE | ✓ | ✓ |
| N | 19,026 | 19,022 |
| F-value | 24.47 | 8.54 |
| R-squared | 0.162 | 0.068 |

*Notes: Regressions are estimated using OLS. The 'Income level 18-25' is log-linearized. Demographic variables include controls for gender, race, father and mother education, marital status, number of children, household size (squared), and household income at the age of 16. In parentheses, heteroskedasticity robust standard errors are reported. In brackets, p-values are reported estimated using the wild bootstrap procedure suggested by Cameron et al. (2008), by clustering standard errors at the level of the region at age 16. Since the number of clusters is small, the more conservative Webb weights are used (Webb, 2013), implemented using the boottest estimator developed by Roodman et al. (2019), with 5000 replications. Sample re-weighted using the wtssall population weights in the GSS. Significance levels: *** p<.01, ** p<.05, * p<.1.*

# Summary

There is substantial diversity in the way work is organized and in the pecuniary and non-pecuniary job attributes that employers offer. This variation is partly driven by the fact that individuals have heterogeneous work preferences and as a consequence respond differently to incentive schemes. As a result, firms are faced with the complex problem of identifying the most effective ways to motivate their employees, given the composition of their workforce. Monetary incentives such as higher wages and performance bonuses are costly and sometimes fall short of achieving the desired effects. In response, a growing number of firms and non-profit organizations also offer non-monetary incentives in order to attract workers and motivate them to increase performance. However, little is known about why workers desire different job attributes, why they respond differently to incentives, and how their environment growing up shapes their skills and work preferences. This thesis has put forward three chapters to fill this gap by studying how preferences for different job attributes form, how and why workers respond to non-monetary incentives, and how the nature of the educational environment growing up can shape the cognitive and non-cognitive skills of children.

In Chapter 2 I designed a large-scale a field experiment with 900 teachers in 39 schools, to study how repeated public praise for the best teachers impacts their performance, over the course of an entire academic year. Teachers in the treatment group who were praised in the first round performed significantly better in subsequent months, while teachers who were not praised in the first round performed significantly worse following the intervention. Repeating

the intervention did not have any additional effect on teacher performance. I investigated whether teachers attempted to game the performance measure following the intervention, by manipulating the grades they gave their own students. I did so by comparing this subjective measure (grades teachers gave their own students) with an objective one; namely, student performance on high-stake standardized and anonymously graded final exams. There was no evidence that the grading of teachers in the treated group became more noisy following the intervention.

The positive effects of unannounced praised were persistent and reflected real increases in student exam performance, emphasizing the importance of teacher effort in shaping student educational outcomes. On the other hand, the negative effects of unannounced praise faded away over time and did not significantly affect the exam performance of students. The results were best explained by a mechanism where public praise sends a comparative message about performance. Updating their beliefs, teachers became more motivated if they received good news through praise, and became discouraged when the news was bad. This suggested that in settings where workers are intrinsically motivated the trade-off between motivation and discouragement will be less pronounced, because in the long-run it will be easier to motivate than to demotivate with public praise.

Chapter 3 estimated the effect of assigning children to academic tracks at early ages, by following a representative sample of children in the Dutch province of Limburg. We studied the effects of assignment to the academic track at the achievement margin on both cognitive and non-cognitive skills, and across relative age. It is well established in the literature that younger students in class have a lower probability of assignment to higher tracks. That is because they are less mature at the time of track assignment, and these differences in maturity are often mistaken as underlying differences in ability. We used a fuzzy regression discontinuity design that exploited school-specific admission thresholds to estimate the effect of attending the academic track at the achievement margin. Moreover, we identified interac-

tions with relative age to see whether relatively older students who were 'over-tracked' were hurt in the long-run.

We found no effect of tracking on cognitive outcomes, across relative age. However, we did find that attending the higher track increased perseverance, need for achievement, and emotional stability for the older students. The results showed that placing relatively more mature students in a learning environment that is challenging given their cognitive ability can have positive spillovers on their non-cognitive skills and on the effort they put into learning. We argued that these positive spillovers on the non-cognitive skills of relatively older students mediate the expected complementarity between ability and academic track attendance and explain why older students do no perform worse on cognitive tests, despite being more often tracked above their ability level. Given the increased importance of non-cognitive skills in determining labor-market outcomes, we argue that effects on both cognitive and non-cognitive skills should be taken into account when assessing the effectiveness of early tracking.

Chapter 4 investigated how experienced macroeconomic conditions when young shape job preferences later in life by using variation across regions and across time for a representative sample of US respondents. More specifically, we used variation in experienced regional income per capita dating back to the 1920s to explain the variation in preferences for having a high income and for doing work that is meaningful at the time of the survey. We focused on experiences during the so-called 'Impressionable Years', as research in both Economics and Psychology has shown that this period is particularly important for the formation of beliefs and preferences.

We found that job preferences vary in systematic ways with macroeconomic conditions: those who enter adulthood during 'good times' and experience relatively high levels of income per capita when young are much more likely to rate the importance of having a high income lower, and the importance of meaningful work higher. Our results were particularly

pronounced for young people, but we showed that macroeconomic conditions during the impressionable years (18-25 years old) have permanent effects on job preferences throughout an individual's life. Our results suggested that recessions and booms permanently shape the job preferences of different cohorts, and so introduced a framework for explaining generational differences in attitudes towards work.

This collection of essays has shown that individuals value non-pecuniary work aspects, and that non-monetary incentives can be effective in settings where employees perform cognitively complex tasks and where increasing performance is difficult. While these effects can be very persistent and result in sustained increases in effort for some workers, they also impose a trade-off for managers and should be considered carefully. However, the evidence suggests that non-monetary incentives can be used more effectively when employees have certain traits. For example, Chapter 2 indicates that intrinsically motivated workers are less likely to try to manipulate the performance measure in order to get the reward and more likely to overcompensate in the long run even if they are demotivated in the short-run by not being rewarded. This over-compensation appears to be driven by the pro-social aspects of a job and suggests that workers internalize the negative externalities that a decrease in their effort will impose on third parties.

The findings of this thesis also speak to a growing interest in using behavioral and experimental methods to better understand how the labor market outcomes of individuals are influenced by their preferences and personality traits. To this end, I have argued that because important events when young can change preferences and personality traits permanently it is important to study individuals throughout their life-cycle in order to understand all the factors that determine economic outcomes later in life. I believe that researchers can learn more about this process by exploiting variation in the exposure to economic and social shocks, as they can be remarkably powerful in shaping individuals. This is particularly true if these shocks occur when said preferences and personality traits are still malleable because individ-

uals have not yet reached full maturity. Future research should assess the magnitude of the social inefficiencies caused by shocks with persistent effects and the extent to which policy makers can intervene to mediate such issues.

# Nederlandse Samenvatting (Summary in Dutch)

Er is een grote diversiteit in de manier waarop werk is georganiseerd en in de financiële en niet-financiële functie-eigenschappen die werkgevers aanbieden. Deze variatie wordt gedeeltelijk gedreven door het feit dat individuen heterogene werkvoorkeuren hebben en daardoor anders reageren op stimuleringsregelingen. Als gevolg hiervan worden bedrijven geconfronteerd met het complexe probleem om te identificeren wat de meest effectieve manieren zijn om hun werknemers te motiveren, gezien de samenstelling van hun personeel. Monetaire prikkels zoals hogere lonen en prestatiebonussen zijn duur en schieten soms tekort bij het bereiken van de gewenste effecten. In reactie daarop bieden een groeiend aantal bedrijven en non-profit organisaties ook niet-monetaire prikkels om werknemers aan te trekken en te motiveren om de prestaties te verbeteren. Er is echter weinig bekend over waarom werknemers verschillende functie-eigenschappen wensen, waarom ze anders reageren op prikkels en hoe hun omgeving tijdens opgroeien hun vaardigheden en werkvoorkeuren vormt. Dit proefschrift heeft drie hoofdstukken naar voren gebracht om dit gat op te vullen door te bestuderen hoe voorkeuren voor verschillende functie-eigenschappen ontstaan, hoe en waarom werknemers reageren op niet-monetaire prikkels en hoe de aard van de onderwijsomgeving tijdens opgroeien de cognitieve en niet-cognitieve vaardigheden van kinderen kan vormen.

In hoofdstuk 2 ontwierp ik een grootschalig veldexperiment met 900 leraren in 39 sc-

holen om te bestuderen hoe herhaaldelijk publiekelijk lof voor de beste leraren hun prestaties beïnvloedt, gedurende een heel academisch jaar. Leraren in de behandelingsgroep die in de eerste ronde werden geprezen, presteerden significant beter in de daaropvolgende maanden, terwijl leraren die niet werden geprezen in de eerste ronde significant slechter presteerden na de interventie. Het herhalen van de interventie had geen extra effect op de prestaties van de leraar. Ik onderzocht of leraren probeerden de prestatiemaatstaf te manipuleren na de interventie, door de cijfers te manipuleren die ze hun eigen studenten gaven. Ik deed dit door deze subjectieve maatstaf (cijfers die leraren hun eigen studenten gaven) te vergelijken met een objectieve maatstaf; namelijk de prestaties van studenten op gestandaardiseerde en anoniem beoordeelde eindexamens van groot belang. Er was geen bewijs dat de beoordeling van leraren in de behandelde groep onvoorspelbaarder werd na de interventie. Bovendien waren de positieve effecten van onaangekondigd lof aanhoudend en weerspiegelden ze een reële toename van de examenprestaties van studenten, wat het belang van de inspanningen van de leraar bij het vormgeven van de onderwijsresultaten van studenten benadrukte. Anderzijds vervaagden de negatieve effecten van onaangekondigde lof in de loop van de tijd en hadden ze geen significante invloed op de examenprestaties van studenten. De resultaten werden het best verklaard door een mechanisme waarbij publiekelijke lof een vergelijkende boodschap over de prestaties uitzendt. Door hun opvattingen bij te werken, werden leraren gemotiveerder als ze goed nieuws ontvingen door lof en werden ze ontmoedigd als het nieuws slecht was. Dit suggereerde dat in situaties waarin werknemers intrinsiek gemotiveerd zijn, de afweging tussen motivatie en ontmoediging minder uitgesproken zal zijn, omdat het op de lange termijn gemakkelijker te motiveren is dan te demotiveren met publieke lof.

Hoofdstuk 3 schatte het effect van het op jonge leeftijd toewijzen van kinderen aan academische trajecten door een representatieve steekproef van kinderen in de Nederlandse provincie Limburg te volgen. We bestudeerden de effecten van toewijzing aan het academische traject bij de prestatiemarge op zowel cognitieve als niet-cognitieve vaardigheden, en

over de relatieve leeftijd. Het is in de literatuur bekend dat jongere studenten in de klas een lagere kans hebben op toewijzing aan hogere trajecten. Dat komt omdat ze minder volwassen zijn op het moment van toewijzing van traject en deze verschillen in volwassenheid vaak worden aangezien voor onderliggende verschillen in vermogen. We gebruikten een vaag regressie-discontinuïteitontwerp dat gebruik maakte van schoolspecifieke toelatingsdrempels om het effect te schatten van het volgen van het academische trajcet bij de prestatiemarge. Bovendien identificeerden we interacties met de relatieve leeftijd om te zien of relatief oudere studenten die 'overgetrajecteerd' waren op de lange termijn last ondervonden. We vonden geen effect van trajectering op cognitieve resultaten, over de relatieve leeftijd. We hebben echter geconstateerd dat het volgen van het hogere traject doorzettingsvermogen, behoefte aan prestatie en emotionele stabiliteit voor de oudere studenten verhoogde. De resultaten toonden aan dat het plaatsen van relatief meer volwassen studenten in een leeromgeving die een uitdaging vormt gezien hun cognitieve vaardigheden, positieve overloopeffecten kan hebben op hun niet-cognitieve vaardigheden en op de moeite die ze doen om te leren. We voerden aan dat deze positieve overloopeffecten op de niet-cognitieve vaardigheden van relatief oudere studenten de verwachte complementariteit tussen bekwaamheid en aanwezigheid op academisch niveau mediëren en verklaren waarom oudere studenten niet slechter presteren op cognitieve tests, ondanks dat ze vaker boven hun bekwaamheidsniveau worden ingeschaald. Gezien het toegenomen belang van niet-cognitieve vaardigheden bij het bepalen van de resultaten op de arbeidsmarkt, stellen wij dat bij de beoordeling van de effectiviteit van vroege trajectering rekening moet worden gehouden met effecten op zowel cognitieve als niet-cognitieve vaardigheden.

Hoofdstuk 4 onderzocht hoe de ervaren macro-economische omstandigheden van jongeren van invloed kunnen zijn op werkvoorkeuren door variatie tussen regio's en tijd te gebruiken in een representatieve steekproef van Amerikaanse respondenten. Meer specifiek gebruikten we variatie in ervaren regionaal inkomen per hoofd van de bevolking daterend

terug tot aan de jaren 1920 om de variatie in voorkeuren voor het hebben van een hoog inkomen en voor het doen van werk dat op het moment van de enquête zinvol is te verklaren. We hebben ons gericht op ervaringen tijdens de zogenaamde 'Vormbare Jaren', omdat onderzoek in zowel Economie als Psychologie heeft aangetoond dat deze periode bijzonder belangrijk is voor het vormen van overtuigingen en voorkeuren. We ontdekten dat werkvoorkeuren systematisch variëren met macro-economische omstandigheden: degenen die de volwassenheid betreden tijdens 'goede tijden' en relatief hoge inkomens per hoofd van de bevolking ervaren wanneer jong schatten het belang van een hoger inkomen veel vaker lager in, en het belang van zinvol werk hoger. Onze resultaten waren vooral zichtbaar voor jongeren, maar we toonden aan dat macro-economische omstandigheden tijdens de vormbare jaren (18-25 jaar oud) permanente effecten hebben op de werkvoorkeuren gedurende het hele leven van een individu. Onze resultaten suggereerden dat recessies en hoogconjunctuur permanent de werkvoorkeuren van verschillende cohorten vormden, en zo een kader introduceerden voor het verklaren van generatieverschillen in attitudes ten opzichte van werk.

Deze verzameling essays heeft aangetoond dat individuen waarde hechten aan niet-financiële werkaspecten en dat niet-financiële prikkels effectief kunnen zijn in situaties waarin werknemers cognitief complexe taken uitvoeren en waar het moeilijk is om prestaties te verbeteren. Hoewel deze effecten zeer hardnekkig kunnen zijn en kunnen resulteren in een langdurige toename van de inspanningen voor sommige werknemers, vormen ze ook een afweging voor managers en moeten ze zorgvuldig worden overwogen. Het bewijs suggereert echter dat niet-financiële prikkels effectiever kunnen worden gebruikt wanneer werknemers bepaalde eigenschappen hebben. Hoofdstuk 2 geeft bijvoorbeeld aan dat intrinsiek gemotiveerde werknemers minder snel proberen de prestatiemaatstaf te manipuleren om de beloning te krijgen en meer kans hebben om op de lange termijn te overcompenseren, zelfs als ze op korte termijn worden gedemotiveerd door niet te worden beloond. Deze overcompensatie lijkt te worden aangedreven door de pro-sociale aspecten van een baan en suggereert dat werknemers

de negatieve externe effecten internaliseren die een vermindering van hun inspanningen aan derden zal opleggen.

De bevindingen van dit proefschrift relateren ook aan een groeiende interesse in het gebruik van gedrags- en experimentele methoden om beter te begrijpen hoe de arbeidsmarktresultaten van individuen worden beïnvloed door hun voorkeuren en persoonlijkheidskenmerken. Hiertoe heb ik betoogd dat, omdat belangrijke gebeurtenissen wanneer jong voorkeuren en persoonlijkheidskenmerken permanent kunnen veranderen, het belangrijk is om individuen gedurende hun hele levenscyclus te bestuderen om alle factoren te begrijpen die de economische resultaten op latere leeftijd bepalen. Ik geloof dat onderzoekers meer over dit proces kunnen leren door variatie in de blootstelling aan economische en sociale schokken te benutten, omdat ze opmerkelijk krachtig kunnen zijn in het vormen van individuen. Dit is met name het geval als deze schokken optreden wanneer genoemde voorkeuren en persoonlijkheidskenmerken nog steeds vervormbaar zijn omdat individuen nog niet volledig volwassen zijn. Toekomstig onderzoek moet de omvang beoordelen van de sociale inefficiënties die worden veroorzaakt door schokken met aanhoudende effecten, alsmede de mate waarin beleidsmakers kunnen ingrijpen om dergelijke inefficiënties te bemiddelen.

# Bibliography

**Akerlof, George A**, "A theory of social custom, of which unemployment may be one consequence," *The quarterly journal of economics*, 1980, *94* (4), 749–775.

**Alesina, Alberto and Nicola Fuchs-Schündeln**, "Goodbye Lenin (or Not?): The Effect of Communism on People's Preferences," *American Economic Review*, September 2007, *97* (4), 1507–1528.

**Almlund, Mathilde, Angela Lee Duckworth, James Heckman, and Tim Kautz**, "Personality psychology and economics," in "Handbook of the Economics of Education," Vol. 4, Elsevier, 2011, pp. 1–181.

**Anderson, Ashton, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec**, "Steering user behavior with badges," in "Proceedings of the 22nd international conference on World Wide Web" ACM 2013, pp. 95–106.

**Ashraf, Nava, Oriana Bandiera, and Scott S Lee**, "Awards unbundled: Evidence from a natural field experiment," *Journal of Economic Behavior & Organization*, 2014, *100*, 44–63.

**Bandiera, Oriana, Iwan Barankay, and Imran Rasul**, "Team incentives: Evidence from a firm level experiment," *Journal of the European Economic Association*, 2013, *11* (5), 1079–1114.

**Baumeister, Roy F**, "The self (In DT Gilbert, ST Fiske, & G. Lindzey (Eds.). The handbook of social psychology (Vol. 1, pp. 680–740)," *NY: McGraw-Hill*, 1998.

**Bedard, Kelly and Elizabeth Dhuey**, "The Persistence of Early Childhood Maturity: International Evidence of Long-Run Age Effects," *Quarterly Journal of Economics*, 2006, *121* (4), 1437–1472.

**Bellé, Nicola**, "Performance-related pay and the crowding out of motivation in the public sector: A randomized field experiment," *Public Administration Review*, 2015, *75* (2), 230–241.

**Benabou, Roland and Jean Tirole**, "Intrinsic and extrinsic motivation," *The review of economic studies*, 2003, *70* (3), 489–520.

**Bénabou, Roland and Jean Tirole**, "Incentives and prosocial behavior," *American economic review*, 2006, *96* (5), 1652–1678.

**Bernheim, Douglas B**, "A theory of conformity," *Journal of political Economy*, 1994, *102* (5), 841–877.

**Besley, Timothy and Maitreesh Ghatak**, "Status incentives," *American Economic Review*, 2008, *98* (2), 206–11.

_ **and** _ , "Prosocial motivation and incentives," *Annual Review of Economics*, 2018, *10*, 411–438.

**Bianchi, Emily C**, "The bright side of bad times: The affective advantages of entering the workforce in a recession," *Administrative Science Quarterly*, 2013, *58* (4), 587–623.

_ , "Entering adulthood in a recession tempers later narcissism," *Psychological Science*, 2014, *25* (7), 1429–1437.

_ , "American individualism rises and falls with the economy: Cross-temporal evidence that individualism declines when the economy falters.," *Journal of Personality and Social Psychology*, 2016, *111* (4), 567.

**Blader, Steven, Claudine Madras Gartenberg, and Andrea Prat**, "The contingent effect of management practices," *Columbia Business School Research Paper*, 2016, (15-48).

**Booij, Adam, Ferry Haan, and Erik Plug**, "Enriching students pays off: Evidence from an individualized gifted and talented program in secondary education," IZA Discussion Paper, no. 9757 2016.

**Borghans, Lex, Roxanne Korthals, and Trudie Schils**, "Track placement and the development of cognitive and noncognitive skills," 2018. Unpublished manuscript.

**Bradler, Christiane, Robert Dur, Susanne Neckermann, and Arjan Non**, "Employee recognition and performance: A field experiment," *Management Science*, 2016, *62* (11), 3085–3099.

**Brunello, Giorgio, Massimo Giannini, and Kenn Ariga**, "The optimal timing of school tracking: A general model with calibration for Germany," in Ludger Woessmann and Paul E. Peterson, eds., *Schools and the equal opportunity problem*, MIT Press, 2007, pp. 129–156.

**Buckles, Kasey and Daniel M Hungerman**, "Season of Birth and Later Outcomes: Old Questions, New Answers," *The Review of Economics and Statistics*, 2013, *95* (3), 711–724.

**Buelens, Marc and Herman Van den Broeck**, "An analysis of differences in work motivation between public and private sector organizations," *Public administration review*, 2007, *67* (1), 65–74.

**Buurman, Margaretha, Josse Delfgaauw, Robert Dur, and Seth Van den Bossche**, "Public sector employees: Risk averse and altruistic?," *Journal of Economic Behavior & Organization*, 2012, *83* (3), 279–291.

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

**_ , Jonah B Gelbach, and Douglas L Miller**, "Bootstrap-based improvements for inference with clustered errors," *The Review of Economics and Statistics*, 2008, *90* (3), 414–427.

**Cassar, Lea and Stephan Meier**, "Nonmonetary Incentives and the Implications of Work as a Source of Meaning," *Journal of Economic Perspectives*, 2018, *32* (3), 215–38.

**_ and _** , "Nonmonetary Incentives and the Implications of Work as a Source of Meaning," *Journal of Economic Perspectives*, 2018, *32* (3), 215–38.

**Chen, Mingliang, Qingguo Ma, Minle Li, Hongxia Lai, Xiaoyi Wang, and Liangchao Shu**, "Cognitive and emotional conflicts of counter-conformity choice in purchasing books online: an event-related potentials study," *Biological Psychology*, 2010, *85* (3), 437–445.

**Chetty, Raj, John N Friedman, and Jonah E Rockoff**, "Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood," *American economic review*, 2014, *104* (9), 2633–79.

**Cooley, Charles Horton**, "Looking-glass self," *The production of reality: Essays and readings on social interaction*, 1902, *6*.

**Crewson, Philip E**, "Public-service motivation: Building empirical evidence of incidence and effect," *Journal of public administration research and theory*, 1997, *7* (4), 499–518.

**Crutzen, Benoît SY, Otto H Swank, and Bauke Visser**, "Confidence management: on interpersonal comparisons in teams," *Journal of Economics & Management Strategy*, 2013, *22* (4), 744–767.

**Deci, Edward L**, "Effects of externally mediated rewards on intrinsic motivation.," *Journal of personality and Social Psychology*, 1971, *18* (1), 105.

**Delfgaauw, Josse, Robert Dur, Joeri Sol, and Willem Verbeke**, "Tournament incentives in the field: Gender differences in the workplace," *Journal of Labor Economics*, 2013, *31* (2), 305–326.

**Dohmen, Thomas and Armin Falk**, "You get what you pay for: Incentives and selection in the education system," *The Economic Journal*, 2010, *120* (546), F256–F271.

**Duckworth, Angela L, Christopher Peterson, Michael D Matthews, and Dennis R Kelly**, "Grit: Perseverance and Passion for Long-Term Goals," *Journal of Personality and Social Psychology*, 2007, *92* (6), 1087–1101.

**Duflo, Esther and Rema Hanna**, "Monitoring works: Getting teachers to come to school," Technical Report, National Bureau of Economic Research 2005.

**Dustmann, Christian, Patrick A Puhani, and Uta Schönberg**, "The Long-term Effects of Early Track Choice," *The Economic Journal*, 2017, *127* (603), 1348–1380.

**Eberts, Randall W, Kevin Hollenbeck, and Joe A Stone**, "Teacher performance incentives and student outcomes," 2000.

**Elder, Todd E and Darren H Lubotsky**, "Kindergarten Entrance Age and Children's Achievement Impacts of State Policies, Family Background, and Peers," *Journal of Human Resources*, 2009, *44* (3), 641–683.

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

**Ertas, Nevbahar**, "Millennials and volunteering: Sector differences and implications for public service motivation theory," *Public Administration Quarterly*, 2016, *40* (3), 517–558.

**Fehr, Ernst and Armin Falk**, "Wage rigidity in a competitive incomplete contract market," *Journal of political Economy*, 1999, *107* (1), 106–134.

**Felson, Richard B**, "The (somewhat) social self: how others affect self-appraisals', in (J. Suis, ed.), Psycho-logical Perspectives on the Self: The Self in Social Perspective, vol. 4," 1993.

**Firestone, William A and James R Pennell**, "Teacher commitment, working conditions, and differential incentive policies," *Review of educational research*, 1993, *63* (4), 489–525.

**Fischer, Paul and Steven Huddart**, "Optimal contracting with endogenous social norms," *American Economic Review*, 2008, *98* (4), 1459–75.

**Frey, Bruno S**, "Not Just for They Money: An Economic Theory of Personal Motivation," 1997.

**Fryer, Roland G**, "Teacher incentives and student achievement: Evidence from New York City public schools," *Journal of Labor Economics*, 2013, *31* (2), 373–407.

**Fuchs-Schündeln, Nicola and Matthias Schündeln**, "On the endogeneity of political preferences: Evidence from individual experience with democracy," *Science*, 2015, *347* (6226), 1145–1148.

**Gallus, Jana**, "Fostering voluntary contributions to a public good: a large-scale natural field experiment at Wikipedia," *Management Science*, 2016, p. Forthcoming.

_ **and Bruno S Frey**, "Awards as non-monetary incentives," in "Evidence-based HRM: a Global Forum for Empirical Scholarship," Vol. 4 Emerald Group Publishing Limited 2016, pp. 81–91.

**Georgellis, Yannis, Elisabetta Iossa, and Vurain Tabvuma**, "Crowding out intrinsic motivation in the public sector," *Journal of Public Administration Research and Theory*, 2010, *21* (3), 473–493.

**Gibbons, Robert**, "Incentives in organizations," *Journal of economic perspectives*, 1998, *12* (4), 115–132.

**Giuliano, Paola and Antonio Spilimbergo**, "Growing up in a Recession," *Review of Economic Studies*, 2014, *81* (2), 787–817.

**Glewwe, Paul, Nauman Ilias, and Michael Kremer**, "Teacher incentives," *American Economic Journal: Applied Economics*, 2010, *2* (3), 205–27.

**Gneezy, Uri and Aldo Rustichini**, "A fine is a price," *The Journal of Legal Studies*, 2000, *29* (1), 1–17.

_ **and** _ , "Pay enough or don't pay at all," *The Quarterly Journal of Economics*, 2000, *115* (3), 791–810.

**Goldberg, Lewis R**, "The development of markers for the Big-Five factor structure.," *Psychological assessment*, 1992, *4* (1), 26.

**Goodhart, Charles AE**, "Problems of monetary management: the UK experience," in "Monetary Theory and Practice," Springer, 1984, pp. 91–121.

**Goodman, Sarena F and Lesley J Turner**, "The design of teacher incentive pay and educational outcomes: Evidence from the New York City bonus program," *Journal of Labor Economics*, 2013, *31* (2), 409–420.

**Grant, Adam M and Francesca Gino**, "A little thanks goes a long way: Explaining why gratitude expressions motivate prosocial behavior.," *Journal of personality and social psychology*, 2010, *98* (6), 946.

**Gubler, Timothy, Ian Larkin, and Lamar Pierce**, "Motivational spillovers from awards: Crowding out in a multitasking environment," *Organization Science*, 2016, *27* (2), 286–303.

**Guyon, Nina, Eric Maurin, and Sandra McNally**, "The Effect of Tracking Students by Ability into Different Schools: A Natural Experiment," *Journal of Human Resources*, 2012, *47* (3), 684–721.

**Hall, Caroline**, "The Effects of Tracking in Upper Secondary School: Evidence from a Large-Scale Pilot Scheme," *Journal of Human Resources*, 2012, *47* (1), 237–269.

**Hanushek, Eric A**, "Teacher characteristics and gains in student achievement: Estimation using micro data," *The American Economic Review*, 1971, *61* (2), 280–288.

**_ , Susanne Link, and Ludger Woessmann**, "Does school autonomy make sense everywhere? Panel estimates from PISA," *Journal of Development Economics*, 2013, *104*, 212–232.

**Hanushek, Erik A and Ludger Woessmann**, "Does Educational Tracking Affect Performance and Inequality? Differences-in-Differences Evidence Across Countries," *Economic Journal*, 2006, *116* (510), 63–76.

**Harrison, Glenn W and John A List**, "Field experiments," *Journal of Economic literature*, 2004, *42* (4), 1009–1055.

**Hart, Oliver and Luigi Zingales**, "Serving shareholders doesn't mean putting profit above all else," *Harvard Business Review*, 2017, *12*, 2–6.

**Heckman, James J, Jora Stixrud, and Sergio Urzua**, "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior," *Journal of Labor Economics*, 2006, *24* (3), 411–482.

**_ , Seong Hyeok Moon, Rodrigo Pinto, Peter A Savelyev, and Adam Q Yavitz**, "Analyzing Social Experiments as Implemented: A Reexamination of the Evidence From the HighScope Perry Preschool Program," *Quantitative Economics*, 2010, *1* (1), 1–46.

**Holmstrom, Bengt and Paul Milgrom**, "Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design," *JL Econ. & Org.*, 1991, *7*, 24.

**Hoogveld, Nicky and Nick Zubanov**, "The power of (no) recognition: Experimental evidence from the university classroom," *Journal of Behavioral and Experimental Economics*, 2017, *67*, 75–84.

**Imbens, Guido and Karthik Kalyanaraman**, "Optimal Bandwidth Choice for the Regression Discontinuity Estimator," *Review of Economic Studies*, 2011, *79* (3), 933–959.

**Inspectie van het Onderwijs**, "De Kwaliteit van het Baissschooladvies," Technical Report, Utrecht: Inspectie van het Onderwijs 2014.

**Jacob, Brian A and Steven D Levitt**, "Rotten apples: An investigation of the prevalence and predictors of teacher cheating," *The Quarterly Journal of Economics*, 2003, *118* (3), 843–877.

**Jaeger, Hans**, "Generations in history: Reflections on a controversial concept," *History and Theory*, 1985, *24* (3), 273–292.

**Kahn, Lisa B**, "The long-term labor market consequences of graduating from college in a bad economy," *Labour Economics*, 2010, *17* (2), 303–316.

**Kamphorst, Jurjen JA and Otto H Swank**, "Don't demotivate, discriminate," *American Economic Journal: Microeconomics*, 2016, *8* (1), 140–65.

**Kautz, Tim, James J Heckman, Ron Diris, Bas Ter Weel, and Lex Borghans**, "Fostering and measuring skills: Improving cognitive and non-cognitive skills to promote lifetime success," Technical Report 19656, National Bureau of Economic Research 2014.

**Knittel, Christopher R and Elizabeth Murphy**, "Generational Trends in Vehicle Ownership and Use: Are Millennials Any Different?," *NBER Working Paper, No. 25674*, March 2019, (25674).

**Koczanski, Peter and Harvey S Rosen**, "Are Millennials Really So Selfish? Preliminary Evidence from the Philanthropy Panel Study," *NBER Working Paper, No. 25813*, 2019.

**Kohn, Alfie**, *Punished by Rewards:: The Trouble with Gold Stars, Incentive Plans, A's, Praise, and Other Bribes*, Houghton Mifflin Harcourt, 1999.

**Korthals, Roxanne, Olivier Marie, and Dinand Webbink**, "Does Early Educational Tracking Increase Inequality? Short and Long Term International Evidence," 2016. Paper presented at ESPE 2016 Berlin.

**Kosfeld, Michael and Susanne Neckermann**, "Getting more work for nothing? Symbolic awards and worker performance," *American Economic Journal: Microeconomics*, 2011, *3* (3), 86–99.

**Krosnick, Jon A and Duane F Alwin**, "Aging and susceptibility to attitude change.," *Journal of Personality and Social Psychology*, 1989, *57* (3), 416.

**Laudenbach, Christine, Ulrike Malmendier, and Alexandra Niessen-Ruenzi**, "Emotional tagging and belief formation-The long-lasting effects of experiencing communism," *American Economic Review*, 2019, *109* (May), 567–571.

**Lazear, Edward P**, "The power of incentives," *American Economic Review*, 2000, *90* (2), 410–414.

**Leigh, Andrew**, "Teacher pay and teacher aptitude," *Economics of education review*, 2012, *31* (3), 41–53.

**Lourenço, Sofia M**, "Monetary incentives, feedback, and recognition—Complements or substitutes? Evidence from a field experiment in a retail services company," *The Accounting Review*, 2015, *91* (1), 279–297.

**Luechinger, Simon, Stephan Meier, and Alois Stutzer**, "Why does unemployment hurt the employed? Evidence from the life satisfaction gap between the public and the private sector," *Journal of Human Resources*, 2010, *45* (4), 998–1045.

**Maestas, Nicole, Kathleen J Mullen, David Powell, Till Von Wachter, and Jeffrey B Wenger**, "The Value of Working Conditions in the United States and Implications for the Structure of Wages," *National Bureau of Economic Research Working Paper*, 2018.

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

_ **and** _ , "School tracking and access to higher education among disadvantaged groups," *Journal of Public Economics*, 2011, *95* (11-12), 1538–1549.

**Malmendier, Ulrike and Stefan Nagel**, "Depression babies: do macroeconomic experiences affect risk taking?," *The Quarterly Journal of Economics*, 2011, *126* (1), 373–416.

_ **and** _ , "Learning from inflation experiences," *The Quarterly Journal of Economics*, 2015, *131* (1), 53–87.

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

**Moldovanu, Benny, Aner Sela, and Xianwen Shi**, "Contests for status," *Journal of political Economy*, 2007, *115* (2), 338–363.

**Mühlenweg, Andrea, Dorothea Blomeyer, Holger Stichnoth, and Manfred Laucht**, "Effects of age at school entry (ASE) on the development of non-cognitive skills: Evidence from psychometric data," *Economics of Education Review*, 2012, *31* (3), 68–76.

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

**Muralidharan, Karthik and Venkatesh Sundararaman**, "Teacher performance pay: Experimental evidence from India," *Journal of political Economy*, 2011, *119* (1), 39–77.

**Murphy, Richard and Felix Weinhardt**, "Top of the class: The importance of ordinal rank," Working Paper 24958, NBER 2018.

**Nelson, Bob**, *1001 ways to energize employees*, Workman Publishing, 1997.

**Oreopoulos, Philip, Till Von Wachter, and Andrew Heisz**, "The short-and long-term career effects of graduating in a recession," *American Economic Journal: Applied Economics*, 2012, *4* (1), 1–29.

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

**Pop-Eleches, Cristian and Miguel Urquiola**, "Going to a better school: Effects and behavioral responses," *American Economic Review*, 2013, *103* (4), 1289–1324.

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

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

**Rogers, Todd and Erin Frey**, "Changing behavior beyond the here and now," 2016.

**Roodman, David, Morten Arregaard Nielsen, James G MacKinnon, and Matthew . Webb**, "Fast and wild: Bootstrap inference in Stata using boottest," *The Stata Journal*, 2019, *19* (1), 4–60.

**Rooney, Patrick M, Xiaoyun Wang, and Mark Ottoni-Wilhelm**, "Generational Succession in American Giving: Donors Down, Dollars Per Donor Holding Steady But Signs That It Is Starting to Slip," *Nonprofit and Voluntary Sector Quarterly*, 2018, *47* (5), 918–938.

**Sanders, Michael, Aisling Ni Chonaire et al.**, ""Powered to Detect Small Effect Sizes": You keep saying that. I do not think it means what you think it means.," Technical Report, Department of Economics, University of Bristol, UK 2015.

**Sliwka, Dirk**, "Trust as a signal of a social norm and the hidden costs of incentive schemes," *American Economic Review*, 2007, *97* (3), 999–1012.

**Slotwinski, Michaela and Alois Stutzer**, "Women Leaving the Playpen: The Emancipating Role of Female Suffrage," *Working Paper*, 2018.

**Springer, Matthew G, Dale Ballou, Laura Hamilton, Vi-Nhuan Le, JR Lockwood, Daniel F McCaffrey, Matthew Pepper, and Brian M Stecher**, "Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching (POINT).," *Society for Research on Educational Effectiveness*, 2011.

**Stajkovic, Alexander D and Fred Luthans**, "Behavioral management and task performance in organizations: conceptual background, meta-analysis, and test of alternative models," *Personnel Psychology*, 2003, *56* (1), 155–194.

**Swank, Otto H and Bauke Visser**, "Motivating through delegating tasks or giving attention," *The Journal of Law, Economics, & Organization*, 2006, *23* (3), 731–742.

**Tella, Rafael Di, Robert J MacCulloch, and Andrew J Oswald**, "Preferences over inflation and unemployment: Evidence from surveys of happiness," *American Economic Review*, 2001, *91* (1), 335–341.

**_ , _ , and _** , "The Macroeconomics of Happiness," *Review of Economics and Statistics*, 2003, *85* (4), 809–827.

**Twenge, Jean M, Stacy M Campbell, Brian J Hoffman, and Charles E Lance**, "Generational differences in work values: Leisure and extrinsic values increasing, social and intrinsic values decreasing," *Journal of Management*, 2010, *36* (5), 1117–1142.

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

**Webb, Matthew D**, "Reworking wild bootstrap based inference for clustered errors," *Queen's Economics Department Working Paper, No. 1315*, 2013.

**Zijsling, Djurre, J Keuning, H Kuyper, T van Batenburg, and B Hemker**, "Cohortonderzoek COOL5 18. Technisch rapport eerste meting in het derde leerjaar van het voortgezet onderwijs," Technical Report, Groningen/Arnhem, the Netherlands: GION/Cito 2009.

The Tinbergen Institute is the Institute for Economic Research, which was founded in 1987 by the Faculties of Economics and Econometrics of the Erasmus University Rotterdam, University of Amsterdam and VU University Amsterdam. The Institute is named after the late Professor Jan Tinbergen, Dutch Nobel Prize laureate in economics in 1969. The Tinbergen Institute is located in Amsterdam and Rotterdam. The following books recently appeared in the Tinbergen Institute Research Series:

711 H.FANG, Multivariate Density Forecast Evaluation and Nonparametric Granger Causality Testing

712 Y. KANTOR, Urban Form and the Labor Market

713 R.M. TEULINGS, Untangling Gravity

714 K.J.VAN WILGENBURG, Beliefs, Preferences and Health Insurance Behavior

715 L. SWART, Less Now or More Later? Essays on the Measurement of Time Preferences in Economic Experiments

716 D. NIBBERING, The Gains from Dimensionality

717 V. HOORNWEG, A Tradeoff in Econometrics

718 S. KUCINSKAS, Essays in Financial Economics

719 O. FURTUNA, Fiscal Austerity and Risk Sharing in Advanced Economies

720 E. JAKUCIONYTE, The Macroeconomic Consequences of Carry Trade Gone Wrong and Borrower Protection

721 M. LI, Essays on Time Series Models with Unobserved Components and Their Applications

722 N. CIURILĂ, Risk Sharing Properties and Labor Supply Disincentives of Pay-As-You-

Go Pension Systems

723 N.M. BOSCH, Empirical Studies on Tax Incentives and Labour Market Behaviour

724 S.D. JAGAU, Listen to the Sirens: Understanding Psychological Mechanisms with Theory and Experimental Tests

725 S. ALBRECHT, Empirical Studies in Labour and Migration Economics

726 Y.ZHU, On the Effects of CEO Compensation

727 S. XIA, Essays on Markets for CEOs and Financial Analysts

728 I. SAKALAUSKAITE, Essays on Malpractice in Finance

729 M.M. GARDBERG, Financial Integration and Global Imbalances.

730 U. THŰMMEL, Of Machines and Men: Optimal Redistributive Policies under Technological Change

731 B.J.L. KEIJSERS, Essays in Applied Time Series Analysis

732 G. CIMINELLI, Essays on Macroeconomic Policies after the Crisis

733 Z.M. LI, Econometric Analysis of High-frequency Market Microstructure

734 C.M. OOSTERVEEN, Education Design Matters

735 S.C. BARENDSE, In and Outside the Tails: Making and Evaluating Forecasts

736 S. SÓVÁGÓ, Where to Go Next? Essays on the Economics of School Choice

737 M. HENNEQUIN, Expectations and Bubbles in Asset Market Experiments

738 M.W. ADLER, The Economics of Roads: Congestion, Public Transit and Accident Management

739 R.J. DÖTTLING, Essays in Financial Economics

740 E.S. ZWIERS, About Family and Fate: Childhood Circumstances and Human Capital

Formation

741 Y.M. KUTLUAY, The Value of (Avoiding) Malaria

742 A. BOROWSKA, Methods for Accurate and Efficient Bayesian Analysis of Time Series

743 B. HU, The Amazon Business Model, the Platform Economy and Executive Compensation: Three Essays in Search Theory

744 R.C. SPERNA WEILAND, Essays on Macro-Financial Risks

745 P.M. GOLEC, Essays in Financial Economics

746 M.N. SOUVERIJN, Incentives at work

747 M.H. COVENEY, Modern Imperatives: Essays on Education and Health Policy

748 P. VAN BRUGGEN, On Measuring Preferences

749 M.H.C. NIENTKER, On the Stability of Stochastic Dynamic Systems and their use in Econometrics

750 S. GARCIA MANDICÓ, Social Insurance, Labor Supply and Intra-Household Spillovers

751 Y. SUN, Consumer Search and Quality

752 I. KERKEMEZOS, On the Dynamics of (Anti) Competitive Behaviour in the Airline Industry

753 G.W. GOY, Modern Challenges to Monetary Policy

754 A.C. VAN VLODROP, Essays on Modeling Time-Varying Parameters

755 J. SUN, Tell Me How To Vote, Understanding the Role of Media in Modern Elections

756 J.H. THIEL, Competition, Dynamic Pricing and Advice in Frictional Markets: Theory and Evidence from the Dutch Market for Mortgages

757 A. NEGRIU, On the Economics of Institutions and Technology: a Computational Approach

758 F. GRESNIGT, Identifying and Predicting Financial Earth Quakes using Hawkes Processes

759 A. EMIRMAHMUTOGLU, Misperceptions of Uncertainty and Their Applications to Prevention

760 A. RUSU, Essays in Public Economics