

C.M. Oosterveen

Education Design Matters

Education Design Matters

ISBN: 978 90 361 0549 1

© Christian Matthijs Oosterveen, 2019

All rights reserved. Save exceptions stated by the law, no part of this publication may be reproduced, stored in a retrieval system of any nature, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, included a complete or partial transcription, without the prior written permission of the author, application for which should be addressed to the author.

This book is no. 734 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.

Education Design Matters

Het Belang van Onderwijs Design

Thesis

to obtain the degree of Doctor from the
Erasmus University 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, March 29, 2019 at 13:30 hours

by

CHRISTIAN MATTHIJS OOSTERVEEN

born in Gouda, The Netherlands

Doctorate Committee

Promotor: Prof. dr. H.D. Webbink

Other members: Prof. dr. L. Borghans

Prof. dr. O.R. Marie

Prof. dr. H. Oosterbeek

Copromotor: Dr. A.C. Gielen

Acknowledgments

Although any errors and omissions are my own, several people played an invaluable role in the development of this dissertation.

First off all, I want to thank my main supervisor; Dinand Webbink. During our first joint projects, you basically taught me how to do research. In the projects that you were not directly involved in, you were always supportive and more than willing to share ideas, suggestions, and provide feedback. I can truly say you were the best supervisor I could have hoped for. Also a major thanks to Sacha Kapoor, you showed me that once you start to believe in your own results, you are probably not critical enough.

I am deeply indebted to my paronyms, Max Coveney and Arash Yazdiha. Max, doing research together on peer effects was the most exciting time of my PhD. I look forward towards our future projects. Arash, without your help during the coursework I most likely would have dropped out after two months in the PhD.

Thanks to my colleagues on the 8th floor. This thesis undoubtedly benefited from the open atmosphere, the discussions, and the many shared lunches. Thank you Albert Jan Hummel for your endless interest and great feedback, on every topic or level of detail. I feel lucky we found shared (research) interests in the economics of alcohol. Thank you Esmée Zwiers, for being the best office mate ever.

I am also most grateful to my friends outside the university. Dino Bektesevic, Lex de Koning, Didier Nibbering, and Guido Vermeer deserve a special shout out. Though I can only address a few of you personally, all of you kept reminding me that doing research is not a substitute for the other joyful things in life.

Finally, I want to thank my mother. After having studied for so long, I have learned that your support is my most valuable asset.

Contents

1	Introduction	1
1.1	Economics (of Education)	1
1.2	Outline	3
1.3	Summary	4
2	What Drives Ability Peer Effects?	7
2.1	Introduction	7
2.1.1	Related Literature and Channels	10
2.2	Context	12
2.2.1	Institutional Setting	12
2.2.2	Close and Distant Peers	14
2.2.3	Assignment of Students to Groups	15
2.3	Data	16
2.3.1	Attendance and Student Evaluations	17
2.3.2	Descriptive Statistics	18
2.4	Empirical Specification	18
2.4.1	Reduced-Form Peer Effects	20
2.4.2	Balancing Tests	21
2.5	Baseline Results	24
2.5.1	First-Year Grades and Passing Rates	24
2.5.2	Randomization Inference	26
2.5.3	Additional Outcomes	27
2.5.4	Robustness	27
2.5.5	Heterogeneity	30
2.5.6	Group Assignment Policies	32
2.6	Nature of Social Interactions	34

2.7	Voluntary Sorting and Potential Implications for Group Assignment Policies	38
2.7.1	Diminishing Peer Effects	38
2.7.2	First-Year Tutorial Attendance	39
2.7.3	Second-Year Tutorial Choice	41
2.7.4	Long-Term First-Year Bonds	45
2.7.5	Implications of Voluntary Sorting for Peer Effects	47
2.8	Conclusion	47
2.A	Appendix	48
3	The Price of Forced Attendance	71
3.1	Introduction	71
3.2	Context	75
3.2.1	University Policy	76
3.2.2	Course Policies	77
3.2.3	Abolition	79
3.3	Data	79
3.3.1	Basic Descriptives	80
3.3.2	Preview of Baseline Results	81
3.3.3	Abolition Results	83
3.4	Empirical Specification	83
3.4.1	Continuity Near the Cutoff	85
3.4.2	Sample Attrition	85
3.4.3	Estimation and Inference	87
3.5	Baseline Results	88
3.5.1	Course-Level Attendance Policies	89
3.5.2	Robustness	92
3.5.3	External Validity	93
3.6	Baseline Mechanisms	93
3.6.1	Peer Effects	94
3.6.2	Attendance is Useful in Some Courses, but not Others?	94
3.6.3	It's About Time	95
3.6.4	Less Time for Leisure	97
3.6.5	Self-Study Time and Efficiency	98
3.7	Spillovers Across Courses	99

Contents

3.8	Long-Run Performance	101
3.9	Conclusion	102
3.A	Appendix	103
4	Wait and See: Gender Differences in Performance on Cognitive Tests	121
4.1	Introduction	121
4.2	The PISA Test	124
4.3	Baseline Results	125
4.3.1	Gender Differences	126
4.3.2	Gender Differences per Topic	126
4.4	Potential Determinants of the Gender Difference in Ability to Sustain Performance	129
4.5	Longer Tests and the Math Gender Gap	131
4.6	Conclusion	132
4.7	Supplementary Material	133
4.7.1	Data and Methodology	133
4.7.2	Baseline Results for the Different PISA Waves	138
4.7.3	Potential Determinants of the Gender Difference in Ability to Sustain Performance	140
4.7.4	Longer Tests and the Math Gender Gap	145
4.7.5	Low Stakes versus High Stakes	147
4.7.6	Robustness	148
4.A	Appendix	153
5	Test Scores, Noncognitive Skills and Economic Growth	177
5.1	Introduction	177
5.2	Previous Studies	180
5.2.1	The Relationship between Cognitive Test Scores and Economic Growth	180
5.2.2	Noncognitive skills, Long-term Individual Outcomes and Cognitive Test Scores	181
5.3	The Test Score Decomposition	182
5.3.1	Interpretation of the Two Components	184
5.3.2	Differences between Countries and Years	188
5.4	Estimation of the Relationship between Skills and Economic Growth	189
5.5	Data	191
5.6	Main Estimation Results	192
5.6.1	Replication of Previous Cross-Country Growth Regressions using PISA	192

5.6.2 The Relationship of the Starting Performance and the Performance Decline with Economic Growth	194
5.7 Robustness Checks	197
5.7.1 Stricter Measures of the Performance Decline	197
5.7.2 From Skills to Growth or From Growth to Skills	200
5.8 Conclusion	204
5.A Appendix	206
Nederlandse Samenvatting (Summary in Dutch)	219
Bibliography	221

Chapter 1

Introduction

Education is an investment in human capital which has large positive impacts upon individual and social outcomes. In particular, education causes individuals to earn higher wages (Harmon et al., 2003), to have better health (Oreopoulos, 2007), to commit fewer crimes (Webbink et al., 2012), and is positively related to the prosperity of regions and countries (Barro, 2001; Gennaioli et al., 2012). It is for this reason that the Netherlands spent 43.8 billion euro on the education system in 2017, which amounts to roughly 6.5 percent of its gross domestic product. Education design involves using these scarce resources to implement a series of policies that characterize the education system. The stakes are high while designing the system; it is a crucial opportunity to influence important outcomes such as wages and health. Policymakers therefore like to make informed decisions while designing the education system. However, this requires knowledge about the consequences of intended policies. To illustrate, one of the most debated policies in education is to decrease the number of students in class. Does such a policy yield benefits in terms of student outcomes? If so, do they outweigh the costs of hiring additional teachers? Another example is that the Netherlands tracks students at a relatively early age into different types of education levels and schools. What are the consequences of early tracking? One of the main goals of Economics (of Education) is to answer such questions. More generally, it aims to provide policymakers with knowledge to make informed decisions when designing the education system. This thesis aims at contributing to this knowledge by means of four self-contained chapters on the impact of design features in education. To put it differently, this thesis investigates education design matters, and finds that education design matters.

1.1 Economics (of Education)

In order to inform policymakers, education economists focus on estimating the *causal* relationship between inputs (*e.g.* an education policy such as smaller classes) and outputs (*e.g.* student outcomes

such as test scores). A causal relationship indicates what would happen to the output in the absence of the input and is, therefore, crucial to inform policymakers about the consequences of education policies. Estimating such causal relationships is, however, far from easy. There are multiple observed and unobserved inputs that often change at the same time, making it difficult to isolate the causal impact of one single input, such as a small class size.

To better understand this difficulty, assume that each individual has two potential outcomes (*e.g.* test scores); one when exposed to the policy (*e.g.* small class size), also referred to as the treatment, and one when not exposed to the policy (*e.g.* large class size), also referred to as the control. Then the causal impact of the policy is the difference between the two potential outcomes. We observe, however, only one potential outcome of each individual. Rubin (1974) denotes this as the fundamental problem of causal inference. In practice, we thus have to compare individuals who received the treatment with individuals in the control. The fundamental assumption is that the observed outcome of the individuals in the control group is identical to the potential outcome of the individuals in the treatment group had they not been treated.

Individuals often select themselves into treatments, making this fundamental assumption unlikely to hold. Imagine that you would observe that pupils in small classes have higher test scores than pupils in large classes. This might reflect a causal impact of class size, but it might also be the case that higher ability pupils have selected themselves into smaller classes. If the latter is the case, then at least part of the difference in test scores between pupils in large and small classes cannot be ascribed to differences in class size. Indeed, this difference rather reflects a violation of the fundamental assumption; the observed test score of pupils in large classes is lower than the potential test score of the pupils in smaller classes had they been exposed to larger classes. Subsequently, we cannot reliably inform policymakers about the consequences of a class-size reduction.

To overcome this problem (*i.e.* selection bias), in several settings economists started to take control over the mechanisms that assign individuals into treatments since the early 1990s. Angrist and Pischke (2010) refer to this as the credibility revolution in empirical economics. The most credible assignment mechanism is randomization. The reason behind this is not difficult to understand; randomization ensures that the treatment and control group are similar except for exposure to the treatment. This makes the fundamental assumption stated above likely to hold. Hence, randomization allows us to uncover causal effects. Krueger (1999) uses random assignment of students to class size and finds that pupils in small classes score better than those in larger classes. However, randomized experiments are often not possible due to financial or ethical constraints. Fortunately, bureaucratic rules or natural forces sometimes allow for indirect control over the assignment mechanism. These are often referred to as natural experiments rather than randomized experiments. For example, Fredriksson

et al. (2013) exploit the fact that classes in Swedish primary schools were formed in multiples of 30; 30 students in a grade level in a school yielded one class, while 31 students yielded two classes. They find that smaller classes have a positive impact on long term outcomes, such as completed education and wages.

1.2 Outline

The first three chapters of this thesis consist of studies that exploit *randomization* of some sorts to estimate the *causal* impact of three separate design features in education. Chapter 2 uses random assignment of students to classes to study ability peer effects. In the presence of peer effects, alternative group assignment policies might have important consequences for student outcomes. The third chapter exploits a bureaucratic feature of a university policy to study the causal impact of additional structure for student performance; something recent scholarship argues may be good for academic performance. Chapter 4 uses the random order of questions in cognitive tests to document that females are better able to sustain their performance during a test than males. This finding suggests that test design could play a role in further promoting gender equality in participation in math and sciences.

Conditional on the question of interest being one for which randomized experiments are feasible, randomized experiments are clearly superior. For some questions in economics, however, randomization is difficult or even conceptually impossible. One example is the impact of macroeconomic policies; experimenting with countries seems quite impossible. In line with Imbens (2010), I would argue that this should not discourage researchers from asking questions concerning the effects from macroeconomic policies. Imbens (2010) writes that history abounds with examples where causality found general acceptance without any experimental evidence. With this in mind, chapter 5 somewhat deviates from the previous chapters, as its focus is not on credible inference. In the fifth chapter we analyze to what extent the well-studied relationship between the performance on international cognitive tests and economic growth should be interpreted as evidence on the importance of cognitive versus noncognitive skills. Given the differences in policy interventions required to foster cognitive and noncognitive skills (Cunha et al., 2010), it is important to gain a better understanding of their respective roles in fostering economic growth. In what follows, I will describe each of the chapters in more detail.

1.3 Summary

Economists have an ongoing interest in ability peer effects. The possible existence of peer effects generates a big promise; simply by reorganizing peer groups, and without additional resources, it may be possible to increase aggregate student performance. In the second chapter we analyze ability peer effects at a large European university across six cohorts of undergraduate economic students. We exploit that students are randomly assigned a tutorial group and one of two subgroups within their tutorial group. The university encourages peer bonding within, and not between, these subgroups via a series of informal meetings. Hence, each student can divide her tutorial peers into *close* and *distant* peers. We find the existence of positive peer effects on student grades and passing rates that originate from close peers only. We take this as evidence that spillovers arise due to social proximity rather than via classroom-level effects. Through the use of supplementary data we provide suggestive evidence that students with better close peers change their study behavior by substituting lecture attendance for collaborative self-study with their close peers at university.

Examining heterogeneity in spillovers by own and close peer ability, we document that high and low ability students benefit (suffer) from social proximity with high (low) ability peers. Alternative group assignment policies - such as tracking high ability students - entail a transfer from one student group to the other.

In the second part of this chapter, we use detailed data on first-year tutorial attendance and second-year tutorial registration and find that students voluntarily sort into new peer groups over time. This sorting behavior leads to an erosion of the social proximity between close peers, and we argue that this erosion provides an intuitive explanation for our finding that the spillovers from assigned close peers diminish over time. Apart from its importance for policies aiming to exploit peer effects, our findings on voluntary sorting provide a rare insight into the degree to which friendship groups can be institutionally manipulated against the formation of homogeneous subgroups based on gender, ethnicity, and prior bonds. This might have implications for promoting diversity in higher education, something that policymakers in both the U.S. and Europe have recently emphasized.

In chapter 3 we turn to the debate on additional structure in higher education. Recent scholarship argues that structure, which amounts to constraining the choices of students, may be good for academic performance. The arguments usually focus on student predispositions towards non-academic activities, emanating from behavioral biases such as impatience, or imperfect information about behaviors that engender success at university. We investigate the impact of an attendance policy that imposed greater structure on students at a large European university. At this university, students who average less than 7 (out of 10) in their first year are forced to attend at least 70 percent of their tu-

torials in second year. Conversely, students above 7 had the freedom in choosing their attendance. This allows for a comparison of students near 7 to estimate the causal impact of a full year of forced, frequent, and regular attendance.

Our findings suggest that additional structure has no positive impact on student performance. Instead, we show forced students have lower grades and lower passing rates. This average effect on second-year student performance aggregates differential effects across courses. The largest effects are in courses where the attendance advantage of above-7 students was greatest, where they had full discretion over their attendance. We argue that for these courses the university policy forces below-7 students to spend a substantial number of hours in a specific way, leaving them with less time for other activities, including activities which are important for grades. Grades decrease because the grade loss from spending less time on other academic activities outweighs the grade gains from additional attendance. Reports of total study time suggest further that forced students spend less time on nonacademic activities such as leisure. Overall, our evidence suggests that this forced attendance policy makes students worse off.

Chapter 4 studies gender gaps in cognitive test scores. An abundance of research has shown that, on average, females outperform males in verbal and reading tests, while males perform better than females in math and science (see *e.g.* Cornwell et al. (2013)). In turn, math-science classes have been found to be important for college attendance, college completion, occupational choices, and wages (Goldin et al., 2006; Joensen and Nielsen, 2009). Chapter 4 provides new insights on gender gaps in the test scores of 15 to 16 year-old students participating in the (low-stakes) PISA test (Programme for International Student Assessment). It studies how gender gaps in test scores change throughout the test. Countries from around the world participate in the PISA, which varies the order of test questions among test booklets and randomly allocates the booklets to students.

We compare, by country, the performance of males and females on the same test question at different positions in the test booklet. Overall, we find females are better able to sustain their performance during tests. This result is present across PISA waves and holds for a vast majority of countries. The pattern is independent of the topic being assessed and provides new insights into gender gaps in test scores. At the beginning of the test, males score better in math and science and females score better in reading. After two hours of test taking, the gender gap in math and science is completely offset and is even reversed in roughly one-third of the countries considered. In more than half of the countries, females decrease their initial disadvantage in math and science by at least 50 percent by the end of the test. At the same time, the advantage that females have in reading grows larger as the test goes on.

In chapter 5 we delve into a question that is of great importance, but where randomization is extremely difficult; what is the relationship between cognitive test scores and economic growth?

Many studies have already found a strong association between economic outcomes of nations and their performance on international cognitive tests. Hanushek and Woessmann (2012) find evidence consistent with a causal interpretation of this relationship. However, noncognitive skills also affect performance on cognitive tests. This raises the question whether noncognitive skills are (partly) responsible for the well-studied relationship between cognitive test scores and economic growth.

In the first part of this chapter, we use a similar method as in chapter 4 to decompose the student performance in the PISA test into two components: the starting performance and the decline in performance during the test. The latter component is interpreted to be closely related to noncognitive skills, whereas the first component is a cleaned measure for cognitive skills. Students from different countries exhibit differences in performance at the start of the test and in their rates of deterioration in performance during the test. In the second part of this chapter, we document that both components have a positive and statistically significant association with economic growth. The estimated effects for both components are quite similar and robust. Our results suggest that noncognitive skills are also important for the relationship between test scores and economic growth.

Reverse causality and omitted variable bias is an obvious concern if one is interested in putting a causal interpretation upon the results of macroeconomic growth regressions. We try to address these issues by applying the decomposition method to an early test in 1991 and via a tentative IV-analysis. We find that our results are consistent with an effect of skills on growth and not vice versa.

Chapter 2

What Drives Ability Peer Effects?

Joint work with Max Coveney

2.1 Introduction

Economists' ongoing interest in classroom peer effects is not hard to justify; simply by reorganizing peer groups, and without additional resources, it may be possible to increase aggregate student performance. Taking into account important methodological advances (Manski, 1993), the past decade of empirical research includes many well-identified studies in primary, secondary, and tertiary education (Sacerdote, 2014). While these studies have to a large extent confirmed the existence of small peer effects in the classroom, little to no credible evidence exists on the mechanisms through which these effects operate. For instance, it remains unclear whether students benefit from better peers because of social interaction with these peers, or because the quality of teacher instruction improves in a classroom with better students, or through another potential mechanism.

This paper is the first to exploit random group assignment to empirically test between two exhaustive and policy-relevant channels driving ability peer effects. Based on the current literature, we distinguish between the following two channels; social proximity and classroom-level effects. Social proximity relates to the degree of familiarity between classroom peers (Foster, 2006), and this channel captures spillovers that arise due to friendship, bonding, and student-to-student interaction between classroom peers. Classroom-level effects capture spillovers that stem from the classroom environment, which are independent of the social proximity between students, *e.g.* teacher response to the ability composition of the classroom. The context in which we study these two channels is the first year of an economics undergraduate program across six cohorts at a large public university in the Netherlands.

We exploit the institutional manipulation of the social proximity between students and their classroom peers. Students are randomly assigned to a tutorial group of approximately 26 students and one of two subgroups of 13 students within their tutorial group. The university encourages interaction, bonding, and friendship within, and not between, these subgroups during the first weeks of the academic year via several informal meetings. From the perspective of one student, the *close* peers are the subset of their tutorial peers with whom social proximity is encouraged, whereas their *distant* peers belong to the adjacent subset with whom social proximity is not encouraged. For each student, her close and distant peers together form her tutorial group whom she follows classes with throughout the first year. By exploiting the differences between these two types of peers, we are able to disentangle the two broad mechanisms driving ability peer effects. We use high school GPA - which includes the nationwide final exams before entering university - as a pre-treatment indicator of own and peer ability. This allows us to avoid problems related to reflection and common shocks.

Exploiting the novel within-classroom random assignment we find that peer effects are solely driven by a student's close peers; the subset of peers within the classroom with whom students are socially proximate. We find no role for distant peers. This implies that meaningful social interaction drives peer effects, whereas classroom-level effects are unimportant. The point estimate from our linear model implies that a one standard deviation increase in close peer GPA causes student performance to increase with 0.026 standard deviations. Using student evaluations we provide suggestive evidence that students with better close peers change their study behavior by substituting lecture attendance for collaborative self-study with their close peers at university. Examining heterogeneity in spillovers by ability, we find that high and low ability students benefit (suffer) from social proximity with high (low) ability close peers. These spillovers, however, diminish over time, and are completely absent by the end of the first year.

Having shown that peer effects arise due to social proximity, the evolution of the social proximity between students and their assigned close peers, and the degree to which new friendship are formed, is of major importance to group assignment policies. We study how students cluster by daily tutorial attendance in first year and find some evidence that the social proximity between assigned close peers gradually diminishes. Analyzing tutorial choice in second year we confirm that students largely sort themselves out of their close peer groups. We also show that they sort into new self-chosen peer groups, which are based on shared characteristics such as gender and ethnicity. We do not find evidence that students sort on ability, though our estimates suggest this could be academically beneficial. Overall, we believe this sorting behavior shows that students have strong preferences dictating with whom they become socially proximate. The erosion of social proximity between assigned close peers

provides an intuitive explanation for the short-lived spillovers on student performance, though we cannot provide causal evidence to confirm this intuition.

Our study has three main implications for group assignment policies aiming to exploit spillovers. First, our results suggest that such policies should focus on fostering social proximity within student groups. As it stands, attempts to implement alternative group assignment policies using estimates of peer effects under one particular assignment policy do not lead to predictable results. A well-known example of this is the study by Carrell et al. (2013), in which the authors use credible estimates of spillovers to construct “optimal” peer groups at the United States Air Force Academy. They find that low ability students whom they intended to help with this group assignment policy actually performed worse than untreated low ability students.¹ The importance of social proximity and the absence of classroom-level effects implies that it may be insufficient to simply place students together in a classroom. Our results suggest group assignment policies could be more successful if social proximity within peer groups was fostered. Additionally, such fostering could result in larger spillovers than those previously observed. Our estimated spillovers in the linear-in-means model are more than twice the size of those found in very similar contexts, where manipulation of social proximity is absent (Booij et al., 2017; Feld and Zöllitz, 2017).

Second, our results imply that social proximity between diverse assigned peers can indeed be manipulated by a relatively simple intervention, consisting of several informal meetings.² However, the persistence of these bonds in the longer run, especially among students of different backgrounds, may be low.

Third, given the importance of social proximity to ability peer effects, our results imply that long-run effects on student performance from group assignment policies may be difficult to sustain. Individuals have strong homophilic preferences, and over time tend to experience diminishing social proximity with their assigned peers as they sort into new peer groups based on these preferences.

With respect to the literature on peer effects more broadly, Sacerdote (2014) highlights the large degree of heterogeneity in the magnitudes of spillovers across the current studies. The findings of this paper may to some extent help explain this heterogeneity. Given that peer effects crucially depend on the degree of social proximity, the study-to-study variation in peer spillovers may partly be explained by the degree that social proximity was present, or perhaps even encouraged.

¹In Carrell et al. (2009), data based on ability mixing (natural random variation) suggested that low ability students would benefit from being mixed with high ability students, were high ability students would not suffer from being paired with low ability students. Carrell et al. (2013) then create optimal squadrons that consisted of low- and high ability students (bimodal squadrons) and squadrons with middle ability students only (homogeneous squadrons).

²The analysis on voluntary sorting shows that a student’s close peers are more strongly related to her first-year tutorial attendance and second-year tutorial registration than distant peers.

Our results may also provide some suggestions for the literature on theoretical models of peer effects, which in turn might generate new insights for empirical work. Most of the well-known models of educational peer effects imply that they take place at the classroom level. Lazear (2001) argues that a classroom can be considered as a public good, where one disruptive student may impose negative externalities on *all* students. The taxonomy of models on peer effects by Hoxby and Weingarth (2005) also encapsulates this idea, whereby *e.g.* one superstar student can increase the grades for the rest of the class. Our results imply more nuanced versions of these existing models; a model which focuses on social interaction would more realistically capture the processes driving peer effects in tertiary education.

Apart from their importance for understanding peer effects, the patterns on voluntary sorting behavior of students also provide a rare insight into how friendship formation occurs at university, a question that has been asked independently by Marmaros and Sacerdote (2006) using data on email exchanges between students. The exogenous allocation of first year students to close peer groups allows us to analyze the importance of “manipulated social proximity” against other factors like ethnicity and gender. These results are of interest because of the recent emphasis on the importance of diversity in the education process both by European and American universities.³ To this end, our results show that the intervention did little to promote long-lasting diversity on campus. We cannot rule out, however, that a more sustained and focused intervention would deliver larger effects.

2.1.1 Related Literature and Channels

Based on the empirical literature, we distinguish between two broad and exhaustive channels driving peer effects; social proximity and classroom-level effects.

- **Social Proximity:** peer effects driven by meaningful social interactions between classroom peers. Peer effects from this channel are restricted to peers who are socially proximate; those for whom bonds exist and social interactions occur.
- **Classroom-Level Effects:** peer effects that stem from the overall classroom environment and are independent of the social proximity between students. They potentially originate from and have an impact on all students in a classroom, even between students that do not explicitly interact.

³In the U.K., the former Prime Minister David Cameron and the Universities and Colleges Admissions Service (UCAS) announced applications to be name-blind from 2017 onward, after which several institutions introduced pilots. In the U.S., many leading American institutions, such as MIT and University of Chicago, filed an amicus brief in November 2015 with the U.S. Supreme Court in *Fisher v. University of Texas*. This brief stressed the role of government in diversity of higher education, of which race and ethnicity are components.

The social-proximity channel would, for instance, include having a high ability peer in the classroom with whom a student discusses material. This could potentially happen both inside or outside class. Alternatively, an example of a classroom-level effect is teachers responding to the composition of students in the classroom. Having many high ability students in a class might induce teachers to change the level of their instruction. A student posing an insightful question in class that benefits all other students is another example of a classroom-level effect.⁴

Several papers rely on social proximity, and thus interaction between peers, as the main explanation for spillovers. Booij et al. (2017) and Feld and Zöllitz (2017) use voluntary course evaluation data and find that students with better tutorial peers reported better interactions with other students. In attributing the negative results of their experiment to voluntary sorting, Carrell et al. (2013) implicitly argue that peer effects are generated via the social proximity of peers.⁵

Other researchers attribute their findings to classroom-level effects. Duflo et al. (2011) argue that the resulting peer effects of a student tracking experiment can be explained by changes in teaching behavior based on the ability composition of the class. Lavy et al. (2012a) and Lavy and Schlosser (2011) explore potential channels using a student survey and find that a higher proportion of low ability students has negative effects on the quality of student-teacher relationships, on teachers' pedagogical practices, and increases classroom disruptions.⁶

The strategies used in the empirical literature thus far to explore potential channels is to (i) search for heterogeneity in the data that supports or refutes certain peer effect channels or (ii) look at additional outcomes using secondary data sources, such as student evaluations.⁷ The results using the first strategy are, however, mostly circumstantial and unable to definitively rule out other competing explanations. An example of this is Carrell et al. (2009), who looks at the heterogeneity of peer effects between courses to find suggestive evidence of study partnerships as a driver of peer effects. With the second strategy researchers must often attribute their results to other unobserved factors (see *e.g.* Feld and Zöllitz (2017)). In both cases, these strategies involve looking for an explanation after the fact. Researchers have rightly been cautious in interpreting the findings derived from these strategies.

⁴Because classroom-level effects are defined as the complement of social proximity, together they are exhaustive. Though our main distinction is between these two broad channels, we also use supplementary data to hint at finer channels such as those listed by Sacerdote (2011). We find suggestive evidence that spillovers revolve around collaborative self-study and peer-to-peer teaching.

⁵Other papers that attribute their results to the social-proximity channel include Garlick (2018); Brunello et al. (2010); Carrell et al. (2009); Stinebrickner and Stinebrickner (2006); Arcidiacono and Nicholson (2005).

⁶Other research relying on a classroom-level explanation are Oosterbeek and Van Ewijk (2014); Burke and Sass (2013); Lyle (2009); Foster (2006); Hoxby and Weingarth (2005).

⁷For strategy (i) see, among others, Garlick (2018); Oosterbeek and Van Ewijk (2014); Duflo et al. (2011); Brunello et al. (2010); Carrell et al. (2009); Lyle (2009); Foster (2006); Arcidiacono and Nicholson (2005); Hoxby and Weingarth (2005); Hoxby (2000). For strategy (ii) see, for example, Booij et al. (2017); Feld and Zöllitz (2017); Lavy et al. (2012a); Lavy and Schlosser (2011); Stinebrickner and Stinebrickner (2006).

The definition of what constitutes a peer group varies substantially in the literature. It includes entire schools (Lavy and Schlosser, 2011), classes (Feld and Zöllitz, 2017), dorms (Garlick, 2018) and dorm roommates (Sacerdote, 2001; Zimmerman, 2003), students in the same group during university orientation week (Thiemann, 2017), students that share more than a certain number of classes (De Giorgi et al., 2010), and students who sit next to each other in class (Lu and Anderson, 2014; Hong and Lee, 2017). It may be that different types of peers deliver spillovers via different mechanisms. The manipulation of social proximity allows us to cleanly separate the two broad channels in the same context. Furthermore, our results may be of more general interest than many of the studies mentioned above, as opportunities to manipulate classroom peers arise in almost every educational setting, while contexts where universities or schools can assign dorm mates or students' seating arrangements are far more infrequent.

Finally, it is worth noting that the relative importance of the two different channels might vary across different levels of education. Our focus is on university students and tutorial peer groups, which are mostly taught by senior students and PhDs. Because of the inexperience of these teachers, one might reason that teacher response is unlikely. However, evidence from a similar public Dutch university suggests academic rank of instructors is unrelated to student performance; Feld et al. (2018) show that full professors are not significantly more effective in tutorial teaching than students or PhDs. Moreover, since future employment at the university depends largely on their performance in student evaluations, teaching assistants (TAs) have incentives to teach well and put forth effort. Similarly, one might argue that disruptive students are not present at the university level. However, personal experience and interviews with TAs suggest otherwise. Notably, every TA at the university of our study undergoes a one-day training, part of which teaches them to deal with disruptive student behavior through role-playing.⁸ Thus, we believe that there is a priori little reason to dismiss the presence of either channel in the university setting, and that our results are not necessarily uninformative for other education contexts.

2.2 Context

2.2.1 Institutional Setting

Our setting for studying peer effects is the economics undergraduate program at a large public university in the Netherlands. Every year the economics program experiences approximately 400 newly

⁸A web search reveals that many other universities also provide advice to their teaching staff on how to deal with disruptive students, indicating that the phenomenon is not absent in higher education. For example, see the following resource page from Stanford University: <https://teachingcommons.stanford.edu/resources/teaching-resources/interacting-students/classroom-challenges>.

enrolled first-year students. During the first two undergraduate years the program is identical for every student, as they follow the same twenty courses across the two years, covering basic economics, business economics, and econometrics. Come the third year, students must choose their own courses. The program only admits Dutch students. The admission requirement is based on having a pre-scientific high school diploma.

The three academic years are divided into five blocks of eight weeks each (seven weeks of teaching and one week of exams).⁹ Students in the first- and second year have one light and one heavy course per block, for which they can earn four and eight credits respectively. Sixty credits account for a full year of study.¹⁰ In the first- and second year, courses consist of both lectures and tutorial sessions. The heavy courses have three large-scale lectures per week, while light courses have two. Heavy courses have two small-scale tutorials per week, while light courses have one. Lectures and tutorials both last for 1 hour and 45 minutes. While attendance at lectures is voluntary, first-year students have to attend at least 70 percent of the tutorials per course. Students who fail to meet the attendance requirement are not allowed to take the final exam for their course and must wait a full academic year before they can take the course again.

During tutorial sessions a teaching assistant (TA) typically works through question sets based on the materials covered in the lectures. Roughly 10 percent of the TAs are PhDs, with some exceptions the remaining 90 percent are senior students. Unlike lectures, the tutorial sessions often require preparation and active participation from the student, *e.g.* via discussion of assignments or related materials. First-year students follow the tutorials with the same group throughout the whole first year. To verify whether the 70 percent attendance requirement is met, TAs register attendance at the start of each session. The requirement ensures that students experience a sizable degree of exposure to tutorials and their tutorial peers, and are not able to voluntarily attend different groups during the first year. Appendix Table A.2.1 gives an overview of the first-year courses, their characteristics, and an accompanying tutorial description. We investigate peer effects originating from these first-year tutorial peer groups.

Grading is done on a scale that ranges from 1 to 10. Students fail a course if their grade is below 5.5. Most of the courses in first- and second year are (partly) multiple choice and therefore graded without interference by the instructor or TAs. For exams with open questions, instructors disallow TAs from grading their own groups.

⁹At the end of the academic year, at the start of summer, there is a resit period. During two weeks first- and second-year students have the opportunity to resit a maximum of three courses.

¹⁰In this institution credits are measured through ECTS, which is an abbreviation for European Transfer Credit System. This measure for student performance is used throughout Europe to accommodate the transfer of students and grades between universities. The guidelines are that one ECTS is equivalent to 28 hours of studying.

2.2.2 Close and Distant Peers

A key institutional feature of the economics program is that each first-year tutorial group is divided into two subgroups. The university induces social proximity, and thus student-to-student interaction, only within these subgroups of students. For a student we term *close* peers to be the group with whom bonds are encouraged, where *distant* peers are the adjacent group of peers in the tutorial group with whom interaction is not encouraged. This means that if student $S1$ and $S2$ are in the same tutorial group but in different subgroups, the close peer group of student $S1$ will be the distant peer group for student $S2$ and vice versa.

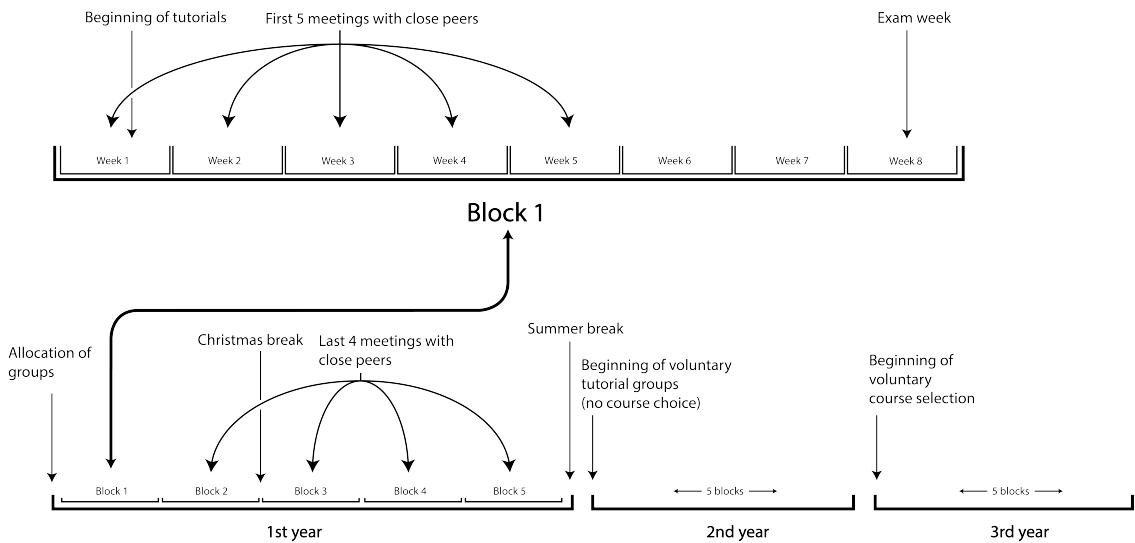
The main purpose of the close peer group is to facilitate the formation of social ties to help students adjust to, and get acquainted with, life at university. These ties are primarily facilitated via five compulsory close peer group meetings during the first block.¹¹ As discussed in more detail below, these meetings revolve around discussion and active student participation, which the university aims to foster via the smaller subgroups. The first close peer group meeting is in the first week of university, before any lectures or tutorials have taken place. As well as meeting each other in the subsequent tutorial sessions, which also include the set of distant peers, there are weekly close peer group meetings up until week five. During the first five weeks close peers see each other 20 times; 5 times at the close peer meetings and 15 times at the regular tutorials. There are four remaining meetings with the close peer groups that are evenly spread out across the year (one per block). An overview of the first block and the whole undergraduate program can be found in Figure 2.1.

The university assigns senior students as discussion leaders to guide the close peer meetings. The subjects and the setting of these meetings are less formal than the tutorial groups. The first close peer meeting is a get-to-know-you session, where students have to introduce themselves to the group. The subsequent four sessions in the the first block consist of group discussions of the use of study timetables, exam preparation, fraud and plagiarism, teamwork, and plans concerning the future of their studies, among other topics. There is an emphasis on active participation of all students during these discussions. Importantly, course material is not discussed during these meetings.

Given the timing and the nature of their introduction, the close peer groups serve as the first plausible group of fellow students that a new student will interact with and form friendships with. Our empirical evidence presented later on implies that the close peer meetings resulted in substantial social proximity between close peers, at least initially. Conversely, the structure of the program resulted in comparatively much less, if any, meaningful bonding with members of distant peer groups.

¹¹While the students do not get any credits for these meetings, according to the Teaching and Examination Regulations students must attend all of these meetings in order to pass the first year. Our administrative attendance data reveals students attend on average 94 percent of the sessions of the group they have been assigned to.

Figure 2.1: An overview of the characteristics of the undergraduate Economics program relevant to our study



2.2.3 Assignment of Students to Groups

During the final year of students' pre-scientific education, and before the start of the academic year, students must preregister for the economics program. Those who have done so are requested to come to campus on the first day of the academic year to confirm their registration.¹² This is done by means of approximately 10 to 15 administrative personnel, who add students' numbers and names to an electronic register.

A list containing the information of all students who confirmed their registration is sent to an administrative worker. This list is then sorted by a randomly assigned ID and group membership is determined on a rotating basis. The first student on the list is allocated to tutorial group 1, close peer group 1A; the second student is allocated to tutorial group 2, close peer group 2A; the third student is allocated to tutorial group 3, close peer group 3A, and so forth. The allocation continues until the maximum tutorial group has been reached, after which the rotation begins again by allocating the next unassigned student to tutorial group 1, close peer group 1B, the next student to tutorial group 2, close peer group 2B, and so forth. The university uses this allocation method to ensure that students are exposed to new peers and that the groups are roughly of equal size.¹³

¹²In this way the university avoids, to a large extent, taking into account no-shows when forming the first-year groups.

¹³We conducted numerous interviews with the administrative worker and university administrators, and received accompanying documentation, in order to confirm that the allocation process occurred as described. The same administrative worker has been in charge of this process across the six cohorts we study. The allocation process is done with BusinessObjects BI and Microsoft Excel software.

Figure 2.2: A graphical representation of the allocation to tutorial and close peer groups for a hypothetical cohort

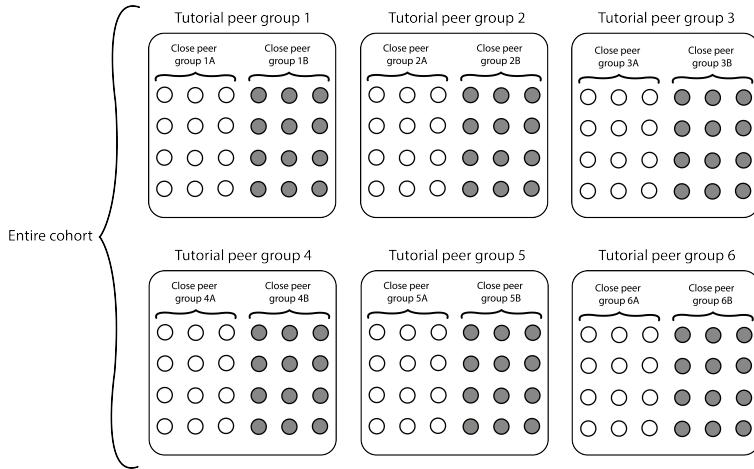


Figure 2.2 clarifies the structure of the tutorial and close peer groups for a hypothetical cohort. The 144 students, represented by dots, are distributed across 6 tutorial groups and 12 close peer groups. For a student in close peer group 1A, her distant peers are those students belonging to close peer group 1B, and vice versa.

A student who wants to follow the program, but did not show up at the first day of the year, is allocated to a group at the discretion of the administrative worker. Reallocating a student to a different group only happens in case of special circumstances, such as when a student practices top sports, has special needs, or has some otherwise unresolvable scheduling conflicts. Again, the groups to which these students are reallocated to is at the discretion of the administrator. Our data does not allow us to observe which student registered late or ended up in their group via a reallocation. According to the administrative worker these cases are rare, but may result in slightly different variation in peer ability and class size than would have been observed when strictly following the allocation procedure described above. We present balancing tests in Section 2.4 that cannot reject the final allocation results in a random assignment of students to tutorial, close, and distant peer groups.

2.3 Data

Our main source of data is the administrative database of the university between the academic years 2009-10 and 2014-15. This database includes the complete history of student outcomes and choices at university; grades of all courses followed by the student, first-year tutorial attendance, and second-year tutorial choice. Additionally we observe a rich set of student characteristics; gender, age, resi-

dential address, high school GPA and zip code, and the groups students have been assigned to in their first year. Our baseline results are based on almost 19,000 first-year grades from 2,300 students.¹⁴ This sample only includes a student's first attempt at completing a course. Although we also observe resits, which are taken at the end of the academic year at start of summer, we do not include them in our analysis as they do not require preparation via tutorials.

High school GPA is a 50-50 weighted average of grades obtained during the last three years of high school and on the nationwide standardized exams at the end of high school (before entering university) across all courses. We use high school GPA as a comprehensive proxy for the latent ability of students and their peers. In case of classical measurement error, our estimate for spillovers would be attenuated as students are randomized into groups (Feld and Zöllitz, 2017).¹⁵

2.3.1 Attendance and Student Evaluations

In the first year all students are required to attend at least 70 percent of the tutorials per course. To verify whether the attendance requirements are met, TAs register attendance at the start of each tutorial. This attendance is then uploaded to the university portal and verified at the end of the block by the exam administration. We merge this attendance data with the administrative database, which allows us to observe attendance at the student-tutorial-course level for 98.5 percent of the student-course observations.¹⁶

At the end of the course, students are invited by email to fill in student evaluations. A set of 20 questions are asked covering 9 characteristics of the course, which are detailed in Appendix Table A.2.2. Merging the student evaluations to the administrative data gives a response rate of roughly 30 percent. Column (1) of Appendix Table A.2.8 reveals that participating in the course evaluation is selective. Students with a better high school GPA are more likely to respond. However, column (1) also shows the absence of a relationship between the high school GPA of a student's close peers and their response rate. Results using the course evaluations should be interpreted with caution, and we use them to provide supplementary evidence on the channels of peer influence.

¹⁴This sample excludes some students. For 227 students we do not observe high school GPA (225 students) or one of the main control variables (2 students). Furthermore, to ensure that peer GPA consists of an appropriate number of students, we dropped fourteen tutorial groups (215 students) for whom we observe less than ten students' GPA in at least one of the two close peer groups. Our results are completely robust to the inclusion of these groups. Note that these groups occurred because of missing data on high school GPA and because some students were reallocated after the initial assignment.

¹⁵There are two potential sources of measurement error in our measure of ability. First, for 50 percent high school GPA is determined via unstandardized school exams. It should be noted, however, that the Dutch Inspectorate of Education pays strong attention to schools where the grades on school exams deviate more than 0.5 points from grades on the nationwide standardized exams (DUO, 2014). Second, although students have followed the same level of education in high school (pre-scientific), entering the last three years of high school students must choose one of four tracks. Though these tracks share compulsory courses (such as Dutch), some courses between tracks differ. For a subsample we can show that over 70 percent of our students followed the same track.

¹⁶For our grade-analysis we use the whole sample. Results are identical for the sample that is matched to the attendance data. We verified that peer high school GPA cannot explain whether a student is matched.

2.3.2 Descriptive Statistics

Table 2.1 shows the descriptive statistics by cohort. Panel A provides an overview of the student characteristics. Panel B does the same for student outcomes. All student characteristics show similar values across cohorts. The percentage of women fluctuates somewhat around 20 percent, the students are on average 19.5 years old halfway into their first year, and their high-school GPA is close to the nationwide average of 6.7 (scale from 1 to 10, a 5.5 is sufficient). Appendix Figure A.2.1 shows histograms of student's own high-school GPA, the leave-out mean for the tutorial- and close peer group, and the mean for the distant peer group. Notice that, in contrast to the leave-out mean for the close peer group, the mean for the distant peer group takes upon identical values for everybody in the same subgroup. This explains the somewhat more discrete nature of this figure. A histogram of the leave-in mean for the close peer group is similar to the mean for the distant peer group.¹⁷

Table 2.1 further shows that the size of the close peer group fluctuates between 12 and 14 students. In 2009 the groups were somewhat larger due to an unexpectedly high number of enrolled students. University grades seem to gradually increase, also reflected by the increase in the number of credits earned. This is most likely the consequence of stricter academic dismissal policies introduced halfway in our sample. Course dropout occurs if a student does not attend the final exam for that particular course. Across cohorts, 8 to 19 percent of the students dropped out of both courses in block 5, the final block of the first year. We refer to this as student dropout.

2.4 Empirical Specification

To derive our empirical model we start with the canonical specification for peer effects as laid out by Manski (1993):

$$Y_{igct} = \alpha_0 + \alpha_1 \bar{Y}_{(-i)gc} + \alpha_2 \bar{GPA}_{(-i)g} + \alpha_3 GPA_i + \mu_{gct} + \epsilon_{igct}$$

Where Y_{igct} is the grade at university of student i in tutorial group g on course c of cohort t . GPA_i is the average grade obtained in high school and the variables $\bar{Y}_{(-i)gc}$ and $\bar{GPA}_{(-i)g}$ are leave-out means for tutorial group g for student i of university grades and high school GPA respectively. Everything else that is common to tutorial group g is captured by μ_{gct} .

In the terminology of Manski (1993), α_1 measures the endogenous effect of peers' outcomes on the outcome of student i , α_2 captures the exogenous effect of pre-determined peer characteristics, and

¹⁷ Angrist (2014) shows that using leave-in means, rather than leave-out means, would only change the peer-effects estimate for close peer high school GPA by a factor of $N_g/(N_g - 1)$, where N_g is the size of close peer group g . Therefore, this distinction has little to no importance for our results.

Table 2.1: Descriptive Statistics per Cohort

	2009-10	2010-11	2011-12	2012-13	2013-14	2014-15
	Mean (SD)					
Panel A: Student Characteristics						
Female	0.21 (0.41)	0.21 (0.41)	0.22 (0.42)	0.22 (0.42)	0.21 (0.40)	0.23 (0.42)
Age	19.54 (1.65)	19.62 (1.29)	19.61 (1.28)	19.57 (1.57)	19.67 (1.34)	19.48 (1.42)
Distance to University (km)	21.77 (26.37)	24.03 (30.96)	21.62 (26.20)	22.56 (26.32)	26.39 (31.96)	18.08 (20.32)
Own High School GPA	6.72 (0.54)	6.60 (0.48)	6.63 (0.49)	6.62 (0.47)	6.68 (0.56)	6.68 (0.47)
Tutorial High School GPA	6.72 (0.09)	6.60 (0.10)	6.63 (0.10)	6.62 (0.10)	6.68 (0.13)	6.68 (0.09)
Close Peer High School GPA	6.72 (0.12)	6.60 (0.12)	6.63 (0.16)	6.62 (0.13)	6.68 (0.17)	6.68 (0.14)
Distant Peer High School GPA	6.73 (0.11)	6.60 (0.11)	6.63 (0.15)	6.62 (0.13)	6.69 (0.17)	6.68 (0.14)
35.30 (1.52)	26.84 (2.80)	22.31 (1.15)	22.08 (1.32)	26.12 (1.29)	24.17 (1.63)	
Tutorial Group Size	17.72 (1.35)	13.51 (1.85)	11.19 (0.88)	11.08 (0.94)	13.12 (1.10)	12.12 (1.06)
Close Peer Group Size	458	371	356	308	442	361
Number of Students						
Panel B: Student Outcomes						
Grades	5.98 (1.76)	5.91 (1.71)	6.38 (1.55)	6.06 (1.64)	6.21 (1.68)	6.35 (1.41)
Attendance	0.89 (0.16)	0.89 (0.12)	0.89 (0.10)	0.88 (0.11)	0.88 (0.11)	0.89 (0.10)
Number of Student-Grades Obs.	3598	2999	3098	2580	3462	2999
Number of Student-Att. Obs.	3433	2955	3094	2577	3436	2950
Number of Credits per Student	29.74 (1.62)	30.06 (18.88)	37.96 (18.12)	33.53 (20.59)	32.39 (21.61)	40.00 (20.69)
Number of Courses per Student	8.49 (2.48)	8.79 (2.26)	9.29 (1.84)	8.76 (2.25)	8.43 (2.55)	8.94 (2.24)
Dropout	0.18 (0.38)	0.16 (0.37)	0.08 (0.27)	0.15 (0.36)	0.19 (0.39)	0.12 (0.33)

Notes:

1. Table shows the mean and standard deviation per cohort of student characteristics (Panel A) and student outcomes (Panel B). Panel B is further divided into student-course level outcomes (first section) and student level outcomes (second section).
2. Age is evaluated on January 1st in the academic year that the cohort started. Distance to University refers to the number of kilometers from a student's registered address to the university. High school GPA and university grades are unstandardized, measured on a scale from 1 to 10.
3. Dropout is the fraction of students who did not write an exam in the last block of the first year (block 5).

μ measures the correlated effects capturing, for example, common shocks such as a good TA. The distinction between α_1 and α_2 reveals little about the channels, but it does have different implications for policy, as endogenous effects might generate a social multiplier.¹⁸ However, identification of α_1 is obscured, mostly due to the well-known reflection problem; did the peers affect student i , or did student i affect her peers? As such we follow most of the previous peer effects literature and solve for the reduced form.

2.4.1 Reduced-Form Peer Effects

Assuming that the number of peers within tutorial group g approaches infinity we arrive at the standard linear-in-means specification:

$$\overline{Y_{igct}} = \beta_0 + \beta_1 \overline{GPA}_{(-i)g} + \alpha_3 GPA_i + \beta_2 \mu_{gct} + \tilde{\epsilon}_{igct} \quad (2.1)$$

Where $\beta_1 = \frac{\alpha_2 + \alpha_1 \alpha_3}{1 - \alpha_1}$. Subsequently a test for whether β_1 is different from zero is a test for the presence of peer effects, may they be exogenous and/or endogenous.

The institutional manipulation of the social proximity between students and their tutorial peers allows us to extend this standard model. We make a distinction between the leave-out mean of the close peer group $\overline{GPA}_{Close_{(-i)g}}$ and the mean of the distant peer group $\overline{GPA}_{Distant_g}$. To identify the separate potential channels we replace $\overline{GPA}_{(-i)g}$ in Equation (2.1) by the following expression:

$$\overline{GPA}_{(-i)g} = \frac{N^C - 1}{N^C + N^D - 1} \overline{GPA}_{Close_{(-i)g}} + \frac{N^D}{N^C + N^D - 1} \overline{GPA}_{Distant_g}$$

Where N^C and N^D are the total number of students in the two subgroups within a tutorial group. In practice, $N^C = N^D = 13$. This substitution allows us to arrive at the following specification:

$$\overline{Y_{igct}} = \beta_0 + \beta_1^C \overline{GPA}_{Close_{(-i)g}} + \beta_1^D \overline{GPA}_{Distant_g} + \alpha_3 GPA_i + \beta_2 \mu_{gct} + \tilde{\epsilon}_{igct} \quad (2.2)$$

Estimates of this equation allow us to separate the two peer effect channels possibly at work. Equation (2.2) tests the restriction of Equation (2.1) that the spillovers β_1 from close and distant peers are identical. Recall that the only distinction between an individual's close and distant peers is that social

¹⁸When referring to the social multiplier, Manski (1993) uses the example of a tutoring program. If such a program is provided to only one half of the student population, it might indirectly help the other half of the students as well, as peers' outcomes affect each other.

proximity was induced with the former, whereas no social proximity exists with the latter.¹⁹ Hence, the difference between β_1^C and β_1^D captures peer effects through the social proximity channel. If β_1^C and β_1^D are approximately equal, this indicates that peer effects work solely through classroom-level effects.²⁰

Consistent with their definitions, the two channels are presented as being substitutes in the production of student grades. However, to capture possible complementarities between social proximity and classroom-level effects, some specifications will also include an interaction between close and distant peer ability.

The peer group meeting intervention that encouraged social proximity permits the investigation of the mechanisms underlying peer effects. In order for our results to be generalizable however, we must assume that the intervention itself does not alter the nature of the mechanisms through which peer effects operate in the classroom. In the counterfactual scenario in which social proximity between close peers was not encouraged, we think our finding of no classroom-level effects would hold. It seems unlikely that a non-invasive intervention of little duration would comprehensively change the nature of classroom peer effect channels. Instead, our findings suggest that without the intervention the spillovers from tutorial peers would be smaller than what we observe, and would diminish at a faster rate.

2.4.2 Balancing Tests

As the average high school grade is a predefined measure, we avoid the reflection problem and the estimates for β_1 are unlikely to be biased by common shocks. The main identifying assumption, however, is that peer high school GPA is uncorrelated with other characteristics that might determine a student's grade. As we are not able to observe all other characteristics that might be important for grades, we need the covariance between $\overline{GPA}_{(-i)g}$ and $(\mu_{gct}, \tilde{\epsilon}_{igct})$ to be zero. Random assignment of students to groups makes this identifying assumption likely to hold.

We test this identifying assumption in several ways. First, we analyze whether the treatment, in the form of assigned peer ability, can be explained by background characteristics (X_i) or high school

¹⁹In practice, we cannot rule out ex-ante that some social proximity exists between a student and her distant peers. If this was the case, we would overestimate the importance of classroom-level effects and underestimate the importance of social proximity. Our finding of zero for β_1^D implies that there was no meaningful social proximity between students and their distant peers.

²⁰In fact, because the mean GPA from the distant peer group contains one more student than the leave-out mean of the close peer group, if the spillovers from close and distant peers are identical then $\beta_1^C = \beta_1^D(\frac{12}{13})$. We confirm this in a simulation, in which we arbitrary re-allocate existing tutorial peer groups into placebo close peer groups 1,000 times. Estimating Equation (2.2) and taking the average of the estimates we verify that $\hat{\beta}_1^C \approx \hat{\beta}_1^D(\frac{12}{13})$. For practical testing purposes we deem this as sufficiently close to equality.

GPA:

$$\overline{GPA}_{(-i)g} = \gamma_0 + \gamma_1 X_i + \gamma_2 GPA_i + T_t + \epsilon_{igt}$$

We include cohort fixed effects (T_t) as randomization into groups takes place cohort-by-cohort. Estimates of γ_1 or γ_2 that are different from zero most likely violate the identifying assumption mentioned above. Table 2.2 shows the results of this test, where column (1) to (3) take tutorial, close, and distant peer high school GPA as outcome variables respectively. Across the three specifications we find all student characteristics to be individually and jointly insignificant.²¹ This stands in stark contrast to the joint significance of student characteristics in a regression where first-year GPA at university is taken as an outcome variable (p -value < 0.000).

Our second balancing test is more flexible. We regress background characteristics - student number, gender, age, and distance to university - and high school GPA on close peer group dummies and cohort fixed effects. Next, in a separate model we regress the student characteristics upon cohort fixed effects only and perform a F-test on the small versus big model. This test would reveal if students with certain characteristics cluster together in certain groups. Appendix Table A.2.3 shows the F-test does not reject the null hypothesis for all student characteristics. In other words, a small model with cohort fixed effects only is favored above a model that also includes close peer group dummies.

We perform a similar analysis per cohort. We regress each student characteristic on a set of close peer group dummies separately for each cohort. Appendix Figure A.2.2a plots the histogram of the p -values of the close peer group dummies obtained from these regressions. As expected under randomization, the p -values are roughly uniformly distributed, where for instance roughly 10 percent of the p -values are below 0.10. Figure A.2.2b shows the results for this analysis are identical if close peer group dummies are replaced with tutorial group dummies. A Kolmogorov-Smirnov equality of distribution test does not reject the null-hypothesis of a uniform distribution in both cases; the p -values are equal to 0.86 and 0.60 for the histograms belonging to the close- and tutorial peer group dummies respectively.

Allocation of teaching assistants to tutorial groups is done for each course by the instructor of that specific course. Our analysis would still be compromised if instructors base the TA assignment upon tutorial group ability. Instructors are unaware of the GPA composition of the tutorial groups and base the assignment of the TAs upon scheduling restrictions. To confirm this, we code the gender of the TA and whether he or she was a PhD. If coordinators base their decisions on the difficulty of groups,

²¹If we regress student high school GPA on peer high school GPA we reach identical conclusions. Guryan et al. (2009) argue this balancing test should also control for the mean high school GPA of all peers that can be matched with student i in group g . In our case this control would be the leave-me-out mean GPA of her cohort. This is infeasible as there is no variation in the group that student i can be matched too. Indeed, GPA_i is related to the mean GPA of her cohort \overline{GPA}_t and the leave-me-out mean GPA of her cohort, $\overline{GPA}_{(-i)t}$, by the following identity: $GPA_i = N \times \overline{GPA}_t - (N - 1) \times \overline{GPA}_{(-i)t}$.

Table 2.2: Balancing Tests for Peer Ability

	Tutorial Peer GPA	Close Peer GPA	Distant Peer GPA
	(1)	(2)	(3)
Student Number	-0.0157 (0.0410)	-0.0187 (0.0451)	-0.0077 (0.0401)
Female	-0.0339 (0.0376)	-0.0319 (0.0457)	-0.0212 (0.0504)
Age	-0.0081 (0.0220)	-0.0024 (0.0232)	-0.0100 (0.0191)
Distance to University	-0.0132 (0.0145)	0.0022 (0.0173)	-0.0227 (0.0151)
Own GPA	0.0076 (0.0281)	-0.0171 (0.0283)	0.0285 (0.0255)
Observations	2296	2296	2296
Adjusted R^2	0.151	0.085	0.098
F-test	0.25	0.26	0.77
p -value	0.938	0.933	0.570

Notes:

1. All regressions also include cohort fixed effects.
2. Peer GPA refers to the leave-out mean of high school GPA for the tutorial- and close peers, and to the mean for distant peers. All dependent and independent variables are standardized except for the female dummy.
3. The F-test, and corresponding p -value, refer to a test for the joint significance of all the independent variables shown in the table.
4. Standard errors in parentheses, clustered on the tutorial level.
5. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

they might, for example, assign PhD's to low GPA groups. Regressing TA type on tutorial peer GPA, however, shows that coordinators do not base TA assignment on class composition (see Appendix Table A.2.4). The same assignment method is used for the discussion leaders that guide the close peer group, though we cannot confirm this empirically as we do not observe these discussion leaders in our data.

We conclude that we are able to identify reduced-form peer effects and estimate Equation (2.1) and (2.2) without controlling for μ_{gct} . Throughout all specifications we will, however, include course-cohort fixed effects and background characteristics; student number, gender, age, and distance to university. The baseline results are identical when we do not control for background characteristics. We cluster standard errors at the tutorial level, which nests the close-peer-group level cluster. Own

GPA, peer GPA, and the outcome variables (when suitable) are standardized over the estimation sample, such that the estimates can be interpreted in terms of standard deviations.

2.5 Baseline Results

Before presenting the baseline results for grades and passing rates, we discuss the extent to which course- and student dropout could potentially bias our estimates. Table 2.1 shows that the student dropout rate at the end of first year is relatively low; between 8 and 19 percent across the six cohorts. In Section 2.5.3 we will show that average peer high school GPA has no impact on the number of courses a student attends the final exam for nor on whether the student dropped out by the end of first year. We can show, but omit for brevity, that these null-results for number of courses and student dropout extend to the non-linear model used in Section 2.5.5. Selection bias therefore does not contaminate the following baseline peer effects estimates.

2.5.1 First-Year Grades and Passing Rates

Table 2.3 shows our baseline results, where panel A regresses first-year grades upon average peer high school GPA. Column (1) shows the estimated effect of tutorial peers. The positive coefficient has a *p*-value of 0.11 and shows that a one standard deviation increase in tutorial peer high school GPA increases a students' first year grade by 0.019 standard deviations. Columns (2) and (3) show the effect while separating the tutorial group by one's close- and distant peers. This reveals that the positive spillovers are entirely driven by close peers. The estimate for peer GPA when moving from tutorial to close peers in column (2) increases somewhat in magnitude and precision. It is statistically significant at the 5%-level. The estimate for distant peers in column (3) is economically and statistically indistinguishable from zero. Column (4) shows the estimates for close- and distant peer high school GPA are identical to (2) and (3) respectively when including both peer measures in one regression. These results imply that peer effects are entirely driven by social proximity.

In terms of the Dutch grading scale, columns (2) and (4) imply that increasing the close peers' high school GPA from 6.5 to 7 increases a student's grade from 7 to 7.14. This is economically small, but 2.1 times the size of Feld and Zöllitz (2017), while Booij et al. (2017) find no peer spillovers in their linear-in-means specification. Both of these studies investigate spillovers in a similar context as ours; classroom peer effects at a public university in the Netherlands. This suggests that fostering social proximity has the capacity to generate larger spillovers than previously found in the literature.

Whereas students with good peers obtain higher grades, they are not necessarily better off if the only goal is to pass courses. We study the probability of passing a first-year course in panel B

Table 2.3: Peer Effects on First-Year Course Grades (Panel A) and Pass or Fail (Panel B)

	(1)	(2)	(3)	(4)	(5)
Panel A: Grades (Standardized)					
Tutorial Peer GPA	0.0191 (0.0118)				
Close Peer GPA		0.0255** (0.0104)		0.0254** (0.0106)	0.0256** (0.0109)
Distant Peer GPA			0.0034 (0.0131)	0.0008 (0.0130)	0.0010 (0.0131)
Close \times Distant Peer GPA					-0.0150 (0.0122)
Own GPA	0.3427*** (0.0119)	0.3434*** (0.0119)	0.3427*** (0.0118)	0.3433*** (0.0119)	0.3433*** (0.0120)
Observations	18736	18736	18736	18736	18736
Adjusted R^2	0.323	0.323	0.322	0.323	0.323
Panel B: Pass (1) or Fail (0)					
Tutorial Peer GPA	0.0090** (0.0043)				
Close Peer GPA		0.0080* (0.0042)		0.0075* (0.0043)	0.0075* (0.0043)
Distant Peer GPA			0.0056 (0.0049)	0.0048 (0.0049)	0.0048 (0.0049)
Close \times Distant Peer GPA					-0.0005 (0.0046)
Own GPA	0.1186*** (0.0048)	0.1189*** (0.0048)	0.1183*** (0.0047)	0.1187*** (0.0048)	0.1187*** (0.0048)
Observations	18736	18736	18736	18736	18736
Pseudo R^2	0.187	0.187	0.187	0.187	0.187

Notes:

1. All regressions include course-cohort fixed effects and controls; student number, gender, age, and distance to university.
2. Peer GPA refers to the leave-out mean of high school GPA for the tutorial- and close peers, and to the mean for distant peers. Own GPA refers to own high school GPA. All GPA measures are standardized.
3. Standard errors in parentheses, clustered on the tutorial level.
4. Panel A is estimated with OLS, Panel B uses Probit. Marginal effects are reported.
5. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

of Table 2.3, where the outcome variable is replaced with a pass-fail indicator. Column (1) shows that a one standard deviation increase in the high school GPA of one's tutorial peers increases the probability of obtaining a sufficient grade by 0.9 percentage points. This effect is significant at the 5%-level. Again, columns (2) to (4) show that these spillovers originate entirely from close peers.

Column (5) of both panel A and B includes an interaction effect between high school GPA of the close and distant peers. This interaction effect tests for possible complementarities between social proximity and classroom-level effects. For instance, having a superstar student in class posing insightful questions may only increase grades if one has high ability close peers to discuss the questions with. We find this interaction term is negative for grades and the probability of passing, but insignificant in both cases. We interpret this as showing that complementarities between both channels are unlikely to play a role.

2.5.2 Randomization Inference

The results above use analytic standard errors. In this section we present *p*-values based on randomization inference for the baseline results on first-year grades, an alternative inference approach that does not rely on large sample approximations. This method involves re-drawing a large number (10,000) of randomly assigned hypothetical tutorial and close peer groups, respecting the size of the original groups. For each of these hypothetical groups, we re-run the models presented in Panel A of Table 2.3 in order to assess the effect of the hypothetical peers' high school GPA on students' first-year grades. Comparing the actual estimate to the estimates from the simulated groups allows us to test the sharp null hypothesis that peer effects are equal to zero (Athey and Imbens, 2017). The results for the corresponding exact *p*-values are presented in Appendix Table A.2.5 and Figure A.2.3, which are nearly identical to those presented in Table 2.3. Given the similarity between the two approaches, the remainder of this paper uses analytic standard errors.

Additionally, these results address one of the concerns of Angrist (2014). He shows that the peer effects estimate is identical to the (scaled) difference between a 2SLS estimator using peer group dummies as instruments for individual high school GPA and an OLS estimator of individual GPA. In some settings this may lead to a spurious, mechanically driven finding of peer effects. In our setting, however, with random assignment of students to many small groups, there is little reason for this estimate to be different from zero in the absence of spillovers (Angrist, 2014). This is confirmed by the fact that the peer effect coefficients from the 10,000 hypothetical groups, containing unconnected students, are centered around zero.

2.5.3 Additional Outcomes

In this section we turn our attention to five additional first-year outcomes: credit weighted GPA, number of credits, number of courses taken, student dropout, and tutorial attendance. We analyze the first four of these outcomes by estimating our baseline equations on the student level.

Table 2.4 shows the results, where columns (1) and (2) reveal that the positive effects on grades and passing rates have a cumulative effect on a student's GPA (p -value<0.01) and the number of credits she collects (p -value=0.13). The estimates indicate that a one standard deviation increase of close peer high school GPA increases a student's credit weighted GPA (total first-year credits) by roughly 0.04 standard deviations (0.52 credits). Column (3) and (4) reveal this increase in student performance is not due to the fact that peers impact the number of courses a student writes the final exam for.²² Column (5) shows that peer GPA does not change the probability of student dropout, which is measured by an indicator variable that takes the value one if a student was no longer active in block five of their first year.

Appendix Table A.2.6 shows the results when analyzing the impact of peer high school GPA on the percentage of tutorials attended per course in the first year. These estimates show that peers do not have an effect on average tutorial attendance. Recall, however, that students are required to attend 70 percent of the tutorials per course, so the scope for any improvement would be limited.

2.5.4 Robustness

The results above show that peer GPA does not affect dropout, which implies our results are not contaminated by selection bias. However, the estimate for own GPA in Table 2.4 reveals that low GPA students take fewer courses and have a higher probability of dropping out by the end of their first year. This means that a student randomized into a tutorial group with many low ability rather than high ability students will experience a larger amount of course dropout among her peers, and thus have a smaller actual class size. This results in a positive correlation between peer GPA and class size, which could partly explain our baseline results if class size also impacts grades. Appendix Figure A.2.4 plots the number of students writing the final exam as a fraction of the initial students per block and separately for high, average, and low GPA close peer groups. This reveals that dropout increases during the year, being 15 to 20 percent at the end of the first year. It also reveals that dropout is somewhat larger for low GPA close peer groups.

We investigate whether our results are robust to class size and course dropout in Table 2.5, which presents the results of our baseline equation while including variables measuring class size and course

²²Note that this also implies that course dropout is not influenced by peer GPA.

Table 2.4: Peer Effects on Additional Outcomes

	GPA Weighted by Credits	Number of Credits	Number of Courses	Followed the Course?	Dropout Balanced Panel
				(4)	
Close Peer GPA	0.0450*** (0.0146)	0.5230 (0.3462)	0.0269 (0.0497)	0.0034 (0.0051)	0.0009 (0.0083)
Own GPA	0.5073*** (0.0168)	8.7081*** (0.3560)	0.4747*** (0.0438)	0.0539*** (0.0054)	-0.0693*** (0.0082)
Observations	2218	2218	2218	22180	2218
R^2	0.300	0.241	0.062	0.048	0.056
Binary Outcome	No	No	No	Yes	Yes

Notes:

1. All regressions include cohort fixed effects and controls; student number, gender, age, and distance to university.
2. Peer GPA refers to the leave-out mean of high school GPA for the close peer group. Own GPA refers to own high school GPA. Both GPA measures are standardized.
3. Column (1), (2), (3) and (5) are estimated on the student level. Column (4) creates a balanced panel on the student-course level, where the outcome variable takes the value one if a student wrote the final exam for that course and zero otherwise.
4. Column (1) has first-year credit weighted GPA as outcome variable and is based on the number of courses that the student took. Column (2), (3) and (4) refer to the number of credits obtained or the number of courses a student wrote the final exam for. Dropout in column (5) is one if a student did not write an exam in the last block of the first year and zero otherwise. Credit weighted GPA in column (1) is standardized, all other outcomes are unstandardized. Number of credits range from 1 to 60. Number of courses range from 1 to 10.
5. Across the six cohorts there are 78 students (3.4%) who confirmed their registration on the first day but for whom we observe no valid grade. These students dropped out before the first exam week. As we cannot calculate a GPA for them, these students are dropped from this analysis. Results do not change when we include these students.
6. Column (1), (2) and (3) are estimated with OLS, column (4) and (5) with Probit. Marginal effects are reported. The R^2 refers to the Adjusted and Pseudo R^2 respectively.
7. Standard errors in parentheses, clustered on the tutorial level.

8. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

dropout as explanatory variables. Column (1) includes a dummy for the assigned number of students to the close peer group at the start of the first year, column (3) for the actual number of students that wrote the exam for the course, and column (6) for the difference between the two. The latter is a measure for dropout per course. All three columns reveal a stable estimate for close peer GPA, suggesting that class size and course dropout are unlikely to explain our baseline results.

Columns (2), (4), and (7) of Table 2.5 include the assigned class size, actual class size, and course dropout as continuous variables, while also including their interaction with close peer GPA. The measures for original and actual class size in column (2) and (4) are standardized, while the difference between the two in column (7) is unstandardized. As such, the estimate for close peer GPA

Table 2.5: Robustness of Baseline Peer Effects

	Grades (Standardized)						
	OLS (1)	OLS (2)	OLS (3)	OLS (4)	IV (5)	OLS (6)	OLS (7)
Close Peer GPA	0.0254** (0.0112)	0.0253** (0.0107)	0.0275** (0.0107)	0.0289*** (0.0108)	0.0291*** (0.0113)	0.0261** (0.0103)	0.0387*** (0.0146)
Peer GPA × Assigned Class Size		-0.0059 (0.0093)					
Peer GPA × Actual Class Size				0.0042 (0.0076)	-0.0043 (0.0112)		
Peer GPA × (Assigned-Actual)							-0.0056 (0.0052)
Own GPA	0.3435*** (0.0120)	0.3436*** (0.0119)	0.3436*** (0.0119)	0.3438*** (0.0119)	0.3438*** (0.0119)	0.3433*** (0.0119)	0.3433*** (0.0119)
Observations	18736	18736	18736	18736	18736	18736	18736
Adjusted R^2	0.324	0.323	0.323	0.323	0.323	0.323	0.323
F-tests on Excl. Instruments			Assigned Class Size: Its Interaction with Peer GPA:		141.85 475.24		
Class-Size Dummies	Yes	No	Yes	No	No	Yes	No
Robustness Check	Assigned Class Size		Actual Class Size			Assigned-Actual)	

Notes:

1. All regressions include course-cohort fixed effects and controls; student number, gender, age, and distance to university.
2. Peer GPA refers to the leave-out mean of high school GPA for the close peer group. Own GPA refers to own high school GPA. Both GPA measures are standardized.
3. Column (1) includes dummies for the number of students at the beginning of the year in the close peer group (assigned class size), column (3) for the number of students that wrote the exam for the course (actual class size on the course-cohort level), and column (6) for the difference between these two (assigned-actual). The latter is a measure for course dropout.
4. Assigned and actual class size are standardized in column (2), (4) and (5). The difference between the two in column (7) is not standardized. The coefficient on close peer GPA in column (7) measures spillovers in classes where there has been no course dropout (assigned-actual=0). Roughly 20 percent of the groups experience no course dropout and have a value of zero, where the average is 2.19.
5. Column (5) uses the assigned number of students and its interaction with close peer GPA as instruments for the actual number of students and for its interaction with close peer GPA.
6. Standard errors in parentheses, clustered on the tutorial level.
7. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

in (7) measures the peer effect for groups where there has been no course dropout. Again we find stable estimates for close peer GPA across all three columns. Moreover, we find the estimates for the interaction terms between peer GPA, class size, and course dropout to be unimportant. From this we conclude that the social proximity, and the corresponding nature of spillovers, is not different between classes of different size.

Whereas assigned class size is exogenous, one may have remaining concerns that actual class size is an outcome of close peer GPA. Therefore we report an additional specification in column (5) of Table 2.5, where we use assigned class size as an instrument for actual class size.²³ Using only the variation in actual class size that originates from the original assignment, we find our results to be virtually unchanged.

2.5.5 Heterogeneity

Do the baseline estimates of Section 2.5.1 hide heterogeneity by own and peer ability? This question has important implications for policy. It is only when peer effects are non-linear that aggregate gains can be generated by reorganizing peer groups.

Following Carrell et al. (2013) we test for heterogeneity using a two-way interaction model. We define low and high ability students to be in the bottom and top quartiles of high school GPA across the six cohorts. The remaining 50 percent of students are defined as being of average ability. For every student we calculate the (leave-out) proportion of low, middle, and high ability students separately for their close and distant peer groups. We estimate models with interactions of student's own ability type with the fraction of high and low ability peers. For each ability type, these interactions show the impact of increasing the proportion of high or low ability students by decreasing the proportion of average ability students. For example, Own Low \times Peer High shows the estimated effect on student performance for low ability students of increasing the proportion of high ability students by decreasing the proportion of average ability students in the relevant peer group.²⁴

Table 2.6 presents our results. Column (1) and (2) first document that our baseline results from the linear-in-means specifications carry over to a model where we use the share of high and low ability students, rather than the mean of peer high school GPA, to measure peer ability. Next, column (3) and (4) show the heterogeneity results on first-year grades for the close and distant peer groups respectively, where column (5) and (6) do this for a pass-fail indicator. The results in column (3) and (5) reveal spillovers that are roughly linear in close peer ability, implying that the estimates of the linear-in-means model are insightful. Specifically, the columns show that the observed close peer spillovers are driven primarily by low and high ability students benefiting from social interactions with high ability students. Both high and low ability students are negatively affected by increasing

²³The variation in assigned class size comes partly from the original allocation and partly from the cases in which the administrator reallocates students across tutorial and close peer groups (see Section 2.2).

²⁴Appendix Table A.2.7 repeats the first balancing test described in Section 2.4 while replacing average peer GPA as the outcome variable separately with the (leave-out) share of low, average, and high ability peers in the close and distant peer group. We find that student characteristics cannot explain the share of peers by ability type; only two out of the 35 estimated coefficients (γ_1 and γ_2) are significant, and the characteristics are always jointly insignificant. This result holds if we perform this balancing test separately for low, average, and high ability students.

Table 2.6: Heterogeneity by High School GPA of Peer Effects

	Grades (Standardized)				Pass (1) or Fail (0)	
	Close	Distant	Close	Distant	Close	Distant
	(1)	(2)	(3)	(4)	(5)	(6)
Share Peer High	0.1774*	0.0096				
	(0.0915)	(0.1104)				
Share Peer Low	-0.0483	0.0171				
	(0.1064)	(0.1108)				
Own High × Peer High			0.3659**	0.0007	0.2258***	0.1069
			(0.1456)	(0.1701)	(0.0730)	(0.0789)
Own High × Peer Low			-0.3036**	0.1024	-0.1141*	0.0329
			(0.1507)	(0.1506)	(0.0614)	(0.0666)
Own Avg × Peer High			-0.0257	-0.0687	-0.0081	-0.0153
			(0.1086)	(0.1239)	(0.0434)	(0.0508)
Own Avg × Peer Low			0.1063	0.0474	0.0698	0.0506
			(0.1196)	(0.1272)	(0.0463)	(0.0509)
Own Low × Peer High			0.3510**	0.1987	0.1146*	0.0884
			(0.1503)	(0.1624)	(0.0593)	(0.0655)
Own Low × Peer Low			-0.1492	-0.1654	-0.0400	-0.0689
			(0.2212)	(0.2002)	(0.0794)	(0.0692)
Observations	18736	18736	18736	18736	18736	18736
R^2	0.323	0.322	0.324	0.323	0.188	0.187
Binary Outcome	No	No	No	No	Yes	Yes

Notes:

1. All regressions include course-cohort fixed effects, controls; student number, gender, age, and distance to university, and own high school GPA.
2. Students are classified into dummies that refer to the bottom 25 percent (low), middle 25 to 75 percent (average), and top 25 percent (high) of high school GPA. The peer measures are the (leave-out) shares of students in the close (distant) peer group belonging to each category. The shares are unstandardized.
3. Odd columns include the shares for the close peer group and even columns for the distant peer group.
4. Column (1) to (4) are estimated with OLS, column (5) and (6) with Probit. Marginal effects are reported. The R^2 refers to the Adjusted and Pseudo R^2 respectively.
5. Standard errors in parentheses, clustered on the tutorial level.
6. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

the share of low ability students, insignificantly so for low ability students. Increasing the share of either high or low ability students appears to have no impact on average ability students. Conversely, column (4) and (6) show that the proportion of high and low ability types in one's distant peer group has no significant effect on grades or passing rates for any ability type, further supporting the lack of classroom-level effects.

The coefficient for Own High \times Peer High in column (3) reveals that increasing the share of high ability students by 25 percent, the equivalent of replacing 3 out of 12 average ability students with 3 high ability students, increases the grade of a high GPA student by almost 0.1 standard deviation. To get a sense of the size of this effect, we follow Marie and Zöllitz (2017) and compare it to other treatments known to have an impact on student performance in higher education. An estimate of 0.1 standard deviation is roughly twice the size of having a same-sex instructor (Hoffmann and Oreopoulos, 2009), resembles the effect of increasing professor quality by one standard deviation (Carrell et al., 2009), and is similar to the impact of a temporary restriction of legal cannabis access (Marie and Zöllitz, 2017). It is perhaps useful to remark that 0.1 standard deviation corresponds to approximately half of the math gender gap in the fifth grade in the U.S. (Fryer Jr and Levitt, 2010).

In an additional analysis, we considered more restrictive definitions of high and low ability students to better reflect the concept of having superstar students or bad apples in the classroom. In particular, we defined superstar students as those having a GPA above 8.25 (cum laude) and bad apples as those having a GPA below 5.75; both categories form roughly one percent of our sample. Subsequently we constructed close and distant peer group dummies which are equal to one if the group contained such a student. Replicating the regression in column (1) of Table 2.6, while replacing the shares with a close-peer-group superstar and bad-apple dummy, we find the first is significantly positive and the latter to be insignificantly negative. Similar to column (2), both the superstar and bad-apples dummy for the distant peer group are smaller and statistically insignificant. Separating these effects by students' own ability, we find a similar pattern for low, average, and high ability students as documented in column (3) and (4) of Table 2.6. These results further support peer effects revolve around meaningful social interaction between peers, rather than classroom-level effects (results available upon request).

2.5.6 Group Assignment Policies

The previous results imply that alternative assignment policies entail a transfer from one student group to the other. Therefore it is not possible to provide a Pareto-ranking of different policies. However, we can use the results in Table 2.6 to estimate the effects of alternative assignment policies. University administrators that want to maximize student grades can use such an exercise to weigh the grade benefits of one group against the costs of another.

Following Booij et al. (2017) we consider five alternative group assignment policies; low, average, high, three-way, and two-way ability tracking. Table 2.7 summarizes, for the average student as well as per ability type, the estimated change in a first-year course grade when switching from the current ability mixing regime to one of the five tracking policies. According to these estimates,

Table 2.7: Estimated Effects of Alternative Group Assignments Compared to Mixing

		Effect For Student With			
		Average Effect	Low GPA [L]	Avg. GPA [A]	High GPA [H]
		(1)	(2)	(3)	(4)
Track Low	[L],[A,H]	-0.0400 (0.0512)	-0.2030 (0.1829)	-0.0290 (0.0300)	0.1009** (0.0396)
Track Average	[A],[L,H]	0.0062 (0.0115)	0.0476 (0.0589)	-0.0202 (0.0432)	0.0177 (0.0502)
Track High	[H],[L,A]	0.0686** (0.0335)	-0.1036** (0.0486)	0.0148 (0.0271)	0.3484*** (0.1167)
Three-way Tracking	[L],[A],[H]	0.0262 (0.0556)	-0.2030 (0.1829)	-0.0202 (0.0432)	0.3484*** (0.1167)
Two-way Tracking	[L,A],[A,H]	0.0123 (0.0255)	-0.1247* (0.0730)	0.0011 (0.0012)	0.1717*** (0.0551)

Notes:

1. For each alternative group assignment, we randomly allocate students depending on their ability type to groups of 14 to 15 students. The student types are low ability [L], average ability [A], and high ability [H] defined by the bottom quartile, two middle quartiles, and top quartile of high school GPA respectively.
2. Low (average or high) tracking involves grouping low (average or high) ability students together, while mixing the remaining students. Three-way tracking involves separate groups for each ability type. Two-way tracking involves defining students as either high or low ability, depending on whether their high school GPA is above or below the median. Groups are then composed of only high or low ability students.
3. For each student we subtract the actual leave-out ability shares (mixing) from the ability leave-out shares obtained via the alternative group assignments, denoted by $(\bar{x}_{track} - \bar{x}_{mixing})'$. Then the average tracking effects are equal to $(\bar{x}_{track} - \bar{x}_{mixing})'\hat{\beta}$. Note that nearly identical estimates can be derived directly from column (3) of Table 2.6.
4. Standard errors are equal to $\sqrt{(\bar{x}_{track} - \bar{x}_{mixing})'V(\hat{\beta})(\bar{x}_{track} - \bar{x}_{mixing})}$, and shown in parentheses.
5. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

the policy that will deliver the largest increase in student performance is the high tracking policy, whereby high ability students are grouped together and low and average students are mixed to form the remaining groups. Note however that this policy is predicted to decrease grades for low ability students compared to mixing.

A potential concern with using estimates based on ability mixing to inform alternative group assignment policies is that some peer configurations will not be covered by the data.²⁵ Such a problem was encountered by Carrell et al. (2013), who found that extrapolating from estimates based on ability mixing failed to predict the results of alternative group assignments. Given that social proximity is

²⁵This point has recently been made by Booij et al. (2017). In their study they manipulate the composition of groups to achieve a wider range of support in peer ability than observed under ability mixing.

vital for the existence of peer effects, it may well be that such failures can be attributed to social proximity breaking down in more extreme group configurations. Our results are based on a setting in which social proximity has been fostered between close peers. If such fostering is achieved in more extreme group configurations then the results presented here may actually provide an accurate description of what will occur under alternative assignment policies. Given the lack of support they should still be treated with caution, however.

2.6 Nature of Social Interactions

Our results indicate better peers have small positive implications for a student's grades, passing rates, cumulative GPA, and credits in first year. These spillovers originate from peers with whom students are socially proximate and interact with. What is the nature of these social interactions? A possible answer to this question allows us to speak to the finer categorization of possible peer effect mechanisms listed by Sacerdote (2011), including peer-to-peer teaching or effects on student motivation or preferences.

We start with the use of course evaluations. Recall that the response rate, which is roughly 30 percent, is unrelated to a student's close peer GPA (see Appendix Table A.2.8). Hence, we worry little about sample selection when interpreting the following set of results. Table 2.8 uses data on self-reported lecture attendance and total study time to investigate whether the beneficial social interactions changed the inputs regarding the study process. Column (1) reports the effect of close peer high school GPA on an indicator for whether the student attended lectures. Column (2) does this for total study time (tutorials + lectures + self study). The estimates reveal that a student with better close peers is less likely to attend lectures, while reported total study time is not impacted. The estimate in column (1) suggests that a one standard deviation increase in close peer high school GPA decreases the probability to attend lectures by 1.8 percentage points. Due to the rough (binary) nature of the question, however, we are inclined to interpret only its sign and significance (p -value=0.019).²⁶ Recall that Appendix Table A.2.6 showed that tutorial attendance is unaffected by close peer high school GPA. Taken together, the estimates in column (1) and (2) suggest that students with better close peers substituted lecture attendance for additional self study.

Next we investigate the impact of close peer high school GPA on perceived lecturer and TA quality, and the perceived usefulness of lectures and tutorials. Column (3) and (4) indicate that having better close peers significantly decreases the perceived quality of the lecturer and usefulness of the

²⁶For this question students are asked only about the extensive margin of their lecture attendance: "Have you attended lectures?". Even students who attended a few lectures may answer this question with yes (1) instead of no (0). As such, it may well be that these results understate the true reduction in lecture attendance.

Table 2.8: Peer Effects on Time Use and Additional Outcomes using Course Evaluations

	Attended Lectures	Total Study Time	Lecturer Quality	Usefulness Lectures	TA Quality	Usefulness Tutorials
	(1)	(2)	(3)	(4)	(5)	(6)
Close Peer GPA	-0.0180** (0.0077)	-0.1935 (0.1877)	-0.0574*** (0.0195)	-0.0585** (0.0285)	-0.0332 (0.0241)	-0.0064 (0.0281)
Own GPA	-0.0139 (0.0089)	-0.5414*** (0.1484)	0.0348* (0.0204)	0.0268 (0.0201)	0.0029 (0.0191)	0.0024 (0.0308)
Observations	4361	4361	3560	2178	3560	2178
R^2	0.147	0.268	0.245	0.251	0.079	0.124
Binary Outcome	Yes	No	No	No	No	No

Notes:

1. All regressions include course-cohort fixed effects and controls; student number, gender, age, and distance to university.
2. Peer GPA refers to the leave-out mean of high school GPA for the close peer group. Own GPA refers to own high school GPA. Both GPA measures are standardized.
3. The dependent variable in column (1) is the answer to the question “Have you attended lectures?”. The dependent variable in column (2) is the answer to the question “Average study time (hours) for this course per week (lectures+tutorials+self study)?” where we used the maximum for the interval to convert the categories into hours. The dependent variables in column (3) and (5) are the mean of the answers to the questions that evaluate the Lecturer/TA. The dependent variables in column (4) and (6) are the answers to the questions “Were the lectures/tutorials useful?”. The dependent variables in column (3) until (6) are standardized.
4. Column (1) is estimated with Probit, the other columns with OLS. Marginal effects are reported. The R^2 refers to the Pseudo and Adjusted R^2 respectively.
5. Standard errors in parentheses, clustered on the tutorial level.

6. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

lectures. This is consistent with, and further reinforces that, students substitute lecture attendance for additional self study. It seems most likely that this increase in self study involves close peers studying together. However, an alternative explanation might be that the beneficial student-to-student interactions only take place during the tutorials, after which individual self study takes place. If this is the case, we would expect students’ perception of the quality of their TA and the usefulness of tutorials to increase when having better close peers. Column (5) shows that close peer high school GPA is unrelated to students’ perceptions of the quality of the TA. Column (6) shows that there are no effects on the perception of the usefulness of the tutorials. Combining these results suggest that students exposed to better close peers substitute lecture attendance for collaborative self study.²⁷

Next we turn to our data on student gender, high school location, and residence to shed additional light on the nature of the social interactions. We calculate for each student the leave-out proportion of females in their close peer group, the number of peers in their close peer group that attended the same

²⁷Appendix Table A.2.8 reveals no effect of students’ close peers on the remaining questions regarding their perceptions of the course.

high school as the student, the distance of the student's residence to the residence of every student in their close peer group, and the leave-out proportion of a student's close peers that live in the city in which the university is located.

Column (1) of Table 2.9 adds the leave-out share of females in the close peer group while interacting it with the female dummy in our baseline equation. The results replicate the finding of Oosterbeek and Van Ewijk (2014), who also find that the gender composition does not have an effect on student performance. Column (1) also documents an unchanged estimate for close peer high school GPA. This implies that the meaningful social interactions do not only take place in certain high ability groups with a high share of males or females. Column (2) shows that being assigned peers from one's former high school in the close peer group does not have implications for spillovers. Given that high school peers are most likely acquainted before university, this points to spillovers also being generated between formerly unknown peers.

If collaborative study meetings would take place outside university, we would expect to observe larger peer effects for students who live closer to their high ability peers. In column (3) we include the median distance of a student's residence to her close peers and interact this with close peer high school GPA. We do not find that a student who lives closer to her peers enjoys larger spillovers. This suggests that the study meetings take place on the university campus.

The notion that students benefit from collaborative self-study outside class implies that students would fail to benefit from having better close peers if these peers have other commitments that prevent such studying. We attempt at investigating this in column (4), which includes a dummy for whether the student lives in the city of the university and the leave-out share of students within their close peer group living in the city. First, notice column (4) documents that city students perform significantly worse in their first year, scoring on average 0.11 standard deviations lower. Moreover, with our administrative tutorial attendance data we can show that the percentage of first-year tutorials attended per course is 0.07 standard deviations lower for city students (p -value=0.001). We conjecture that these findings partly reflect the large range of extra-curricular activities available to these students, most of whom are living outside of their parent's home for the first time.²⁸ The coefficient for the interaction of close peer ability with the proportion of close peers living in the city implies that peer effects vanish if all of one's close peers live in the city. Although we cannot definitively rule out other

²⁸For example, these activities could include a fraternity membership, which is common among our student population living in the city. From the Dutch student survey "Studenten Monitor" we observe that students living outside of their parent's home spend in total roughly twice as much money on fraternity memberships and roughly 1.5 as much money on leisure activities than students living with their parents (<http://www.studentenmonitor.nl/>).

Table 2.9: Peer Effects by Gender, Prior Bonds and Location

	Grades (Standardized)			
	(1)	(2)	(3)	(4)
Close Peer GPA	0.0250** (0.0104)	0.0286** (0.0110)	0.0254** (0.0104)	0.0190* (0.0099)
Share of Female Peers	-0.0087 (0.0120)			
Female × Share of Female Peers	0.0290 (0.0232)			
Peer Same High School × Peer GPA		-0.0155 (0.0229)		
Distance of Peers to Your Residence × Peer GPA			-0.0091 (0.0089)	
Live in City				-0.1113*** (0.0294)
Share of Peers that Live in City × Peer GPA				-0.0246** (0.0110)
Own GPA	0.3430*** (0.0119)	0.3426*** (0.0116)	0.3435*** (0.0120)	0.3408*** (0.0122)
Observations	18736	18229	18736	18736
Adjusted R^2	0.323	0.324	0.324	0.325

Notes:

1. All regressions include course-cohort fixed effects and controls; student number, gender, age, and distance to university.

2. Peer GPA refers to the leave-out mean of high school GPA for the close peer group. Own GPA refers to own high school GPA. Both GPA measures are standardized.

3. Column (1) includes the leave-out share of females in the close peer group (standardized) and its interaction with the gender dummy. Column (2) includes the number of students that attended the same high school (unstandardized) and its interaction with close peer GPA. For some students we do not observe their high school address, explaining the somewhat fewer number of observations. Column (3) includes the median distance of a students' peers to his or her residence (standardized) and its interaction with close peer GPA. Column (4) includes the leave-out share of peers that live in the city where the university is located (standardized) and its interaction with close peer GPA.

4. Standard errors in parentheses, clustered on the tutorial level.

5. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

competing explanations, we believe this result is most consistent with the existence of spillovers depending upon peers not having busy social lives or other distractions outside of class.²⁹

²⁹Notice that this result is unlikely to be explained by non-city peers studying together in public transport. Column (3) in Table 2.9 has shown that spillovers are unrelated to having high ability peers closer to ones' residence.

To summarize, these results suggest meaningful social interaction between close peers takes place on campus, where in place of attending lectures, students study together with their close peers. It seems that social interaction with high ability peers increases grades by increasing the productivity of (collaborative) self study. While only suggestive, these findings are consistent with laboratory evidence examining peer effects mechanisms. Kimbrough et al. (2017) find that low ability participants were able to solve more logic puzzles when allowed to socially interact with high ability participants, and audio recording revealed that these social interactions generated spillovers via peer-to-peer teaching.

2.7 Voluntary Sorting and Potential Implications for Group Assignment Policies

Our results indicate that peer effects in the classroom work through social proximity. The extent to which students sort out of their close peer groups over time, and become socially proximate with other, self-chosen peers, is therefore crucial to the evolution of peer effects from assigned close peers. For example, interventions aiming to help low ability students by matching them with high ability students may not be sustainable if the social proximity between these students wanes over time. Had classroom-level effects driven peer effects, any changes in social proximity would be of no concern.

In this section we analyze voluntary sorting and discuss its potential implications for group assignment policies. First, we track the effect of close peers on grades during the first year. We find that peer effects from close peers diminish over time; they are strongest in the first block and vanish by the fourth block of the first year. Second, we use detailed tutorial attendance data and present some evidence that the social proximity between close peers diminishes in a similar fashion during the first year. Third, we use second-year tutorial registration and confirm that students largely sort out of their assigned close peer group. Concurrently, students sort into new peer groups based on prior bonds, ethnicity, and gender, but not on ability. While we cannot know with certainty the reason that academic spillovers from close peers vanishes during the first year, we believe the degree and type of voluntary sorting behavior provides an intuitive explanation.

2.7.1 Diminishing Peer Effects

To study how peer effects evolve over time we repeat the analysis of close peer GPA on grades per block of the first year. The results are presented in Table 2.10 where the column number refers to the block being analyzed. Columns (1) and (2) reveal that during the first two blocks the estimates for close peer GPA are comparatively large (p -value < 0.05). The magnitude slightly drops in block 3,

Table 2.10: Peer Effects per Block

	Grades (Standardized)				
	Block 1	Block 2	Block 3	Block 4	Block 5
	(1)	(2)	(3)	(4)	(5)
Close Peer GPA	0.0404** (0.0177)	0.0361** (0.0145)	0.0318** (0.0145)	0.0080 (0.0135)	0.0062 (0.0178)
Own GPA	0.4139*** (0.0159)	0.3451*** (0.0145)	0.3849*** (0.0176)	0.2537*** (0.0151)	0.3026*** (0.0160)
Observations	4271	4024	3650	3462	3329
Adjusted R^2	0.280	0.474	0.264	0.191	0.301

Notes:

1. All regressions include course-cohort fixed effects and controls; student number, gender, age, and distance to university.
2. Peer GPA refers to the leave out mean of high school GPA for the close peer group. Own GPA refers to own high school GPA. Both GPA measures are standardized.
3. Standard errors in parentheses, clustered on the tutorial level.
4. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

while still being significant at the 5%-level. In blocks 4 and 5 spillovers become statistically indistinguishable from zero. Appendix Table A.2.9 shows that distant peers are unimportant throughout all five blocks. We investigate several potential explanations for the diminishing peer effects in the Appendix, such as differences in course types across blocks, direct effects of dropout, and measurement error in peer ability due to dropout. The results imply these explanations are unimportant.

2.7.2 First-Year Tutorial Attendance

To study whether a reduction in social proximity could be a potential explanation, we analyze whether a student's tutorial attendance is associated with the attendance of their close and distant peers. Given a preference to attend tutorials with ones' friends, we interpret coordination of tutorial attendance among students as being indicative of social proximity. Let $Attendance_{isgct}$ be a binary variable taking the value one if student i attended tutorial session s , in group g , for course c , of cohort t . We run the following regression:

$$Attendance_{isgct} = \delta_0 + \delta_1 \overline{Att\ Close_{(-i)sgc}} + \delta_2 \overline{Att\ Distant_{sgc}} + G_{gct} + \delta_3 X_i + \epsilon_{isgct} \quad (2.3)$$

Where $\overline{Att\ Close_{(-i)sgc}}$ and $\overline{Att\ Distant_{sgc}}$ are the proportions of individual i 's close and distant peers who attend session s of course c . By running this regression per block, δ_1 and δ_2 detect any

changes in attendance coordination as the first year progresses. Recall that across the seven weeks there are fourteen and seven tutorial sessions for large and small courses respectively.

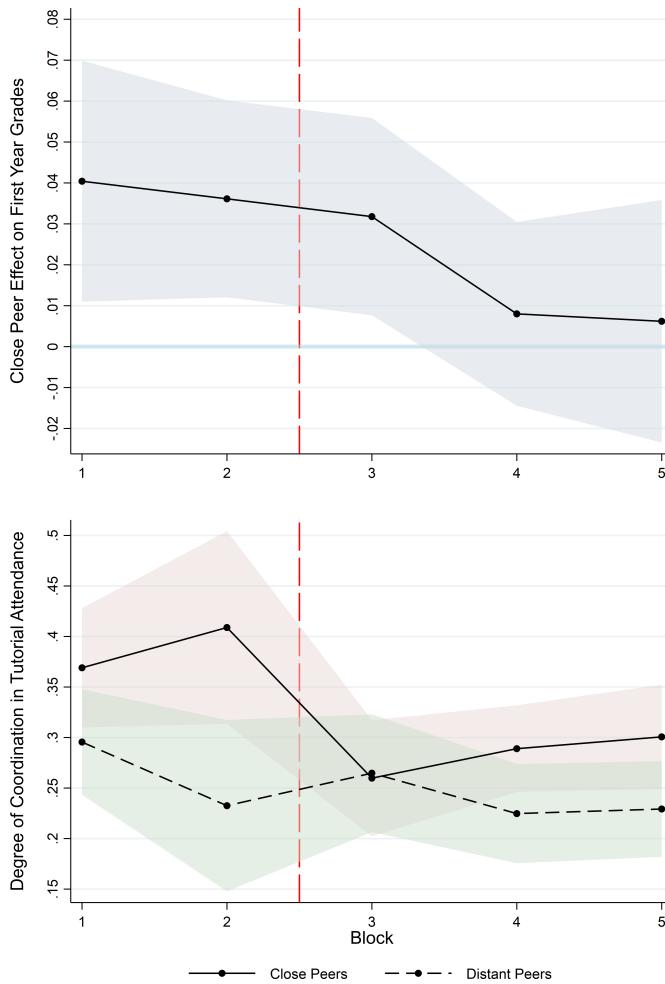
Equation (2.3) regresses attendance on its own group leave-out mean. If one is trying to detect causal peer effects this model would suffer from the reflection problem. Our goal, however, is to detect the degree of attendance coordination. Is a student more likely to go to tutorials with her close than distant peers, and does this change over time? The reflection problem poses no threat to answering this question. Another concern with such models is that group-level attendance shocks, such as bad weather, can result in coefficients that suggest peer coordination even if peers do not deliberately coordinate. Given that such shocks will take place at the tutorial level, both δ_1 and δ_2 are affected by these shocks. We will only compare their relative sizes and changes across blocks. Moreover, note that Equation (2.3) includes course-tutorial fixed effects (G_{gct}) to capture common shocks. The remaining control variables (X_i) are identical to the baseline regressions.

The results of the δ_1 and δ_2 coordination coefficients from these regressions per block are presented visually in Figure 2.3.³⁰ To highlight their potential relevance for the diminishing peer effects, the figure also contains a similar representation of the close group peer effect on grades during the first year.

We identify three main patterns. First, the degree of coordination in attendance between a student and her close peers is higher than between a student and her distant peers. This supports the notion that the close peer group meetings induced social proximity and further reinforces our results in Section 2.5. Second, the attendance coordination with close peers falls over time. Notably, the largest drop occurs after the second block, at which point there is a Christmas break. The timing of the break is indicated by the dashed vertical line in the figure. This drop is relatively large, significant (p -value=0.028), and stands in stark contrast to all other changes in coordination across blocks, which are relatively small and insignificant. Third, this drop in coordination after the second block is not visible between students and their distant peers. We take this as evidence that, while there was initially a difference in the degree of social proximity between a student and her close and distant peers, this diminished over time. The Christmas break might have resulted in a severing of the bonds between close peers.

The results above provide some evidence that the social proximity between assigned close peers diminishes as time progresses. Are students sorting out of their close peer group into other groups? Second-year tutorial registration, which by then is under the purview of the students, provides us with an opportunity to analyze exactly this.

³⁰The full regression results are presented in Appendix Table A.2.10. The table also presents p -values of a test for the equality of coefficients between adjacent blocks for close and distant peers separately and p -values of a test for the equality of the coefficients between close and distant peers within a block.

Figure 2.3: Diminishing Peer Effects in Grades (top graph) and Tutorial Attendance Coordination (bottom graph)

Notes:

1. Top graph shows the point estimates of close-peer effects on first year grades per block and the corresponding 90% confidence interval. The precise estimates can be found in Table 2.10.
2. Bottom graph shows the point estimates of first year tutorial attendance coordination for both close and distant peers and the corresponding 90% confidence interval. The precise estimates can be found in Appendix Table A.2.10.
3. The vertical dashed line indicates the timing of the first two-week break that occurs during the students' first year.

2.7.3 Second-Year Tutorial Choice

All students in second year have to register for the tutorials a few weeks before the start of the course.

If we assume that students prefer to be in a tutorial group with one's friends, then observing joint

tutorial registration allows us to analyze peer group formation. We look for evidence of students co-registering based on shared characteristics; a phenomenon referred to as homophily. In particular, we use the following six characteristics: close and distant peer groups, ethnicity, gender, former bonds based on a student's high school, and ability (measured by high school GPA). Recall the program only admits Dutch students, and so students with a different ethnicity than Dutch are either first- or second generation immigrants. In the Dutch context, the categories European (81%, including Dutch), Arabic (5%), and Asian (14%) are ex-ante most relevant.³¹

Similar to the strategy of Marmaros and Sacerdote (2006), we first form all possible pairs of students who are observed to take course c in the second year of cohort t . Given N_{ct} students this procedure generates $(N_{ct} \times N_{ct} - 1)/2$ pairings of students.³² Let $SecondYearTutorial(i, j)_{ct}$ be an indicator variable taking the value of one if both student i and j registered to the same second-year tutorial group and zero otherwise. We define a similar set of indicator variables for each of the characteristics listed above, taking the value of one if students i and j share that particular characteristic and zero otherwise. We then run the following regression per block:

$$SecondYearTutorial(i, j)_{ct} = \pi_0 + \pi_1 SharedCharacteristic(i, j) + C_{ct} + \epsilon(i, j)_{ct} \quad (2.4)$$

π_1 captures the change in the probability of two students sharing the same tutorial group in second year if they *e.g.* share the same gender. Equation (2.4) includes course-cohort fixed effects (C_{ct}), but, as the unit of observation is a student pair, it does not include other control variables. We cluster standard errors based on a variable that takes upon unique values for every combination of first-year tutorial groups of each student pair (i, j) .

Table 2.11 reports the results per block of the second year.³³ The last row reports the unconditional mean for the outcome variable, which is approximately 7 percent. The estimates for the shared characteristics can be compared to this mean, which reflects the probability of any two students registering together, independent of shared characteristics.³⁴

The results reveal four main patterns. First, the Close Peer Group coefficient indicates that only some bonds from the close peer groups have remained up until the second year. The coefficient in block 1 is 0.06, which indicates that sharing a close peer group increases the probability of co-

³¹We determine ethnicity using the surname-based classification algorithm NamePrism (Ye et al., 2017).

³²This is done by crossing the relevant list of student numbers with itself, removing all duplicate pairs (i, i) , and keeping only one instance of the same pairing $((i, j) \text{ and } (j, i))$.

³³Notice that the last year we observe is 2014-15, and we therefore do not observe the second-year tutorial registration for the 2014 cohort.

³⁴This unconditional mean coincides with our student-level data, where we observe roughly 200 students and 14 tutorials of 14 students each per course-cohort combination. As such, there is roughly a probability of 1/14 of registering in the same tutorial with any other student.

registration by 6 percent. As any two students have a 7 percent probability to co-register, a student registers together with 1 out of every 14 students. This becomes 2 out of 14 when the students originate from the same close peer group. Though this coefficient is significantly different from 0, it is far away from 1. This confirms the attendance results above and shows that by second year students have sorted out of their close peer groups to a large extent.

Second, a comparison of the Close Peer Group and Distant Peer Group coefficients reveals that, across blocks, the former is roughly 1.5 to 2 times larger than the later. The differences are statistically significant (p -values < 0.05) and provide further evidence that the close peer group meetings indeed manipulated social proximity. Notice, however, that the Distant Peer Group estimates are also positive and statistically significant. Thus, while distant peers remain less important than close peers, a student is more likely to form bonds with her distant peers than with someone in a different first-year tutorial group altogether. Based on the estimate from block 1 of 0.035, on average students co-register with approximately 1 out of every 10 students from their distant peer group.

Third, Table 2.11 reveals that students sort into more homogeneous peer groups. Especially minorities, such as Arabic and female students are significantly more likely to appear in the same tutorial groups. The marginal effect of *e.g.* the coefficient for Both Arabic in block 2 indicates that two ethnically Arabic students are roughly 2 percentage points more likely to register together than *e.g.* an ethnic Arab with any other student. The largest predictor of co-registration is having shared the same high school, which increases the probability of being in the same second-year tutorial by roughly 8 percentage points across blocks. This supports our assumption that students seek to co-register to tutorial groups with existing friends, and rejects an explanation where student clustering is observed only due to shared preferences on the exact time at which the second-year tutorials are held.³⁵ Comparing the coefficients of the shared characteristic with the baseline unconditional mean of the outcome variable reveals that these effects are large. Sharing an Arabic ethnicity, or having shared the same high school, increases the baseline probability to register for the same tutorial by 33 and 110 percent respectively.

³⁵To this end, it is useful to note that across courses there are approximately two to three tutorial groups (of in total fourteen) taught at identical times. Thus, students with similar preferences regarding tutorial times could still register in different tutorial groups. We do not, however, observe the time of the second-year tutorial groups.

Table 2.11: Voluntary Sorting in Second-Year Tutorials (All Blocks)

	Same Tutorial in Second Year? Yes (1) or No (0)				
	Block 1	Block 2	Block 3	Block 4	Block 5
	(1)	(2)	(3)	(4)	(5)
Close Peer Group	0.0591*** (0.0063)	0.0570*** (0.0052)	0.0487*** (0.0050)	0.0512*** (0.0054)	0.0486*** (0.0067)
Both Asian	0.0031 (0.0049)	0.0029 (0.0050)	0.0007 (0.0053)	-0.0044 (0.0058)	0.0127** (0.0055)
Both Arabic	0.0128 (0.0108)	0.0227* (0.0119)	0.0258* (0.0136)	0.0578*** (0.0137)	0.0273* (0.0160)
Both Europe	0.0043** (0.0019)	0.0044** (0.0021)	0.0020 (0.0019)	0.0033 (0.0021)	0.0056** (0.0023)
Both Female	0.0376*** (0.0044)	0.0311*** (0.0044)	0.0270*** (0.0039)	0.0216*** (0.0040)	0.0284*** (0.0052)
Both Male	0.0010 (0.0021)	0.0018 (0.0023)	-0.0007 (0.0022)	-0.0011 (0.0022)	-0.0004 (0.0028)
Same High School	0.0844*** (0.0064)	0.0867*** (0.0072)	0.0884*** (0.0065)	0.0912*** (0.0066)	0.0886*** (0.0077)
Distant Peer Group	0.0350*** (0.0063)	0.0396*** (0.0059)	0.0385*** (0.0056)	0.0342*** (0.0065)	0.0309*** (0.0067)
Both High GPA	-0.0009 (0.0019)	-0.0023 (0.0025)	0.0034 (0.0025)	0.0039* (0.0024)	0.0030 (0.0029)
Both Average GPA	0.0009 (0.0016)	0.0030* (0.0018)	0.0015 (0.0016)	-0.0017 (0.0019)	0.0040 (0.0025)
Both Low GPA	-0.0015 (0.0039)	0.0000 (0.0045)	-0.0012 (0.0045)	0.0003 (0.0043)	0.0038 (0.0061)
Unconditional Mean	0.0756*** (0.0006)	0.0755*** (0.0006)	0.0764*** (0.0006)	0.0766*** (0.0006)	0.0746*** (0.0008)
Observations	229428	214691	218183	188896	106630
Pseudo R^2	0.010	0.010	0.008	0.009	0.008

Notes:

1. All regressions include course-cohort fixed effects, other controls are excluded.
2. Block 5 has half the number of observations as one course does not have tutorials.
3. The unit of analysis is a student-pair. The outcome variable is one if both students in the pair registered for the same tutorial and zero otherwise. The explanatory variables are one if both students in the pair share the given characteristic.
4. Models are estimated with Probit. Marginal effects are reported.
5. Standard errors in parentheses, clustered based on a variable that takes upon unique values for every combination of first-year tutorial groups of each student pair.
6. Unconditional mean refers to the mean of the outcome variable in that particular block. Standard error reported in parentheses.
7. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Fourth, the estimates for high school GPA reveal little to no co-registration of students with similar ability. The small degree of ability clustering is in line with the findings of Marmaros and Sacerdote (2006).³⁶

2.7.4 Long-Term First-Year Bonds

Which characteristics determine the long-term bonds that persist from the first-year close peer group? To investigate this, Table 2.12 shows results for a specification which includes interaction terms for each shared characteristic and the indicator for shared first-year close peer group. It appears that long-term first-year bonds are especially prevalent among close peers of the same gender; across the second year a pair of female (male) close peers are roughly 5 (3.5) percentage points more likely to form a long-term bond than a mixed gender pair. The estimates also reveal that long-term first-year bonds do not seem to be based on ability, which is consistent with our evidence of little to no clustering by ability presented above.

Note that in this specification the Close Peer Group coefficient provides a rare insight into the degree to which friendship groups can be institutionally manipulated against the formation of homogeneous subgroups based on gender, ethnicity, and prior bonds. More specifically, the coefficient measures the probability of co-registration among first-year close peers who share no observable characteristics. While relatively large and statistically significant in the first two blocks, as the second year progresses the coefficient diminishes in size and significance. This suggests that the manipulated social proximity further decreases in the long-term among students who differ on a wide range of characteristics.

Policy makers and university administrators in both the U.S. and Europe have recently emphasized the importance of diversity in higher education. Table 2.12 implies that the group intervention mainly formed long-term bonds among students with similar characteristics and did little to promote long-lasting diversity on campus. We cannot rule out, however, that a more sustained or focused intervention would be more successful.

³⁶We have performed a similar analysis for third-year course choice. The results of the regressions, per characteristic and for the ten most popular courses, are presented visually in Appendix Figure A.2.5. The conclusions are threefold. First, we find no evidence that close or distant peers choose the same courses in third year. Second, we find strong evidence of third-year course clustering based on shared high school, gender, and ethnicity. Third, in contrast to our results with the second-year tutorial registration, we find strong clustering based on ability. Taken together, this suggests that course choice also captures that students with some characteristics have preferences for certain topics, rather than reflecting bonding. For instance, high ability students sort into difficult courses.

Table 2.12: Characteristics of Long-Term First-Year Bonds (All Blocks)

	Same Tutorial in Second Year? Yes (1) or No (0)				
	Block 1	Block 2	Block 3	Block 4	Block 5
	(1)	(2)	(3)	(4)	(5)
Close Peer Group	0.0409*** (0.0107)	0.0332*** (0.0093)	0.0090 (0.0095)	0.0190** (0.0091)	0.0129 (0.0115)
Both Asian × Close Peer Group	-0.0299 (0.0242)	-0.0711* (0.0387)	0.0122 (0.0236)	-0.0069 (0.0262)	-0.0808** (0.0399)
Both Arabic × Close Peer Group	0.0034 (0.0407)	-0.0264 (0.0530)	0.0453 (0.0563)	-0.0531 (0.0693)	-0.0013 (0.0688)
Both European × Close Peer Group	0.0175* (0.0097)	0.0039 (0.0093)	0.0116 (0.0079)	0.0117 (0.0095)	0.0028 (0.0106)
Both Female × Close Peer Group	0.0480*** (0.0149)	0.0615*** (0.0172)	0.0562*** (0.0150)	0.0534*** (0.0170)	0.0536*** (0.0180)
Both Male × Close Peer Group	0.0122 (0.0109)	0.0283*** (0.0109)	0.0391*** (0.0104)	0.0351*** (0.0099)	0.0416*** (0.0128)
Same High School × Close Peer Group	0.0221 (0.0254)	0.0377 (0.0289)	0.0143 (0.0324)	0.0239 (0.0316)	0.0537 (0.0343)
Both High GPA × Close Peer Group	-0.0175* (0.0089)	-0.0005 (0.0107)	0.0006 (0.0106)	-0.0155 (0.0103)	-0.0209* (0.0116)
Both Average GPA × Close Peer Group	-0.0078 (0.0086)	-0.0023 (0.0084)	0.0061 (0.0089)	0.0025 (0.0090)	0.0254** (0.0102)
Both Low GPA × Close Peer Group	-0.0538** (0.0273)	-0.0308 (0.0195)	0.0089 (0.0252)	-0.0209 (0.0237)	-0.0046 (0.0237)
Unconditional Mean	0.0756*** (0.0006)	0.0755*** (0.0006)	0.0764*** (0.0006)	0.0766*** (0.0006)	0.0746*** (0.0008)
Observations	229428	214691	218183	188896	106630
Pseudo R^2	0.010	0.010	0.008	0.009	0.008

Notes:

1. Table shows results of a regression including all observable shared characteristics as predictors of shared second year tutorial, and with interactions between the shared characteristics and an indicator for shared first year close-peer group. Only results of shared first year close-peer group and the interaction terms shown.
2. All regressions include course-cohort fixed effects, other controls are excluded.
3. Block 5 has half the number of observations as one course does not have tutorials.
4. The unit of analysis is a student-pair. The outcome variable is one if both students in the pair registered for the same tutorial and zero otherwise. The explanatory variables are one if both students in the pair share the given characteristic.
5. Models are estimated with Probit. Marginal effects are reported.
6. Standard errors in parentheses, clustered based on a variable that takes upon unique values for every combination of first-year tutorial groups of each student pair.
7. Unconditional mean refers to the mean of the outcome variable in that particular block. Standard error reported in parentheses.
8. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

2.7.5 Implications of Voluntary Sorting for Peer Effects

We have provided some evidence that the social proximity between close peers decreases during the first year. In turn, we have shown that by second year students' chosen peer groups hardly resemble their first-year assigned groups; they prefer to become socially proximate with others based on shared characteristics, such as gender and ethnicity. While we cannot know with certainty the reason that academic spillovers from close peers vanished during the first year, the voluntary sorting behavior provides an intuitive explanation. The social proximity between assigned close peers waned over time, which might have led to the corresponding decline in spillovers.

Another result is that students do not appear to choose their peers based on whether they are beneficial to their performance at university. We find no evidence of sizable sorting by ability and close peers that stay together do not appear to base this choice on high school GPA. For instance, we do not find evidence that high (low) ability students sort into (out of) study partnerships with other high (low) ability students, though according to our peer effect estimates this would be academically beneficial. This stands in stark contrast to sorting based on other characteristics. Students are seemingly willing to trade off potentially higher grades in order to satisfy other preferences when choosing peers. Consistent with this, column (6) of Appendix Table A.2.13 shows that second-year chosen tutorial peers do not generate spillovers on student performance in second year.³⁷

This might have further implications for group assignment policies. Policy makers may hope that targeted students would form new friendships with academically beneficial peers, thereby enjoying persistent peer effects despite sorting away from their assigned peers. Apparently this is not the case.

2.8 Conclusion

The promise of the peer effects literature is that simply reorganizing students among classes could increase overall student performance. Despite an abundance of papers aiming to properly identify spillovers, the literature has not yet delivered on this promise. A primary reason for this is our inability to understand the channels at work behind the various reduced-form estimates.

Our first set of results address this shortcoming. We focus on first-year student performance across six cohorts of economics undergraduate students at a large public university in the Netherlands. Students are randomly assigned to a tutorial group and one of two subgroups within their tutorial group. We take advantage of a university policy that stimulates social proximity within, and not between, these subgroups via a series of informal meetings at the start of the first year. We find the

³⁷Column (6) of Appendix Table A.2.13 shows that high school GPA of second-year chosen tutorial peers, while being instrumented with first-year assigned close peer GPA, has an insignificant effect on second-year grades.

existence of spillovers on student performance that originate from students' socially proximate peers only. This implies that social proximity between peers, and the corresponding meaningful social interactions, are the driving force behind peer effects. Supplementary data suggests that these social interactions involve collaborative studying outside of class that occurs at university. Our non-linear estimates imply that alternative group assignment policies may result in aggregate, but not Pareto, improvements in performance.

The second part of this paper investigates the implications of voluntary sorting for group assignment policies. Given that peer effects arise due to social proximity, who students choose to become socially proximity with and how this evolves over time is crucially important. We first document that peer effects from assigned close peers diminish over time, and are completely absent by the end of first year. Using administrative data on daily tutorial attendance in first year and tutorial choice in second year we find that students increasingly sort out of their assigned close peer group into more homogeneous groups. This voluntary sorting behavior foreshadows, and we argue provides an intuitive explanation for, the short-lived spillovers on student performance.

Similar to our analysis in Section 2.5, some researchers have used their reduced-form estimates on student performance to predict the effects of alternative group assignment policies (Booij et al., 2017) or to estimate the effects of optimal group assignment policies (Carrell et al., 2013). Such a practice usually assumes that there are no costs accompanying these effects. As our results imply that spillovers work solely through improving the productivity of (collaborative) self study, rather than through increasing teacher effort or decreasing leisure time, this assumption may be justifiable.

2.A Appendix

Potential Explanations for the Decline in Peer Effects

Why do we observe that spillovers gradually diminish during, and become absent at the end of, the first year? Two overarching factors that vary during the first year, and could potentially drive the diminishing peer effect, are changes in the type of courses and dropout. Below we explore the evidence for each of these competing explanations.

Changes in the content, structure, and other characteristics of the courses during the first year could potentially drive the diminishing peer effect estimates. To explore this possibility, we look at heterogeneity in peer effects by course type. Following the classification of the university we group the ten first year courses into three categories: economics, business economics, and econometrics courses (see Appendix Table A.2.1). Appendix Table A.2.11 replicates our baseline specification

while including an interaction between close peer ability and an indicator for course type, where economics courses are the baseline. The estimates reveal spillovers are statistically indistinguishable between the different types of courses. Feld and Zöllitz (2017) reach similar conclusions, also at a public university in the Netherlands. In Lavy et al. (2012b) identification of peer effects is obtained via individual fixed effects together with the assumption that spillovers are the same across English, mathematics, and science courses. Note that Appendix Table A.2.11 does reveal the estimate of own high school GPA differs for the different types of courses. It appears that the returns to peer ability are disconnected from the returns to students' own ability.

It may be that the nature of the tutorial sessions changes from course to course, and that this has consequences for the existence of peer effects. Appendix Table A.2.1 reveals that the nature of the tutorial sessions is unrelated to the presence of peer effects. For example, tutorial descriptions are identical in Accounting and Microeconomics situated in block 1 and 2 and Marketing and Organisation & Strategy situated in block 4 and 5, while spillovers are only found in the former courses.

If courses get progressively easier during the first year, then the potential for peers to improve students grades could also diminish. To investigate this possibility Appendix Figure A.2.6 displays the coefficients on own and peer high school GPA per block, separately for small and big courses. Apart from a drop for the estimate on own high school GPA in block 4 for the big course (Marketing), the estimate for own GPA does not show a diminishing pattern across blocks. For instance, from block 2 to 3 this coefficient slightly increases, whereas the estimate for peer high school GPA decreases. Note that this evidence coincides with the results in Appendix Table A.2.11, where the returns to peer GPA were detached from the returns to own GPA. Based on the three pieces of evidence presented above, we conclude changes in course type is an unlikely explanation for the diminishing peer effects.

A second potential explanation of diminishing peer effects is dropout. Indeed, Appendix Figure A.2.4 shows that dropout gradually increases as the blocks progress. Dropout could potentially reduce our peer effects estimates for at least two reasons; dropouts might be more responsive to peer high school GPA, and dropouts change the composition of the actual peer group for the remaining students. To investigate whether dropout interacts with the decline in peer spillovers, we repeat our robustness analysis of Section 2.5.4 and interact peer GPA with the number of course dropouts per close peer group for blocks 1 to 3 and block 4 to 5 separately. The results are presented in Appendix Table A.2.12. The estimates for close peer GPA imply that the decline in spillovers is present even in groups that did not experience any course dropout.

A consequence of dropout is that high school GPA of the initial close peer group becomes a worse measure of the actual ability of close peers. We overcome this potential problem by using an instrumental variable approach. For each student, per course, we calculate the average close peer

GPA of only those close peers who are also observed to write the final exam for the course. We then repeat the regression of close peer GPA on first-year grades per block, while instrumenting the actual close peer GPA with the initially assigned close peer GPA. The first and second stage results of these regressions are presented in Appendix Table A.2.13. Panel A shows that assigned peer GPA is a strong instrument for actual peer GPA throughout the first year. Panel B shows that the decline in spillovers remains when using the variation in actual close peer GPA that originates from the assigned close peer GPA. From these results, we conclude that dropout is unlikely to be responsible for the diminishing peer effects during the first year.

Finally, Appendix Table A.2.14 repeats the analysis on lecture attendance and total study time for blocks 1 to 3 and blocks 4 to 5 separately. The negative estimate for close peer GPA on lecture attendance is only present in blocks 1 to 3. This suggests that the channel put forward in Section 2.6 - collaborative self study - is only present in the period for which we find significant peer effects.

Table A.2.1: Overview of the First Year

Course	Block	ECTS	Course Type	Tutorial Description	Exam Qs.
Accounting	1	8	Business Eco.	TA explains the assignments that students are supposed to make in preparation of the tutorial.	MC
Math I	1	4	Econometrics	TA briefly explains material of the week, after which the students make assignments in (unobserved) groups that count for a small percentage to the final grade.	Open
Microeconomics	2	8	Economics	TA explains the assignments that students are supposed to make in preparation of the tutorial.	MC (2009-11) and Open (2012-14)
ICT	2	4	Econometrics	TA briefly explains material of the week through predefined assignments, after which the students make assignments that count for a small percentage to the final grade.	Both MC and Open Qs.
Macroeconomics	3	8	Economics	TA explains the assignments that students are supposed to make in preparation of the tutorial.	MC
Math II	3	4	Econometrics	TA briefly explains material of the week, after which the students make assignments in (unobserved) groups that count for a small percentage to the final grade.	Open
Marketing	4	8	Business Eco.	TA explains the assignments that students are supposed to make in preparation of the tutorial.	Both MC and Open Qs.
Applied Statistics	4	4	Econometrics	TA briefly explains material of the week, after which the students make assignments in (unobserved) groups that count for a small percentage to the final grade.	MC
Organisation and Strategy	5	8	Economics	TA explains the assignments that students are supposed to make in preparation of the tutorial.	Both MC and Open Qs.
Financial Accounting	5	4	Business Eco.	TA explains the assignments that students are supposed to make in preparation of the tutorial.	MC

Notes: This is an overview of the year 2013-14. The courses are identical in the other years, with the only exception being that in 2014-15 the course "Skills" replaced "Financial Accounting". Course type refers to the categorization made by the university. For a given course the use of tutorials only varies minor between the years. For all years and courses, the percentage of weekly tutorial assignments that count for the final grade is smaller than 20 percent. Mostly they count for 10 percent or they allow a student to make a bonus question on the final exam.

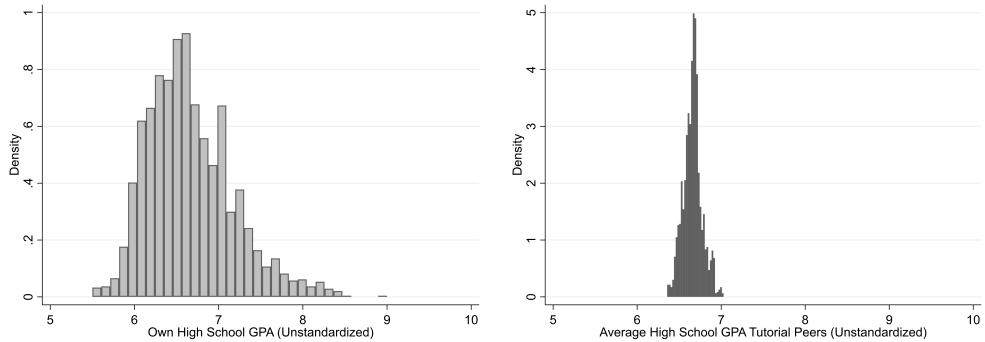
Table A.2.2: Overview of Categories and Questions in Course Evaluations

Question	Measurement scale	Category	Std.?
Objectives of course are clear	1-5	General	
Course is relevant for my studies	1-5	General	Yes
Course is interesting	1-5	General	
Course is well organized	1-5	Structure	Yes
Course material is understandable	1-5	Structure	
Can be completed within allocated study points	1-5	Fairness	
Time needed to complete exam is enough	1-5	Fairness	Yes
Exam reflects course content	1-5	Fairness	
Exam questions are clearly defined	1-5	Fairness	
Total study time (lectures+tutorials+self study)	1-10	Total study time	No
Have you attended lectures?	0-1	Lecture attendance	No
Lectures are useful	1-5	Lectures useful	Yes
Tutorials are useful	1-5	Tutorials useful	Yes
Lecturer is competent	1-5	Quality lecturer(s)	
Lecturer makes you enthusiastic	1-5	Quality lecturer(s)	Yes
Lecturer can be easily contacted	1-5	Quality lecturer(s)	
Lecturer provides sufficient assistance	1-5	Quality lecturer(s)	
TA gives good tutorials	1-5	Quality TA	
TA can be easily contacted	1-5	Quality TA	Yes
TA provides sufficient assistance	1-5	Quality TA	

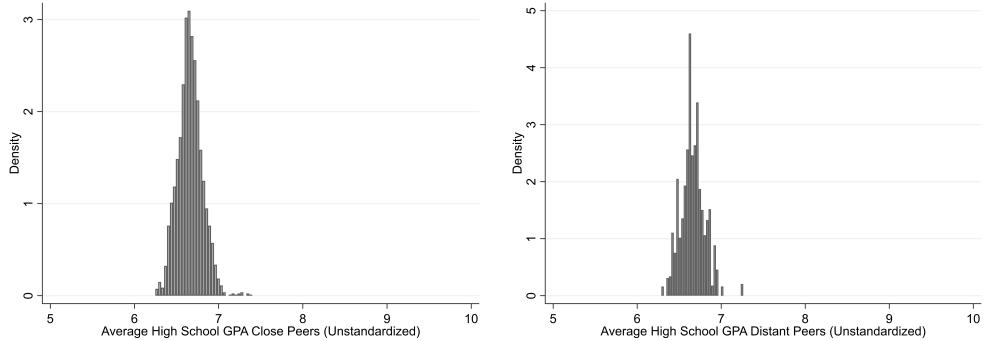
Notes: Questions are measured on a Likert scale, where 1 equals strongly disagree and 5 equals strongly agree, with the two exceptions being total study time (1 being 0 hours, 2 being [1 – 5] hours, 3 being [6 – 10] hours and 10 being ≥ 40 hours) and lecture attendance (1 being yes and 0 being no). We take the mean for questions within a category, ignoring potential missing values within a category. Std. refers to whether the (mean of a) category was standardized before the analysis.

Figure A.2.1: Histograms of High School GPA (Unstandardized)

(a) Own High School GPA (Left) and Tutorial Peer High School GPA (Right)



(b) Close Peer High School GPA (Left) and Distant Peer High School GPA (Right)



Notes:

1. Figure shows histograms of student's own high school GPA, the leave-out mean for the tutorial- and close peer group, and the mean for the distant peer group.
2. In contrast to the leave-out mean for the close peer group, the mean for the distant peer group takes upon identical values for everybody in the same subgroup. This explains the somewhat more discrete nature of this figure. A histogram of the leave-in mean for the close peer group is similar to the mean for the distant peer group, where it would only change the peer-effects estimate on close peer high school GPA by a factor of $N_g/(N_g - 1)$, where N_g is the size of close peer group g (Angrist, 2014).

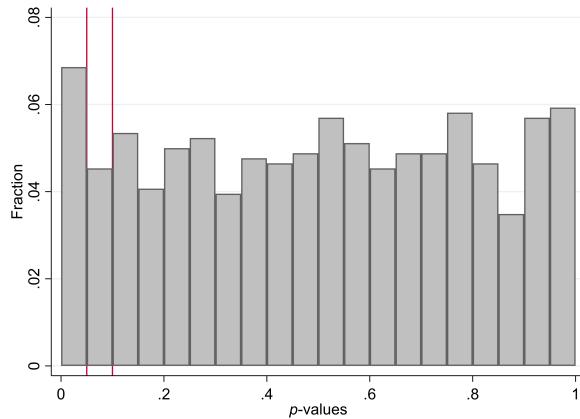
Table A.2.3: Balancing Tests for Background Characteristics

	Student Number	Gender	Age	Distance to University	High School GPA
	(1)	(2)	(3)	(4)	(5)
Close Peer Group 1	855.7164 (3904.3123)	0.0117 (0.1363)	-0.7190 (0.4737)	-5.8792 (9.0781)	0.1693 (0.1668)
Close Peer Group 2	1578.1608 (3904.3123)	0.0117 (0.1363)	-0.1796 (0.4737)	-2.5438 (9.0781)	0.0331 (0.1668)
Close Peer Group 3	-2206.2697 (4027.6709)	-0.0855 (0.1406)	-0.3141 (0.4887)	-7.1964 (9.3649)	0.0785 (0.1721)
Close Peer Group 4	2209.5719 (4099.9048)	-0.0772 (0.1431)	-0.5452 (0.4975)	-3.3456 (9.5329)	0.0247 (0.1752)
Close Peer Group 5	683.0497 (3904.3123)	0.0117 (0.1363)	0.4360 (0.4737)	13.8007 (9.0781)	-0.0804 (0.1668)
Close Peer Group 6	257.0553 (3802.7454)	-0.1105 (0.1327)	1.0621** (0.4614)	-1.9330 (8.8419)	0.2936* (0.1625)
Close Peer Group 7	-598.4830 (3962.8418)	0.0248 (0.1383)	-0.3997 (0.4808)	-1.4188 (9.2142)	0.2320 (0.1693)
Close Peer Group 8	2902.1579 (3851.1900)	-0.0000 (0.1344)	-0.5621 (0.4673)	1.2335 (8.9546)	0.3371** (0.1646)
Close Peer Group 9	1121.8830 (3904.3123)	0.0117 (0.1363)	-0.6279 (0.4737)	0.0155 (9.0781)	0.2005 (0.1668)
	⋮	⋮	⋮	⋮	⋮
Observations	2296	2296	2296	2296	2296
Adjusted R^2	0.832	-0.013	-0.005	-0.003	0.001
F-test	0.75	0.85	0.93	0.87	0.94
p-value	0.993	0.921	0.728	0.878	0.687

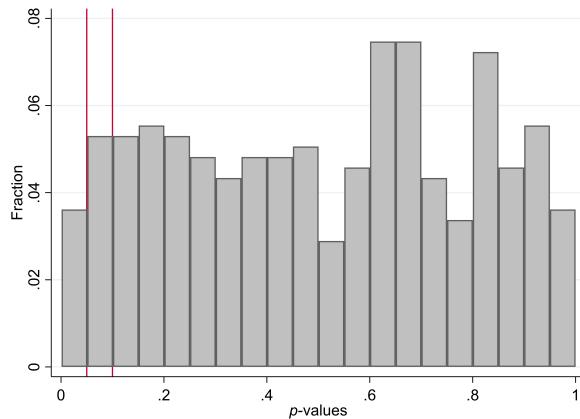
Notes:

1. Regressions include cohort fixed effects and dummies for the close peer group. No further controls are included.
2. The dependent variable is shown at the top of each column.
3. The F-test, and corresponding p-value, refer to a test for the joint insignificance of the close peer group dummies. It tests whether a large model with both cohort dummies and close peer group dummies can explain the background characteristics better than a small model with only cohort dummies.
4. Standard errors in parentheses.
5. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure A.2.2: Histograms of p -values of Balancing Tests



(a) p -values of Close Peer Group Dummies



(b) p -values of Tutorial Group Dummies

Notes:

1. Figures display histograms of the p -values of group dummies that originate from regressions where student characteristics are explained by group dummies.
2. The regressions were estimated for all student characteristics (student number, gender, age, distance to university, and high school GPA) separately for each cohort. The histograms include the p -values of all years and student characteristics combined.

Table A.2.4: Balancing Tests for TA Characteristics

	Is TA a PhD?	Is TA Female?
	Yes (1) or No (0)	
	(1)	(2)
Tutorial Peer GPA	-0.0041 (0.0120)	-0.0148 (0.0199)
Own GPA	0.0005 (0.0024)	-0.0042 (0.0036)
Observations	17535	6921
Adjusted R^2	0.254	0.345

Notes:

1. All regressions include course-cohort fixed effects and controls; student number, gender, age, and distance to university.
2. Peer GPA refers to the leave-out mean of high school GPA for the tutorial peers. Own GPA refers to own high school GPA. Both GPA measures are standardized.
3. Despite the two binary outcomes, we estimate the models with OLS. In some cases the course-cohort dummies predict the outcome variable perfectly, which means the Probit estimates for these dummies must be (minus) infinity.
4. Standard errors in parentheses, clustered on the tutorial level.
5. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.2.5: Randomization Inference and Exact *p*-values

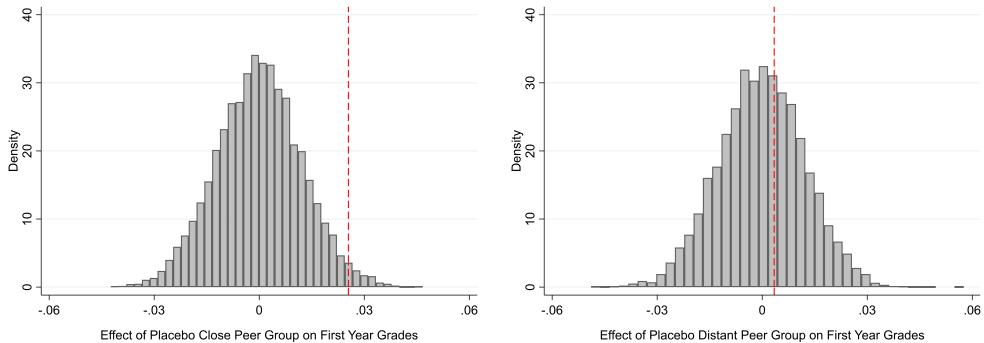
	Simulated Mean (SD)	Estimated Value	Exact <i>p</i> -value
Panel A: Separate Models			
Tutorial Peer GPA	-0.0002 (0.0128)	0.0191	0.1387
Close Peer GPA	-0.0001 (0.0123)	0.0255	0.0392
Distant Peer GPA	-0.0002 (0.0123)	0.0034	0.7852
Panel B: Simultaneous Model			
Close Peer GPA	-0.0001 (0.0123)	0.0254	0.0408
Distant Peer GPA	-0.0002 (0.0123)	0.0008	0.9489
Panel C: Simultaneous Model with Interaction			
Close Peer GPA	-0.00003 (0.0124)	0.0256	0.0412
Distant Peer GPA	-0.0002 (0.0124)	0.0010	0.9341
Close \times Distant Peer GPA	-0.00003 (0.0126)	0.0150	0.2326

Notes:

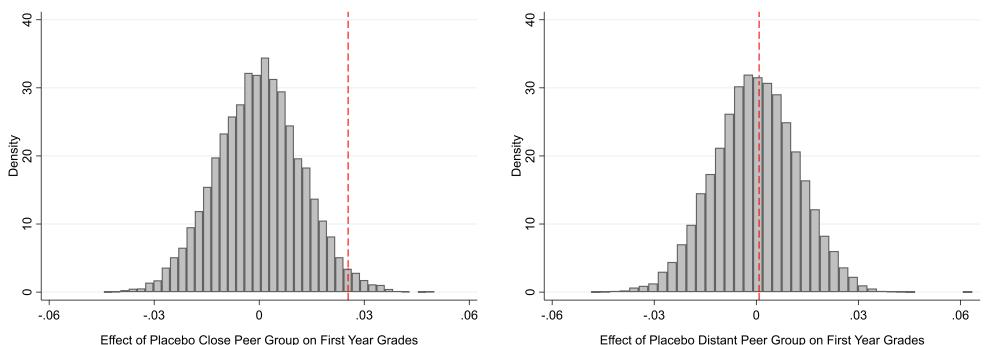
1. Table summarizes the results of a randomization inference analysis of our baseline results presented in Table 2.3, in which we re-draw 10,000 alternative close- and tutorial peer group assignments. The table presents the mean and standard deviation of the coefficients under the 10,000 re-draws, the coefficient values under the actual assignment, and the exact *p*-value based on the randomization inference.
2. Panel A displays the results for models in which the peer GPA measures have been included separately. Panel B displays the results for a model in which the close and distant peer GPA measures have been included simultaneously. Panel C shows the results for a model in which the close and distant peer GPA measures, as well as their interaction, have been included simultaneously. Panels A, B and C correspond to columns (1) to (3), (4) and (5) of Panel A of Table 2.3, respectively.
3. The exact *p*-value shows the proportion of coefficients under the 10,000 re-draws for which a value at least as extreme as the actual value is observed.

Figure A.2.3: Histograms of Estimates Under 10,000 Group Assignment Re-draws

(a) Close Peer GPA (Left) and Distant Peer GPA (Right) on First Year Grades, Separate Models



(b) Close Peer GPA (Left) and Distant Peer GPA (Right) on First Year Grades, Simultaneous Model



Notes:

1. Figures show histograms of the estimates of close and distant peer GPA on first-year grades under 10,000 alternative group assignments.
2. Top figures (a) show results for models in which the peer GPA measures have been included separately. Bottom figures (b) show results for a model in which peer GPA measures have been included simultaneously.
3. Red dashed lines indicate the observed estimate under the actual assignment.

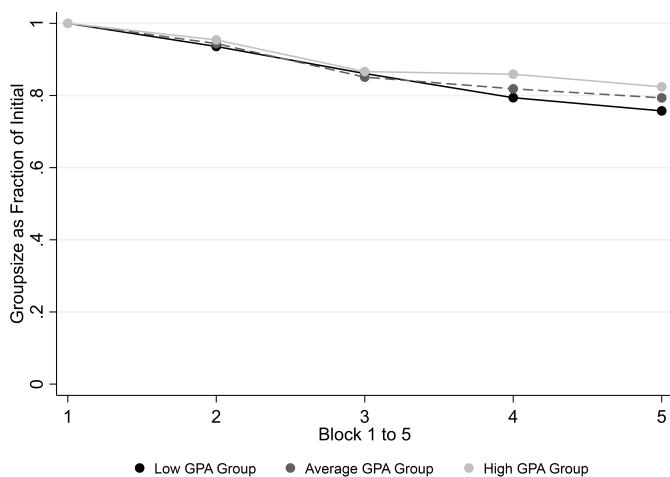
Table A.2.6: Peer Effects on First-Year Tutorial Attendance

Attendance (% Tutorials Attended; Standardized)					
	(1)	(2)	(3)	(4)	(5)
Tutorial Peer GPA	-0.0122 (0.0176)				
Close Peer GPA		-0.0030 (0.0131)		-0.0017 (0.0123)	-0.0016 (0.0121)
Distant Peer GPA			-0.0128 (0.0149)	-0.0126 (0.0143)	-0.0124 (0.0143)
Close \times Distant Peer GPA					-0.0146 (0.0189)
Own GPA	0.0378*** (0.0090)	0.0377*** (0.0092)	0.0381*** (0.0090)	0.0381*** (0.0091)	0.0380*** (0.0092)
Observations	18445	18445	18445	18445	18445
Adjusted R^2	0.122	0.122	0.122	0.122	0.123

Notes:

1. All regressions include course-cohort fixed effects and controls; student number, gender, age, and distance to university.
2. Peer GPA refers to the leave-out mean of high school GPA for the tutorial- and close peers, and to the mean for distant peers. Own GPA refers to own high school GPA. All GPA measures are standardized. The outcome variable is the standardized percentage of tutorials attended per course.
3. Standard errors in parentheses, clustered on the tutorial level.
4. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure A.2.4: Course Dropout per Block



Notes:

1. This figure plots the number of students writing the final exam as a fraction of the initial students per block, separately for high, average, and low GPA close peer groups.
2. Low and high ability groups are in the bottom and top quartiles of close peer high school GPA. The average group refers to the middle 50 percent.

Table A.2.7: Balancing Tests for Non-linear Peer Ability

	Close Peer Group			Distant Peer Group		
	Share Low	Share Avg	Share High	Share Low	Share Avg	Share High
	(1)	(2)	(3)	(4)	(5)	(6)
Student Number	0.0017 (0.0050)	-0.0007 (0.0071)	-0.0010 (0.0064)	0.0031 (0.0047)	-0.0076 (0.0060)	0.0045 (0.0052)
Female	0.0027 (0.0059)	-0.0056 (0.0074)	0.0029 (0.0059)	0.0105** (0.0048)	-0.0085 (0.0067)	-0.0021 (0.0063)
Age	-0.0008 (0.0027)	-0.0001 (0.0034)	0.0008 (0.0029)	0.0007 (0.0026)	-0.0025 (0.0030)	0.0018 (0.0022)
Distance to University	0.0025 (0.0021)	-0.0027 (0.0027)	0.0002 (0.0023)	0.0029 (0.0019)	-0.0019 (0.0024)	-0.0010 (0.0021)
Own GPA	0.0058** (0.0024)	-0.0050 (0.0032)	-0.0008 (0.0036)	-0.0032 (0.0026)	0.0037 (0.0025)	-0.0006 (0.0029)
Observations	2296	2296	2296	2296	2296	2296
Adjusted R^2	0.076	0.071	0.070	0.073	0.086	0.093
F-test	1.61	0.87	0.10	1.44	0.92	0.18
p -value	0.154	0.502	0.992	0.207	0.468	0.972

Notes:

1. All regressions also include cohort fixed effects.
2. The outcome variables are the (leave-out) proportion of low, middle, and high ability students separately for close and distant peer groups. Low and high ability students are defined as students in the bottom and top quartiles of high school GPA across the six cohorts, the remaining 50 percent is referred to as average ability. The dependent variables are unstandardized, where the independent variables are standardized except for the female dummy.
3. The F-test, and corresponding p -value, refer to a test for the joint significance of all the independent variables shown in the table.
4. Standard errors in parentheses, clustered on the tutorial level.
5. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.2.8: Peer Effects on Perceptions of Course using Course Evaluations

	Completed the Evaluation?	General Structure	Fairness	
	(1)	(2)	(3)	(4)
Close Peer GPA	0.0021 (0.0084)	-0.0198 (0.0183)	-0.0327 (0.0254)	-0.0244 (0.0179)
Own GPA	0.0494*** (0.0059)	0.0620*** (0.0151)	0.0657*** (0.0171)	0.1087*** (0.0165)
Observations	18736	3352	3352	3352
R^2	0.058	0.156	0.147	0.272
Binary Outcome	Yes	No	No	No

Notes:

1. All regressions include course-cohort fixed effects and controls; student number, gender, age, and distance to university.
2. Peer GPA refers to the leave-out mean of high school GPA for the close peer group. Own GPA refers to own high school GPA. Both GPA measures are standardized.
3. The dependent variable in column (1) equals one if the student completed the course evaluation and zero otherwise. The dependent variables in column (2) until (4) are the means of the answers to questions that embody the course characteristic showed in the top of the column. The dependent variables in column (2) until (4) are standardized.
4. Column (1) is estimated with Probit, the other columns with OLS. Marginal effects are reported. The R^2 refers to the Pseudo and Adjusted R^2 respectively.
5. Standard errors in parentheses, clustered on the tutorial level.
6. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.2.9: Peer Effects for Distant Peers per Block

	Grades (Standardized)				
	Block 1 (1)	Block 2 (2)	Block 3 (3)	Block 4 (4)	Block 5 (5)
Distant Peer GPA	-0.0159 (0.0199)	-0.0014 (0.0179)	0.0167 (0.0168)	0.0064 (0.0147)	0.0180 (0.0184)
Own GPA	0.4133*** (0.0159)	0.3442*** (0.0143)	0.3836*** (0.0176)	0.2534*** (0.0151)	0.3021*** (0.0158)
Observations	4271	4024	3650	3462	3329
Adjusted R^2	0.279	0.473	0.263	0.191	0.301

Notes:

1. All regressions include course-cohort fixed effects and controls; student number, gender, age, and distance to university.
2. Peer GPA refers to the mean of high school GPA for the distant peer group. Own GPA refers to own high school GPA. Both GPA measures are standardized.
3. Standard errors in parentheses, clustered on the tutorial level.
4. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

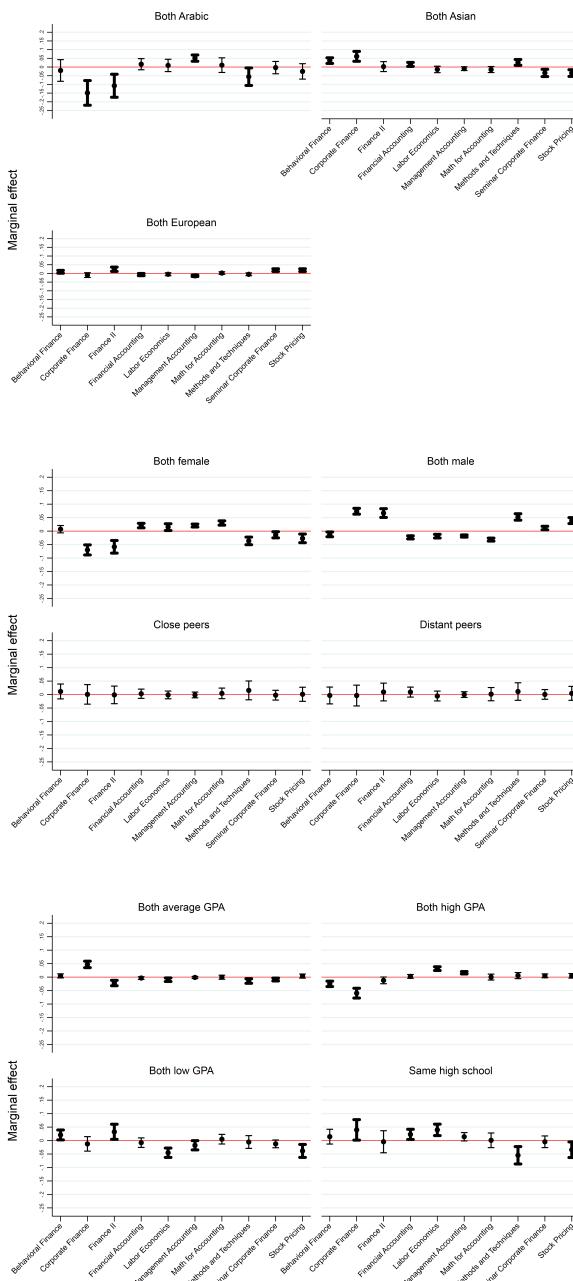
Table A.2.10: Coordination of First-Year Tutorial Attendance

	Attended Tutorial? Yes (1) or No (0)				
	Block 1	Block 2	Block 3	Block 4	Block 5
	(1)	(2)	(3)	(4)	(5)
Mean Attendance Close Peers	0.3690*** (0.0355)	0.4089*** (0.0573)	0.2598*** (0.0345)	0.2890*** (0.0258)	0.3007*** (0.0310)
Mean Attendance Distant Peers	0.2956*** (0.0314)	0.2326*** (0.0510)	0.2647*** (0.0350)	0.2247*** (0.0295)	0.2293*** (0.0285)
Observations	40321	40045	32920	33882	19654
Adjusted R^2	0.079	0.086	0.136	0.059	0.060
p -value Block $t = (t - 1)$ Close		0.556	0.028	0.486	0.776
p -value Block $t = (t - 1)$ Distant		0.287	0.578	0.335	0.908
p -value Close = Distant	0.173	0.078	0.889	0.107	0.133

Notes:

1. All regressions include course-tutorial fixed effects and controls; student number, gender, age, and distance to university.
2. Mean attendance refers to leave-out mean attendance per tutorial session for close peers and to the mean attendance per tutorial session for distant peers. The unit of analysis is on the student-tutorial-course level.
3. Block 5 contains somewhat less observations because the big course has 6 tutorials (one every week) instead of 13 to 14 tutorials (two every week).
4. The p -value “Block $t = (t - 1)$ ” refers to a test for the equality of coefficients between adjacent blocks for close and distant peers separately. The p -value “Close = Distant” tests the equality of the coefficients between close and distant peers within a block.
5. The outcome is a binary variable, where the regressions are estimated with OLS. Our goal is to detect coordination in first-year attendance by relating the attendance of a student to her peers, we do not aim to estimate a causal peer effects regression. Probit estimates, and corresponding marginal effects, show qualitatively similar results.
6. Standard errors for the coefficients in parentheses, clustered on the tutorial level.
7. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure A.2.5: Voluntary Sorting in Third-Year Courses



Notes:

1. Figures display marginal effects and 90% confidence intervals of a Probit model that explains whether a student pair enrolled in the same course with their shared characteristics (e.g. both students in the pair are female).
2. The models are identical to the ones displayed in Table 2.11, only the binary outcome variable in this model is equal to one if a student pair enrolled in the same course and zero otherwise.
3. Significant marginal effects are made bold.

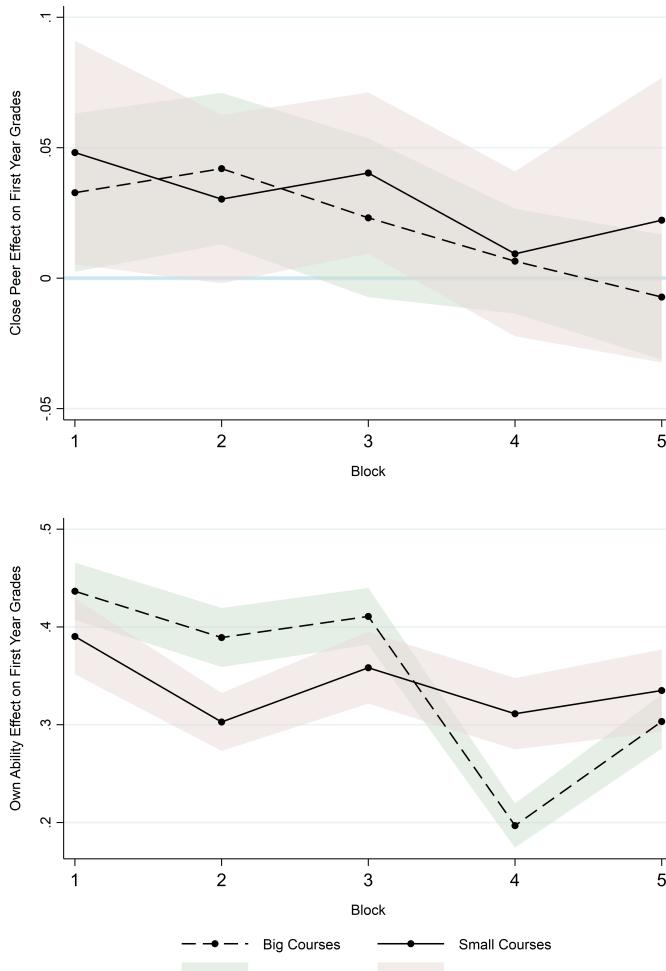
Table A.2.11: Peer Effects by Course Type

	Grades (Standardized)
	(1)
Close Peer GPA	0.0205* (0.0115)
Business Economics \times Peer GPA	0.0002 (0.0116)
Econometrics \times Peer GPA	0.0120 (0.0147)
Own GPA	0.3712*** (0.0136)
Business Economics \times Own GPA	-0.0535*** (0.0115)
Econometrics \times Own GPA	-0.0292** (0.0123)
Observations	18736
Adjusted R^2	0.323

Notes:

1. The regression includes course-cohort fixed effects and controls; student number, gender, age, and distance to university.
2. Peer GPA refers to the leave-out mean of high school GPA for the close peer group. Own GPA refers to own high school GPA. Both GPA measures are standardized.
3. The dummy business economics is one for business-economics courses and the dummy econometrics is one for econometrics courses. The baseline consists of economics courses. Appendix Table A.2.1 shows which courses belong to which category.
4. The course dummies are not included as separate variables as they are a linear combination of the course-cohort dummies.
5. Standard errors in parentheses, clustered on the tutorial level.
6. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure A.2.6: Effect of Peer GPA and Own GPA per Block



Notes:

1. Top graph shows the point estimates of close peer high school GPA on first year grades per block for big (8 ECTS) and small (4 ECTS) courses separately, and the corresponding 90% confidence intervals.
2. Bottom graph shows the point estimates of own high school GPA on first year grades per block for big and small courses separately, and the corresponding 90% confidence intervals.

Table A.2.12: Robustness of Peer Effects to Dropout Per Blocks

	Grades (Standardized)	
	Block 1-3 Block 4-5	
	(1)	(2)
Close Peer GPA	0.0514*** (0.0162)	0.0007 (0.0203)
Peer GPA \times (Assigned-Actual)	-0.0092 (0.0061)	0.0027 (0.0068)
Own GPA	0.3819*** (0.0130)	0.2779*** (0.0138)
Observations	11945	6791
Adjusted R^2	0.352	0.257

Notes:

1. All regressions include course-cohort fixed effects and controls; student number, gender, age, and distance to university.
2. Peer GPA refers to the leave-out mean of high school GPA for the close peer group. Own GPA refers to own high school GPA. Both GPA measures are standardized.
3. The regressions include a measure for the difference between the number of students at the beginning of the year in the close peer group (assigned class size) and the number of students that wrote the exam for the course per close peer group (actual class size). This is a measure for course dropout and is not standardized.
4. The coefficient on close peer GPA measures spillovers in classes where there has been no course dropout (assigned-actual=0).
5. Standard errors in parentheses, clustered on the tutorial level.
6. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.2.13: Instrumental Variable Analysis. Assigned Peer GPA is used as an Instrument for Actual Peer GPA

	First Year						Second Year		
	Block 1		Block 2		Block 3		Block 4	Block 5	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)			
Panel A: Actual Peer High School GPA (First Stage)									
Assigned Close Peer GPA	0.8520*** (0.0229)	0.8647*** (0.0223)	0.8996*** (0.0321)	0.8985*** (0.0360)	0.9032*** (0.0395)	0.0660*** (0.0214)			
Own GPA	-0.0028 (0.0067)	-0.0013 (0.0063)	0.0006 (0.0110)	0.0046 (0.0129)	0.0043 (0.0143)	0.0251* (0.0139)			
Adjusted R^2	0.918	0.865	0.801	0.745	0.723	0.166			
F-test on Excl. Instrument	1386.07	1498.46	784.00	622.00	521.67	9.49			
Panel B: Grades (Standardized; Second Stage)									
Actual Peer GPA	0.0475** (0.0205)	0.0418** (0.0165)	0.0353** (0.0162)	0.0089 (0.0149)	0.0069 (0.0196)	0.1850 (0.3411)			
Own GPA	0.4140*** (0.0158)	0.3451*** (0.0144)	0.3849*** (0.0176)	0.2537*** (0.0150)	0.3026*** (0.0158)	0.3091*** (0.0203)			
Observations	4271	4024	3650	3462	3329	10470			
Adjusted R^2	0.280	0.474	0.263	0.190	0.301	0.196			

Notes:

1. All regressions include course-cohort fixed effects and controls; student number, gender, age, and distance to university.
2. Own GPA refers to own high school GPA. All GPA measures are standardized.
3. In Panel A, the independent variable (Assigned Peer GPA) refers to the leave-out mean of high school GPA for the close peer group at the start of the first block in first year. This variable is used as an instrument for Actual Peer GPA, which is calculated on the course-cohort level and is equal to the leave-out mean of high school GPA for the close peer group in column (1) to (5) or for the tutorial peer group in column (6) while only taking into account the students who wrote the final exam of that course.
4. Panel B shows the results for the second stage, where Actual Peer GPA is the independent variable and has been instrumented with Assigned Peer GPA. The outcome variables are the standardized course grades for the first year per block in column (1) to (5) and for the second year pooled in column (6).
5. The number of observations for the second year in column (6) is lower than the baseline results for the first year. This is for three reasons; we do not observe the second-year grades of the 2014 cohort, students do not take all second-year courses in their second year, and for a small percentage we do not observe second-year tutorial choice.
6. Standard errors in parentheses, clustered on the tutorial level.
7. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.2.14: Peer Effects on Time Use per Blocks using Course Evaluations

	Attended Lectures		Total Study Time	
	Block 1-3	Block 4-5	Block 1-3	Block 4-5
	(1)	(2)	(3)	(4)
Close Peer GPA	-0.0223*** (0.0083)	-0.0072 (0.0112)	-0.1239 (0.1922)	-0.3454 (0.2989)
Own GPA	-0.0118 (0.0098)	-0.0176 (0.0124)	-0.4963*** (0.1730)	-0.6343*** (0.2050)
Observations	2995	1366	2995	1366
R^2	0.192	0.048	0.297	0.204
Binary Outcome	Yes	Yes	No	No

Notes:

1. All regressions include course-cohort fixed effects and controls; student number, gender, age, and distance to university.
2. Peer GPA refers to the leave-out mean of high school GPA for the close peer group. Own GPA refers to own high school GPA. Both GPA measures are standardized.
3. The dependent variable in column (1) and (2) is the answer to the question “Have you attended lectures?”. The dependent variable in column (3) and (4) is the answer to the question “Average study time (hours) for this course per week (lectures+tutorials+self study)?” where we used the maximum for the interval to convert the categories into hours.
4. Column (1) and (2) are estimated with Probit, column (3) and (4) with OLS. Marginal effects are reported. The R^2 refers to the Pseudo and Adjusted R^2 respectively.
5. Standard errors in parentheses, clustered on the tutorial level.
6. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Chapter 3

The Price of Forced Attendance

Joint work with Sacha Kapoor and Dinand Webbink

3.1 Introduction

For many people their first real encounter with autonomy happens at college or university. Out from under the roofs of their parents and high school teachers, how they manage their lives is now largely up to them. Many students use their newfound autonomy to skip class, especially in the early years of their undergraduate education, choosing instead to focus on extracurricular activities, such as student government, watching March Madness, or chasing other young men and women. To combat the rampant absenteeism this newfound autonomy begets,¹ and because of the substantial returns to college performance and graduation (Oreopoulos and Petronijevic, 2013; Cunha, Karahan and Soares, 2011; Jones and Jackson, 1990), university administrators and instructors often mandate frequent and regular class attendance among their students.²³ These attendance policies provide students with structure, helping them to circumvent behavioral predispositions towards non-academic activities, and ultimately to avoid decisions that can be bad for their lifetime utility (Lavecchia, Liu and Oreopoulos, 2014). By this token, and as long as attendance is valuable, additional structure should be good for academic performance. At the same time, however, additional structure constrains choices (*e.g.* time on self study) which are important for grades and, by doing so, precludes sensible students from choices that best serve their own self interest. This can be bad for academic performance.

¹Student absenteeism can be upwards of 60 percent of classes (Romer, 1993; Kottasz et al., 2005; Desalegn, Berhan and Berhan, 2014).

²An early discussion of mandatory attendance can be found in the correspondence section of the Journal of Economic Perspectives in 1994 (Correspondence, 1994). Motivated by Romer (1993), it consists of short letters by economics professors detailing their use of mandatory attendance.

³American universities spend 33 percent of their total budget on student instruction. This amounts to 56.7 billion dollars (for private nonprofit universities, years 2013-2014). Obtained via NCES: https://nces.ed.gov/programs/digest/d15/tables/dt15_334_40.asp, retrieved on 15-02-2017.

In this article we argue that additional structure is, in fact, bad for the performance of relatively good students. To make this argument, we draw on a natural experiment at a large European University to estimate the causal effects of a full year of forced, frequent, and regular attendance. The experiment requires students who average less than 7 (out of 10) in their first year to attend 70 percent of tutorials in each of their second-year courses. It imposes heavy time costs on students, as they can expect to spend 250 additional hours traveling and attending tutorials over a full academic year, amounting to approximately 7 additional hours per week. Students who fail to meet the attendance requirement face a stiff penalty, as they are not allowed to write the final exam for their course and must wait a full academic year before they can take the course again. Because students have imprecise control over their *average* grade in the first year, the experiment facilitates a regression discontinuity design (Lee, 2008; Lee and Lemieux, 2010) for identifying the various effects of forced attendance.

What does it mean to be forced? Our working definition is that a person is forced if a higher authority unilaterally takes away some of their potential choices. Or, more formally, if the authority imposes a heavy sometimes infinite penalty on a particular choice.⁴ The policy we study is well within confines of this definition.⁵ The policy asks students to come to campus frequently and regularly, choices which are normally under the purview of the student, and imposes a heavy penalty when they fail to do so. In addition to fitting well with a natural definition for economists, students perceived the policy as one where their attendance was forced, as this was how it was communicated to them by the university. Our data supports the notion that attendance was forced, as below-7 students collectively failed to meet the 70 percent criteria in less than one half of one percent of their courses. A more severe penalty, death *e.g.*, would have increased participation by less than half a percent, in other words.

Our estimates imply that forced students can expect a GPA decrease of 0.20 standard deviations over the remainder of their undergraduate degree. They can expect a decline of 0.15 standard deviations in their second year, when their attendance is forced, and a decline of 0.25 deviations in their third and last year, when they regain the right to decide their attendance. While the negative effects on second and third year performance are marginally significant or insignificant at conventional levels, we are able to rule out positive effects in the ranges of 0.05 and 0.1 standard deviations.

The average effect on second-year performance aggregates differential effects across all courses. While the university required all students below 7 to attend 70 percent of tutorials in all their second-

⁴Our paper is about more than just the role of sticks versus carrots in university education. A stick is typically defined as a penalty on performance, which itself is determined by choices and luck. Sticks constrain choices only implicitly, as the decision maker still has the freedom to make “bad” choices, and can simply hope that good luck helps them avoid penalties for poor performance.

⁵Our definition differs from the notion of labor coercion, which focuses on how physical force or the threat of it influences labor market institutions and outcomes.

year courses, it had no policy on how students above 7 should be treated. Unsurprisingly, several courses overlaid their own attendance initiatives onto the university policy, each differing in the intensity of the attendance constraint they imposed on students who scored above 7 in first year. Some courses penalized absenteeism by any student (absence-penalized courses), others strongly intimated and explained why all students should attend (attendance-encouraged courses), while a third group of courses followed the university policy and left the attendance decision up to above-7 students (attendance-voluntary courses). We observe the same students in all three scenarios because students have no discretion over course choice in second year.

The university policy had its largest effects in courses where attendance was voluntary for above-7 students. In these attendance-voluntary courses, it increased attendance by more than 50 percent, significantly decreased grades by 0.35 standard deviations, and significantly decreased the chances of passing by more than 10 percentage points. Self reports of total study time suggest further that forced students spend less time on non-academic activities such as leisure.

We delve into mechanisms behind the significant grade decreases in attendance-voluntary courses. We argue that the university policy forces students to spend a substantial number of hours in a specific way, leaving them less time for other activities, including activities which are important for grades. Grades decrease because the grade loss from spending less time on other academic activities outweighs the grade gains from additional attendance. What we observe fits with a model where students care about their grades and make informed decisions about their attendance. The latter is reinforced by our complier analysis, which identifies the most affected students, and shows that the largest policy effects are on the attendance of students who live far from campus or had a greater propensity to miss tutorials in first year.

In addition to aggregating the differential effects across all courses, the average second-year effect aggregates spillovers across courses taken concurrently. Forced students have slightly lower grades and passing rates in absence-penalized courses, even though they are not disadvantaged in their attendance decisions, having the same attendance rates as above-7 students. We explain that absence-penalized courses are always taken concurrently with a course where forced students are at an attendance disadvantage, arguing in turn that the grade decreases are consistent with negative spillovers from these courses. The spillovers, together with our results for activities other than attendance (*e.g.* leisure), suggest that the policy effect on grades does not operate through attendance alone.⁶

⁶We use all three courses, including the attendance-encouraged courses, to investigate other mechanisms, such as direct policy effects on self-perception or identity or stigmatization by other students, general discontent with the policy itself, negative peer effects, or course-level differences in the usefulness of tutorials. Our results imply these other mechanisms are unimportant.

The university policy was abolished in the last year of our sample. The abolition came as a surprise, as students only learned of it after the start of their second year. We show there was no grade difference near 7 for the abolition cohort. No grade difference for this cohort provides additional evidence against differential sorting of forced students into second year. More generally, it helps us rule out shocks other than the policy as drivers of worse performance just below the threshold. It also helps us show our results are driven by worse performance among forced students rather than better performance among above-7 students. Finally, it supports the presence of spillovers during years when the policy was in place.

Our study contributes to an expanding literature on incentives in education. A good deal of recent work analyzes the effects of interventions that reward students financially for “good” choices or better academic performance (Angrist, Oreopoulos and Williams, 2014; Castleman, 2014; Cohodes and Goodman, 2014; De Paola, Scoppa and Nistico, 2012; Leuven, Oosterbeek and van der Klaauw, 2010; Dynarski, 2008).⁷ We instead analyze the effect of an intervention which penalizes students heavily for “bad” choices, where the penalty is in terms of time rather than money.

Our findings contribute to debates over the merits of mandatory attendance in higher education (Romer, 1993).⁸ The argument for mandatory attendance is based on a robust positive correlation between grades and attendance.⁹ The argument has been reinforced by studies showing positive correlations between mandatory attendance and grades (see *e.g.* Marburger (2006) and Snyder et al. (2014)). We estimate the causal effects of a large-scale mandatory attendance policy and find negative effects.

One explanation for the discrepancy relates to the weight of the constraint imposed by the policy we study. A hefty constraint, spanning a full academic year, makes a negative finding more plausible. Another explanation relates to identification concerns in other studies. Previous research has relied on either year-over-year comparisons of students from different cohorts, or a discontinuity that allocates students to mandatory attendance later on in the same course. These strategies are problematic because the estimated effects of mandatory attendance may instead be attributable to heterogeneity across cohorts or students’ initial efforts to avoid mandatory attendance later on. Our context allows for within-cohort comparisons and lets us deal with anticipation effects.

⁷For more comprehensive lists, at all levels of education, see Lavecchia, Liu and Oreopoulos (2014) and Gneezy, Meier and Rey-Biel (2011).

⁸Our study has an indirect link with the compulsory schooling literature (Angrist and Krueger, 1991; Oreopoulos, 2007). We also examine the effect of a policy that penalizes people for specific choices. We differ in that our focus is on attendance at university, with steep and enforced penalties for absenteeism, and that we show that such policies can be very costly for students.

⁹For some of the many examples, see Romer (1993), Durden and Ellis (1995), Kirby and McElroy (2003), Stanca (2006), Lin and Chen (2006), Marburger (2001), Martins and Walker (2006), Chen and Lin (2008), and Latif and Miles (2013).

This article contributes, more generally, to debates over the role of structure in higher education (Lavecchia, Liu and Oreopoulos, 2014; Scott-Clayton, 2011). Arguments for additional structure usually focus on student predispositions towards non-academic activities, emanating from behavioral biases such as impatience, or imperfect information about behaviors that engender success at university. Our findings imply structure is detrimental to students with a GPA of 7, as well as to students with a GPA around 7 (Cerulli et al., 2017); above-average students at a prominent university in the Netherlands. From this perspective, our contribution is in showing that the cost of structure in higher education is lower academic performance among relatively good students.

3.2 Context

Our venue is the economics undergraduate program at a large public university in the Netherlands. The economics program itself is large - in the 2013-14 academic year alone, the program saw an influx of approximately 700 students. Students have no discretion over the courses they take in the first two years of the program, as all students follow the same ten courses per year, covering basic economics, business economics, and econometrics (See Table A.3.1 in the Appendix). Students have discretion over their courses in third year and, in line with this, declare a minor and major specialization (*e.g.* Accounting and Finance) which they can subsequently continue through to a Masters program.¹⁰ The economics program is given in both Dutch and English. The only difference between the programs is that the Dutch program has approximately 2.5 times more students.

Academic years are divided into five blocks, of eight weeks each (seven weeks of teaching and one week of exams). First- and second-year students have one light and one heavy course in each block, where they get four credits for the light course, and eight for the heavy course.¹¹ Heavy courses have three large-scale lectures per week, while light courses have two. Lecture attendance is always voluntary. Heavy courses have two small-scale tutorials (≈ 30 students) per week, while light courses have one. Lectures and tutorials both last for 1 hour and 45 minutes. Unlike lectures, but much like what may be found in structured college programs, tutorials require preparation and active participation of the student, via *e.g.* discussions of assignments and related materials.

Second year courses each have several time slots for tutorials and students can choose the one they wish to attend. Students register for slots a few weeks before the block begins. At the time of

¹⁰The Dutch and North American systems differ in two important ways. First, majors are defined more narrowly, as students decide to pursue economics, political science, sociology, and other social sciences before entering university. Second, they do three rather than four years of bachelors before a Masters.

¹¹In Europe study credits are denoted by ECTS, which is an abbreviation for European Transfer Credit System. This is a common measure for student performance to accommodate the transfer of students and grades between European Universities. One ECTS is supposed to be equivalent to roughly 28 hours of studying. 60 ECTS account for one year of study.

Completed first year	GPA < 7	GPA ≥ 7
Yes No	Forced Forced	Free Forced

registration, students are unaware of the teaching assistant (TA) that will teach each tutorial group, which are mostly senior-undergraduate and PhD students. Students cannot switch their tutorial group after the registration period ends. All students must register for a tutorial, including the ones scored above 7 in the first year. We observe for which group and at which time the student registered and can evaluate whether there were systematic differences in registration patterns for forced and voluntary students.

Grading is done on a scale that ranges from 1 to 10. Students fail a course if their grade is below 5.5. The average grade in the first year is weighted by the amount of credits the student gets for completing the course.

3.2.1 University Policy

Second-year students must attend 70 percent of tutorials for all of their second-year courses if they:

1. had an average grade (weighted by course credits) of less than 7 in first year;
2. failed at least one of their 10 first year courses.¹²

The table summarizes the students who had to comply with the policy. Students were not allowed to write the final exam and had to wait a full year before retaking the course if they failed to fulfill the 70 percent attendance requirement.

Our analysis focuses on the sample of students who completed the first year on time because first year completion rates for students around the cutoff is 92 percent.¹³ In this our primary estimation sample, the mean and standard deviation of first-year GPA are 6.99 and 0.70. The analogues in the unrestricted sample are 6.65 and 0.79. The means imply that the university policy assigns above-average students to forced attendance and, because the university is one of the more prominent universities in the Netherlands, that our findings apply to populations of relatively good students.

¹²Courses are grouped (Table A.3.1) such that a student can compensate a failing grade of between 4.5 and 5.4 from one course with a passing grade from another. This applies to all students, whether they are above or below the threshold of 7. A student who receives an 8 in microeconomics and 4.5 in macroeconomics can, in effect, take 1 point from their micro grade and use it towards their macro grade.

¹³In principle, one could estimate a local difference-in-difference, comparing changes in the grades of these students, around the cutoff, with changes in the grades of students who did not complete the first year. We did not do this because completion rates were so high near seven.

The policy imposes sizeable time costs on students. Forced students must spend 26 hours per block (3.5 hours per week) in tutorials.¹⁴ Once we account for the travel time of the average student, about 45 minutes each way,¹⁵ forced students must spend 50 hours per block traveling to and attending tutorials.¹⁶ All costs are in terms of time rather than money because student travel is fully subsidized in the Netherlands.

Students were made aware of the policy in their first year. Incoming students are assigned to tutors who, among other things, explain the policy to them. Student awareness facilitates adjustments in anticipation of forced attendance in the second year. As we explain later, our identification strategy is robust to anticipation effects as long as the average grade in first year is somewhat outside the student's control.

The introduction of the policy had nothing to do with the historical grade distribution of first-year students. The policy was introduced as part of a university-wide initiative to personalize education via small-scale tutorials. The initiative came about for three reasons: (i) the university had grown to a scale that made education impersonal; (ii) tutorials encourage active participation; (iii) the tutorials facilitate student involvement in the university community. Forced attendance was made part of the initiative to ensure a return on the university's sizeable investment in small-scale tutorials.

3.2.2 Course Policies

While the university forced the attendance of all below-7 students in all their second-year courses, courses differed in how they dealt with above-7 students. Table 3.1 provides a detailed overview on the courses and, in particular, on how they dealt with these students. Attendance was voluntary in two of the courses. Three courses strongly encouraged these students to attend. Three courses penalized them, and in fact also the below-7 students, for not attending. In this last set of courses, students had to complete assignments at the tutorials that made up five to thirty percent of their final grade. By not attending, students received a zero on this part of the course, meaning that at most they could obtain a 7 to 9.5 (rather than 10). The remaining two courses had no tutorials, and the final grade (mostly) consist of writing a research report in groups. Accordingly, these two courses are excluded from our analysis.¹⁷

¹⁴This is based on the fact that there are 3 tutorials of 1.75 hour per week, 7 non-exam weeks in a block, and that students must attend 70 percent of tutorials.

¹⁵The average student lives 22.9 kilometers from campus. From the Dutch student survey "Studenten Monitor" we observe that more than 70 percent of university students travel by public transport (<http://www.studentenmonitor.nl/>). To get an idea of the travel time, we used the Dutch public transport website (<http://9292.nl/>) to check travel times between the university and the few larger cities within a radius of 20 and 30 kilometers of the university.

¹⁶50 hours is a lower bound, as it ignores the preparation time for active participation in tutorials.

¹⁷There is no difference in grades near 7 for these two courses. Note that they do not provide credible placebo tests as final grades are largely determined via group work.

Table 3.1: Attendance Policies of Second-Year Courses

Course	ECTS	Tutorials	Policy	Years	Tutorial Description	Exam Qs.	Block
International Economics	8	Yes	Encourage	2009/13	Students explicitly told to attend 10 of 13 tutorials. Discussion of exercises that are hand in before tutorial. No direct influence on final grade.	MC	1
Ageing or Fiscal Economics	4	Yes	Penalize	2010/13	Economics of Ageing: Exercises + Presentations. Accounts for (roughly) 30 percent of their final grade. Fiscal Economics: Exercises, Accounts for 25 percent of final grade. Absence implies a 0 out of respectively 30 and 25.	MC	1
Finance I	8	Yes	Encourage	2009/13	Exercises, Outside tutorials there are weekly quizzes that account for 20 percent of final grade.	MC	2
Applied Statistics II	4	Yes	Penalize	2009/13	Exercises, Accounts for 15 percent of final grade. Absence implies a 0 out of 15.	Open	2
Applied Microeconomics	8	Yes	Encourage	2009/13	Draws on tutorial exercises for two interim tests which account for 20 percent of the final grade.	MC	3
History of Economic Thought	4	No	None	2009/13	Group and individual research projects.	3	
Methods & Techniques	8	Yes	Penalize	2009/13	Exercises in Computer Lab, Accounts for 5 percent of final grade. Absence implies a 0 out of 5.	MC	4
Behavioral Economics	4	Yes	Voluntary	2010/13	Exercises, Actual Experiments, No direct influence on final grade.	MC	4
Intermediate Accounting	8	Yes	Voluntary	2009/13	Exercises, No direct influence on final grade.	MC	5
Research Project	4	No	None	2009/13	Group research projects.	5	

Notes:

1. The description is extracted from course guides.
2. The Policy column indicates whether the course layered their own attendance policy over top of the forced attendance policy of the university. Voluntary indicates that attendance was voluntary for above-7 students. Penalized indicates that above-7 students (and all students) were penalized for missing tutorials. Encouraged indicates that attendance was strongly encouraged for above-7 students.

Note that because second-year students have no discretion over course choice, the pool of treated (and control) students is the same across the three types of courses. The lack of choice leaves no room for differential selection of voluntary students into one type of course or another.¹⁸ Ultimately, the course policies provide us with three counterfactuals: the grades of students whose attendance is voluntary, strongly-encouraged, and penalized. The three counterfactuals help us sort through various mechanisms which can generate and foster a relationship between forced attendance and academic performance.

3.2.3 Abolition

The policy lasted five years, starting in 2009-10 and ending 2013-14. Thus, the 2008-09 cohort was the first to be subjected to the policy in their second year, while the 2012-13 cohort was the last. The policy was abolished in 2014-15 because the student body and faculty, rightfully, as this paper shows, lobbied against it. The abolition came as a surprise to the 2013-14 cohort, as they were only made aware of it *after* their second-year had started, in the first block of the academic year 2014-15. They had the same incentive to score above 7 in first year as earlier cohorts, even though below-7 students were ultimately given discretion over their attendance in second year.

3.3 Data

Our main information source is the administrative data of the university. Our sample ranges from the 2008-09 academic year until 2014-15. We observe grades at the level of the student for all three of their undergraduate years, tutorial attendance for the first two years, course evaluations, and various personal characteristics. After restricting the sample to be within 0.5 grade points of 7, our main estimation sample, we have 5000 course-student observations based on more than 700 students.

The university uses attendance lists to track the attendance of students at tutorials. Students must sign in and teaching assistants must upload the attendance data to the university's online portal. The uploaded data is then used by the exam administration to verify that the attendance requirement is met.¹⁹

Our attendance variable is the percentage of tutorials the student attends (per course). It was measured quite accurately because teaching assistants were tasked with preventing fraudulent sign-

¹⁸Table 3.1 also shows multiple choice questions are used on the exams of all but one course. This precludes TAs from having a direct effect on grades.

¹⁹While matching attendance with the administrative data (*e.g.* grades and demographics), we experienced a match rate of 93 percent (in our main sample). We compared the matched observations with the non-matched observations and find that: (i) grades do not differ between the two groups; (ii) the treatment effect on grades is not different between the two groups; (iii) scoring below a seven in the first year could not explain whether or not a record is matched (See Table A.3.2 in the Appendix). Therefore we work with this 93-percent sample throughout the paper.

ins, as instructors required them to count the number of students present. The attendance statistics for voluntary students reinforces this point. On average these students attend tutorials 55 percent of the time. We can show that they also attend roughly 55 percent of their lectures. The match between tutorial and lecture attendance, together with the idea that students incur sunk costs of visiting campus, suggests that tutorial attendance is measured accurately.

Our data includes information from course evaluations. One week before the exam, students are invited by email to anonymously evaluate the course online at the university portal. They are reminded of the evaluations shortly after the exam. All evaluations contain the same set of 21 core questions, which are grouped into the general opinion of the course, structure, fairness, quality of lecturer and tutor, and usefulness of the lectures. Importantly, students are asked about their attendance at lectures, as well as the time they spend on their studies in total. Together with the data on tutorial attendance, we can infer how students adjust their time use between classes and studying on their own in response to forced attendance.²⁰ Note that the evaluations are filled out by 20 percent of the students. Later we will show that the response rate is the same just to the left and right of a first-year GPA of 7.

Our data on the personal characteristics of students includes information on their gender, age, distance from their residence to the university (in kilometers), and whether they are from the European Economic Area (EEA).²¹ For Dutch students we also have information on their performance in high school. Their grade for each of their high school courses is a 50-50 weighted average of the grade they earned in the course and the grade they earned on a nationwide exam for that course.

3.3.1 Basic Descriptives

Table 3.2 provides a basic summary of the data. The table compares students with an average first-year grade between 6.5 and 7 to students whose average grade was between 7 and 7.5. The unit of observation in the top panel is the student-course combination. The unit of observation in the bottom panel is the student. Second-year grades are measured in standard deviations.

The top panel shows forced students score 0.42 standard deviations worse than their peers. This is despite the fact that they attend tutorials 14 percentage points more of the time. The bottom panel implies students on one side of the cutoff are roughly similar to students on the other. The main difference being that poor performing students are likely to be over-represented to the left of 7 as visualized by their GPA in high school. Accordingly, we will account for this in our more flexible regression specifications by focusing on changes near 7.

²⁰For comprehensive details of the course evaluations see Table A.3.3 in the Appendix.

²¹Tuition fees are based on the student's EEA classification. Students who enroll in 2017-18, for example, pay €2,006 if they are from inside the EEA and €8,900 if not.

Table 3.2: Basic Descriptives (All 8 Eligible Courses)

	Grade Range	
	[6.5-7)	[7-7.5]
Course level (second year)		
Observations	2610	2291
Grade (<i>s.d.</i>)	-0.23	*** 0.19
Attendance tutorials	0.90	*** 0.76
Student level		
Observations	386	331
Distance to university (km)	24.13	22.04
Age	20.28	20.23
Gender (1=female)	0.30	0.31
European Economic Area (1=yes)	0.93	0.93
High-School Grade (<i>s.d.</i>)	-0.10	*** 0.12

Notes:

1. Each high-school grade is a 50-50 weighted average of the grade the high school assigned and the grade the student received on a national exam for the course.
2. *s.d.* denotes measurement in standard deviations.
3. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$
4. Stars denote the statistical significance for the difference in means, standard errors are clustered on the student level.

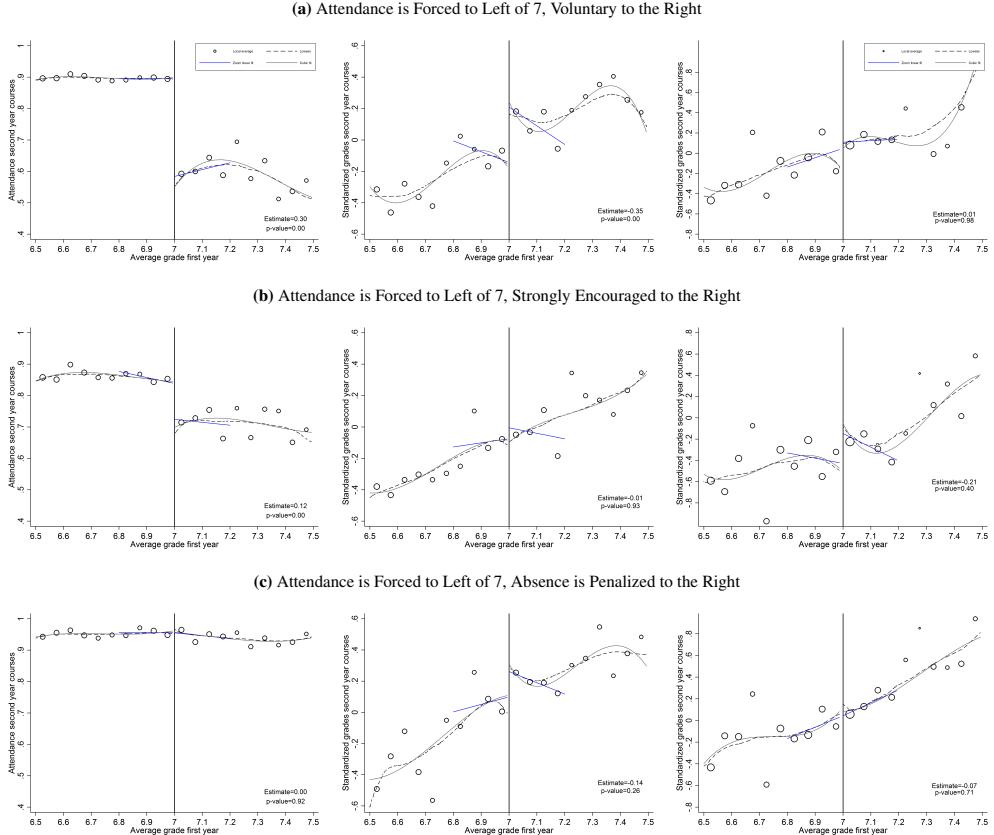
3.3.2 Preview of Baseline Results

The leftmost column of Figure 3.1 examines the effect on attendance for the three types of courses. In courses where above-7 students were given the option to attend, the difference in attendance between students above and below 7 was more than 30 percentage points. This translates into five extra tutorials for an eight credits course (three for a four credit course), or about 13 hours of extra schooling per block. In courses where the above-7 students were encouraged to attend, the difference was 12 percentage points. There was no attendance difference in courses that had their own penalty for being absent.

The middle column of Figure 3.1 examines the unconditional effect on grades. In courses where above-7 students had the option to attend, grades decreased by 0.35 of a standard deviation. In the other courses the effect on grades is a statistical zero. A grade comparison for attendance-voluntary and attendance-encouraged courses suggests that grades might only decrease if the additional time constraint is severe.

The figure at the bottom of the middle column shows a small grade decrease in absence-penalized courses. Although this grade difference across the cutoff is small and statistically insignificant, it appears to be puzzling at first sight as there is no attendance difference in these courses. Section 3.6 presents evidence against the decrease reflecting a direct effect of the university policy on grades

Figure 3.1: Second Year Attendance and Grades, by Course Type. The Left and Middle Panel show Attendance and Grades during the Policy (2009-14) respectively, Panel on the Right shows Grades after the Policy is Abolished (2014-15).



Notes:

1. Locally linear, cubic and weighted scatterplots (lowess) for attendance or 2nd-year grade against average 1st-year grade.
2. Dots are based on local averages for a binsize of 0.05. Dot sizes reflect the number of observations used to calculate the average.
3. Linear and cubic fits are chosen according to our preferred specifications (see Section 3.4). Lowess makes no assumption on functional form (estimated with a bandwidth of $0.8N$).
4. Binsizes for local averages are selected via F-tests for regressions of 2nd-year grades on K bin dummies and $2K$ bin dummies for the average 1st year grade.

(e.g. via self perception of the student) together with across-course differences in the value of tutorial attendance. Section 3.7 provides evidence that this difference reflects negative spillovers from adjacent courses where above-7 students have discretion over their attendance. This suggests that the university policy does not operate through attendance alone.

3.3.3 Abolition Results

The rightmost column of Figure 3.1 plots grade distributions in the abolition year, 2014-15. The figures show little to no difference in grades around the cutoff for the three types of courses. We observe that both the direct (top right) and spillover (bottom right) effects of the university policy have disappeared. Appendix Table A.3.4 shows formally that the differences are all statistically insignificant.²²

Appendix Table A.3.5 compares mean grades above and below 7, before and after the abolition, in attendance-voluntary courses alone. The table shows that the unstandardized grades of students who expected to be forced in 2014-15, but ultimately were not, are 0.35 points (on a 10-point scale) higher than the grades of forced students from earlier cohorts. It also shows that this across-cohort difference is similar to the within-cohort difference of 0.37. In addition to providing further evidence that it is the forced attendance which decreases grades, Table A.3.5 implies that our estimates are being generated by lower performance of forced students, rather than by better performance of unforced students.

3.4 Empirical Specification

The second-year grade $g_{ijc}^{(2)}$ of student i in course j and cohort c is given by

$$g_{ijc}^{(2)} = \beta_0 + \beta_1 D_{ic} + f(\bar{g}_{ic}^{(1)} - 7) + f(\bar{g}_{ic}^{(1)} - 7)D_{ic} + C_{jc}^{(2)} + \mathbf{X}_i \boldsymbol{\Gamma} + \varepsilon_{ijc}^{(2)} \quad (3.1)$$

where D_{ic} equals 1 if first-year GPA is below 7, $\bar{g}_{ic}^{(1)}$ is their GPA in first year, $f(\cdot)$ is some polynomial expansion in $\bar{g}_{ic}^{(1)}$, $C_{jc}^{(2)}$ are course-cohort fixed effects, and \mathbf{X}_i includes personal characteristics such as age. We allow the polynomial to differ from the left to the right of 7 (see the discussion in Lee and Lemieux (2010)), in part because it allows us to later analyze the external validity of our estimates (Cerulli et al., 2017). Our primary interest is β_1 , the effect of forced attendance near 7. The adoption and use of the forced attendance policy suggests $\beta_1 > 0$. The constraint it imposes on choices suggests $\beta_1 < 0$.

We can interpret estimates of β_1 causally if (Lee, 2008):

Identifying Assumption: Students have imprecise control over their average grade in the first year, meaning that conditional on their characteristics, the distribution for average grades is continuous around 7.

Because students were made aware of the policy early on and throughout their first year, they could try to take actions to avoid forced attendance in the second year. Our identification strategy will still work as long as first-year grades are at least somewhat outside of the student's control.

²²We cannot plot attendance because the university stopped registering attendance in the abolition year.

The above is generally a weak identifying assumption (Lee, 2008) and is reasonable in our setting. The assignment to forced attendance is based on the student's *average* grade. As students accumulate grades they lose control over the average. Importantly, first-year adjustments to the threat of second-year forced attendance, such as the practice of asking professors for grade increases,²³ have less of an effect on first-year GPA than on the grade of any one course.²⁴ The lack of control, together with the presence of aggregate shocks to the first-year performance of the individual student, should be enough for generating random assignment around 7.

To gain intuition for the identification argument, let

$$g_{ijc}^{(1)} = e_{ijc}^{(1)} + a_{ijc} + \delta_{jc}^{(1)} + \eta_{ijc}^{(1)}$$

denote the student's grade in first-year course j , a_{ijc} is their ability, $\delta_{jc}^{(1)}$ is something particular about the course-cohort combination (such as the professor or teaching assistant), and $\eta_{ijc}^{(1)}$ is the idiosyncratic component of the first-year grade. $e_{ijc}^{(1)}$ encapsulates any choice that affects grades, including the intensity of effort, study hours, tutorial and class attendance, or requests for grade increases. Second-year tutorial attendance is mandatory if:

$$\bar{e}_{ic}^{(1)} + \bar{a}_{ic} + \bar{\delta}_c^{(1)} + \bar{\eta}_{ic}^{(1)} < 7$$

where the bars indicate that the variable is averaged over all first-year courses j .

The argument has three parts. The first is the student has limited control, $\bar{e}_{ic}^{(1)}$, over their average performance, as the effect of *e.g.* grade manipulation is smaller in the aggregate. The second is that there are aggregate shocks to first-year performance, $\bar{\eta}_{ic}^{(1)}$, such as bad luck across the exams they wrote that year. Shocks like these ensure that two students, with similar ability and average effort, end up on either side of the cutoff. As a result, the conditional distribution of first-year GPA is continuous and the variation in treatment status will be random in a neighborhood of 7. The third is that randomization near the cutoff takes place cohort by cohort. The student pool near 7 in one cohort may differ from the student pool near 7 in another. The presence of $\bar{\delta}_c^{(1)}$ suggests we should control for differences across cohorts.

²³ Asking professors for grade increases, or any other such practice, can effect treatment assignment only when cumulative GPA is very close to 7.

²⁴We are developing a companion article that studies adjustments to the threat of forced attendance. Our evidence shows that the threat does elicit a response but that, as expected, the response is almost never enough to get out of forced attendance. This claim is supported by various randomization and McCrary tests, as well as the null effects for abolition year. Nonetheless, because of this concern, we will use models that exclude potentially problematic neighbourhoods around 7 (donut-hole RD models) to demonstrate the robustness of our results.

3.4.1 Continuity Near the Cutoff

Local randomization of the treatment near the cutoff gives us two testable implications: (i) observed characteristics are identical from one side of the cutoff to the other; (ii) the probability density for GPA is continuous. We evaluate the implications one by one.

Table 3.3 presents estimates of our main empirical specification (Equation (3.1)) where instead of grades the dependent variables are student characteristics. The table presents results for local linear regressions (panel A) and a third order polynomial for $f(\cdot)$, with our main estimation sample (panel B) and the full sample (panel C).

Students to the left and right of the cutoff are similar in whether they come from the European Economic Area, age, distance from the university (in kilometers), and in their performance in high school (level, track, and average grade).²⁵ This conclusion holds if we select the bandwidth optimally for each background characteristic (Appendix Table A.3.6). It also holds if we consider grade differences for various high school courses (Appendix Table A.3.7). Although much of the evidence supports the local randomization interpretation, in two of the three specifications of Table 3.3 the estimates indicate that women are underrepresented just to the right of the cutoff, consistent with the idea that women are manipulating grades less than men. The gender imbalance near 7, and residual concerns for grade manipulation more generally, further motivates estimation of donut-hole RD models.

We examine whether the probability density for GPA is continuous around 7 (McCrary, 2008). If students can manipulate their GPA here, then we could observe bunching just above 7. To check we estimated Equation (3.1) using normalized counts of the number of students as the dependent variable.²⁶ Figure 3.2 summarizes the results, showing no evidence of bunching above the threshold. Table A.3.8 in the Appendix verifies this, formally showing that we are unable to reject the null of continuity near the cutoff.

3.4.2 Sample Attrition

The policy may have incentivized students to drop courses if and once they fail the 70 percent attendance requirement. Attrition of this sort could threaten identification because dropouts are not graded. Accordingly, we test for a policy effect on the number of second year courses for which a student obtained a valid grade. The results in Appendix Table A.3.9 (Columns 1-3) imply the policy

²⁵A Dutch high school student might have followed two different levels before enrolling at university (easy=0, difficult=1). They might have followed one of 4 tracks within each level (1=least prestigious, 4=most prestigious). For the latter track variable, the results are unchanged if we account for the ordered nature of the variable.

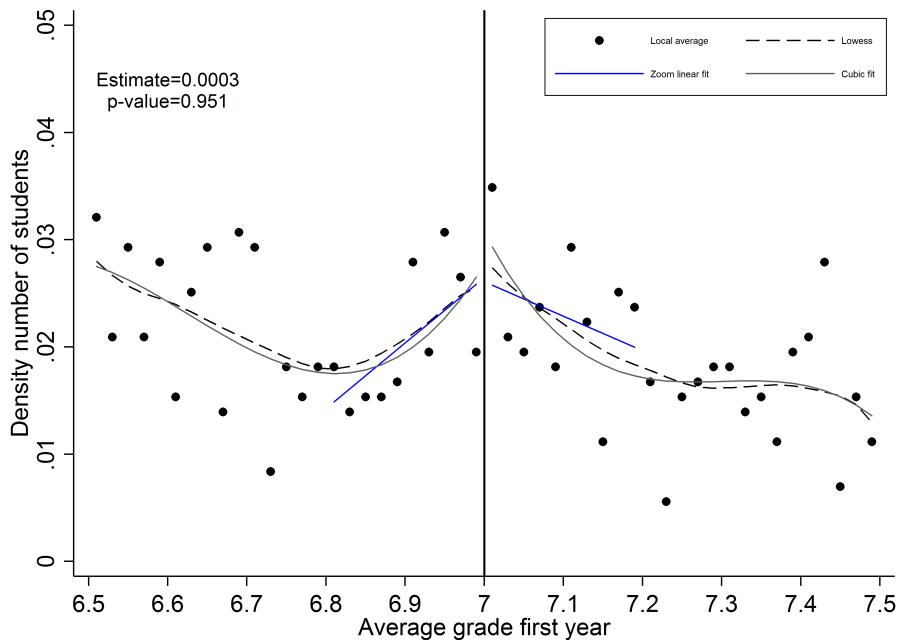
²⁶To count the number of students we select bin sizes in accordance with the proposed strategy of McCrary (2008). The results are robust to the bin size.

Table 3.3: Balancing Tests around the Cutoff

	Distance to University	Age	Gender	European Economic Area	High School Level	High School Track	High School Grade
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A: Local linear regression							
Average 1 st -year Grade is Below 7	1.796 (0.24)	0.257 (1.22)	0.257*** (2.50)	-0.032 (-0.53)	0.021 (1.14)	0.241 (1.14)	-0.209 (-0.83)
Observations	310	310	310	310	248	248	248
Adjusted R^2	0.000	-0.000	0.001	-0.004	0.006	0.274	0.003
B: Third order polynomial							
Average 1 st -year Grade is Below 7	-3.999 (-0.42)	0.357 (1.27)	0.278** (2.13)	0.024 (0.31)	0.004 (0.21)	0.170 (0.63)	-0.348 (-0.82)
Observations	717	717	717	717	592	592	592
Adjusted R^2	0.008	-0.005	-0.000	0.005	0.017	0.258	0.008
C: Third order polynomial on whole sample							
Average 1 st -year Grade is Below 7	1.349 (0.24)	0.248 (1.37)	0.084 (1.06)	-0.005 (-0.12)	0.042* (1.72)	0.245 (1.47)	-0.313 (-1.29)
Observations	1420	1420	1420	1420	1176	1176	1176
Adjusted R^2	0.009	-0.001	0.002	0.008	0.011	0.350	0.065
Mean Outcome Var.	22.978	20.289	0.290	0.938	0.979	2.501	6.882

Notes:

1. The unit of observation is the student.
2. Regressions include cohort fixed effects.
3. Top panel displays local linear regressions with the optimal bandwidth of 0.2 around first-year grade of 7. Middle panel shows regressions for the optimal bandwidth of 0.5 with the third order polynomial. Bottom panel includes all observations. Polynomial is interacted with the treatment.
4. Column (5) until (7) only use the sample of Dutch students.
5. t -statistics in parentheses, standard errors are robust.
6. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure 3.2: No Bunching Just Above 7. RD plot of the density for the number of students.

has no effect on the number of completed courses. The intercepts support this conclusion, as they show students near the cutoff complete almost every course (nine out of ten).

Students near 7 may differ in their propensity to complete course evaluations and thus compromise the use of course evaluations in our analysis. Appendix Table A.3.9 (Columns 4-7) reports estimates of the policy effect on an indicator for whether students completed the course evaluation, for all eight courses, and separately for attendance-voluntary courses. We find no statistical differences in the propensity to complete the evaluation near 7. As with course completion, our evidence suggests no differential selection into course evaluations.

3.4.3 Estimation and Inference

While discussing the results we present two specifications for Equation (3.1): a linear and third order polynomial for $f(\cdot)$ with bandwidths of 0.2 (6.8 to 7.2) and 0.5 (6.5 to 7.5) respectively. For all specifications we cluster standard errors at the level of the student.²⁷

²⁷We do not cluster on the tutorial group because peer composition differs from course to course. However, we show that our results are robust to including tutorial fixed-effects.

We elaborate on how we settled on our preferred specifications. Because local randomization implies that local comparisons provide an unbiased estimate of β_1 , our starting point will almost always be specifications with a narrow bandwidth, where $f(\cdot)$ is taken to be linear. We follow Imbens and Lemieux (2008) and refer to these as locally linear regressions. For these specifications, we use a bandwidth of 0.2 because it is optimal according to the cross-validation method, and confirmed via the various bandwidth selectors provided by Calonico et al. (2016). Our second set of specifications use a wider bandwidth of 0.5 together with the third-order polynomial for $f(\cdot)$. A wider bandwidth and flexible functional form lets us use more data and approximate what a locally-randomized experiment would have shown (Van der Klaauw, 2002). To select $f(\cdot)$ for this larger estimation sample, we estimated Equation (3.1) while adding equal-sized bin dummies of GPA and including higher-order polynomials until the bin dummies were jointly insignificant.²⁸ We did this for multiple bandwidth choices. To select the 0.5 bandwidth, we again made use of the various bandwidth selectors of Calonico et al. (2016).²⁹

One remaining concern relates to whether GPA has enough mass points to warrant a continuity-based RD design, which allows for the possibility that average potential outcomes vary with the running variable (GPA). To this end, note that there are 228 unique GPA values for the 717 students in our estimation sample of 6.5 to 7.5, amounting to approximately one GPA value for every 3 students. This amount of coverage of the support for GPA is usually sufficient for a continuity-based design.³⁰

3.5 Baseline Results

Table 3.4 reports estimates that are based on pooled data from the 8 affected courses. Basing estimates on the pooled data allows us to account for across-course error correlation within students. Estimates of the average effect are found in Columns (1) and (2). Panels A and B report the estimated effects for attendance and grades.

The university-wide policy increases the attendance of forced students by 15 percentage points ($p < 0.01$). It decreases their grades by 0.15 standard deviations. While we are unable to reject the null hypothesis of no average effect on grades, we are able to reject null hypotheses of positive effects of 0.05 and 0.1 standard deviations, with p -values of 0.09 and 0.03, respectively.

²⁸We ran various regressions while changing the number of bins, but our preferred specification includes the number of bins (8) for which we first stopped rejecting the small (few dummies) versus the big model while choosing the binsize for the local averages for the RD graphs (see Figure 3.1).

²⁹See Appendix Figure A.3.1 and Table A.3.10 for more details on the optimal bandwidth selection. Note that we use the equation between student grades and first-year GPA for selecting the bandwidth and polynomial order. This seems reasonable as the relationship between attendance and first-year GPA is relatively flat to left and right of 7. In the latter case we would expect the polynomial to be linear and the optimal bandwidth to be wide.

³⁰Cattaneo, Idrobo and Titiunik (2018) analyze an example where for every 110 observations one unique value for the running variable is observed. They conclude that continuity-based analysis might be possible in this context.

Table 3.4: RD for All 8 Eligible Courses

	Average Effect		Marginal Effects by Course Type	
	(1)	(2)	(3)	(4)
A: Attendance (% Tutorials Attended)				
Average 1 st -year Grade is Below 7	0.151*** (4.30)	0.147*** (4.28)	0.151*** (2.96)	0.146*** (2.94)
Attendance is Voluntary × Treatment			0.193*** (3.91)	0.195*** (3.96)
Absence is Penalized × Treatment			-0.151*** (-2.99)	-0.149*** (-2.97)
Adjusted R^2	0.306	0.311	0.365	0.370
B: Grade (Standardized)				
Average 1 st -year Grade is Below 7	-0.153 (-1.26)	-0.154 (-1.28)	0.0293 (0.18)	0.0262 (0.16)
Attendance is Voluntary × Treatment			-0.451** (-2.36)	-0.447** (-2.35)
Absence is Penalized × Treatment			-0.188 (-1.07)	-0.185 (-1.06)
Observations	4901	4901	4901	4901
Adjusted R^2	0.210	0.210	0.210	0.210
Controls	No	Yes	No	Yes

Notes:

1. Regressions include course-cohort fixed effects.
2. Controls include distance to the university, age, gender, and European Economic Area.
3. All regressions use a third-order polynomial, as well as their interactions with the treatment, with a bandwidth of 0.5.
4. t -statistics in parentheses, standard errors are clustered on the student level.
5. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

3.5.1 Course-Level Attendance Policies

Table 3.5 evaluates the policy effect on attendance for the three types of courses. Columns (1) and (2) of Table 3.5 report estimated effects on attendance in courses which gave above-7 students the option to attend, (3) and (4) report effects for courses where attendance was strongly encouraged, and (5) and (6) report effects for courses where everyone was penalized for being absent.

Forced students attended 29 to 34 percentage points more tutorials than above-7 students in attendance-voluntary courses ($p < 0.01$). They attended 11 to 15 percentage points more in attendance-

Table 3.5: Forced Students Attend More Often

Attendance (% Tutorials Attended)									
Courses where Attendance is Forced to the Left of 7 and where to the Right									
	Attendance is Voluntary	Attendance is Encouraged	Absence is Penalized	(1)	(2)	(3)	(4)	(5)	(6)
A: Local linear regression									
Average 1 st -year Grade is Below 7	0.296*** (6.25)	0.285*** (6.21)	0.122*** (3.01)	0.107*** (2.78)	0.002 (0.10)	0.000 (0.01)			
Observations	547	547	847	847	742	742			
Adjusted <i>R</i> ²	0.366	0.376	0.153	0.174	0.154	0.180			
B: Third order polynomial									
Average 1 st -year Grade is Below 7	0.344*** (5.79)	0.335*** (5.72)	0.151*** (2.97)	0.145*** (2.92)	0.000 (0.01)	-0.000 (-0.01)			
Observations	1275	1275	1965	1965	1661	1661			
Adjusted <i>R</i> ²	0.408	0.412	0.172	0.184	0.146	0.151			
Controls	No	Yes	No	Yes	No	Yes			

Notes:

1. Regressions include course-cohort fixed effects.
2. Controls include distance to the university, age, gender, and European Economic Area.
3. Top panel uses a bandwidth of 0.2 around a first-year grade of 7. Bottom panel uses a bandwidth of 0.5. Polynomial is interacted with the treatment.
4. *t*-statistics in parentheses, standard errors are clustered on the student level.
5. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

encouraged courses ($p < 0.01$). They had the same attendance as above-7 students in absence-penalized courses.

Analogous estimates for grades are found in Table 3.6. Columns (1) and (2) show forced students have grades which are 0.34 to 0.43 standard deviations lower in attendance-voluntary courses ($p < 0.01$). Columns (3) and (4) shows little to no grade difference in attendance-encouraged courses. Columns (5) and (6) show the grades of forced students are 0.14 to 0.17 standard deviations lower in absence-penalized courses, though these differences are statistically insignificant at the 10 percent level. Note that columns (3) and (4) of Table 3.4 show the estimates in Tables 3.5 and 3.6 are similar to the estimates we would obtain with pooled data and interactions between the treatment variable and course type.

Students can actually be better off with lower grades if their goal is to pass and forced attendance makes passing equally or more likely, perhaps because tutorials give students a better overview of the

Table 3.6: Forced Students Perform Worse in courses where attendance is voluntary for students scoring above 7 in first year.

	Grade (Standardized)					
	Courses where Attendance is Forced to the Left of 7 and where to the Right					
	Attendance is Voluntary	Attendance is Encouraged	Absence is Penalized			
	(1)	(2)	(3)	(4)	(5)	(6)
A: Local linear regression						
Average 1 st -year Grade is Below 7	-0.349*** (-2.80)	-0.342*** (-2.77)	-0.011 (-0.09)	0.010 (0.08)	-0.143 (-1.12)	-0.164 (-1.27)
Observations	547	547	847	847	742	742
Adjusted R^2	0.177	0.174	0.201	0.200	0.096	0.099
B: Third order polynomial						
Average 1 st -year Grade is Below 7	-0.422*** (-2.65)	-0.426*** (-2.74)	0.029 (0.18)	0.041 (0.26)	-0.158 (-0.97)	-0.169 (-1.02)
Observations	1275	1275	1965	1965	1661	1661
Adjusted R^2	0.216	0.216	0.251	0.250	0.156	0.158
Controls	No	Yes	No	Yes	No	Yes

Notes:

1. Regressions include course-cohort fixed effects.
2. Controls include distance to the university, age, gender, and European Economic Area.
3. Top panel uses a bandwidth of 0.2 around a first-year grade of 7. Bottom panel uses a bandwidth of 0.5. Polynomial is interacted with the treatment.
4. t -statistics in parentheses, standard errors are clustered on the student level.
5. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

minimum they need to know. They may, in other words, achieve their desired result (passing) with less effort.

Columns (1) and (2) of Table 3.7 suggest this is not the case. Forced attendance decreases the probability of passing by 10 to 13 percentage points. The narrow-bandwidth estimates are statistically significant at the 5 percent level. The wide-bandwidth estimates have p -values which are a bit above 10 percent.³¹ Columns (3) and (4) show there is effectively no difference in passing rates for attendance-encouraged courses. Columns (5) and (6) show passing rates which are 7 percentage points lower, with p -values which fluctuate around 10 percent.

³¹A probit analysis with a third-order polynomial yields similar but stronger (statistically) results.

Table 3.7: Forced Students are Less Likely to Pass

Passes the Course						
Courses where Attendance is Forced to the Left of 7 and where to the Right						
	Attendance is Voluntary	Attendance is Encouraged	Absence is Penalized			
	(1)	(2)	(3)	(4)	(5)	(6)
A: Local linear regression						
Average 1 st -year Grade is Below 7	-0.125** (-2.16)	-0.117** (-2.02)	0.000 (0.01)	0.003 (0.05)	-0.066 (-1.56)	-0.072* (-1.74)
Observations	547	547	847	847	742	742
Adjusted <i>R</i> ²	0.074	0.070	0.093	0.088	0.050	0.051
B: Third order polynomial						
Average 1 st -year Grade is Below 7	-0.106 (-1.47)	-0.103 (-1.45)	0.029 (0.40)	0.034 (0.47)	-0.070 (-1.29)	-0.071 (-1.33)
Observations	1275	1275	1965	1965	1661	1661
Adjusted <i>R</i> ²	0.082	0.082	0.120	0.118	0.092	0.097
Controls	No	Yes	No	Yes	No	Yes

Notes:

1. Regressions include course-cohort fixed effects.
2. Controls include distance to the university, age, gender, and European Economic Area.
3. Top panel uses a bandwidth of 0.2 around a first-year grade of 7. Bottom panel uses a bandwidth of 0.5. Polynomial is interacted with the treatment.
4. *t*-statistics in parentheses, standard errors are clustered on the student level.
5. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

3.5.2 Robustness

We analyzed the robustness of the result that forced attendance lowers grades in courses where attendance was voluntary for students to the right of 7. We estimated Equation (3.1) with the third order polynomial while varying the size of the bandwidth from 0.3 until 1.0. Appendix Figure A.3.2 shows the estimates hover around -0.4 and -0.3 and are significant across the whole range of optimal bandwidths using the various bandwidth selectors of Calonico et al. (2016).

We tested for significance at fake cutoffs. We estimated our main specification using the third order polynomial and a bandwidth of 0.5, while implementing fake cutoffs at every 0.005 points for GPA between 6.5 and 7.5, where the true cutoff is at 7. Appendix Figure A.3.3 presents a histogram and probability density of the β_1 estimates. The distribution mean is 0.02. The estimate at the true cutoff is extreme relative to the mean, having an empirical *p*-value that ranges between 3 and 6 percent (depending on whether normality is assumed).

We used donut-hole RD models to address concerns about the potential for manipulation and gender imbalance near the cutoff. Appendix Figure A.3.4 shows the effect on grades is more negative as observations near 7 are removed. Note that this is consistent with forced attendance being relatively more costly to students who try to avoid it.³²

We tested whether our results change if we restrict the linear polynomial $f(\cdot)$ to be the same on both sides of the cutoff. Appendix Table A.3.11 shows our results, where the estimates are virtually unchanged for all groups of courses. Finally, the robustness is supported by negligible effects in the abolition year. See the right panel of Figure 3.1 and Appendix Table A.3.4 for details.

3.5.3 External Validity

Our RD estimates apply to students with a GPA of 7. A valid question relates to the applicability of the estimates to students with a GPA other than 7. To speak to this question, Cerulli et al. (2017) recommend examining the coefficient on the interaction of the linear polynomial term and the treatment ($(\bar{g}_{ic}^{(1)} - 7)^1 D_{ic}$ in Equation (3.1)). A non-zero coefficient implies that students with a GPA just around, but not equal to, 7 can expect different treatment effects. It is effectively the Treatment Effect Derivative (TED) at 7.

We examine the TEDs for attendance and grades in Appendix Table A.3.12, for all 8 courses, as well as for attendance-voluntary courses, where the effect sizes are largest. We find that the TED estimates are all statistically insignificant. Similar implications follow from Figure 3.1, as it shows similar curves to the left and right of 7, especially in the case where attendance is the dependent variable. We also follow the suggestion of Cerulli et al. (2017) to consider the relative TED, *i.e.* the treatment effect divided by the TED and multiplied by the bandwidth. If the relative TED is less than one in absolute value, then the treatment effect changes sign somewhere in the estimation sample defined by the bandwidth. Appendix Table A.3.12 shows the relative TEDs have values which are above one for five out of the eight specifications. For the remaining three specifications we cannot reject the null hypothesis that the relative TED is equal to one. This suggests that our RD estimates apply to students whose GPA differs slightly from 7.

3.6 Baseline Mechanisms

We explore several potential mechanisms behind lower grades and passing rates in courses where attendance was voluntary for above-7 students.

³²We consider donut holes with a maximum size of 6.95-7.05. To see why, suppose the GPA of the student is 6.95. To get to 7 they would need to receive a grade increase of more than 0.376 (0.752) for an eight (four) credits course. These sorts of increases are large and unlikely.

3.6.1 Peer Effects

If the performance decline is driven by lower quality peers and TAs, then we would expect a more moderate or negligible decline if our estimates were based on comparisons of forced and voluntary students who attend the same tutorial. Appendix Table A.3.13 considers these comparisons, presenting treatment effect estimates for grades which are conditional on fixed effects for the tutorial group. The estimates are similar to our baseline estimates, suggesting that peer and TA quality are relatively unimportant for the effect of forced tutorial attendance on performance.

Appendix Table A.3.13 also evaluates peer effects more specifically, using the most common peer effects specifications in the literature (Booij, Leuven and Oosterbeek, 2017), and focusing on whether the policy effect differs depending on the peer. The table reports effects of treatment interactions with the average 1st-year grade for the peer group, as well as interaction effects for the average peer registration time for tutorials, measured in differences in days from the course mean registration time. The interaction effects account for the possibility that students coordinate tutorial times with their most preferred peers, which for forced students might very well include other low achievers. It also helps with the possibility that weak students coalesce simply because registration is left to the last minute.

The effects of treatment interactions with peer quality are modest. All the estimates are statistically insignificant at conventional significance levels, while the main treatment estimate is unchanged compared to our baseline specifications. Negligible peer effects are unsurprising given recent discussions and results in the literature (Sacerdote, 2014).³³

3.6.2 Attendance is Useful in Some Courses, but not Others?

The effectiveness of tutorials provides another alternative explanation for why performance is worse in attendance-voluntary courses. To justify our thought process on this, we will draw on estimates from the attendance-encouraged and absence-penalized courses. Notice that the grades of forced students were about 0.15 standard deviations lower in courses where all students were penalized for missing tutorials. There was no grade difference in attendance-encouraged courses. The patterns may reflect the combined influence of a direct effect of the university policy (*e.g.* via self perception) and course-specific heterogeneity in the usefulness of attendance. Grades may be similar in attendance-encouraged courses because attendance is useful which cancels out the 0.15 reduction in grades. Students may have lower grades in attendance-voluntary courses because attendance is use-

³³Feld and Zöllitz (2017) is especially relevant. They estimate positive but small peer effects in tutorials for economics students at another Dutch university.

less. Useless attendance can reinforce the 0.15 reduction from the direct effect, taking it down to the 0.35 reduction we observe in the data.

If our results are driven by differences in the usefulness of attendance, then TA and Lecturer quality should be highest in attendance-encouraged courses. Appendix Table A.3.14 uses data from the abolition year to investigate this possibility, reporting estimates of the relationship between perceived TA/Lecturer quality and fixed effects for the different types of courses, the baseline being the courses where students to right of 7 were encouraged. Data from the abolition year circumvents concerns about whether the course evaluations are contaminated by participation in forced attendance.

Appendix Table A.3.14 suggests, if anything, that TA quality is lowest in courses where attendance was encouraged (the base group). It also shows no statistical difference in lecturer quality across the three types of courses. The evidence suggests our results are not explained by a direct negative effect combined with course-specific heterogeneity in the usefulness of attendance.

3.6.3 It's About Time

If the policy has its largest effects on students who pay a high price for or derive little additional utility from attendance, then our results would be consistent with a model where students care about grades, where they think carefully about their attendance, and, importantly, where their time is being constrained by the policy. These students are really forced, being pushed further away from the choices they would make in the absence of the policy.

We estimate

$$A_{ijc}^{(2)} = \gamma_0 + \gamma_{1ic} D_{ic} + \varepsilon_{ijc}^{(2)} \quad (3.2)$$

where $A_{ijc}^{(2)}$ is the percentage of tutorials attended. If γ_{1ic} is large then the student's desired attendance is low, such that they would have attended far fewer tutorials in the absence of forced attendance. Alternatively, a small γ_{1ic} implies attendance is desirable, such that the student attends the same number of tutorials with or without forced attendance. In the parlance of the treatment effects literature (Angrist and Pischke, 2008), students who otherwise prefer not to attend (large γ_{1ic}) are compliers. Students who would attend anyways (small γ_{1ic}) are always takers. There are no never takers or defiers by the very definition of the policy, as it leaves students with no choice but to attend tutorials when their first-year GPA is below 7. Indeed, of the courses from students with a first-year GPA below 7, we observe only 0.44 percent with an attendance rate below 70 percent.³⁴

³⁴One might argue that the grade for never takers are never observed, as they cannot write the exam. However, in Section 3.4.2 we showed students generally participate in every second-year course, and that their near-perfect course participation is unaffected by the treatment (leaving no room for never takers).

Table 3.8: Differential Effects on Attendance

	Attendance (% Tutorials Attended)		
	(1)	(2)	(3)
Average 1 st -year Grade is Below 7	0.337*** (5.83)	0.389*** (7.15)	0.390*** (7.27)
Distance to University	-0.040** (-2.43)	-0.013 (-1.59)	-0.036** (-2.26)
Distance \times Treatment	0.044*** (2.61)		0.041** (2.51)
Attendance in First Year (Standardized)		0.152*** (8.39)	0.151*** (8.43)
Attendance in First Year \times Treatment		-0.133*** (-7.24)	-0.130*** (-7.21)
Observations	1275	1275	1275
Adjusted R^2	0.417	0.485	0.490

Notes:

1. Courses where attendance was voluntary for students scoring above 7 in first year.
2. Regressions include course-cohort fixed effects, a polynomial in first-year grade, its interaction with the treatment, distance to university, age, gender, and European Economic Area.
3. Distance and attendance in first year are standardized, where the standard deviations are 30.9 kilometers for distance and 0.102 for attendance (on a scale from 0 to 1).
4. Bandwidth is 0.5.
5. t statistics in parentheses, standard errors are clustered on the student level.
6. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

We operationalize γ_{1ic} via treatment interactions with proxies for the price of and additional utility from attendance. Our price proxy is distance to the university. Distant students pay a higher price for attendance because they have to spend more time travelling to campus. Our proxy for the additional utility of attendance is students' average tutorial attendance in first year. Students with a high propensity to attend in first year presumably derive additional utility from attendance in second year.³⁵

Estimates are found in Table 3.8. From left to right the panel reports interaction effects for distance to the university, average attendance in the first year, and both together. Distance and first-year attendance are standardized, where the standard deviations are 30.9 kilometers for distance and 0.102 for attendance (on a scale from 0 to 1).

³⁵This proxy is implied by the assumption that preferences over tutorial attendance are stable from first to second year. Our results are consistent with the assumption.

Three patterns stand out. First, the direct effect of the characteristic is always opposite, but similar in magnitude, to the effect of its interaction with the treatment. This suggests the interactions pick up the student's counterfactual attendance had the policy not been in place. Second, Column (1) indicates the policy had a larger effect on students who live far from campus. The effect on attendance increases by 4.4 percentage points for students that live one standard deviation further from campus. This suggests distant students have a greater propensity to attend less in the absence of forced attendance. Third, Column (2) shows the policy had a smaller effect on students who have a higher attendance propensity. The effect on attendance decreases by 13 percentage points for students that attended one standard deviation more tutorials in first year. The results are fairly stable when both interactions are included together (Column (3)).

3.6.4 Less Time for Leisure

Table 3.9 uses data from course evaluations to provide more direct evidence of the effect on time use. The left panel reports the effect on an indicator for whether the student attended lectures. The right panel reports the effect on total study time (lectures+tutorials+self study).³⁶ Note that the distance control accounts for direct influences of travel time on student responses.

Columns (1) to (2) show forced students are 28 to 45 percentage points more likely to attend lectures. The estimates, while marginally insignificant, are in line with the increases in tutorial attendance. The intercept and slope in the lecture attendance regressions are similar to the intercepts and slopes in the tutorial attendance regressions. This suggests the policy forces students to pay a time cost that becomes sunk after they arrive at campus, such that lecture attendance is relatively cheap when the student is already there. The sunk cost interpretation is reinforced when analyzing the lecture attendance for courses that penalize all students for absence. The average student, forced or otherwise, will attend 90 percent of lectures and 90 percent of tutorials for this group of courses.

Columns (3) to (4) shows results for total study time. We refrain from interpreting the exact magnitudes of the estimates, as our study-time measure is discrete with bins of 5 hours. The signs suggest, however, that forced attendance increases total study time. While the statistical significance is marginal, the estimates are consistent with reduced time for other courses and leisure. Later, when we investigate spillovers across courses, we will show estimates which are in fact consistent with reduced time for leisure. Lower grades and less leisure implies students are worse off under forced attendance.

³⁶Total study time is measured in 10 categories (1=0 hours, 2=1 to 5 hours, and 10=more than 40 hours). We used the maximum for the interval to convert the categories into hours, where the category 10 is assigned 45 hours. Only the intercepts change if we use the minimum or the mean.

Table 3.9: Less Time for Leisure or Non-Academic Activities?

	Attended Lectures		Total Study Time	
	(1)	(2)	(3)	(4)
Average 1 st -year Grade is Below 7	0.282 (1.21)	0.455 (1.47)	5.170 (1.60)	8.391* (1.81)
Intercept	0.575*** (3.27)	0.435** (2.08)	8.726*** (3.04)	5.720* (1.68)
Polynomial	1 st	3 rd	1 st	3 rd
Bandwidth	0.2	0.5	0.2	0.5
Observations	89	235	89	235
Adjusted <i>R</i> ²	-0.093	0.045	0.404	0.315

Notes:

1. Courses where attendance was voluntary for students scoring above 7 in first year.
2. The dependent variable in the left panel is the answer to the question “Have you attended lectures?”. The dependent variable on the right is the answer to the question “Average study time (hours) for this course per week (lectures+tutorials+self study)?” where we used the maximum for the interval to convert the categories into hours.
3. Regressions include course-cohort fixed effects, a polynomial in first-year grade, its interaction with the treatment, distance to university, age, gender, and European Economic Area.
4. The intercepts are calculated via regressions which exclude course-cohort fixed effects and controls. They approximate the outcome mean near the threshold of students right of seven.
5. *t*-statistics in parentheses, standard errors are clustered on the student level.
6. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

3.6.5 Self-Study Time and Efficiency

Given our results, we feel that a reduction in self-study time or efficiency is the most reasonable explanation for the performance decline in attendance-voluntary courses. If the production function for grades is increasing in attendance, self-study hours, as well as self-study efficiency, and the policy increases attendance, then the only way for grades to decrease is if students spend less time on self study or become less efficient at it. If students were studying less or less efficiently, then our results would be consistent with a model where students care about their grades and leisure, where they think about attendance and self study carefully, and where the policy constrains self study indirectly via the additional constraint on attendance. Note, however, that our data does not allow us to show this explicitly.

Our argument fits well with the discussion of higher-education production in the careful time use study of Stinebrickner and Stinebrickner (2008). They show that one additional hour of study (in the first semester) causes GPA to increase by 0.36 points. Our results are consistent with this mechanism,

but also with a mechanism where there is a decline in study efficiency, and quite possibly with other mechanisms that fall outside the traditional scope of education production.

3.7 Spillovers Across Courses

The estimates for absence-penalized courses support the idea that the policy has effects outside of its direct influence on attendance. There is no attendance difference in these courses, yet there are differences in grades and passing rates. Why would this be the case? The results in Section 3.6 show peer effects and other direct impacts of the policy, such as adverse influences on self perception or identity that discourage students from doing their best, are unimportant for the effect on performance. In addition, Appendix Table A.3.15 uses our data on course evaluations to evaluate the effect of the policy on student perceptions of course attributes such as course structure and fairness. It shows no evidence of differential perceptions around the cutoff.

Another explanation is that the grade decrease in absence-penalized courses reflects negative spillovers from other courses. Being forced to spend extra time on one course could come at the cost of performance in another course situated in the same block.

Recall that the students in our setting all take the same 10 courses, in pairs of two, with one heavy (8 credit) and one light (4 credit) course in each block. While the three absence-penalized courses eliminate the disadvantage of forced students, they are adjacent to courses where forced students are disadvantaged. In Block 1 the heavy course is International Economics and the light course is Ageing and Fiscal Economics. The light course, Ageing and Fiscal Economics, penalizes absenteeism indiscriminately, whereas the heavy course, International Economics, encourages but does not force the attendance of above-7 students. The same scenario plays out in Block 2, but with Finance I as the heavy course, and Applied Statistics II as the light course. In Block 4 the scenario is slightly different. The heavy course is Methods and Techniques and penalizes absenteeism indiscriminately, where the light course is Behavioral Economics in which above-7 attendance is voluntary.³⁷ The extra time forced students spend on adjacent courses (International Economics, Finance I, and Behavioral Economics) should eat into the time they have for absence-penalized courses. This would provide an explanation for the small grade decrease observed in absence-penalized courses.

We look for further evidence of spillovers of this sort. To do this we investigate how self-reported total study hours differs depending on the course and block. If additional time on attendance-encouraged or attendance-voluntary courses crowds out self-study time for adjacent absence-penalized courses, and since absence-penalized courses have the same class attendance across the 7-threshold,

³⁷ Block 3 and 5 both contain one course without tutorials. Grades for these courses cannot be credibly analyzed as they are largely determined via group work.

then we should observe an increase in total study hours for forced students in courses where attendance is voluntary or encouraged and a decrease in total study hours in the adjacent absence-penalized courses.

Estimates are found in Table 3.10. The table organizes the estimates by block (1 to 5) and then by the course weight (light or heavy). Our hypothesis stipulates that (i) we should observe positive effects in Columns (3) and (4) for Blocks 1 and 2, as well as in Columns (1) and (2) for Block 4, and (ii) negative effects in Columns (1) and (2) for Blocks 1 and 2, as well as in Columns (3) and (4) for Block 4.

From Table 3.10 we find evidence supporting the first hypothesis. Columns (3) and (4) of Block 1 and 2 document positive effects on total study hours for forced students. While the estimates are borderline significant, the patterns support an increase in total study hours.³⁸ Columns (1) and (2) of Block 4 are somewhat inconclusive, the estimates are positive but imprecise. Support for the second hypothesis is moderate, as the increase in total study time does not seem to crowd out study hours for the adjacent courses. The estimates in Columns (1) and (2) for Blocks 1 and 2 and Columns (3) and (4) for Block 4 are statistical zeros. While the zeros do not provide strong evidence for a crowding-out of study hours in absence-penalized courses, the contrast with the positive estimates in the adjacent attendance-voluntary and attendance-encouraged courses is notable.

A couple of factors can explain the lack of support for the crowding-out of study hours. One is that our measure of total study hours is too imprecise from the perspective of identifying spillovers from other courses. Another more economically substantive factor is that crowding-out may operate along another margin such as study efficiency. Our view is that this mechanism is consistent with the fairly stable patterns observed in Columns (3) and (4) of Block 1, 2, 3, and 5 and Columns (1) and (2) of Block 4, namely that forced students spend more time on their courses where attendance was voluntary or encouraged for above-7 students. This increase in total study time might make the remaining study time in the adjacent absence-penalized course less effective.

Our analysis here warrants further comment. First, to properly quantify spillover effects it would have been useful to have adjacent courses which are identical, with credit weights of 6 and 6 rather than 8 and 4, apart from their attendance policies for above-7 students. Second, whether the decrease in absence-penalized courses is generated by spillovers or not has no bearing on our capacity to answer our research question, *i.e.* to quantify the effect of a full year of forced attendance on academic performance. Third, the estimates in Table 3.10 imply that forced students enjoy less leisure under the

³⁸To this end we can show that pooling the data for these courses yields statistically significant increases in total study time (for both bandwidths).

Table 3.10: Total Study Time Across Courses

		Total Study Time			
		Light Course		Heavy Course	
		(1)	(2)	(3)	(4)
Block 1: Absence-Penalized Light Course, Attendance-Encouraged Heavy Course	Average 1 st -year Grade is Below 7	2.945 (0.54)	-1.831 (-0.36)	5.874 (1.31)	5.916 (1.30)
	Observations	42	94	59	160
Block 2: Absence-Penalized Light Course, Attendance-Encouraged Heavy Course	Average 1 st -year Grade is Below 7	-1.368 (-0.24)	1.021 (0.20)	8.023 (1.50)	12.04* (1.96)
	Observations	50	130	48	119
Block 3: No Tutorials for Light Course, Attendance-Encouraged Heavy Course	Average 1 st -year Grade is Below 7	NA	NA	1.854 (0.28)	8.995 (1.16)
	Observations			50	121
Block 4: Attendance-Voluntary Light Course, Absence-Penalized Heavy Course	Average 1 st -year Grade is Below 7	0.832 (0.19)	5.117 (0.99)	0.169 (0.03)	4.723 (0.65)
	Observations	43	115	61	146
Block 5: No Tutorials for Light Course, Attendance-Voluntary Heavy Course	Average 1 st -year Grade is Below 7	NA	NA	10.17* (1.91)	10.93 (1.45)
	Observations			46	120
	Polynomial	1 st	3 rd	1 st	3 rd
	Bandwidth	0.2	0.5	0.2	0.5

Notes:

1. The dependent variable is the answer to the question “Average study time (hours) for this course per week (lectures+tutorials+self study)?” where we used the maximum for the interval to convert the categories into hours.
2. Attendance-Encouraged, Absence-Penalized, Attendance-Voluntary refer to how courses treated above-7 students. Below-7 students are forced in all these courses.
2. Regressions include course-cohort fixed effects, a polynomial in first-year grade, its interaction with the treatment, distance to university, age, gender, and European Economic Area.
3. Columns with an odd number use a bandwidth of 0.2 around a first-year grade of 7 and the even columns a bandwidth of 0.5. Polynomial is interacted with the treatment.
4. *t*-statistics in parentheses, standard errors are clustered on the student level.
5. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

forced attendance policy. The estimates reinforce our earlier claim that students are worse off under forced attendance.

3.8 Long-Run Performance

We investigate the effect of forced attendance in second year on performance in third year when, according to the university, tutorial attendance was once again under the purview of the student. Table 3.11 reports the effects of forced attendance on third-year grades. Columns (1) and (3) report estimates without controlling for course-cohort fixed effects, where Columns (2) and (4) include them as controls. Although we realize course-cohort fixed effects might potentially be bad controls, they are informative about why the performance of forced students is worse in third year.

Table 3.11: Performance Decline in Third Year

	Grade (Standardized)			
	Local linear regression		Third order polynomial	
	(1)	(2)	(3)	(4)
Average 1 st -year Grade is Below 7	-0.247* (-1.86)	-0.134 (-1.05)	-0.176 (-1.12)	-0.0750 (-0.52)
Observations	1869	1869	4236	4236
Adjusted R^2	0.003	0.223	0.022	0.254
Course-Cohort FE	No	Yes	No	Yes

Notes:

1. No student is required by the university to attend tutorials in third year.
2. Regressions include a polynomial in first-year grade, its interaction with the treatment, distance to university, age, gender, and European Economic Area.
3. Column (1) and (2) use a bandwidth of 0.2 around 7, whereas column (3) and (4) use a bandwidth of 0.5.
4. t -statistics in parentheses, standard errors are clustered on the student level.

5. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Students who were forced in the past have lower grades on average. Column (1) shows a decline of 0.25 standard deviations ($p < 0.1$) for students near 7. Column (3) shows the decline is 0.18 standard deviations ($p > 0.1$) if the larger bandwidth of 0.5 is used. Column (3) rejects a positive effect of 0.1 with a p -value of 8 percent.

We find evidence that performance is worse even after students regain the right to decide their attendance. Why would this be the case? One explanation relates to course choice and the grades students expect to receive. In addition to retaining decision rights over attendance, students had the right to pick their courses in third year. At the same time, they carry their second-year grades with them. The historical grades in third-year courses and their own historical performance are information they can use to select courses. The courses they select can drive down their third-year performance. The estimates in Columns (2) and (4) are consistent with this, as they show course-cohort fixed effects eliminate roughly half of the negative effect.

3.9 Conclusion

We estimate the causal effects of a full year of forced, frequent, and regular attendance on the academic performance of the above-average student at a large public university. Our estimates imply that forced students, with a first-year GPA at or around (Cerulli et al., 2017) 7, can expect a GPA de-

crease of 0.20 standard deviations over the remainder of their undergraduate degree.³⁹ The aggregate effect consists of a decrease of 0.15 standard deviations in second year, when attendance is forced, and a decrease of 0.25 standard deviations in third year, when they regain discretion over their attendance. While the negative effects on second and third year performance are marginally significant or insignificant, we are able to rule out positive effects in the ranges of 0.05 and 0.1 standard deviations.

The effects on second-year performance are moderated by the attendance policies of individual courses. The largest effects are in courses where the attendance advantage of above-7 students was greatest, where they had full discretion over their attendance. Forced attendance decreases grades in these courses by 0.35 standard deviations and the chances of passing by more than 10 percentage points. The smallest, and statistically negligible, effects are in courses where the attendance of above-7 students was strongly encouraged, suggesting the effects may depend on the degree of the attendance disadvantage of forced students. We find intermediate effects in courses that eliminated the attendance advantage of above-7 students via absenteeism penalties that applied equally to all students. We argue that these intermediate effects reflect negative spillovers from adjacent courses where forced students are disadvantaged.

Our evidence suggests forced students enjoy less leisure in second-year. We also showed that grades are lower in third-year, when all students regain the right to decide their attendance. The decrease in grades in second and third year, together with the reduction in leisure, imply that the university policy makes forced students worse off. The moderating-effects of course-level attendance policies suggests we are underestimating their loss relative to a counterfactual policy that leaves above-7 students with full discretion over attendance in all their courses.

3.A Appendix

³⁹The decrease is a weighted average of the effect on all courses in the second year (Column (1) of Table 3.4) and third year (Column (1) of Table 3.11). This point estimate is statistically significant at the 5%-level.

Table A.3.1: Overview of Program

Group	First Year Courses	Second Year Courses
A	Microeconomics	Applied Microeconomics
	Macroeconomics	International Economics
	Organisation and Strategy	History of Economic Thought
B	Financial Information Systems	Intermediate Accounting
	Marketing	Behavioral Economics
	Financial Accounting	Finance I
C	Mathematics I	Methods & Techniques
	Mathematics II	Research Project
	Applied Statistics I	Applied Statistics II
	ICT	Economics of Ageing (Eng) or Fiscal Economics (Dutch)

Notes:

1. The Economics of Ageing is taught in the English program. The Dutch program substitutes this for Fiscal Economics.
2. Students can compensate an insufficient grade (between a 4.5 and 5.4) with grades from other courses in the same group if: the other grades are sufficient (above 5.5) and the (weighted) average within the cluster is above 5.5. This applies to all students, whether they are above or below the threshold for the forced attendance policy.

Table A.3.2: No Sample Selection when Matching Grades with Attendance

	Grade (standardized)				Matched	
	(1)	(2)	(3)	(4)	(5)	(6)
Matched	-0.0511 (-0.88)	0.0219 (0.56)	0.491 (1.17)	0.628** (2.29)		
Average 1 st -year Grade is Below 7			-0.132 (-0.98)	-0.133 (-0.98)	0.00161 (1.05)	0.00112 (0.47)
Their Interaction (Matched×Treatment)			-0.0276 (-0.26)	-0.0346 (-0.49)		
Polynomial	-	-	1 st	3 rd	1 st	3 rd
Bandwidth	0.2	0.5	0.2	0.5	0.2	0.5
Observations	2298	5297	2298	5297	2298	5297
Adjusted <i>R</i> ²	-0.000	-0.000	0.168	0.211	0.994	0.984

Notes:

1. Matched is a variable which equals 1 if the grade record found a match with the attendance data and 0 otherwise.
2. Columns (1) and (2) regress second year grades on a constant and the matched-variable and shows that grades are similar for matched and nonmatched records.
3. Columns (3) and (4) show the reduced-form effect is not different between matched and nonmatched records (Matched×Treatment). The final two columns regress the matched-variable upon scoring below 7 in the first year and thereby show the policy is unable to explain whether or not a record is matched.
4. Columns (3) until (6) include course-cohort fixed effects.
5. *t*-statistics in parentheses, standard errors are clustered on the student level.
6. * *p* < 0.10, ** *p* < 0.05, *** *p* < 0.01.

Table A.3.3: Overview of Categories and Questions in Course Evaluations

Question	Measurement scale	Category
Objectives of course are clear	1-5	General
Course is relevant for my studies	1-5	General
Course is interesting	1-5	General
Course is well organized	1-5	Structure
Course material is understandable	1-5	Structure
Can be completed within allocated study points	1-5	Fairness
Time needed to complete exam is enough	1-5	Fairness
Exam reflects course content	1-5	Fairness
Exam questions are clearly defined	1-5	Fairness
Total study time (lectures+tutorials+self study)	1-10	Total study time
Have you attended lectures?	0-1	Lecture attendance
Lectures are useful	1-5	Lectures useful
Lecturer is competent	1-5	Quality lecturer(s)
Lecturer makes you enthusiastic	1-5	Quality lecturer(s)
Lecturer has good command of English	1-5	Quality lecturer(s)
Lecturer can be easily contacted	1-5	Quality lecturer(s)
Lecturer provides sufficient assistance	1-5	Quality lecturer(s)
TA gives good tutorials	1-5	Quality TA
TA can be easily contacted	1-5	Quality TA
TA provides sufficient assistance	1-5	Quality TA
TA has good command of English	1-5	Quality TA

Notes:

1. Questions are measured on a Likert scale, where 1 equals strongly disagree and 5 equals strongly agree, with the two exceptions being total study time (1 being 0 hours, 2 being [1 – 5] hours, 3 being [6 – 10] hours and 10 being ≥ 40 hours) and lecture attendance (1 being yes and 0 being no).
2. We take the mean for questions within a category, ignoring potential missing values within a category. The more sophisticated approach of calculating the principal components gives qualitatively similar results.

Table A.3.4: Negligible Effects when Forced Attendance is Abolished

Grade (Standardized)									
Courses where Attendance was Previously Forced to the Left of 7 and where to the Right									
	Attendance was Voluntary	Attendance was Encouraged	Absence was Penalized	(1)	(2)	(3)	(4)	(5)	(6)
A: Local linear regression									
Average 1 st -year Grade is Below 7	0.00815 (0.02)	-0.0355 (-0.10)	-0.210 (-0.85)	-0.299 (-1.28)	-0.0746 (-0.37)	-0.216 (-1.12)			
Observations	190	190	292	292	292	292			
Adjusted R^2	0.177	0.167	0.025	0.060	0.208	0.242			
B: Third order polynomial									
Average 1 st -year 7 Grade is Below 7	-0.121 (-0.28)	-0.141 (-0.31)	-0.403 (-1.30)	-0.428 (-1.43)	-0.0665 (-0.27)	-0.161 (-0.64)			
Observations	384	384	585	585	575	575			
Adjusted R^2	0.236	0.240	0.089	0.106	0.269	0.279			
Controls	No	Yes	No	Yes	No	Yes			

Notes:

1. Regressions include course-cohort fixed effects.
2. Controls include distance to the university, age, gender, and European Economic Area.
3. Top panel uses a bandwidth of 0.2 around a first-year grade of 7. Bottom panel uses a bandwidth of 0.5. Polynomial is interacted with the treatment.
4. t -statistics in parentheses, standard errors are clustered on the student level.
5. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.3.5: It is the Forcing that Worsens Performance

Cohort	GPA $\in [6.9 - 7.0]$	GPA $\in [7.0 - 7.1]$
2009 - 2013	6.40	$p = 0.004^{***}$
	$p = 0.126$	$p = 0.487$
2014	6.75	$p = 0.651$
		6.88

Notes:

1. Local averages of raw grades for a bandwidth of 0.1.
2. Courses where attendance was normally voluntary for students scoring above 7 in first year.
3. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.3.6: Additional Balancing Tests. Optimal Bandwidth Selected per Background Characteristic using Calonico et al. (2016).

	Distance to University	Age	Gender	European Economic Area	High School Level	High School Track	High School Grade
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A: Local linear regression							
Average 1 st -year Grade is Below 7	2.706 (0.52)	0.274 (1.50)	0.168* (1.86)	-0.0328 (-0.77)	0.0317* (1.75)	0.142 (0.99)	-0.236 (-1.24)
Observations	585	643	386	574	523	537	493
Adjusted R^2	0.002	0.003	-0.007	0.005	0.025	0.279	0.011
B: Third order polynomial							
Average 1 st -year Grade is Below 7	-6.912 (-0.64)	0.214 (0.85)	0.267** (2.20)	0.0378 (0.46)	0.00521 (0.25)	0.246 (1.02)	-0.366 (-0.84)
Observations	599	871	817	607	603	724	558
Adjusted R^2	0.010	-0.001	0.002	0.003	0.016	0.301	0.005
Mean Outcome Var.	22.978	20.289	0.290	0.938	0.979	2.501	6.882

Notes:

1. The unit of observation is the student.
2. Regressions include cohort fixed effects.
3. Top panel displays local linear regressions with an optimal bandwidth that is calculated for every background characteristic separately using MSERD of Calonico et al. (2016), where the bottom panel displays this for the third order polynomial. Polynomial is interacted with the treatment.
4. Column (5) until (7) only use the sample of Dutch students.
5. t -statistics in parentheses are robust.
6. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.3.7: Additional Balancing Tests Using Secondary School Grades as Dependent Variable.

Grade (Standardized)							
Dutch	English	Economics	General science	Civic education	History	German	Man. and org.
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A: Local linear regression							
Average 1 st -year Grade is Below 7	-0.247 (-0.96)	-0.179 (-0.67)	0.0918 (0.43)	-0.0128 (-0.05)	0.208 (0.79)	0.0788 (0.26)	0.423 (1.49)
Observations	241	234	228	232	229	187	165
Adjusted R^2	-0.001	0.004	-0.003	-0.018	0.005	-0.030	0.020
B: Third order polynomial							
Average 1 st -year Grade is Below 7	-0.324 (-0.82)	-0.190 (-0.50)	0.166 (0.53)	0.0385 (0.09)	0.424 (0.97)	0.0394 (0.09)	0.574 (1.39)
Observations	578	560	554	552	543	457	393
Adjusted R^2	-0.004	0.004	0.013	-0.007	0.011	-0.006	0.035
C: Third order polynomial on whole sample							
Average 1 st -year Grade is Below 7	-0.260 (-1.21)	-0.131 (-0.63)	-0.0388 (-0.15)	0.0312 (0.13)	0.0244 (0.10)	0.125 (0.63)	0.191 (0.87)
Observations	1143	1112	1094	1086	1074	916	809
Adjusted R^2	0.037	0.035	0.095	0.036	0.046	0.050	0.083

Notes:

1. Regressions use secondary school grades as the dependent variable.
2. Data is at the student level. Regressions include cohort fixed effects. Column (7) controls for track choice via a binary indicator.
3. Panel A displays local linear regressions with the optimal bandwidth of 0.2 around 7. Panel B shows regressions for the optimal bandwidth of 0.5 with the third order polynomial. Panel C includes all observations. Polynomial is interacted with the treatment.
4. Observations differ per column, as not all students have followed the same courses.
5. t -statistics in parentheses, standard errors are robust.
6. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.3.8: No Bunching Just Above 7. Tested through the method proposed by McCrary (2008).

	Counts of Number of Students		
	Local linear regression	Second order polynomial	Third order polynomial
	(1)	(2)	(3)
A: Binsize as suggested by McCrary (2008)			
Average 1 st -year Grade is Below 7	0.000363 (0.06)	-0.00203 (-0.39)	-0.00294 (-0.41)
Observations	20	50	50
Adjusted R^2	0.127	0.211	0.203
B: Bins two times as small			
Average 1 st -year Grade is Below 7	-0.0000178 (-0.01)	-0.00119 (-0.39)	-0.00205 (-0.50)
Observations	40	100	100
Adjusted R^2	0.000	0.088	0.078
C: Bins four times as small			
Average 1 st -year Grade is Below 7	-0.0000519 (-0.03)	-0.000632 (-0.39)	-0.00108 (-0.51)
Observations	80	200	200
Adjusted R^2	-0.009	0.032	0.026

Notes:

1. The local linear regression is estimated on the optimal bandwidth of 0.2 around a first-year grade of 7, whereas the second- and third order polynomial is estimated on the optimal bandwidth of 0.5. Polynomial is interacted with the treatment.
2. The panels refer to the different binsize as to compute the histogram for the number of students. Panel A uses the plug-in estimate of McCrary (2008), panel B and C subsequently undersmooth and compute bins two and four times as small respectively. Results are robust to the binsize.
3. t -statistics in parentheses, standard errors are robust.
4. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.3.9: Sample Selection

	Number of Courses		Completed Course Evaluation				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Average 1 st -year Grade is Below 7	0.126 (0.48)	-0.013 (-0.08)	0.039 (0.11)	-0.091 (-1.51)	-0.059 (-0.80)	-0.079 (-1.19)	-0.073 (-0.93)
Intercept	9.168*** (54.03)	9.199*** (85.74)	9.165*** (40.76)	0.204*** (4.80)	0.185*** (3.88)	0.181*** (4.17)	0.169*** (3.39)
Polynomial	1 st	1 st	3 rd	1 st	3 rd	1 st	3 rd
Bandwidth	0.2	0.5	0.5	0.2	0.5	0.2	0.5
Observations	310	717	717	2136	4901	547	1275
Adjusted R^2	0.040	0.035	0.031	0.055	0.072	-0.010	0.018

Notes:

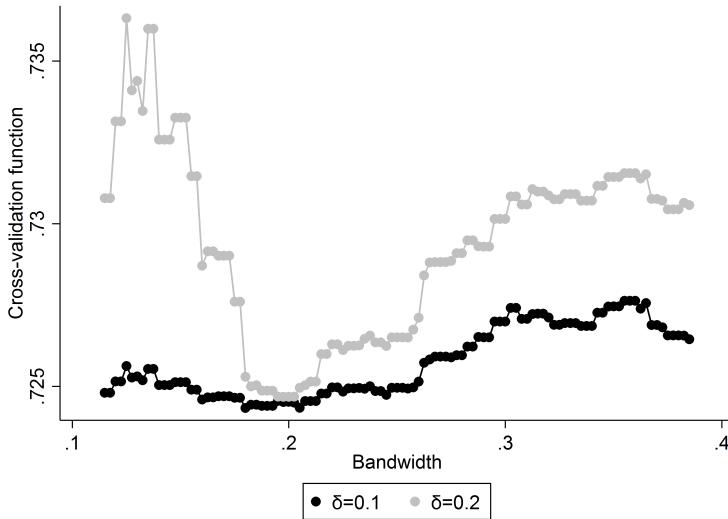
1. Columns (1) until (3) include cohort fixed effects, whereas column (4) until (7) include course-cohort fixed effects. No further controls are included. Polynomial is interacted with the treatment.

2. The intercepts are calculated via regressions which exclude course-cohort fixed effects. They approximate the outcome mean near the threshold of students right of seven.

3. t -statistics in parentheses, standard errors are robust (columns (1) until (3)) or clustered on the student level (columns (4) until (7)).

4. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure A.3.1: Selection of Optimal Bandwidth for the Local Linear Regression



Notes: We follow Imbens and Lemieux (2008) to obtain predicted grades on either side of the cutoff and use the predictions to define a cross-validation criterion for selecting the bandwidth. δ denotes the distance from the grade of the student to cutoff and appears in the criterion function. δ equal to 0.1 and 0.2 roughly correspond to 10 and 20 percent of the observations at both sides of the cutoff. For both values the criterion is minimized at a bandwidth of 0.2.

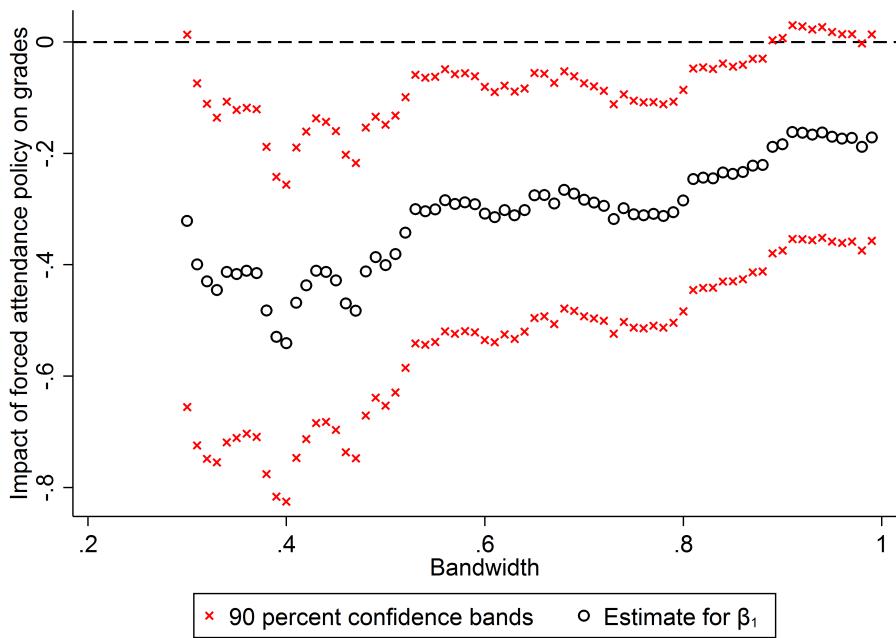
Table A.3.10: Calculations of the Optimal Bandwidth Using Methods of Calonico et al. (2016)

		First order polynomial		Third order polynomial	
		Left of 7	Right of 7	Left of 7	Right of 7
Mean	MSE rd	0.220	0.220	0.413	0.413
	MSE two	0.226	0.226	0.437	0.639
	MSE sum	0.327	0.327	0.491	0.491
	MSE comb1	0.220	0.220	0.423	0.413
	MSE comb2	0.226	0.327	0.437	0.491
Coverage	CER rd	0.139	0.139	0.248	0.248
	CER two	0.168	0.263	0.263	0.384
	CER sum	0.207	0.207	0.295	0.295
	CER comb1	0.139	0.139	0.248	0.248
	CER comb2	0.168	0.207	0.263	0.295

Notes:

1. Optimal bandwidth sizes for both the local linear regressions and the third order polynomial.
2. For the local linear regression the result corresponds with the cross-validation method depicted in Figure A.3.1, the desired bandwidth hovers around 0.2 for both MSE- and CER methods.
3. For the third order polynomial the optimal bandwidth is between 0.4 and 0.6 for the MSE methods, while being significantly smaller for the CER methods. As such, for the third order polynomial we start out with a bandwidth of 0.5, but check for robustness.

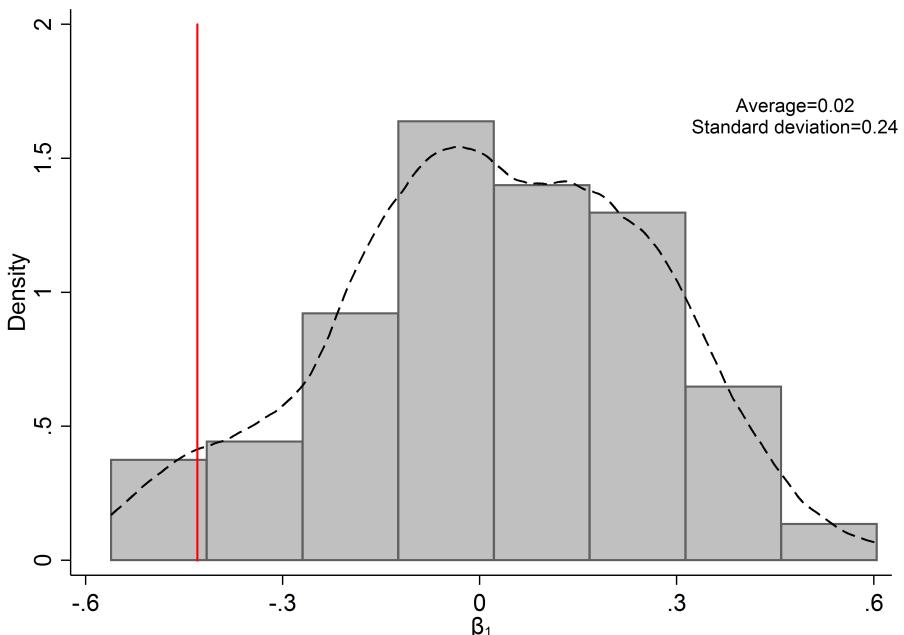
Figure A.3.2: Estimate Insensitive to Bandwidth Choice. Courses where attendance was voluntary above 7.



Notes:

1. The figure plots the estimates of the policy effect on grades for different bandwidths, against the bandwidth used to estimate the treatment effect.
2. The estimates are saddle by their confidence intervals.
3. The bandwidth ranges from 0.3 until 1.0.
4. Estimates based on specifications that control for a third order polynomial in the first year grade, its interactions with a treatment dummy at the cutoff, fixed effects for the course-cohort combination, distance to university, age, gender, and European Economic Area.

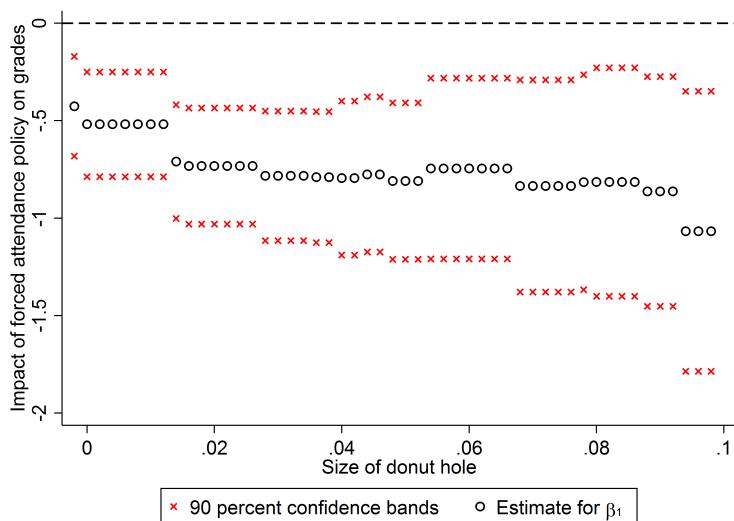
Figure A.3.3: Fake Cutoffs



Notes:

1. Histogram for the estimates of the policy on grades at cutoffs that are arbitrarily assigned by us to every 0.005-points for GPA between 6.5 and 7.5.
2. Estimates use the sample of courses where attendance was voluntary for students scoring above 7 in first year.
3. Estimates based on specifications that control for a third order polynomial in the first year grade, its interactions with a treatment dummy at the fake cutoff, fixed effects for the course-cohort combination, distance to university, age, gender, and European Economic Area.
4. Vertical red line identifies the estimate at the true cutoff of 7.
5. Bandwidth for estimation is 0.5.

Figure A.3.4: Robustness of Estimate Against a Donut Hole RD



Notes:

1. The figure plots the estimates of the policy effect on grades for different ranges of removed observations near the cutoff (the donut hole), against the size of the donut hole.
2. The estimates are saddled by their confidence intervals.
3. The donut hole ranges from 0 unto 0.1.
4. Estimates based on specifications that control for a third order polynomial in the first year grade, its interactions with a treatment dummy at the cutoff, fixed effects for the course-cohort combination, distance to university, age, gender, and European Economic Area.

Table A.3.11: Results of Local Linear Regressions. Restricting the polynomial to be similar on both sides of cutoff.

	Courses where Attendance is Forced to the Left of 7 and where to the Right							
	All Courses		Attendance is Voluntary		Attendance is Encouraged		Absence is Penalized	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A: Forced Attendance Increases Attendance								
Average 1 st -year Grade is Below 7	0.128*** (4.31)	0.120*** (4.26)	0.301*** (5.87)	0.291*** (5.86)	0.123*** (2.86)	0.108*** (2.66)	0.000893 (0.06)	-0.000203 (-0.01)
Adjusted R^2	0.305	0.316	0.366	0.375	0.154	0.175	0.155	0.181
B: Forced Attendance Decreases Grades								
Average 1 st -year Grade is Below 7	-0.154* (-1.66)	-0.150 (-1.63)	-0.357*** (-2.94)	-0.350*** (-2.92)	-0.0196 (-0.16)	0.000265 (0.00)	-0.157 (-1.28)	-0.178 (-1.44)
Observations	2136	2136	547	547	847	847	742	742
Adjusted R^2	0.165	0.166	0.178	0.176	0.202	0.200	0.096	0.099
Controls	No	Yes	No	Yes	No	Yes	No	Yes

Notes:

1. Regressions include course-cohort fixed effects.
2. Controls include distance to the university, age, gender, and European Economic Area.
3. Estimated by local linear regression with a bandwidth of 0.2 around first-year grade of 7, polynomial is restricted to be similar on both sides of the cutoff.
4. t -statistics in parentheses, standard errors are clustered on the student level.
5. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.3.12: Testing the External Validity of the RD Estimate. Using the Method of Cerulli et al. (2017).

	All Courses	Courses Where Attendance to the Right is Voluntary		
	(1)	(2)	(3)	(4)
A: Attendance (% Tutorials Attended)				
Average 1 st -year Grade is Below 7	0.118*** (4.48)	0.147*** (4.28)	0.285*** (6.21)	0.335*** (5.72)
Treatment Effect Derivative (TED)	-0.127 (-0.56)	-0.255 (-0.40)	-0.427 (-1.09)	-1.190 (-1.09)
Relative TED	4.67	1.16	3.33	0.56
P-value of: Relative TED =1				0.338
Observations	2136	4901	547	1275
Adjusted <i>R</i> ²	0.316	0.311	0.376	0.412
B: Grade (Standardized)				
Average 1 st -year Grade is Below 7	-0.139 (-1.49)	-0.154 (-1.28)	-0.342*** (-2.77)	-0.426*** (-2.74)
Treatment Effect Derivative (TED)	0.825 (0.93)	1.112 (0.49)	0.523 (0.47)	0.719 (0.24)
Relative TED	0.84	0.28	3.27	1.19
P-value of: Relative TED =1	0.892	0.272		
Observations	2136	4901	547	1275
Adjusted <i>R</i> ²	0.167	0.210	0.174	0.216
Polynomial	1 st	3 rd	1 st	3 rd
Bandwidth	0.2	0.5	0.2	0.5

Notes:

1. Regressions include course-cohort fixed effects, a polynomial in first-year grade, its interaction with the treatment, distance to the university, age, gender, and European Economic Area.
2. Columns with an odd number use a bandwidth of 0.2 around a first-year grade of 7 and the even columns a bandwidth of 0.5. Polynomial is interacted with the treatment.
3. The TED is defined as the linear term on the running variable that is interacted with the treatment variable. It measures whether the treatment effect changes while moving away from the cutoff.
4. The relative TED divides the treatment effect by the absolute TED, while multiplying the TED with the size of the bandwidth. If the absolute value is smaller than 1, it means that the treatment effect changes sign somewhere in the estimation sample considered by the bandwidth.
5. *t*- statistics in parentheses, standard errors are clustered on the student level.
6. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.3.13: Negligible Effects of Low-Achieving Peers

	Grade (Standardized)		
	(1)	(2)	(3)
Average 1 st -year Grade is Below 7	-0.401** (-2.46)	-0.423** (-2.30)	-0.430*** (-2.75)
Average 1 st -year Grade Among Peers		0.087 (0.81)	
Their Interaction (Treatment \times Peers)		0.008 (0.05)	
Average Registration Time Among Peers		0.002 (0.30)	
Its Interaction with Treatment		0.001 (0.13)	
Observations	1275	1275	1275
Adjusted R^2	0.209	0.215	0.215

Notes:

1. Courses where attendance was voluntary for students scoring above 7 in first year.
2. Column (1) includes tutorial fixed effects. The remaining regressions include course-cohort fixed effects.
3. Regressions use a third order polynomial in first-year grade, as well as their interactions with the treatment and include distance to the university, age, gender, and European Economic Area.
4. The peer group average is the leave-out mean.
5. Bandwidth is 0.5.
6. t -statistics in parentheses, standard errors are clustered on the student level.
7. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.3.14: Attendance is Useful in Some Courses, but Not Others? Evidence from the Abolition Year

	TA Quality		Lecturer Quality	
	(1)	(2)	(3)	(4)
Courses where Attendance was Voluntary (Right of 7)	0.187 (0.63)	-0.021 (-0.08)	-0.122 (-0.46)	-0.012 (-0.08)
Courses where Absence was Penalized (Right of 7)	0.135 (0.84)	0.271* (1.96)	-0.133 (-0.84)	0.024 (0.25)
Intercept	4.165*** (33.90)	4.094*** (36.23)	3.837*** (35.43)	3.826*** (49.80)
Bandwidth	0.2	0.5	0.2	0.5
Observations	94	199	89	184
Adjusted R^2	-0.011	0.015	-0.014	-0.011
P-value for Difference Between Rows 1 and 2	0.866	0.239	0.955	0.777

Notes:

1. Sample is from year when forced attendance was abolished.
2. TA and Lecturer Quality are the averages of questions which are measured on a 5-point likert scale (1 equals strongly disagree and 5 equals strongly agree). Questions include, for example, "Lecturer is competent". See Appendix Table A.3.3 for detailed definitions of the dependent variables.
3. The p -value indicates whether the course dummies are significantly different from each other.
4. t -statistics in parentheses, standard errors are clustered on the student level.
5. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.3.15: Absence of Other Channels. Using All 8 Eligible Courses.

	General	Structure	Fairness	Usefulness Lectures
	(1)	(2)	(3)	(4)
Average 1 st -year Grade is Below 7	-0.118 (-0.58)	-0.340 (-1.53)	-0.334 (-1.39)	-0.0618 (-0.13)
Constant	4.064*** (22.43)	3.963*** (27.76)	3.698*** (17.47)	3.483*** (15.28)
Observations	1003	1005	910	603
Adjusted <i>R</i> ²	0.220	0.243	0.244	0.041

Notes:

1. The dependent variables are drawn from the course evaluations using all 8 eligible courses. See Table A.3.3 for detailed definitions of the dependent variables.
2. Regressions include course-cohort fixed effects, a third order polynomial in first-year grade, its interaction with the treatment, distance to university, age, gender, and European Economic Area.
3. Bandwidth is 0.5 around first-year grade of 7.
4. The intercepts are calculated via regressions which exclude course-cohort fixed effects and controls. They approximate the outcome mean near the threshold of students right of seven.
4. *t*-statistics in parentheses, standard errors are clustered on the student level.
6. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Chapter 4

Wait and See: Gender Differences in Performance on Cognitive Tests

Joint work with Pau Balart

4.1 Introduction

Gender differences in cognitive skills have been studied by psychologists and economists for decades. On average, females outperform males in verbal and reading tasks, while males perform better than females in math and science; see, for example, Hyde and Linn (1988), Hyde et al. (1990), Caplan et al. (1997), Kimura (2004), Dee (2007), Fryer Jr and Levitt (2010), Cornwell et al. (2013), and Quinn and Cooc (2015).¹ Hyde and Mertz (2009) have debated the size of the gender gap in math, claiming that in the U.S. females have reached parity with males in math performance. A similar conclusion was reached in the meta-study by Lindberg et al. (2010), with an exception for high school-aged students.² Math and science classes at high school have been found to be important for predicting college attendance, college completion, occupational choices, and wages (Goldin et al., 2006; Joensen and Nielsen, 2009), and they have been related to the STEM gender gap.³ The results of the Programme for International Student Assessment (PISA) confirm that gender gaps during secondary school persist in many OECD countries, with males performing better in math and science and females performing better in reading.

¹The gender gap in mathematics has received special attention because of its importance for male-female differences in economic outcomes. Its root causes are highly debated. Some explanations include the role of culture (Hyde and Mertz, 2009; Sapienza et al., 2010; Nollenberger et al., 2016), larger male test score variability (Machin and Pekkarinen, 2008), and the stereotype threat (Steele, 1997; Brown and Josephs, 1999; Johns et al., 2005).

²Lindberg et al. (2010) have also shown the presence of larger gender gaps when evaluating deeper levels of mathematical knowledge and in more selective samples (the gender gap persists for white students).

³See Wang and Degol (2017) for a review on competing explanations for the origins of the STEM gender gap.

The present study provides new insights on gender gaps in the test scores of 15 to 16 year-old students participating in the PISA. In this article, we explore how gender gaps in test scores change throughout the test, finding that females are better able to sustain their performance. Extending the approach proposed by Borghans and Schils (2012), we compare the performance of males and females at the beginning of and during the test, and we do so separately for math and science questions (topics favorable to males) and reading questions (the topic favorable to females). Our main finding is that females experience a lower decline in performance during the test regardless of the topic being assessed. Countries from around the world participate in the PISA, which varies the order of test questions among test booklets and randomly allocates the booklets to students. Our results can be considered to come from a worldwide experiment: *by country*, we compare the performance of males and females on the *same test question at different positions* in the test booklet.

Of the 74 countries that participated in the PISA 2009, we first found that in 71 of them, females were better able to sustain their performance during the test than males (statistically significant for 58 countries at the 10% level and for 56 countries at the 5% level).⁴ Secondly, we found that for reading questions at the beginning of the test, the gender gap favored females in 73 countries (statistically significant in 66 countries at the 10% level and in 64 countries at the 5% level) and that during the test, this advantage became more pronounced in 68 countries (40 statistically significant at the 10% level, 37 at the 5% level). In contrast, for math and science questions there was an initial gap in test scores that favored males in 68 countries (62 statistically significant at the 10% level, 59 at the 5% level), but during the test this advantage shrank in 68 countries (46 statistically significant at the 10% level, 42 at the 5% level). Consequently, 20 countries saw the highly studied gender gap in mathematics and science completely offset or even reversed after two hours of test taking. In 42 countries females decreased their initial disadvantage in math and science by at least 50 percent by the end of the test. There was no single country or topic in which males exhibited a statistically significant lower level of decline in performance during the test than females.

These findings stood up to numerous checks. We considered different PISA waves (2006 to 2015), an increase in the number of questions to measure the gender gaps at the start of the test, alternative codification of questions that were not reached, different methods of computing performance during the test, various estimation methods, and different units of analysis. An additional set of analyses showed that testing strategies (e.g., doing the easier questions first or going back and forth between items), being stumped on difficult questions, or the presence of a short break of 5 minutes in the

⁴We use 10% as the conventional minimum level of statistical significance when displaying figures that provide information on our results (two-tailed test). We also discuss the main results at the 5% level throughout the text and report the specific *p*-values of our main estimates in the tables in the Supplementary Material.

middle of the test were not driving our results. These findings are shown in the Supplementary Material.

If the identified gender difference in participants' ability to sustain performance was driven by cognitive skills or by the stereotype threat associated with them (Steele, 1997; Brown and Josephs, 1999; Johns et al., 2005), then it should mimic the gender gaps in topics being assessed. This is in stark contrast to our findings: females are better able to sustain their performance regardless of the topic being tested. Consequently, the observed gender difference could not have been driven by cognitive skills.

We studied the relationship between the gender difference in performance during the test and well-known noncognitive skills that (i) have been shown by previous literature to be important for test scores and (ii) were measured via the PISA student questionnaires. We found that noncognitive skills, such as conscientiousness and locus of control, were unable to explain our findings. Next we drew on the PISA 2015, which was computer based and has detailed data on the number of actions performed (e.g., mouse clicks and key presses) and the time spent on each question, to investigate how these two inputs changed during the test. We found that both declined during the test, but there was no difference in the patterns between males and females. This suggests that our findings are not driven by a difference in effort, but rather by the efficacy of the mental processes that translate these inputs into a correct answer. This explanation of our findings is consistent with gender differences in boredom: males have been found to experience higher levels of boredom when performing activities that have a long duration, and individuals who experience boredom have impaired performance on various tasks (Zuckerman et al., 1978; Vodanovich and Kass, 1990a; Fisher, 1993; Vodanovich et al., 2005; Kass et al., 2010; Eastwood et al., 2012). However, our data does not allow us to empirically test this relationship.

The present study contributes to the literature on gender differences in testing behavior. Willingham and Cole (2013) found that males have an advantage on multiple choice questions. In an experimental setting, Baldiga (2013) have shown that females have a lower willingness to guess on multiple choice tests that penalize wrong answers. As the expected value of guessing is generally positive, this negatively affects females' scores.⁵ This literature concludes that evaluating knowledge with a multiple-choice test is favorable to males. Our study suggests a new implication in terms of test design: shorter tests are favorable to males and longer tests are favorable to females.

We drew on a dataset from Lindberg et al. (2010) and found empirical support for this suggestion. In their meta-analysis, Lindberg et al. (2010) amassed information on male and female performance

⁵Tannenbaum (2012), Pekkarinen (2015), and Akyol et al. (2016) obtained similar results using data on real tests, while Espinosa and Gardeazabal (2013) used data from a field experiment.

on more than 400 different math tests worldwide. We extended upon this dataset using the number of questions on the test as a measure of test length. Regressing the math gender gap on the number of questions, we found that longer tests are significantly associated with females decreasing the gender gap in math. Recently, Oxford University also provided evidence that longer tests decrease the gender gap on math and computer science exams. After extending the exam time by 15 minutes, exam administrators found that the relative performance of females increased. Besides time pressure, our results suggest that an alternative explanation could be that females are more productive during the extra 15 minutes.⁶⁷ Terrier (2016) documented that increasing the relative grades of females, due to teacher favoritism, increases the likelihood of females to select a science track in high school. Our results raise the question of whether test design, and in particular test duration, could play a role in further promoting gender equality in participation in math and sciences.

As the PISA is a low-stakes test, it is ex-ante unclear whether our findings can be extrapolated to high-stakes contexts. Considering country differences in testing cultures and motivation (Gneezy et al., 2017; Sjøberg, 2007), we conclude that females' ability to better sustain their performance could still occur in situations with higher stakes, but perhaps to a lesser degree.

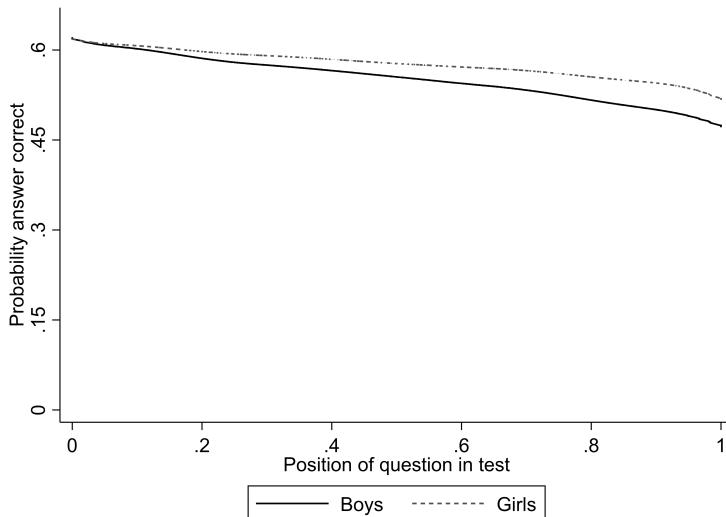
4.2 The PISA Test

The PISA is a triennial international survey that aims to evaluate the skills and knowledge of 15-year-old students worldwide in math, science, and reading. Every three years the PISA focuses on one of these topics, meaning that roughly half of the questions on the test pertain to that specific topic. For our baseline results we use microdata from the 74 countries participating in the PISA 2009. This was the most recent wave that provides a balanced distribution between topics in which females perform better (reading) and topics in which males perform better (mathematics and science). Hereafter, we briefly mention two relevant aspects of the test design for our analysis. Detailed information can be found in the Supplementary Material and the corresponding technical reports (OECD, 2009, 2012, 2014, 2015).

First, the PISA 2009 had 13 different versions of the test (booklets). Each specific question appeared in four different positions of four different booklets. The test lasted for two hours and contained approximately 60 test items. Secondly, these booklets were randomly handed out to students. The random assignment of booklets to students ensures that the variation of the position of a ques-

⁶⁷<http://www.businessinsider.com/oxford-university-gives-students-extra-time-to-finish-exams-2018-1?international=true&r=US&IR=T>

⁷Our results do not distinguish between a scenario where both the time of the test and number of questions increase and a scenario where the time of the test increases while keeping the number of questions fixed. We refer to longer tests as both having more questions and more time.

Figure 4.1: Performance throughout the test for males and females in Ireland

Notes: The figure is based on the PISA 2009 and uses LOWESS to visualize the relationship between the probability to answer a question correct and the position of the question in the test. We use a bandwidth of $0.8N$, the program default.

tion in the test is unrelated to students' characteristics or question difficulty. Balancing tests in the Supplementary Material confirm this random allocation.

4.3 Baseline Results

Figure 4.1 illustrates the main idea of our analysis. It shows the proportion of correct answers against the position of that question in the PISA test for males and females separately. Data are presented for Ireland, because in that country males and females had identical performance at the start of the test. The proportion of males and females that correctly answer the first question was equal, while a higher share of females answered the questions correctly when they were positioned later in the test. Figure 4.1 also shows that for both sexes, questions had a lower probability of being answered correctly when they occupied a position further towards the end of the test. Borghans and Schils (2012) and Torija (2012) have been the pioneers in documenting this pattern, which the former referred to as “the performance decline”.⁸ The key message of Figure 4.1, however, is that the performance decline is less strong for females.

⁸In a similar vein, Sievertsen et al. (2016) documented that test performance decreases if the test takes place later on the school day for pupils aged between 6 and 16 in Denmark.

Formally, we used ordinary least squares to explain whether a student answered a question correctly (1 denoting a correct answer and 0 an incorrect answer) with the position of the question in the test (normalized between 0 and 1 for the first and the last question, respectively). The difference in the regressions' intercept for males and females is informative of the gender gap at the beginning of the test, whereas the difference in the coefficients on the question ordering measures the gender gap in students' ability to sustain performance (the difference in the slope between males and females in Figure 4.1). Our main aim is to analyze these gender differences across topics.

4.3.1 Gender Differences

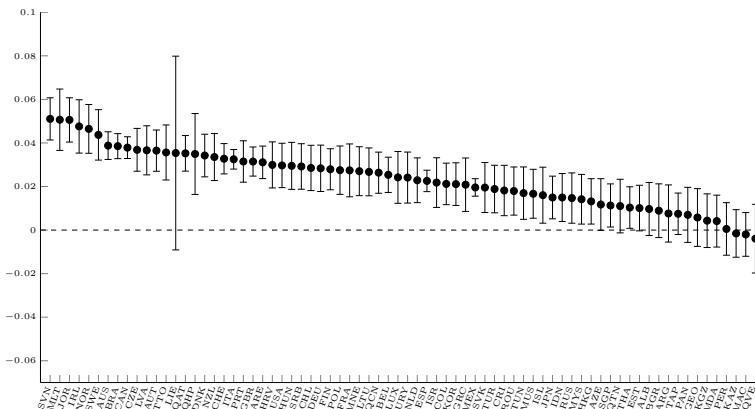
Figure 4.2 shows the first step in our study. It reports the estimated gender differences in ability to sustain performance across countries and their corresponding 90% confidence intervals. In other words, it displays the linear estimates of the gender differences in slopes –which for Ireland are shown in Figure 4.1– for each participating country. Positive values indicate countries in which females were better able to sustain their performance during the test than males. Figure 4.2 shows that this was the case for all participating countries, except for Kazakhstan, Miranda (a state in Venezuela), and Macao (China). In none of these three exceptions was the gender difference statistically significant. In contrast, the lower decline in performance experienced by females was statistically significant at the 10% level in 58 out of the 74 participating countries (in 56 if we set the significance level at 5%).⁹ To illustrate the interpretation of the results, the point estimate of 0.05 for Slovenia implies that, if we assume males and females will perform similarly on the first question of the test, the probability of answering the last question correctly is five percentage point higher for Slovenian females.

4.3.2 Gender Differences per Topic

The second step, and the main aim of our study, was to analyze gender differences in performance at the start of and during the test, both in topics favorable to males (math and science) and topics favorable to females (reading).

The estimates for reading questions are displayed in the top panel of Figure 4.3, while math and science questions are displayed in the bottom panel. We have plotted point estimates as well as the corresponding 90% confidence interval for each country. Grey lines (with squares representing point estimates) represent the female-male gap at the beginning of the test in each country. Black lines (with dots representing point estimates) represent the female-male gap in terms of ability to sustain

⁹The precise estimates for the gender differences per country and their corresponding *p*-values can be found in Table A.4.1 of the Supplementary Material.

Figure 4.2: Gender differences in performance during the test. PISA 2009

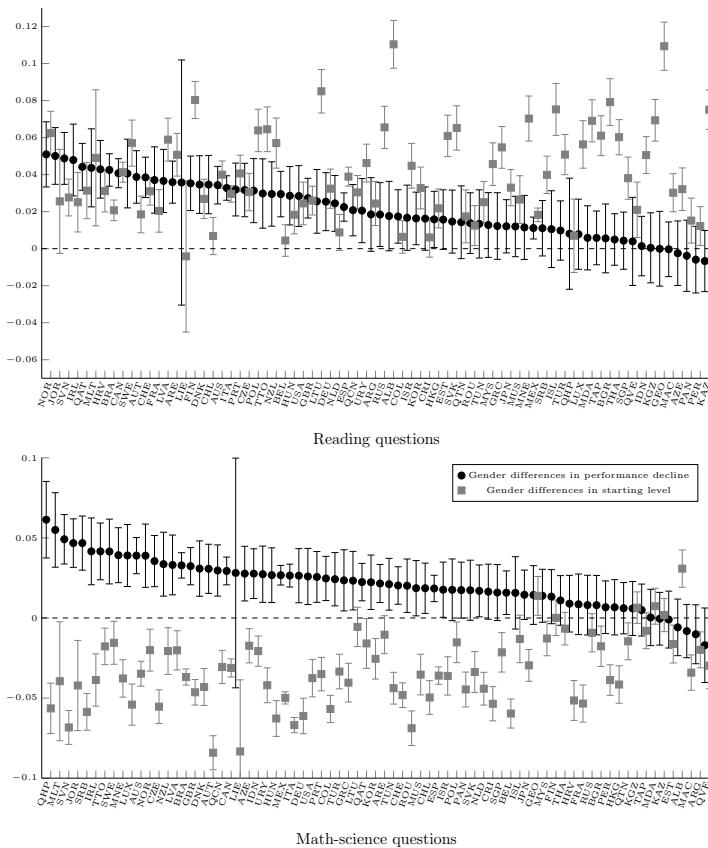
Notes: The figure plots the point estimate, together with its 90% confidence interval, of the gender difference in performance during the test for each country participating in the PISA 2009. Positive values indicate countries in which females experienced a lower decline over the course of the test than males.

performance during the test. Countries are ordered according to the size of the latter metric. Positive values indicate countries in which females showed an advantage.

When looking at the reading questions (top panel in Figure 4.3), 66 out of the 74 grey confidence intervals are strictly positive. Consistent with previous research documenting that females perform better in reading than males, we found that they outperformed males in this topic at the beginning of the test. At the same time, females were better able to sustain their performance in reading in 68 countries. The lower decline in performance for females is statistically significant at the 10% level in 40 countries (and in 37 at the 5% level). On reading questions, females performed better both at the beginning of the test and in sustaining their performance during the test.

In line with the previous literature on gender gaps in math and science, for 62 out of the 74 participating countries, the grey confidence intervals are strictly negative (bottom panel in Figure 4.3), indicating that males outperformed females in initial performance in math and science. In contrast, in most of the countries the black confidence intervals exhibit positive values, indicating that females were better at sustaining performance during the test in math and science. Point estimates have a positive value in 68 countries and are statistically significant at the 10% level in 46 of them (and in 42 at the 5% level). Despite males having an initial advantage in these topics, there was not a single country in which they were significantly better able to sustain this performance during the test.¹⁰

¹⁰The numerical estimates per country and the corresponding *p*-values can be found in Table A.4.2 of the Supplementary Material.

Figure 4.3: Gender differences in initial performance and performance during the test by topic. PISA 2009

Notes: The figure plots the point estimate, along with its 90% confidence interval, of the gender gap in initial performance and the decline in performance for each country participating in the PISA 2009. Positive values indicate the gender gap favors females.

This finding suggests that longer cognitive tests exacerbate the gender gap in reading and shrink it in math and science. In line with the literature on the gender gap in math and science, females scored lower at the beginning of the math and science test by a statistically significant degree in 62 countries. According to our estimates, however, this gender gap was completely offset or even reversed in 20 countries after two hours of test taking. In 42 countries females decreased their initial disadvantage by at least 50 percent at the end of the test. Table A.4.5 of the Supplementary Material provides a country-by-country overview of the point in the test at which females closed the gender gap in math and science.

4.4 Potential Determinants of the Gender Difference in Ability to Sustain Performance

The combination of the two graphs in Figure 4.3 provides evidence of the existence of a gender difference in ability to sustain performance that does not depend on knowledge of or ability in the topic being assessed. Hereafter, we explore other potential determinants.

We started out by investigating the role of well-known noncognitive skills. An increasing body of literature has reported gender differences in noncognitive skills that are favorable to females in various dimensions and are relevant for test performance.¹¹ Most importantly, agreeableness, openness (Duckworth et al., 2010), self-concept (Eklöf, 2007), locus of control, self-discipline (Borghans et al., 2008b), and conscientiousness (Heckman and Kautz, 2012) have been related to test scores.¹² The PISA student questionnaires contain several questions that are direct measures of these specific noncognitive skills.¹³ For example, the PISA 2012 contained the proposition “Sometimes I am just unlucky” as a measure for locus of control. Using various measures for these noncognitive skills, we found that they were unable to mediate the gender difference in ability to sustain performance.

Next, we repeated the analysis in Section 4.3 using data from the most recent PISA wave (2015). The PISA 2015 test was given on the computer and navigation across question units was restricted. This allowed us to investigate whether our results were driven by test taking strategies, which we define as any strategy that leads a student to answer the questions in a different order than the one proposed. The results are shown in the Supplementary Material. We found identical gender differences for this PISA wave and therefore disregarded the possibility that test taking strategies were a determinant for the gender differences in performance during the test.

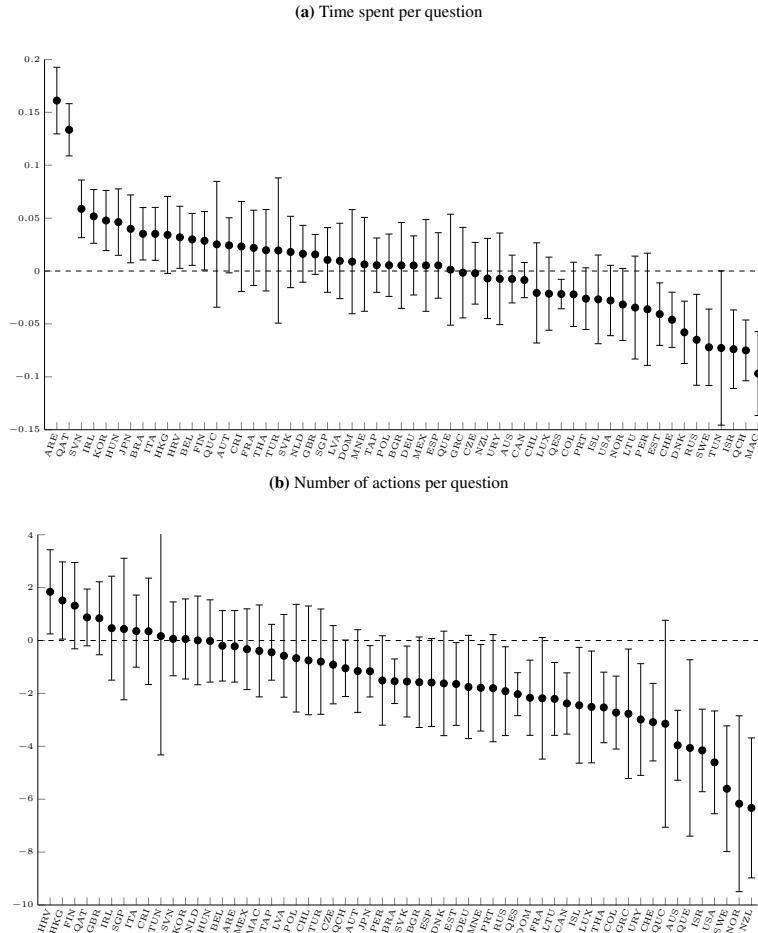
Conceptually, knowledge is a static input that remains fixed during a test. In contrast there could have been other test inputs that may have changed during the test. We refer to the latter as *dynamic inputs*. The PISA 2015 also contained information on two dynamic test inputs: time spent per ques-

¹¹Females have been found to have more self-discipline (Duckworth and Seligman, 2006), have less behavioral problems (Jacob, 2002), be less overconfident (Niederle and Vesterlund, 2007), show more developed attitudes towards learning (Cornwell et al., 2013), and report higher levels of extraversion, agreeableness, and conscientiousness (Schmitt et al., 2008; Chapman et al., 2007)

¹²The main evidence that test scores also depend upon noncognitive skills arises from (i) the finding that under low-stakes testing conditions, scores of low performers can be substantially improved by offering a reward or other forms of extrinsic motivation (Gneezy and Rustichini (2000), Duckworth et al. (2011), and Segal (2012)); (ii) the existence of a direct relationship between measures of noncognitive skills and test scores (Borghans et al. (2008b), Heckman and Kautz (2012), and Borghans et al. (2011)); and (iii) grades and achievement tests being better predictors of life outcomes than “pure” measures of intelligence (Borghans et al., 2016).

¹³The questions included are referred to by the PISA as measures for a specific noncognitive skill or are similar to the validated measures of the Big Five Inventory (John and Srivastava, 1999), the grit scale (Duckworth et al., 2007), Rotter’s locus-of-control scale (Rotter, 1966), the self-control scale (Tangney et al., 2004), and the Motivated Strategies for Learning Questionnaire (MSQL) (Duncan and McKeachie, 2005).

Figure 4.4: Gender differences in the decline of time spent per question and number of actions per question during the test. PISA 2015



Notes: The figure plots the point estimate, together with its 90% confidence interval, of the gender gap in decline of (a) time spent per question and (b) the number of actions per question for each country participating in the PISA 2015. Positive values indicate countries in which females experienced less of a decline than males.

tion and the number of actions per question.¹⁴ We applied the method used in Section 4.3 while considering these two measures as outcome variables to test whether females were better able keep up their test inputs. Figure 4.4a shows that the time spent per question during the test did not show an obvious pattern between sexes. Depending on the country, either females or males decreased the

¹⁴The actions that were counted were clicks, double-clicks, key presses, and drag/drop events. The PISA test contains open questions and its interface has tools to generate an answer, e.g., a calculator, which makes the number of actions taken a good measure of test input. For 48 out of 58 countries we found a statistically significant positive correlation between the number of actions and answering a question correctly. OECD (2015) stated that more able students generally take more time to complete the test. In analyzing these two additional inputs, we found that both the number of actions and time spent per question decreased as the test went on.

amount of time spent per question more quickly, with most of the estimates not being statistically significant. This finding also goes against the possibility that females having better time-management skills (Misra and McKean, 2000; Trueman and Hartley, 1996) could explain our result. Figure 4.4b reveals that for most of the countries the number of actions per question during the test decreased more quickly for females than for males.

In light of these results, the gender difference in ability to sustain performance does not seem to be driven by a difference in the inputs used to provide correct answers (i.e., cognitive ability in a topic, time spent on an item, and actions taken to answer an item), but rather by the efficacy of the mental process that translates these inputs into a correct answer. Although we are not able to empirically test this hypothesis with the available data, it is consistent with the existence of a gender difference that arises when considering the temporal dimension of performance: boredom. Males have been found to experience higher levels of boredom on activities with a long duration, which might cause impaired performance after some time of test taking (Zuckerman et al., 1978; Vodanovich and Kass, 1990a; Fisher, 1993; Vodanovich et al., 2005; Kass et al., 2010; Eastwood et al., 2012).¹⁵

4.5 Longer Tests and the Math Gender Gap

Our baseline finding suggests that longer tests favor females and shorter tests favor males. We tested this implication by making use of a dataset put together by Lindberg et al. (2010), who performed a meta-analysis to investigate gender differences in recent studies of mathematics performance. After extensive identification of literature on mathematics tests, they recorded the performance of males and females on 441 such tests. We expanded upon their dataset and were able to collect the number of questions for 203 of them, which we used as a proxy for test length.

Table 4.1 shows the estimates of regressing the math gender gap on a constant and the number of questions in a test. It confirms that longer tests are associated with a smaller gender gap in math.¹⁶ Column (1) suggests that males perform roughly 0.2 standard deviations better than females on shorter tests. However, females are on par with males if the test reaches 125 questions. Column (2) shows that this result is robust to excluding an extreme test with 240 questions. While these two columns directly use the data from Lindberg et al. (2010), we compiled information on the performance of males and females on the tests. In columns (3) and (4) we can see that the results are robust to our own calculation of the math gender gap and to reducing the weight to one-half for studies

¹⁵See the Supplementary Material for detailed results and further discussion on the two test inputs and noncognitive skills.

¹⁶This analysis simply compares the mean of the gender gap across tests with different numbers of questions. While it explains the low R^2 in the regressions, it is sufficient to study the existing correlation between the two.

Table 4.1: Relationship between the gender gap in math and the number of questions in a specific test

	Whole sample	Exclude outlier	Recalculated gender gap	Weighted regression
	(1)	(2)	(3)	(4)
Number of questions	-0.00159** (-2.06)	-0.00188** (-2.10)	-0.00152* (-1.97)	-0.00149* (-1.94)
Constant	0.200*** (4.59)	0.210*** (4.48)	0.194*** (4.40)	0.205*** (4.33)
<i>N</i>	203	202	203	203
Adj. <i>R</i> ²	0.012	0.015	0.011	0.010

Notes: *t* statistics in parentheses, heteroskedasticity robust standard errors

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The gender gap is measured by subtracting the mean performance of girls from the mean performance of boys and dividing this by the pooled standard deviation. The equations estimated are as follows: $mathgendergap_i = \alpha_0 + \alpha_1 numberofquestions_i + \epsilon_i$, where i is a subscript for test i .

(observations) that we coded differently than Lindberg et al. (2010). The Supplementary Material provides information on the data collection, additional robustness checks, and results when using the maximum time to complete the test as a measure for test length.

4.6 Conclusion

In this article, we present a gender difference in test performance that has been overlooked: females are better able to sustain their performance during test taking. In our preferred specification, for 20 out of the 74 participating countries, this gender difference offsets or even reverses the highly studied gender gap in mathematics and science after two hours of test taking. Our findings suggest that longer tests are favorable to females. This, in turn, raises the question of whether test design could play a role in increasing the propensity of females to take intensive math and science courses. Our study also contributes to the debate on the size and existence of the gender gap in mathematics. For instance, while some studies (Fryer Jr and Levitt, 2010) found that a gender gap in mathematics is present at the elementary school-level, others did not (Hyde et al., 1990; Lindberg et al., 2010). According to our findings, the length of a test may help to explain these differences in previous research.

PISA scores receive an enormous amount of attention from policymakers. In many countries they are considered key indicators for the design and evaluation of educational policies. These facts highlight the importance of our findings despite the low-stakes nature of the tests analyzed. However,

a natural question to ask is whether the gender difference in maintaining performance is also present in settings with higher stakes; a question we feel should drive future experimental research.

In the Supplementary Material we aim to begin to answer this question and provide three pieces of preliminary evidence suggesting that our finding is still present when higher stakes are at play. First, in the data from Lindberg et al. (2010), we found that the significant negative relationship between the math gender gap and length of the test persists even if we only consider tests with stakes. Secondly, we considered country differences in testing culture. Gneezy et al. (2017) found that test takers in Shanghai have higher intrinsic motivation than in the U.S., while Sjøberg (2007) observed that institutional promotion and motivational messages regarding international standardized tests are more prevalent in Asian countries. If higher stakes reduce gender differences when it comes to sustaining performance, we should observe less of a gender difference in Asian countries. We found this is indeed the case, but the gender difference is not entirely eliminated; in 60 percent of the Asian countries it is present and statistically significant. Considering the specific case of Shanghai studied by Gneezy et al. (2017), we found that Shanghai males significantly outperform females at the beginning of the test in math and science by more than 3 percentage points, but females significantly reduce this gender gap as the test goes on, making it negligible by the end of the test. Thirdly, with the PISA data, we constructed a measure of subjective stakes by calculating the average number of unanswered questions per country, which we expected to be high if the test were considered to have low stakes.¹⁷ By doing so, we did not find that the gender difference in ability to sustain performance throughout the test is larger in countries where the incidence of non-response is higher.

4.7 Supplementary Material

4.7.1 Data and Methodology

Data: The PISA Test

The PISA is a triennial international test administered by the OECD and aims to evaluate skills and knowledge of 15-year-old students in math, science, and reading. Every three years the PISA is focused on one of these three topics, implying that around one half of the questions in the test are from that specific topic. For our baseline results we used microdata of the 74 countries participating in the PISA 2009, for which the main topic of evaluation was reading. This provides a quite balanced distribution between topics in which females perform better (reading) and topics in which males

¹⁷The PISA does not penalize incorrect answers. This implies that not giving an answer to a question is a strictly dominated strategy for an individual interested in performing well on the test. Higher (subjective) stakes should reduce this type of careless testing behavior.

perform better (mathematics and science). Therefore, it allowed us to analyze gender differences in performance during the test per topic. We used microdata on each students' answer to every single administered question. Using the codebooks, we could retrieve which question the student had to answer on each position of the test. We also use the PISA 2006, 2012, and 2015, which focus on science, math, and science respectively.¹⁸ All four PISA waves share two main characteristics that are important for investigating the gender difference in performance during the test.

First, the PISA uses multiple versions of the test (booklets). As shown in Table A.4.3, the PISA 2009 has 20 different booklets: 13 “standard” booklets and 13 “easier” booklets, with 6 booklets belonging to both categories. Each country opts for either the set of 13 standard or 13 easier booklets, we included all of them in our analysis.¹⁹ All booklets contain four clusters of questions (test items), where the total test consists of approximately 60 test items. Each cluster of questions represents 30 minutes of testing time, which meant each student undertakes two hours of testing. Students take a short break after one hour, typically of 5 minutes. For both the set of standard and easier booklets, there are 13 clusters of test items (7 reading, 3 science, and 3 mathematics) and they are distributed over the different set of 13 booklets according to a rotation scheme. Each cluster appeared in each of the four possible positions within a booklet once (OECD, 2012). This meant that one specific test item appeared in four different positions of four different booklets.²⁰

Second, these booklets are randomly assigned to students (OECD, 2012). This random assignment ensures that the variation in question numbers, that results from the ordering of clusters, is unrelated to characteristics of students. Balancing tests for the four waves confirm this random allocation of booklets. Table A.4.4 shows results of separate regressions where background characteristics are regressed upon booklet and country dummies for the PISA 2009 (country dummies are included as only within a country the same set of booklets are being randomized). Almost all booklet dummies enter these regressions insignificant and for all regressions the F-test for joint significance of the booklet dummies does not reject the null-hypothesis.²¹ In our estimation we included question fixed effects to exploit the exogenous variation in item ordering *within* a question *across* students.

¹⁸ For the PISA 2012 the codebooks do not contain this information, the OECD provided them to us.

¹⁹ The PISA 2009 is the first wave where countries with a previous PISA score below 450 had the possibility to opt for a subset of seven easier booklets. Similar to countries with standard booklets, the 30 countries that opted for the easier set also got 13 booklets. The subset of seven easier booklets are identical to the first seven standard booklets, but each had a reading cluster replaced (3A and 4A) with an easier reading cluster (3B and 4B). This does not pose any threat to our main analysis, as we are comparing students within a country and an absolute comparison between countries is not our main interest.

²⁰ For the PISA 2015 the rotation scheme is somewhat more complicated, but the two characteristics necessary for identification remained. We will come back to this shortly when using this wave in Section 2.

²¹ Results for the PISA 2006, 2012, and 2015 are similar (joint insignificance of the booklet dummies) and are available upon request.

Two other important characteristics of the PISA are its worldwide country participation and its sampling procedure. First, international country participation allowed us to analyze whether the gender difference is systematically present across countries and to investigate the external validity of our results. Second, the PISA used a two-stage stratified sample design. The first-stage sampling units consisted of individual schools being sampled from a comprehensive national list of all PISA-eligible schools.²² The second-stage sampling units were the students. Once schools are selected, a complete list of all 15-year-old students within the school is prepared. If this list contained more than 35 students, 35 of them are randomly selected.²³ All of them are selected if this list contained less than 35 students (OECD, 2012). Although gender is exogenous by nature, due to the PISA sampling process it might be the case that males and females are not equally represented across schools with similar quality. In our estimations we controlled for the quality of the school via the inclusion of school fixed effects.

Data: Lindberg et al. (2010)

To explore the implication that longer tests favor females, we extended the database by Lindberg et al. (2010). They performed a meta-analysis to investigate the existence of gender differences in mathematics performance. This involved the identification of possible studies that investigated performance at math tests. Using computerized database searches, they generated a pool of potential articles. After careful selection, the final sample of studies included data from 441 mathematics tests (see Lindberg et al. (2010) for further details). They coded these tests according to several characteristics, the most important one being the performance of females and males. For every test they calculated the gender gap according to Cohen's d , which is equal to the difference in the mean performance of males and females divided by the pooled standard deviation ($\frac{X_{males} - X_{females}}{\sigma_p}$). It can be interpreted as a standardized gender gap. They found the gender difference in mathematics to be negligible.

For every test in their dataset we tried to collect the following information from the original articles: number of questions, maximum time allowed to complete the test, and the stakes of the exam. If this information was not available in the original studies, we sent the authors an email asking for the information. The dataset of Lindberg et al. (2010) contains 441 exams. For 243 of these tests we found evidence that they had to be completed within a certain time limit. Only these are of interest, without a limit of time there is no reason a test should measure sustained performance. Tests without a limit of time are, for example, tests that are done at home or out of academic time. For 203 of those 243 tests we were able to collect the number of questions and for 175 exams we

²²Schools are sampled with probabilities proportional to a measure of school size. Prior to sampling, schools on the national list are being assigned to explicit strata where the aim is adequate representation of the 15-year-old population.

²³For the PISA 2015 this number was equal to 42 students.

collected the maximum time allowed to complete the test. Sample attrition does not seem to be a problem as the average size of the gender gap is similar for tests with and without time limit.

Methodology

To better understand the relevance of this article's research contribution, we first discuss the relevant previous research by Torija (2012) and Borghans and Schils (2012). They showed that questions have a lower probability of being answered correctly when they occupy a position further towards the end of the test. Whereas Torija (2012) compared the performance on a single question at different positions for the PISA 2000, Borghans and Schils (2012) used the PISA 2003 and 2006 and estimated the following equation for each country separately (we refrain from using country subscripts in the whole paper):

$$y_{ij} = \alpha_0 + \alpha_1 Q_{ij} + u_{ij} \quad (4.1)$$

Where y_{ij} is a dummy for whether student i answered question j correctly and Q_{ij} is the position of question j in the version of the test answered by student i and is normalized between 0 and 1, respectively denoting the first and last question of the test. α_1 tells us whether the probability to answer a question correctly is affected by the position of that question in the test. They estimated Equation (4.1) and showed that α_1 is negative for each country, which they denoted as the “performance decline” and interpreted as a measure for noncognitive skills. The constant of Equation (4.1) represents the score of the average student at the start of the test ($Q_{ij} = 0$). Borghans and Schils (2012) showed the decline in performance is related to personality traits, mainly agreeableness and motivation towards learning, and predicts later life outcomes above and beyond the pure test score. Previous research has exploited this finding to investigate the relationship between social gender norms and gender gaps in test scores (Rodriguez-Planas and Nollenberger, 2018) and the variation between noncognitive skills and test scores (Zamarro et al., 2016).

Gender Differences

We investigate gender differences in performance during the test while also making a distinction between the topic that favors females (reading) and the topics that favor males (math and science). The PISA 2009 turns out to be an optimal test for doing this, as (i) clusters of questions varied in order between booklets, (ii) booklets are randomly handed out to students, and (iii) it had an equal division between reading questions R_j , females' favorable topic, and math-science questions N_j (non-reading), males' favorable topic (see Section 4.7.1).²⁴ Given these three conditions we propose

²⁴To create the 20 booklets it has in total 130 reading questions and 89 non-reading (math-science) questions. The number of reading questions is less than one third in the PISA waves that focused on mathematics or science.

to estimate the following two models per country:

$$y_{hij} = \underbrace{\beta_0}_{\text{Intercept}} + \underbrace{\beta_1 F_i}_{\text{Gender Dummy}} + \underbrace{\beta_2 Q_{ij}}_{\text{Topic Dummy}} + \underbrace{\beta_3 Q_{ij} F_i}_{\text{Interaction}} + J_j + H_h + \epsilon_{hij} \quad (4.2)$$

$$y_{hij} = \gamma_0^R R_j + \gamma_0^N N_j + \gamma_1^R R_j F_i + \gamma_1^N N_j F_i + \gamma_2^R R_j Q_{ij} + \gamma_2^N N_j Q_{ij} + \gamma_3^R R_j F_i Q_{ij} + \gamma_3^N N_j F_i Q_{ij} + J_j + H_h + v_{hij} \quad (4.3)$$

Where h is a subscript for the school, F_i is a gender dummy which equals 1 if student i is a female and J_j and H_h are question- and school fixed effects respectively. Focusing on Equation (4.2), our estimate of interest is β_3 , which tells us whether the performance of females develops differently during the test than those of males. Figure 4.1 in the main text indicated β_3 is bigger than zero: females are better able to sustain their performance during the test.

Gender Gaps during the Test

Equation (4.3) introduces and interacts topic dummies with the variables for the question order and the gender dummy. It has the exact same interpretation as Equation (4.2) with the coefficients separated by topic R_j and N_j .²⁵ γ_1^R and γ_1^N measure gender differences at the start of the test in reading and non-reading questions respectively, whereas γ_3^R and γ_3^N measure gender differences in the ability to sustain performance per topic. As the gender dummy takes value 1 for females, positive values of γ_1^T and γ_3^T indicate that females, respectively, have an initial advantage and a higher ability to sustain their performance in topic T .

Equation (4.3) delivers the main insights of our paper. It allows us to analyze the impact of gender differences in performance during the test on the widely studied gender gaps. In particular, to evaluate the gender gaps at the beginning, during, and end of the test, we define the following:

- **Gender gap at the start of the test** = $E[y|\text{female, start of test, topic}=T] - E[y|\text{male, start of test, topic}=T] = (\gamma_0^T + \gamma_1^T) - \gamma_0^T = \gamma_1^T$
- **Gender gap at the end of the test** = $E[y|\text{female, end of test, topic}=T] - E[y|\text{male, end of test, topic}=T] = (\gamma_0^T + \gamma_1^T + \gamma_2^T + \gamma_3^T) - (\gamma_0^T + \gamma_2^T) = \gamma_1^T + \gamma_3^T$

Estimation and Inference

Both equations include question- and school fixed effects. As described in Section 4.7.1, the order of (clusters of) questions differ between booklets and booklets are randomly handed out to students. Conditional on question fixed effects, our strategy exploits *within* question variation *across* students.

²⁵Note that Equation (4.3) does not include a constant or a separate coefficient for Q_{ij} , as the variables R_j and N_j include all questions in the PISA test.

As such, the identifying assumption becomes random variation in the position of a question across different students. This assumption is likely to hold due to the random allocation of booklets to students. Moreover, by including school fixed effects we are controlling for school quality. As the PISA first sampled schools and then randomly sampled students within schools, it might be the case that males and females are not equally represented across schools with similar quality. Imagine a country having two schools: the first has 80% males and is of high quality and the second has 80% females and is of low quality. We would find that males performed better in the PISA test. In our view this is actually a school characteristic, which we partial out by the inclusion of school fixed effects.

Our baseline results are estimated on the item level, but we also estimated Equation (4.2) on the cluster level while excluding the question fixed effects. As such, the unit of analysis exactly matches the unit of randomization: y_{hij} represents the average performance within a cluster j and Q_{ij} is the position of the cluster in the test. In our main specification we considered skipped questions as incorrectly answered and unreached questions as missing. We will perform robustness checks concerning the way we deal with unreached questions in Section 4.7.6.²⁶ We used OLS to estimate Equation (4.2) and (4.3) and checked for robustness with probit. Throughout the paper we clustered standard errors at the student level and present the results without using PISA sample weights.²⁷

4.7.2 Baseline Results for the Different PISA Waves

PISA 2009

In Figure 4.2 and 4.3 of the main article we estimated Equation (4.2) on the item level. In this section we estimate it on the cluster level. As such, the unit of analysis exactly matches the unit of randomization: y_{hij} measures the number of correct responses divided by the number of questions within a cluster and Q_{ij} represents the position of the cluster within the test $(0, \frac{1}{3}, \frac{2}{3}, 1)$. Figure A.4.1 shows that the pattern is identical to our baseline results. For 70 out of the 74 participating countries, we found females are better able to sustain their performance, where we did lose some power as for 50 countries it is statistically significant at the 10%-level. This is most likely the consequence of a reduction in both the number of observations and variation in Q_{ij} .

²⁶ Across the PISA waves males have slightly more unreached questions than females, our results are thus not explained by females spending more time on each question trying to provide an accurate answer. Section 4.7.6 confirms this.

²⁷ This immediately corrected for heteroscedasticity that arises due to the binary nature of a dependent variable. For the baseline results we also clustered standard errors on the cluster level of the booklets, the item level (the PISA 2009 contains 220 unique items), and on the school level. Significance is virtually identical for all three levels of clustering (available upon request).

²⁸ Absolute comparisons between countries are not our main interest, as such we did not use the PISA weights to adjust the estimate from the PISA sample to an estimate for the population or the 80 replicates of the weights for its corresponding standard error. Moreover, the weights no longer correctly adjust the sample to the population as our outcome variable is on the item level and not at the student level. We did verify, however, that our baseline results are unchanged when using the weights (available upon request).

Using the definitions in Section 4.7.1 we can make precise statements on the gender gaps at the beginning and end of the test. According to our baseline estimates, in math and science the gender gap favored males in 68 countries at the beginning of the test. This widely studied gender gap, however, was completely offset or even reversed in 20 countries after two hours of test taking. In other words, at the end of the test females had an equal or higher probability of answering math-science questions correctly for these 20 countries, whereas this was not the case at the beginning of the test. In 42 countries females have decreased the gender gap at the start by at least 50 percent at the end of the test. Table A.4.5 provides an overview when in the test females close the gender gap in math and science per country. For reading, the gender gap at the beginning of the test was favorable to females in 73 countries, where at the end females' advantage has increased by at least 50 percent for 39 countries.

PISA 2006 and 2012

In the main text we displayed the results for the PISA 2009. This section continues by applying a similar analysis to the PISA 2006 and 2012. Our purpose is twofold. First, we can test whether our results are robust to the use of different PISA waves. Second, the PISA 2006 focused on science and the PISA 2012 on math, which assured that the distribution of science and non-science questions for the PISA 2006 and math and non-math questions for the PISA 2012 is quite balanced. This allows us to study the gender differences for math and science separately.

Figure A.4.2 first shows the estimates for the complete test, estimated with Equation (4.2). It indicates that our previous results are present across the PISA waves. For all three waves we found that for more than 96 percent of the countries females were better able to sustain their performance than males, being statistically significant for more than 77 percent of the countries (at the 10%-level). There is not a single country for which the gender difference during the test significantly favored males. We can also see that gender differences are quite stable over time. The correlation of β_3 across the three PISA waves is around 0.45.²⁹

Figure A.4.3 separates the analysis per topic. The upper panel displays the results for science (the PISA 2006) and the lower panel displays the estimates for math (the PISA 2012).³⁰ We found that for both topics males performed better at the beginning of the test and females were better able to sustain their performance during the test. For science (math) performance during the test favored females for

²⁹Slovenia is the country for which females' advantage in performance during the test is greatest across all three PISA waves. Two other countries that exhibited large gender differences favoring females are Norway and Poland, whereas Asian countries like Kyrgyzstan, Taiwan or Macao (China) have a smaller females' advantage in performance during the test.

³⁰For the PISA 2006 we combined the math and reading questions into one non-science dummy and for PISA 2012 we combined the science and reading questions into one non-math dummy. We did not show figures for these estimates, which are available upon request.

52 (59) out of 57 (68) countries, being statistically significant for 34 (31) of them at the 10%-level. In contrast, males only provided a statistically significant smaller decline in Montenegro at the 5%-level when evaluating science questions separately. This confirms that, separately for math and science, the gender gap at the beginning of the test favored males, but this advantage was smaller, offset, or reversed at the end of the test.

PISA 2015

We continue our analysis by using the most recent PISA wave (2015). For 58 out of 73 participating countries this test was made on the computer, which provided us with additional information to investigate potential determinants in the next section. Together with the implementation of the computer-based test, the PISA introduced a few other changes to the test design.³¹ The characteristics necessary for identification remained: clusters of questions vary between the booklets and booklets are randomly assigned to students. Using the same strategy as in Section 4.3, Figure A.4.4 and A.4.5 show our baseline results from the previous waves carry over to the PISA 2015. The (seemingly) smaller estimated gender differences in performance during the test for PISA 2015 can be explained by the sample of countries that administered the computer-based test.

4.7.3 Potential Determinants of the Gender Difference in Ability to Sustain Performance

As the gender difference is unrelated to the topic being assessed on a test, we have argued that our findings cannot be driven by cognitive skills. In this section, we investigate various other potential determinants.

Well-known Noncognitive Skills

Previous literature has shown that noncognitive skills are an important determinant of test scores (Gneezy and Rustichini, 2000; Borghans et al., 2008b; Duckworth et al., 2011; Segal, 2012) and that gender differences in noncognitive skills favor females in various dimensions (Jacob, 2002; Duckworth and Seligman, 2006; Chapman et al., 2007; Niederle and Vesterlund, 2007; Schmitt et al., 2008;

³¹Three changes were implemented. First, next to science, math, and reading the PISA 2015 introduced a new domain called collaborative problem solving. As not all countries participated with this new domain and it was not represented in previous waves, we only use in our analysis the booklets that do not contain the clusters related to collaborative problem solving. Second, the PISA used 35 different booklets per country, which is substantially more than the 13 booklets in previous waves. Third, a somewhat more sophisticated rotation design made it possible for a cluster of questions to be at the same position in more than one booklet. A student was randomly assigned to one of the 35 booklets, this determined the position of the science clusters and the position and exact id of the math and reading clusters. A second random number for the student combined with his or her booklet number determined the exact id of the science clusters. See OECD (2015) for more details.

Cornwell et al., 2013). The PISA includes student questionnaires that contain several questions that serve either as direct measures or proxies for various types of noncognitive skills. We tested whether certain specific noncognitive skills could affect our findings by estimating Equation (4.2) while separately including each of these noncognitive skills (NC) and their interaction with the position of the question: $y = \beta_0 + \beta_1 F + \beta_2 Q + \beta_3 QF + \beta_4 NC + \beta_5 QNC + \epsilon$.

We considered questions that were specifically designed to measure noncognitive skills that have been shown to be important for test scores in previous studies: agreeableness, openness (Duckworth et al., 2010), self-concept (Eklöf, 2007), locus of control, self-discipline (Borghans et al., 2008b), and conscientiousness (Heckman and Kautz, 2012). The questions included are referred to by the PISA as measures for a specific noncognitive skill or are similar to the validated measures of the Big Five Inventory (John and Srivastava, 1999), the grit scale (Duckworth et al., 2007), Rotter's locus-of-control scale (Rotter, 1966), the self-control scale (Tangney et al., 2004), and the Motivated Strategies for Learning Questionnaire (MSQL) (Duncan and McKeachie, 2005). We also included several other questions that could be seen as proxies for other well-known noncognitive skills, such as neuroticism. Table A.4.6 provides an overview of the exact questions used, associated noncognitive skills, and for some questions references to similar questions on the validated scales. Two examples are "Sometimes I am just unlucky" and "I continue working on tasks until everything is perfect" to measure locus of control and conscientiousness, respectively.

Our results show that these noncognitive skills cannot parse out the gender difference in performance during the test. Across all the waves and measures we still found that females were significantly better at being able to sustain their performance in at least 70 percent of the countries considered.

Recent research proposes and validates an alternative measure for noncognitive skills; careless answering behavior in surveys (Zamarro et al., 2017, 2018; Cheng et al., 2018). Following this research, we calculated the non-response (number of unanswered questions divided by the total number of questions) in the student questionnaire that accompanies the PISA test to construct a non self-reported measure of noncognitive skills.³² Similarly, we included this measure and its interaction with the question ordering in Equation (4.2) and found that it cannot parse out the gender difference.³³

³²Cheng et al. (2018) found that careless answering patterns are associated with conscientiousness and neuroticism.

³³Similar to our baseline results, we found females were significantly better at being able to sustain their performance in 58 countries. The precise results for the analyses with all the measures of noncognitive skills are available upon request.

Test Taking Strategies

Our baseline results could be explained by gender differences in test taking strategies, which we defined as any deviation in response patterns from the actual ordering of questions. For example, females might be more inclined to first take a quick look at every question on the test and answer the ones that they think are easy. Alternatively, at the end of the test females might have a greater propensity to check their answers.

To investigate these types of explanations, we took advantage of the fact that on the computer-based PISA 2015 test, students were not allowed to go back and forth among units of questions (OECD, 2015). We should note that questions on the PISA tests are organized into units. Reading units contain 3.5 questions on average while math and science units contain on average 1.6 questions. We estimated Equation (4.2) for the PISA 2015 at the unit level, where y_{hij} represents the average performance within a unit j , Q_{ij} is the position of the unit within the test, and question fixed effects are replaced by unit fixed effects. To identify the gender difference, we only used the variation in unit ordering across students. As students could not go back and forth between units, we can be sure that the position of the unit in the test is the actual position in which the unit was answered. Figure A.4.6 shows an identical pattern to our baseline results, removing the possibility that gender differences in test taking strategies drove our results. We did, however, lose significance for six countries, which is most likely the consequence of a reduction in the number of observations and in the variation in Q_{ij} .

Dynamic Test Inputs

Conceptually, students' knowledge is fixed throughout the test. In contrast, there are other test inputs that may change during the test. We refer to the latter as *dynamic inputs*. Because of the computer-based nature of the PISA 2015, we can get information on two of these inputs: time spent per question (T_Q) and actions per question (A_Q).³⁴ We use the subscript Q to highlight that these two inputs may change depending on the position of the question in the test. Consider a production function where these inputs are used to generate correct answers (Y_Q):

$$Y_Q = \theta_Q g(C, T_Q, A_Q)$$

θ_Q is interpreted as a “total factor productivity” parameter, which we view as the efficacy of the mental process that transforms inputs into correct answers. Importantly, it can also vary in Q . This parameter may account for mental fatigue or any other element not fully captured by T_Q and A_Q . Regarding the

³⁴The number of actions in the PISA test does not simply mean filling in an item. The PISA interface contains open questions as well as a calculator which allowed us to consider the number of actions as an input.

role of the two dynamic inputs, the OECD (2015) has stated that better performing students generally take more time to complete the test, and in 48 out of 58 countries we found a statistically significant positive correlation between the number of actions and answering a question correctly. These findings indicate that, consistent with the definition of an input, the first partial derivatives of Y_Q with respect to A_Q and T_Q are positive. The derivative of Y_Q with respect to Q can be expressed as:

$$\frac{\partial Y_Q}{\partial Q} = \frac{\partial \theta_Q}{\partial Q} g(C, T_Q, A_Q) + \theta_Q \frac{\partial g(C, T_Q, A_Q)}{\partial T_Q} \frac{\partial T_Q}{\partial Q} + \theta_Q \frac{\partial g(C, T_Q, A_Q)}{\partial A_Q} \frac{\partial A_Q}{\partial Q}$$

One explanation for the gender difference in performance decline is that females might be better able to keep up their levels of dynamic inputs during the test.³⁵ This would be the case if $\frac{\partial T_Q}{\partial Q}$ and/or $\frac{\partial A_Q}{\partial Q}$ were greater for females than for males. We tested this possibility by estimating Equation (4.2), replacing the outcome variable with T_Q and A_Q . The former is measured in seconds, while the latter is a composite measure of the number of clicks, double-clicks, key presses, and drag/drop events.

We found that the number of actions and time spent per question also declined during the test. On average, students used fewer actions and spent less time per question at the end of the test than at the beginning. This finding is consistent with the existence of a decline in performance. However, can the dynamic inputs explain the gender difference? Figure 4.4a shows that the time spent per question during the test did not show an obvious pattern by gender. Depending on the country, either females or males decrease the amount of time spent per question more quickly, but the estimate is mostly insignificant. This finding also rejects the possibility that females having better time management skills (Misra and McKean, 2000; Trueman and Hartley, 1996) could explain our baseline result. Figure 4.4b reveals that for most countries the number of actions per question during the test dropped faster for females. The pattern was not as strong as our baseline result; we found this in 45 out of 58 countries, being statistically significant for 24 of them at the 10% level.³⁶

We conclude that dynamic inputs cannot explain the gender difference in performance during the test. This is confirmed by augmenting Equation (4.2) with the two measures and estimating $y = \beta_0 + \beta_1 F + \beta_2 Q + \beta_3 QF + \beta_4 T + \beta_5 QT + \beta_6 A + \beta_7 QA + \epsilon$. By doing this, we still found that females were better able to sustain their performance during the test; see Figure A.4.7. One possible interpretation of our findings can be made in terms of effort. The dynamic inputs are interpreted as proxies of observable effort. As shown above, they declined as the test went on but were not capable of capturing gender differences in ability to sustain performance. Thus, we attribute this gender difference to θ_Q . To make this explicit within our framework, we estimated a linear

³⁵By definition, knowledge (C) is constant during the test ($\frac{\partial C}{\partial Q} = 0$).

³⁶For two countries we found that males decreased their number of actions significantly faster during the test at the 10% level.

approximation of the relationship between the gender differences in ability to sustain performance and gender differences in dynamic inputs during the test across countries (c):

$$\underbrace{\frac{\partial Y_Q}{\partial Q} \Big|_f - \frac{\partial Y_Q}{\partial Q} \Big|_m}_{\beta_{3c}^Y} = \underbrace{\left[\frac{\partial \theta_Q}{\partial Q} \Big|_f - \frac{\partial \theta_Q}{\partial Q} \Big|_m \right]}_{\delta_0} + \underbrace{\left[\frac{\partial T_Q}{\partial Q} \Big|_f - \frac{\partial T_Q}{\partial Q} \Big|_m \right]}_{\delta_1 \widehat{\beta_{3c}^T}} + \underbrace{\left[\frac{\partial A_Q}{\partial Q} \Big|_f - \frac{\partial A_Q}{\partial Q} \Big|_m \right]}_{\delta_2 \widehat{\beta_{3c}^A}} + \epsilon_c$$

The intercept of this regression captures the gender differences in performance during the test that could not be explained by the dynamic inputs, but might be related to differences in total factor productivity ($\frac{\partial \theta_Q}{\partial Q}$). A positive intercept is consistent with females being better able to transform inputs into correct answers as the test goes on. Figure A.4.8 displays these two regressions visually and clearly shows that the intercept is positive for both dynamic inputs. Also note that, as expected, the gender differences in performance decline and dynamic inputs during the test show a significant positive relationship.³⁷ Table A.4.7 shows the estimates of the corresponding regressions and confirms the visual results, where columns (1), (3), and (5) display the results for time, number of actions, and both inputs combined. Columns (2), (4), and (6) include an interaction between the gender difference in dynamic inputs during the test and at the start of the test, which controls for the notion that a drop in dynamic inputs might have a larger effect if the starting level of dynamic inputs were lower.³⁸ Consistent with diminishing marginal returns to these dynamic inputs, we found this interaction was negative: females' ability to better keep up their inputs during the test leads to a smaller gender difference in performance during the test for females that have a higher baseline level of dynamic inputs. However, the coefficients on the interaction terms are insignificant and do not change the magnitude or significance of the positive intercept.

We view θ_Q as the efficacy of the mental process that translates test inputs into answers. What does this mental process entail? As it cannot be observed in our dataset, we cannot provide a conclusive answer to this question. However, we speculate that our finding is related to the literature that has documented a gender difference that arises when considering the temporal dimension in performance, i.e., boredom.

Previous research has found that females experience lower levels of boredom when performing activities with a long duration (Vodanovich et al., 2005; Vodanovich and Kass, 1990a; Zuckerman et al., 1978). Eastwood et al. (2012) argued that a definition for boredom could be given in terms

³⁷ Countries for which the decline in the number of actions and time spent per question is stronger for females show a smaller gender difference in ability to sustain performance and vice versa.

³⁸ This notion of nonlinearity is captured by the conceptual framework as the drop in dynamic inputs ($\frac{\partial T_Q}{\partial Q}$) is multiplied by the change in the production function ($\frac{\partial g(\cdot)}{\partial T_Q}$).

of *attention*, as performance on sustained attention tasks (so-called vigilance tasks) associated with common measures of boredom.³⁹ Individuals who experience boredom have impaired performance on various tasks (Eastwood et al., 2012; Kass et al., 2010; Fisher, 1993). This literature argues that the response to boredom is different between people who seek external stimulation (agitated boredom) versus internal stimulation (apathetic boredom) (Malkovsky et al., 2012; Vodanovich and Kass, 1990b). Our results fit well with an agitated type of boredom, where a common response is to force oneself to pay attention to the task at hand (Malkovsky et al., 2012; Eastwood et al., 2012; Harris, 2000; Fisher, 1993). However, our data does not allow us to provide conclusive evidence in favor of this hypothesis or to rule out other competing explanations.

Being Stumped

Alternatively, one might think that the ability to sustain performance is about dealing with difficult questions during the test. Students might get demotivated by certain questions, causing them to perform poorly thereafter. This might eliminate the gender difference if males suffered more greatly from this phenomenon.⁴⁰ Measuring whether a student got stumped on a question is not easy, but by means of the PISA 2015 we can conceptualize it as the question in which a student put forth more effort (the maximum number of actions) while answering it wrong.⁴¹ We re-estimated Equation (4.2) while including a dummy S , which equaled 1 after such a question, and interacted it with Q : $y = \beta_0 + \beta_1 F + \beta_2 Q + \beta_3 QF + \beta_4 S + \beta_5 QS + \epsilon$. We found our results to be unchanged (Figure A.4.9).

4.7.4 Longer Tests and the Math Gender Gap

Our results in the main text reveal a negative association between the math gender gap and test length via the use of an extended version of the dataset from Lindberg et al. (2010). In this section we investigate the robustness of this finding.

Column (1) of Table A.4.8 displays the original estimates of Table 4.1. The constant reveals that on very short tests, males perform 0.2 standard deviations better than females, but females fully close this gap on a test with 125 questions. Table 4.1 has already shown three checks to confirm the robustness of this pattern. First, we collected information on the gender gap of the tests, where the results did not change if we consider the gender gap we calculated. Moreover, we excluded one

³⁹More specifically, the definition has two components: (i) not being able to successfully pay the attention required to participate in a satisfying activity and (ii) being aware of this, resulting in either an attempt to engage with the task at hand or awareness of engagement in matters unrelated to the task.

⁴⁰Buser and Yuan (2016) found that females were more likely to stop competing if they lost, suggesting that females would suffer more from being stumped.

⁴¹We also used our data from the PISA 2009 and measured being stumped as the first question that students were expected to answer but did not give an actual answer to (skipped questions). The student is expected to have read these questions but not to have answered. Results are identical (available upon request).

extreme test with 240 questions⁴² and gave a weight of one-half to studies that we coded differently than Lindberg et al. (2010). This did not change our results.

The statistically significant negative association notwithstanding, there does appear to be a lot of noise in the math gender gap as displayed by the low adjusted R^2 . A substantial part of the tests in our sample contained few questions; they were the ones that also introduced a large part of the unexplained variance. In columns (2) and (3) of Table A.4.8 we trimmed our sample and excluded the exams with fewer than 10 and 40 questions, respectively. The fit increased while excluding shorter tests. Possible explanations for this are that short tests are subject to more noise (when measuring ability) or that the number of questions is a worse proxy for test length when there are fewer questions.

An alternative measure for the length of the test is the maximum time allowed to complete the exam. We preferred to use the number of questions to measure test length for three reasons. First, in our experience (while collecting the data), the maximum time often did not correspond to the actual length of the test. A short test in practice might have a long maximum time limit.⁴³ Secondly, information about the number of questions was available for more tests than the maximum time limit. Thirdly, this measure coincides with our analysis and the PISA data where we used the position of the question in the test to explain whether a question was answered correctly.

Nevertheless, we performed the basic analysis again and regressed the math gender gap on a constant and the maximum time allowed to complete the test in column (4) of Table A.4.8. The maximum time was also negatively associated with the gender gap, but insignificantly so (p -value = 0.30). We redid the analysis in Table 4.1 and used our own recalculated gender gap and gave a weight of one-half to tests that we coded differently (columns (5) and (6)). This did not change our results. Next, we trimmed our sample for the same reasons as with the number of questions: to exclude extremely long tests and to remove short tests that might include more noise concerning the measured gender gap. Column (7) removes five extreme tests that took longer than 170 minutes and reports a significant negative estimate for the coefficient of maximum time allowed to complete the test at the 1% level.⁴⁴ Columns (8) and (9) also exclude tests with a time limit shorter than 5 and 20 minutes, respectively, and also find significant negative effects.

Next, we investigated whether the relationship between the math gender gap and the length of the test was related to gender differences in performance during the test that were found with the PISA. The previous regressions did not exploit exogenous variation in the length of tests. A competing explanation might be that long and short tests simply differ in other characteristics that correlate with

⁴²The second longest test contained 135 questions.

⁴³For example, one article (Murphy and Ross, 1990) noted that on a short exam of 8 questions, students could take a maximum of 50 minutes.

⁴⁴The sixth longest test had a maximum time of 135 minutes.

the gender gap as well. Columns (2) and (3) of Table A.4.9 provide evidence against this competing explanation. By using information on the world region in which the test was given, the columns split the sample into world regions for which the relationship between the math gender gap and number of questions is strongly present (Europe, Australia, and the Middle East) and for which it is not present at all (Asia). Column (1) shows the estimate for the whole sample as a comparison. If gender differences in performance during the test that were found with the PISA drove this relationship, we expect that differences between males and females would be larger in Europe, Australia, and the Middle East than in Asia. Using our baseline results, we observe that the gender difference is indeed two times as small in Asian countries. A regression of the size of the gender difference (β_3) on a dummy that equals 0 if the country is Asian and 1 if its European, Australia, or in the Middle East reveals a significant positive estimate with a t-statistic of 4.00 (robust standard errors).

4.7.5 Low Stakes versus High Stakes

Many studies have found the existence of gender differences in performance when under pressure and in competitive environments (Croson and Gneezy, 2009; Gneezy et al., 2003; Gneezy and Rustichini, 2004). Azmat et al. (2016) and Iribarri and Rey-Biel (2018) showed that females perform relatively worse as the stakes on a test increase. Ors et al. (2013) showed that males get better test scores when they are competing for college seats than what would be predicted by their previous grades, while the opposite is true for females. In contrast to this latter branch of literature, in our setting the test takers did not face competition or pressure. In fact, final scores of the PISA test are not communicated to the test takers.

Is the low stakes nature of the PISA test responsible for the observed gender difference in ability to sustain performance? Given the discussion on the possible determinants in Section 4.7.3, one might expect it to be smaller (or even absent) in a high-stakes context. If this were true, our results might provide an additional explanation to the findings of Azmat et al. (2016) and Iribarri and Rey-Biel (2018), as they found that females perform relatively better on low-stakes tests than on high-stakes ones. In this section we investigate the possible influence of stakes in our results.

First, we tested whether the relationship between the math gender gap and length of the test was also present on tests with stakes. To do so we coded whether tests included in Lindberg et al. (2010) had any stakes. While this information was unavailable for 90 studies, column (4) of Table A.4.9 shows that the same negative relationship is found when restricting the regression to tests with stakes.

Secondly, we took advantage of country differences in testing culture. Gneezy et al. (2017) found that students in Shanghai have higher test motivation than U.S. students. They showed that, in response to financial incentives, performance among Shanghai students did not change, while the test

scores of U.S. students increased substantially. According to Sjøberg (2007) test takers in Singapore saw the PISA test as a relatively high-stakes test compared to, for example, European countries. If the stronger test taking culture found in these two articles extends to other Asian countries, it could explain the lower gender difference that we found in Asian countries. Then, higher stakes may reduce the gender differences in ability to sustain performance throughout the test. Note, however, that for 60 percent of the Asian countries, the gender difference in performance during the test is present and statistically significant. Relating our results to Sjøberg (2007), in both PISA waves in which Singapore participated (2009 and 2012) we found a significantly lower decline for females at the 10% level. With respect to Gneezy et al. (2017), the PISA 2009 only sampled Chinese test takers from Shanghai. Table A.4.5 shows that in Shanghai, males significantly outperform females at the beginning of the test in math and science by more than 3 percentage points, but females significantly reduce the gender gap as the test goes on. The gap is exactly offset at the end of the test. To sum up, by considering cross country differences in testing cultures, we conclude that the gender difference in ability to sustain performance might be lower but not absent in the presence of stakes.

Ultimately, we would like to have a measure of motivation in the PISA test per country and study its association with the size of the gender difference in ability to sustain performance. To construct such a measure, we used the average number of unanswered questions per student as a measure of test motivation. The idea being that, as the PISA test has no penalty for incorrect answers, not giving an answer to a question is a strictly dominated strategy. We expect that this type of careless testing behavior would occur less often if the perceived stakes were high. We regressed the size of the gender difference in ability to sustain performance on our measure for the stakes of the PISA test. By doing so, we did not find that the gender difference is larger in countries where the incidence of non-response is higher. To the contrary, we found that countries with a low non-response rate (i.e. high subjective stakes) had a somewhat larger gender difference in their ability to sustain performance. Similar to before, this result suggests that the gender difference in sustaining performance throughout the test is not absent in a high-stakes context.

4.7.6 Robustness

This section analyzes the robustness of our baseline results. In particular, we consider nonlinearity in three different ways, analyze unreach questions, use a different definition for the performance at the start of the test, and analyze the potential effects of the small break halfway during the PISA test. Unless noted otherwise, we will use the PISA 2009 throughout this section.

Probit

Our main estimates were computed using OLS. As the dependent variable is binary, we replicated our baseline results making use of a probit model. The results for the complete test and per topic are shown in Figure A.4.10 and A.4.11 respectively, which are very similar to the ones obtained by OLS. Despite the sign of these coefficients being sufficient to inform us about the direction of the marginal effects, it does not necessarily do so about its statistical significance. For the complete test we also tested the significance of the marginal effects by using a Welsch t-test, where the results are virtually identical to directly testing the coefficients, see Table A.4.10. We provide technical details on the Welsch t-test below.

Technical Details on the Welsch t-test

We used the following procedure to test for differences in marginal effects. After estimating the probit model, we took the derivative of Equation (4.2) with respect to Q : $f(\cdot)(\beta_2 + \beta_3 F)$ (suppressing subscripts). Subsequently we evaluated this expression for males at $F = 0$, $Q = 0.5$, and $Q * F = 0$ and for females at $F = 1$, $Q = 0.5$, and $Q * F = 0.5$. As such, we had a value for the average marginal effects of males ($\frac{1}{N} \sum_{n=1}^N f(\cdot)|_{males} \beta_2$) and females ($\frac{1}{N} \sum_{n=1}^N f(\cdot)|_{females} (\beta_2 + \beta_3)$). In practice, these marginal effects are tested through standard z -tests. Therefore, we performed a simple Welsch t-test on the significant difference of them. More specifically, we applied the Welsch t-test as follows (omitting the summations):

$$\frac{f(\cdot)|_{males} \beta_2 - (f(\cdot)|_{females} \beta_2 + f(\cdot)|_{females} \beta_3)}{\sqrt{Var[f(\cdot)|_{males} \beta_2] + Var[f(\cdot)|_{females} \beta_2] + Var[f(\cdot)|_{females} \beta_3] + 2Cov[f(\cdot)|_{females} \beta_2, f(\cdot)|_{females} \beta_3]}} \sim t_k \quad (4.4)$$

Where k are the degrees of freedom of a t-distribution using the Satterthwaite approximation. Note that this procedure does not explicitly has two independent samples. Alternatively, we could repeat this procedure and estimate separate probit models for males and females and compare their marginal effects through the Welsch t-test. However, the disadvantage of this is that the marginal effects do not match the coefficients we present in the paper. Therefore, we opted for the procedure above. See Table A.4.10 for the results.

Nonlinear in Q

The models in Equation (4.2) and (4.3) assume a linear relationship between the answer to a question and the position of the question in the test. This is a rather strong assumption, as estimating $y = \beta_0 + \beta_1 Q + \beta_2 Q^2 + \epsilon$ does show the presence of nonlinear effects.⁴⁵

Despite the deviations from linear appear to be small and not homogeneous across countries, we also tested whether allowing for nonlinear effects has consequences for the gender difference.⁴⁶ First, estimating Equation (4.2) while adding a quadratic term, $y = \beta_0 + \beta_1 F + \beta_2 Q + \beta_3 QF + \beta_4 Q^2 + \epsilon$, gave us identical results to Section 4.3.1. Second, we estimated Equation (4.2) while including an interaction between the quadratic term and the female dummy: $y = \beta_0 + \beta_1 F + \beta_2 Q + \beta_3 QF + \beta_4 Q^2 + \beta_5 Q^2 F + \epsilon$. The marginal effect of Q in this case equals $\beta_2 + \beta_3 F + 2\beta_4 Q + 2\beta_5 QF$. As such, the relevant test for the gender difference in performance during the test becomes $\beta_3 + 2\beta_5 Q \neq 0$. The difference between males and females depends on the position of the question in the test. We tested for the presence of a gender difference at every possible value of Q , which also provided insights in the distribution of the gender difference throughout the test. Note that for 36 countries the estimate for β_5 was significantly negative at the 10%-level, which implies that for some countries the gender difference in performance during the test decreases as the test goes on.⁴⁷

Figure A.4.14 graphs the number of countries for which the gender difference ($\beta_3 + 2\beta_5 Q$) is significantly different from zero at the 10%-level at each position of the test. The black bars indicate females were better able to sustain their performance, whereas the grey bars do this for males. Until halfway the test ($Q = 0.5$) there is strong support that females were better able to sustain their performance. Thereafter the gender difference decreases, but up until the end of the test there are more countries for which females were better able to sustain their performance than males (18 versus 12). We found strong evidence for the gender difference, but the size seems to decrease towards the end of the test.

⁴⁵The results for the linear coefficient in Q (β_1) showed the estimate is significantly negative for 63 of the 74 countries at the 10%-level. In zero cases it was significantly positive. The quadratic estimate in Q (β_2) was significantly negative (positive) for 58 (3) countries at the 10%-level. As such, for most countries the decline in performance increased during the test. Note that for all 74 participating countries we found either a significant negative estimate for β_1 or for β_2 .

⁴⁶Figure A.4.12 shows the fitted values for a linear and quadratic estimate of the decline in performance for the median country in terms of the nonlinear effect size (Italy). The linear line seems to approximate the quadratic line relatively well. Figure A.4.13 visualizes that the exact shape of the decline differs per country, by showing the fitted values of the quadratic performance decline for the five countries with the most extreme nonlinear shapes.

⁴⁷For the other 38 countries we could not reject the null hypothesis of $\beta_5 = 0$ at the 10%-level. When we did not reject the null hypothesis of $\beta_3 = 0$ or $\beta_5 = 0$ we did set the estimate equal to zero while calculating the marginal effects.

A Relative Measure

One might argue that a relative version of the decline in performance is a more comprehensive measure. Imagine a simple version of Equation (4.1) represented by $y = \alpha + \beta Q$. Test takers with gender A score 1 at the beginning of the test and $\frac{1}{2}$ at the end of the test, where test takers with gender B score, respectively, $\frac{1}{2}$ and $\frac{1}{4}$ at the beginning and end of the test. The linear equation representing the probability of a correct answer is $y = 1 - \frac{1}{2}Q$ for gender A and $y = \frac{1}{2} - \frac{1}{4}Q$ for gender B , where the decline in performance is $-\frac{1}{2}$ for A and $-\frac{1}{4}$ for B . However, as for both sexes the score at the end of the test is half the score at the beginning of the test, one might prefer a measure that shows a similar decline in performance. In other words, as gender A started off at a higher level compared to gender B , it is also allowed to have a larger absolute deterioration in performance during the test.⁴⁸

Note that such an alternative relative measure does not have qualitative consequences for our results, females' higher ability to sustain their performance is unrelated to whether they score better or worse at the beginning of the test. However, this relative measure might capture why the gender difference in performance during the test was more significant in math and science compared to reading.

A relative measure can be obtained by computing the ratio between the slope and the constant. Note that by implementing this correction, the above example would exhibit the same decline for A and B , that is: $\frac{\beta_A}{\alpha_A} = \frac{\beta_B}{\alpha_B} = -\frac{1}{2}$. The proposed correction for the complete test can be analyzed by the following nonlinear Wald test:

$$H_0 : \frac{\beta_2}{\beta_0} = \frac{\beta_2 + \beta_3}{\beta_0 + \beta_1}$$

$$H_1 : \frac{\beta_2}{\beta_0} \neq \frac{\beta_2 + \beta_3}{\beta_0 + \beta_1}$$

Like in our baseline results, there are only three countries in which the relative decline in performance was smaller for males and in none of these countries the difference is statistically significant.⁴⁹ Moreover, the number of countries for which females were significantly better able to sustain their performance notably increases from the 58 found in our baseline results to 67. This relative measure reinforces our baseline results.

When analyzing gender differences in performance during the test per topic T we implement the following test:

$$H_0 : \frac{\gamma_2^T}{\gamma_0^T} = \frac{\gamma_2^T + \gamma_3^T}{\gamma_0^T + \gamma_1^T}$$

$$H_1 : \frac{\gamma_2^T}{\gamma_0^T} \neq \frac{\gamma_2^T + \gamma_3^T}{\gamma_0^T + \gamma_1^T}$$

⁴⁸A usual way of dealing with this type of concern consists of taking the logarithm of the dependent variable and interpreting the coefficients as a rate rather than as a slope (semi-elasticity). This is not possible in our setup given the presence of zeros in the dependent variable.

⁴⁹The precise results can be found in Table A.4.11.

Using this approach, there are 72 out of 74 countries for which females were better able to sustain their performance in reading, where the statistical significance increases from 40 countries in Section 4.3.2 to 61. This result follows from the fact that in most countries females experienced a higher performance at the start of the test in reading.

In the case of math and science questions males started off from a higher level. Despite this, the results are very similar to the ones obtained in Section 4.3.2. The number of countries in which females were significantly better able to sustain their performance is 39, compared to 46 in our baseline results. There is one additional country for which the performance during the test favored males, seven in total. In none of these seven countries the difference is statistically significant.

Details on Non-Linear Wald Test

As we wanted to test whether students with a higher starting level also have a larger decline in performance, we specified the following nonlinear test (coefficients from Equation (4.2)): $\frac{\beta_2}{\beta_0} = \frac{\beta_2 + \beta_3}{\beta_0 + \beta_1}$. Cameron and Trivedi (2005) explained the nonlinear Wald test is invariant to algebraically equivalent ways of writing the nonlinear combinations of coefficients. As such, they suggested testing the combination in multiple algebraically equivalent ways. Our results did not change if we tested the following mathematically equivalent combination of coefficients: $\beta_2(\beta_0 + \beta_1) = \beta_0(\beta_2 + \beta_3)$.

Note that the nonlinear combination of coefficients can be interpreted in terms of whether the ratio $\frac{P[Q_1=1]}{P[Q_0=1]}$ is equal between males and females. If we follow Equation (4.2), we see that this ratio for males equals $\frac{\beta_0 + \beta_2}{\beta_0}$ and for females this ratio is $\frac{\beta_0 + \beta_1 + \beta_2 + \beta_3}{\beta_0 + \beta_1}$. If we want to test whether these ratios are equal, we test: $1 + \frac{\beta_2}{\beta_0} = 1 + \frac{\beta_2 + \beta_3}{\beta_0 + \beta_1}$. This exactly matches the test that we started with above.

Unreached Questions

In our main specification unreached questions were coded as missing. Although on average males had slightly more unreached questions than females (respectively 0.763 and 0.755 unreached questions on a test with roughly 60 questions), one might be worried that our baseline results partially pick up that females spent more time on each question trying to provide an accurate answer. In this section, we investigate the robustness of our findings by considering the case in which unreached questions are coded as wrong answers. The most suitable PISA wave to carry out this analysis is the PISA 2015,

because it minimizes possible mistakes in the classification of unreached items.⁵⁰ Figure A.4.15 documents our results are unchanged.

Increasing the Number of Questions to Measure Performance at the Start

Using the performance on the first question as a measure for the gender gaps at the beginning of the test might be too restrictive. At the same time, one might think the decline in performance is not severe at the first items of the test. We test for the robustness of our results by increasing the number of questions that are considered to be at the beginning. We re-estimated Equation (4.2) and (4.3) while coding the first five questions as the initial ones by setting a value of $Q_{ij} = 0$ for any item j that was ordered in any of the first five positions in the test. By doing so the results were virtually identical to our baseline results, see Figure A.4.16 and A.4.17.

Short Break after one Hour

The PISA test takers had a short break of typically 5 minutes after one hour of test taking. We tested whether this short break affects the gender difference by making use of the halfway dummy H and re-estimated Equation (4.2) as follows: $y = \beta_0 + \beta_1 F + \beta_2 Q + \beta_3 QF + \beta_4 H + \beta_5 QH + \epsilon$. We do not know at which item the student exactly was when they were allowed to have the short break, therefore we simply conceptualized H to be equal to 1 if the student was halfway during the test ($Q \geq 0.5$) and 0 otherwise. The inclusion of this break does not affect the gender difference, our results were identical to those reported in Section 4.3.1 (see Figure A.4.18).

4.A Appendix

⁵⁰Unreached items are defined as all the successive unanswered questions clustered at the end of a test, except for the first missing answer (OECD, 2012). The possibility of going back and forth across test items makes the pen-and-paper based PISA waves prone to an incorrect categorization of unreached items.

Figure A.4.1: Gender differences in performance during the test on the cluster level. PISA 2009

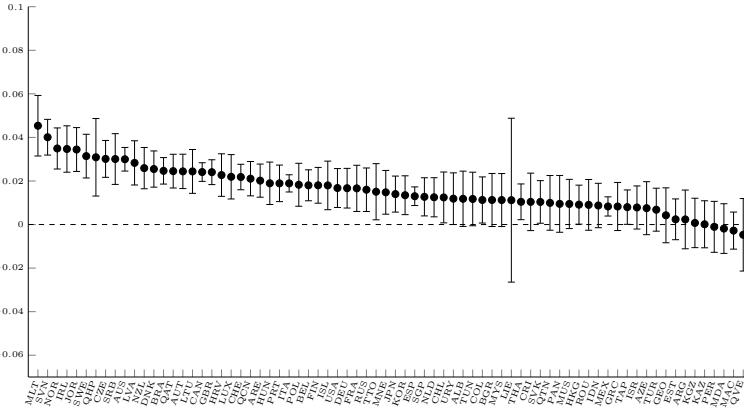


Figure A.4.2: Gender differences in performance during the test. PISA 2006 and PISA 2012

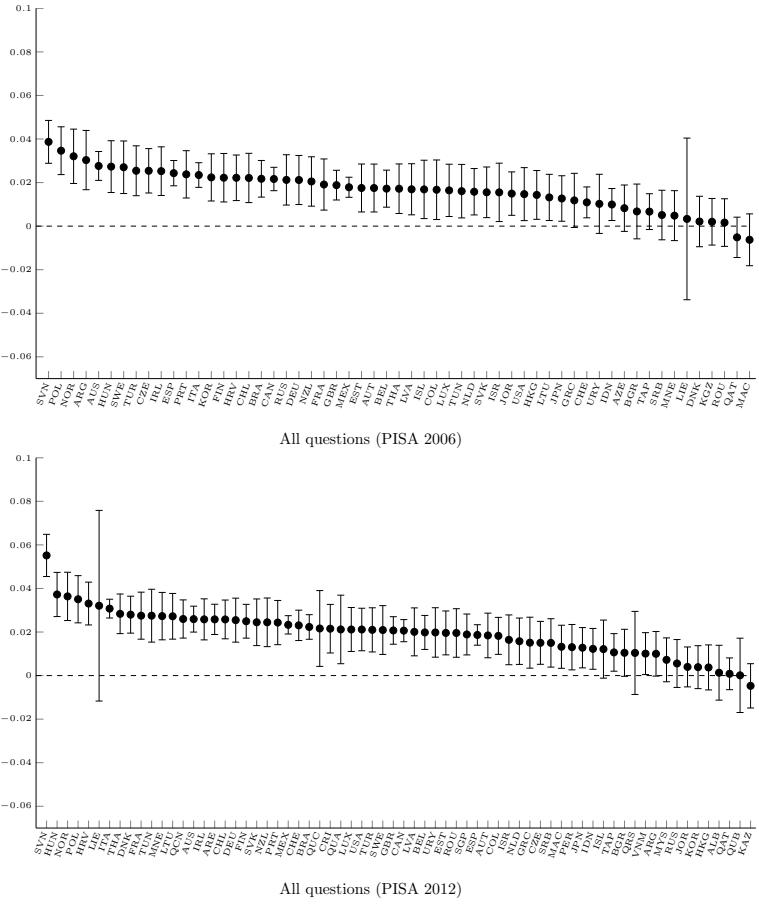


Figure A.4.3: Gender differences in initial performance and performance during the test by topic. PISA 2006 (science) and PISA 2012 (math)

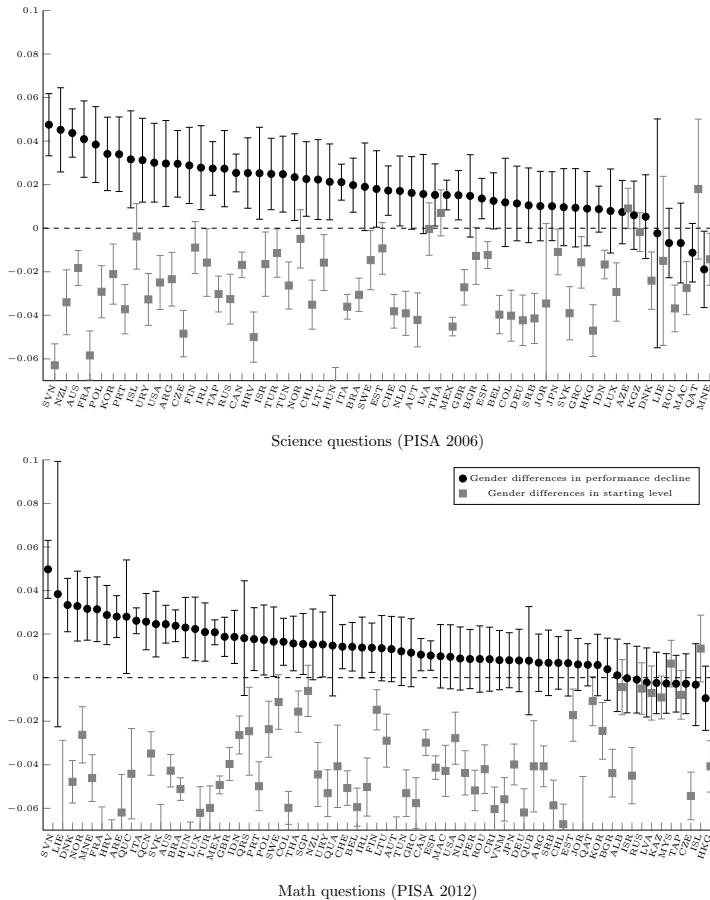


Figure A.4.4: Gender differences in performance during the test. PISA 2015

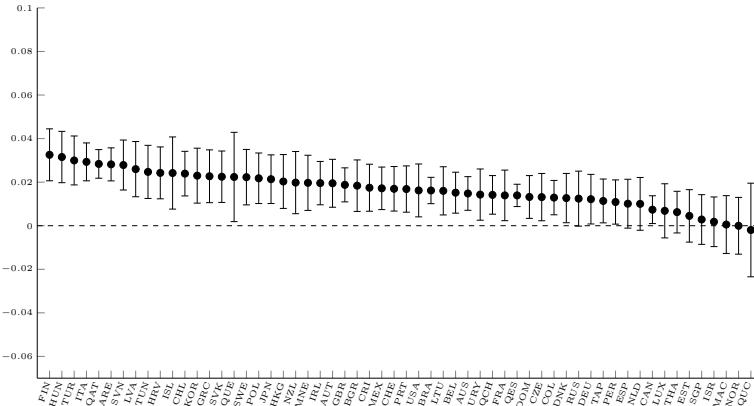


Figure A.4.5: Gender differences in initial performance and performance during the test by topic. PISA 2015 (science)

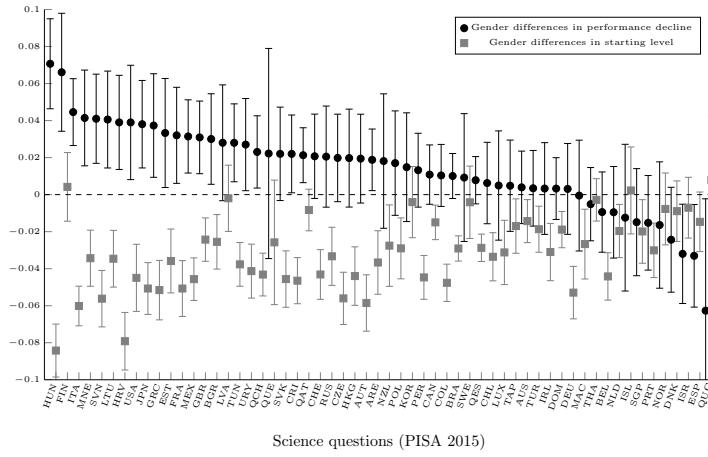


Figure A.4.6: Gender differences in performance during the test on the unit level. PISA 2015

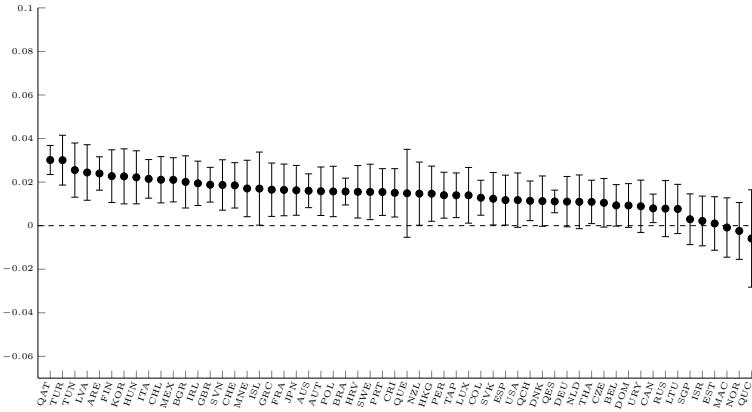


Figure A.4.7: Gender differences in performance during the test while controlling for number of actions and time spent per question. PISA 2015

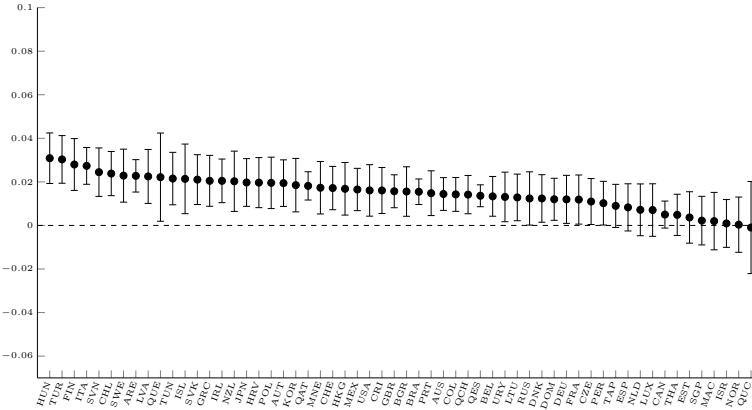
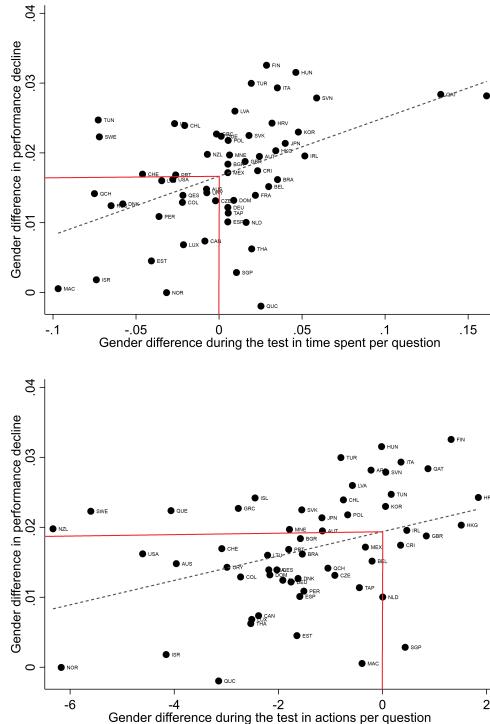


Figure A.4.8: Positive intercept and slope when regressing the gender difference in performance decline upon the gender difference during the test in time spent and number of actions per question. See Table A.4.7 for the regression results.



Notes: The figures are based upon the PISA 2015 and display a scatter plot and a linear OLS regression line between the gender difference in performance decline and the gender difference during the test in time spent per question (upper graph) and number of actions per question (bottom graph).

Figure A.4.9: Gender differences in performance during the test controlling for stumping. PISA 2015

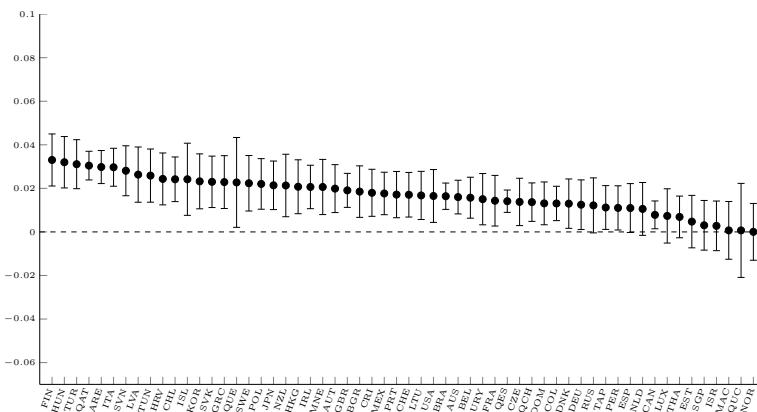


Figure A.4.10: Gender differences in performance during the test, probit estimation. PISA 2009

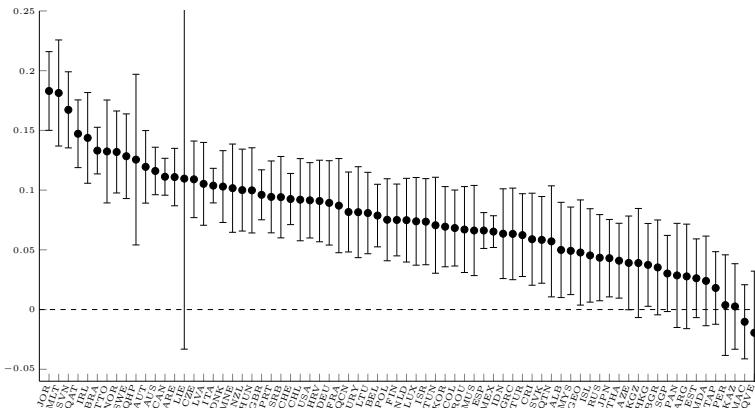


Figure A.4.11: Gender differences in initial performance and performance during the test by topic, probit. PISA 2009

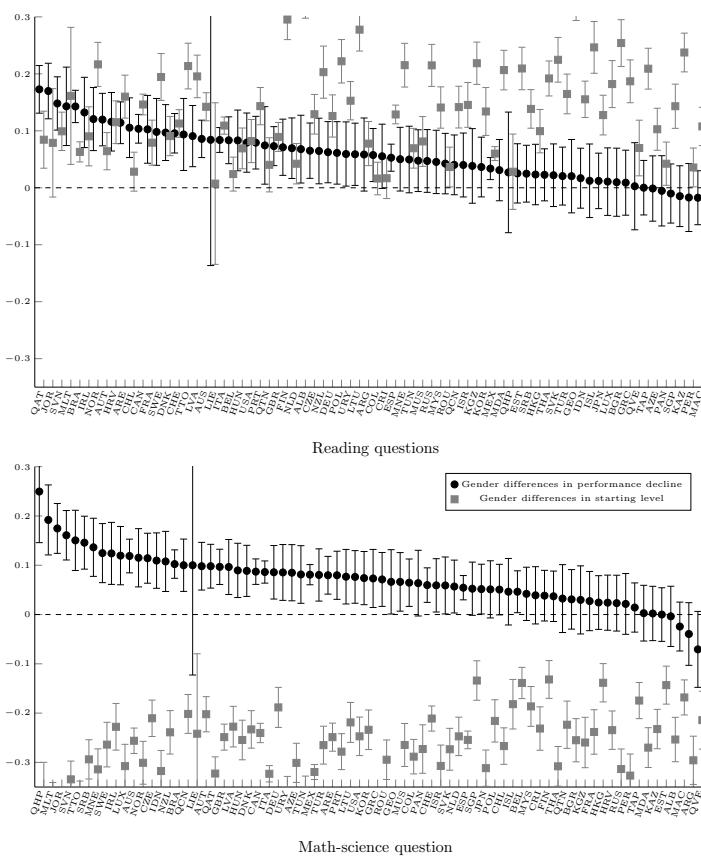
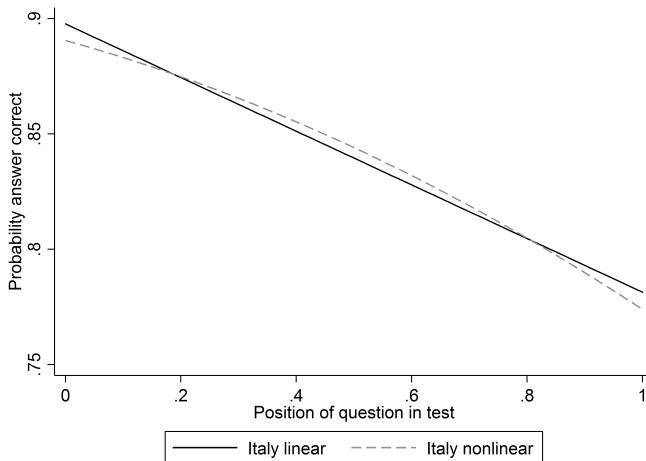
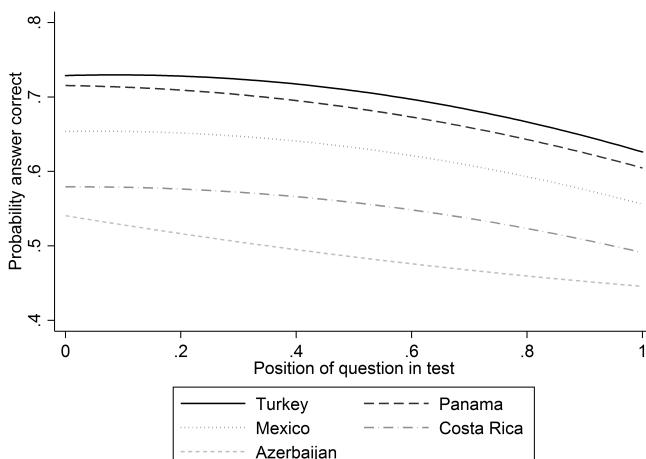


Figure A.4.12: Linear and nonlinear estimate of the decline in performance during the test for Italy



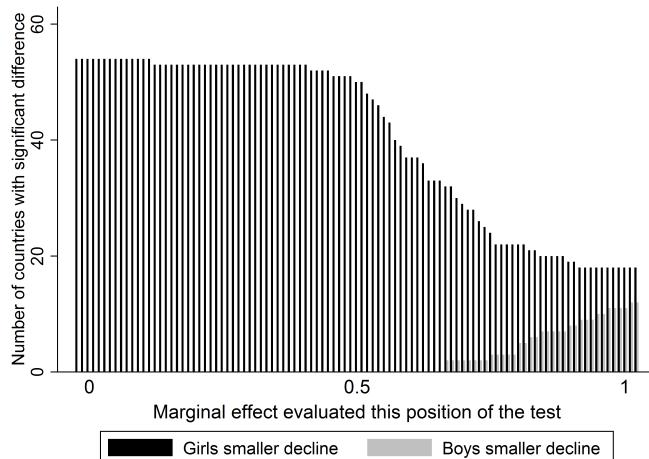
Notes: The figure is based upon the PISA 2009 and displays fitted values for a linear and quadratic estimate of the performance decline.

Figure A.4.13: Nonlinear estimates of the decline in performance during the test for the five countries with the most extreme nonlinear shape



Notes: The figure is based upon the PISA 2009 and displays fitted values for the quadratic estimate of the performance decline. For Turkey, Mexico, Panama, and Costa Rica the decline increases as the test continues, where the opposite is true for Azerbaijan.

Figure A.4.14: Testing for gender differences in the decline in performance at different positions of the test. PISA 2009



Notes: The figure displays the number of countries for which the gender difference in the decline in performance is significantly different at each position of the test.

Figure A.4.15: Gender differences in performance during the test coding unreached questions as wrong. PISA 2015

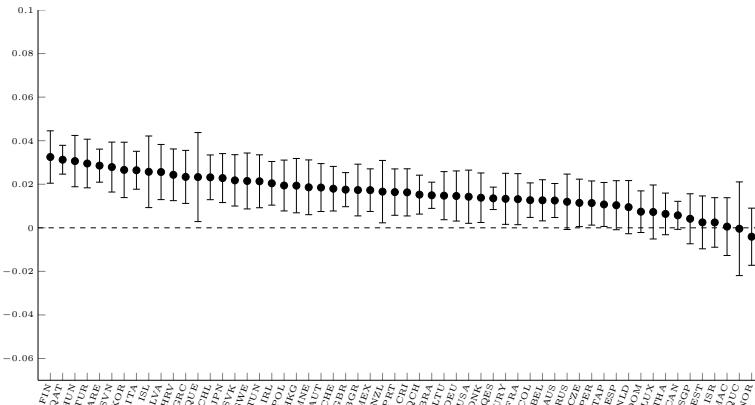


Figure A.4.16: Gender differences in performance during the test, first five questions included in $Q = 0$. PISA 2009

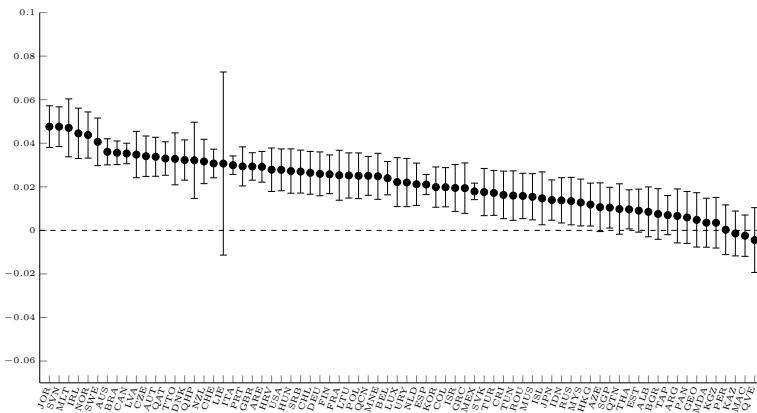


Figure A.4.17: Gender differences in initial performance and performance during the test by topic, first five questions included in $Q = 0$. PISA 2009

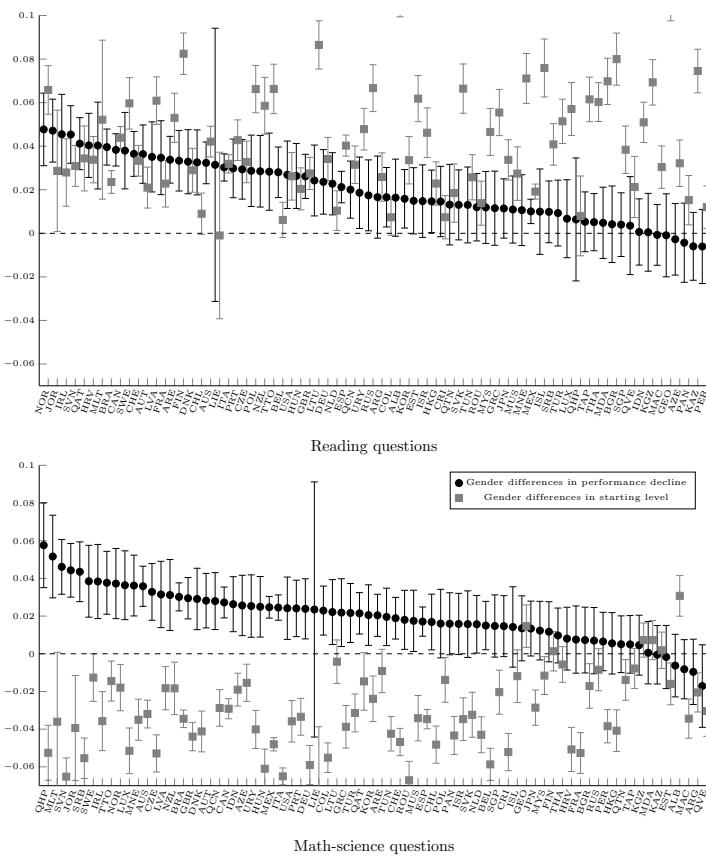


Figure A.4.18: Gender differences in performance during the test controlling for the short break in the middle of the test.
PISA 2009

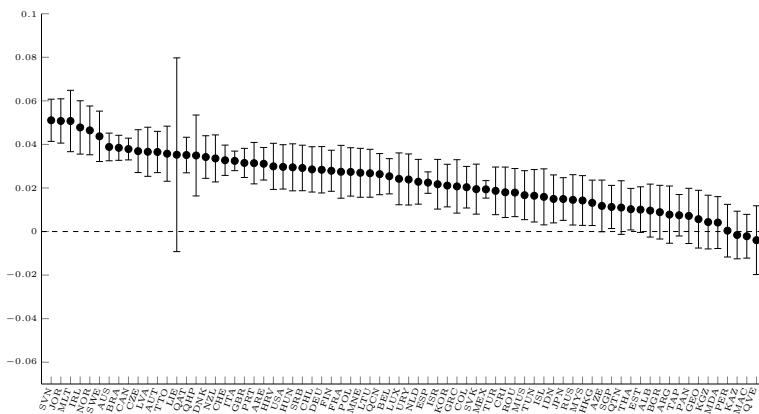


Table A.4.1: Gender differences in performance during the test. PISA 2009

Country	Gender diff. during the test	p-value	Country	Gender diff. during the test	p-value	Country	Gender diff. during the test	p-value
ALB	0.0097 (0.0074)	0.1887	HRV	0.0299 (0.0064)	3.06e-06	NZL	0.0335 (0.0065)	2.96e-07
ARE	0.0311 (0.0045)	6.60e-12	HUN	0.0295 (0.0065)	6.50e-06	PAN	0.0070 (0.0077)	0.3622
ARG	0.0076 (0.0080)	0.3384	IDN	0.0150 (0.0067)	0.0250	PER	0.0005 (0.0073)	0.9437
AUS	0.0388 (0.0038)	0.0000	IRL	0.0476 (0.0074)	1.43e-10	POL	0.0275 (0.0067)	0.0000
AUT	0.0365 (0.0057)	2.09e-10	ISL	0.0160 (0.0078)	0.0398	PRT	0.0315 (0.0058)	4.62e-08
AZE	0.0117 (0.0072)	0.1026	ISR	0.0218 (0.0069)	0.0016	QAT	0.0352 (0.0049)	1.09e-12
BEL	0.0254 (0.0049)	2.30e-07	ITA	0.0325 (0.0027)	0.0000	QCN	0.0263 (0.0057)	4.17e-06
BGR	0.0089 (0.0075)	0.2335	JOR	0.0506 (0.0062)	2.22e-16	QHP	0.0349 (0.0113)	0.0019
BRA	0.0385 (0.0035)	0.0000	JPN	0.0150 (0.0059)	0.0118	QTN	0.0110 (0.0075)	0.1396
CAN	0.0379 (0.0030)	0.0000	KAZ	-0.0016 (0.0066)	0.8148	QVE	-0.0039 (0.0096)	0.6798
CHE	0.0328 (0.0042)	7.99e-15	KGZ	0.0043 (0.0075)	0.5639	ROU	0.0179 (0.0067)	0.0075
CHL	0.0285 (0.0063)	5.97e-06	KOR	0.0211 (0.0059)	0.0004	RUS	0.0147 (0.0070)	0.0358
COL	0.0212 (0.0058)	0.0002	LIE	0.0354 (0.0270)	0.1895	SGP	0.0113 (0.0060)	0.0599
CRI	0.0182 (0.0070)	0.0095	LTU	0.0267 (0.0067)	0.0001	SRB	0.0292 (0.0063)	4.00e-06
CZE	0.0368 (0.0060)	6.17e-10	LUX	0.0242 (0.0072)	0.0008	SVK	0.0195 (0.0070)	0.0050
DEU	0.0284 (0.0065)	0.0000	LVA	0.0366 (0.0068)	8.33e-08	SVN	0.0510 (0.0059)	0.0000
DNK	0.0342 (0.0059)	7.85e-09	MAC	-0.0020 (0.0061)	0.7423	SWE	0.0437 (0.0070)	4.63e-10
ESP	0.0226 (0.0030)	3.77e-14	MDA	0.0041 (0.0072)	0.5685	TAP	0.0075 (0.0058)	0.1944
EST	0.0101 (0.0063)	0.1110	MEX	0.0196 (0.0024)	8.88e-16	THA	0.0103 (0.0058)	0.0738
FIN	0.0279 (0.0057)	1.01e-06	MLT	0.0507 (0.0085)	2.94e-09	TTO	0.0356 (0.0077)	3.28e-06
FRA	0.0274 (0.0074)	0.0002	MNE	0.0270 (0.0068)	0.0001	TUN	0.0170 (0.0073)	0.0200
GBR	0.0315 (0.0041)	8.22e-15	MUS	0.0167 (0.0068)	0.0143	TUR	0.0188 (0.0066)	0.0044
GEO	0.0058 (0.0080)	0.4721	MYS	0.0142 (0.0069)	0.0406	URY	0.0241 (0.0071)	0.0007
GRC	0.0208 (0.0074)	0.0050	NLD	0.0229 (0.0062)	0.0002	USA	0.0297 (0.0062)	1.53e-06
HKG	0.0132 (0.0063)	0.0370	NOR	0.0465 (0.0068)	7.79e-12			

Notes: Obtained by OLS estimations of Equation (4.2). Standard errors (in parentheses) clustered at the student level.

Table A.4.2: Gender differences in initial performance and performance during the test by topic. PISA 2009

Country	Diff. in reading starting level	p-value	Diff. in reading during the test	p-value	Diff. in science and math starting level	p-value	Diff. in science and math during the test	p-value
ALB	0.1104 (0.0078)	0.0000	0.0176 (0.0114)	0.1213	0.0308 (0.0071)	0.0000	-0.0059 (0.0107)	0.5840
ARE	0.0507 (0.0070)	4.28e-13	0.0359 (0.0069)	1.81e-07	-0.0104 (0.0072)	0.1460	0.0215 (0.0072)	0.0028
ARG	0.0244 (0.0072)	0.0007	0.0185 (0.0121)	0.1251	-0.0199 (0.0068)	0.0034	-0.0102 (0.0111)	0.3619
AUS	0.0400 (0.0045)	0.0000	0.0343 (0.0061)	2.42e-08	-0.0348 (0.0047)	1.71e-13	0.0389 (0.0068)	1.29e-08
AUT	0.0185 (0.0060)	0.0021	0.0387 (0.0086)	6.51e-06	-0.0841 (0.0064)	0.0000	0.0307 (0.0093)	0.0009
AZE	0.0321 (0.0070)	3.91e-06	-0.0025 (0.0106)	0.8156	-0.0173 (0.0065)	0.0075	0.0277 (0.0103)	0.0074
BEL	0.0444 (0.0052)	0.3951	0.0295 (0.0074)	0.0001	-0.0597 (0.0055)	0.0000	0.0158 (0.0083)	0.0557
BGR	0.0791 (0.0077)	0.0000	0.0055 (0.0113)	0.6229	-0.0177 (0.0076)	0.0205	0.0080 (0.0112)	0.4755
BRA	0.0207 (0.0034)	8.20e-10	0.0425 (0.0053)	1.11e-15	-0.0368 (0.0031)	0.0000	0.0328 (0.0048)	7.59e-12
CAN	0.0413 (0.0033)	0.0000	0.0407 (0.0047)	0.0000	-0.0312 (0.0034)	0.0000	0.0294 (0.0053)	2.29e-08
CHE	0.0310 (0.0045)	7.83e-12	0.0385 (0.0066)	5.82e-09	-0.0482 (0.0047)	0.0000	0.0203 (0.0071)	0.0043
CHL	0.0068 (0.0061)	0.2626	0.0346 (0.0095)	0.0003	-0.0496 (0.0064)	6.88e-15	0.0186 (0.0095)	0.0509
COL	0.0064 (0.0053)	0.2325	0.0174 (0.0087)	0.0468	-0.0568 (0.0051)	0.0000	0.0247 (0.0084)	0.0032
CRI	0.0061 (0.0064)	0.3468	0.0162 (0.0104)	0.1169	-0.0535 (0.0065)	2.22e-16	0.0164 (0.0104)	0.1152
CZE	0.0306 (0.0061)	5.89e-07	0.0317 (0.0087)	0.0003	-0.0554 (0.0064)	0.0000	0.0355 (0.0097)	0.0002
DEU	0.0324 (0.0064)	4.44e-07	0.0254 (0.0095)	0.0075	-0.0612 (0.0067)	0.0000	0.0264 (0.0102)	0.0100
DNK	0.0269 (0.0064)	0.0000	0.0346 (0.0094)	0.0002	-0.0431 (0.0070)	6.71e-10	0.0308 (0.0105)	0.0032
ESP	0.0389 (0.0031)	0.0000	0.0225 (0.0046)	1.24e-06	-0.0360 (0.0033)	0.0000	0.0184 (0.0050)	0.0002
EST	0.0609 (0.0068)	0.0000	0.0157 (0.0098)	0.1074	-0.0163 (0.0072)	0.0234	-0.0010 (0.0108)	0.9268
FIN	0.0803 (0.0061)	0.0000	0.0353 (0.0089)	0.0001	0.0001 (0.0067)	0.9868	0.0133 (0.0102)	0.1932
FRA	0.0204 (0.0070)	0.0034	0.0371 (0.0109)	0.0007	-0.0534 (0.0070)	3.02e-14	0.0085 (0.0115)	0.4561
GBR	0.0260 (0.0047)	3.44e-08	0.0273 (0.0065)	0.0000	-0.0464 (0.0048)	0.0000	0.0323 (0.0070)	4.45e-06
GEO	0.1093 (0.0079)	0.0000	0.0000 (0.0122)	0.9974	0.0138 (0.0073)	0.0607	0.0143 (0.0111)	0.1971
GRC	0.0547 (0.0068)	3.33e-15	0.0122 (0.0109)	0.2618	-0.0404 (0.0073)	3.64e-08	0.0235 (0.0116)	0.0421
HKG	0.0218 (0.0064)	0.0006	0.0158 (0.0092)	0.0855	-0.0415 (0.0071)	4.60e-09	0.0067 (0.0105)	0.5231
HRV	0.0311 (0.0068)	4.85e-06	0.0429 (0.0094)	5.59e-06	-0.0515 (0.0075)	8.52e-12	0.0089 (0.0107)	0.4046
HUN	0.0183 (0.0062)	0.0033	0.0286 (0.0096)	0.0028	-0.0628 (0.0068)	0.0000	0.0267 (0.0103)	0.0097
IDN	0.0506 (0.0061)	0.0000	0.0014 (0.0097)	0.8868	-0.0207 (0.0057)	0.0003	0.0277 (0.0094)	0.0033
IRL	0.0251 (0.0097)	0.0100	0.0479 (0.0118)	0.0000	-0.0387 (0.0099)	0.0001	0.0416 (0.0126)	0.0010
ISL	0.0752 (0.0085)	0.0000	0.0105 (0.0126)	0.4021	-0.0131 (0.0090)	0.1443	0.0157 (0.0137)	0.2521
ISR	0.0448 (0.0073)	9.50e-10	0.0167 (0.0107)	0.1176	-0.0362 (0.0073)	6.11e-07	0.0176 (0.0104)	0.0912
ITA	0.0296 (0.0027)	0.0000	0.0328 (0.0040)	2.22e-16	-0.0669 (0.0029)	0.0000	0.0265 (0.0043)	9.37e-10
JOR	0.0256 (0.0170)	0.1336	0.0501 (0.0093)	6.36e-08	-0.0421 (0.0170)	0.0134	0.0468 (0.0092)	3.48e-07
JPN	0.0330 (0.0059)	2.48e-08	0.0121 (0.0087)	0.1656	-0.0296 (0.0060)	8.43e-07	0.0145 (0.0093)	0.1209
KAZ	0.0751 (0.0065)	0.0000	-0.0067 (0.0100)	0.5041	0.0018 (0.0063)	0.7692	-0.0005 (0.0100)	0.9617
KGZ	0.0693 (0.0068)	0.0000	0.0005 (0.0115)	0.9685	0.0065 (0.0060)	0.2776	0.0059 (0.0102)	0.5616
KOR	0.0328 (0.0068)	1.63e-06	0.0163 (0.0086)	0.0578	-0.0255 (0.0077)	0.0009	0.0223 (0.0103)	0.0303
LIE	-0.0042 (0.0248)	0.8666	0.0357 (0.0401)	0.3732	-0.0833 (0.0271)	0.0021	0.0281 (0.0435)	0.5178
LTU	0.0850 (0.0071)	0.0000	0.0256 (0.0104)	0.0143	-0.0555 (0.0074)	0.4562	0.0233 (0.0111)	0.0354
LUX	0.0563 (0.0078)	4.40e-13	0.0078 (0.0115)	0.4965	-0.0541 (0.0078)	4.14e-12	0.0390 (0.0117)	0.0009
LVA	0.0589 (0.0070)	0.0000	0.0366 (0.0103)	0.0004	-0.0202 (0.0074)	0.0065	0.0331 (0.0113)	0.0034
MAC	0.0302 (0.0062)	1.26e-04	-0.0003 (0.0089)	0.9710	-0.0341 (0.0067)	3.45e-07	-0.0082 (0.0101)	0.4171
MDA	0.0691 (0.0069)	0.0000	0.0058 (0.0105)	0.5781	0.0073 (0.0068)	0.2805	0.0003 (0.0107)	0.9757
MEX	0.0183 (0.0022)	2.22e-16	0.0111 (0.0036)	0.0018	-0.0499 (0.0023)	0.0000	0.0267 (0.0037)	6.39e-13
MLT	0.0491 (0.0223)	0.0276	0.0437 (0.0128)	0.0006	-0.0395 (0.0025)	0.0793	0.0550 (0.0141)	0.0001
MNE	0.0703 (0.0074)	0.0000	0.0115 (0.0105)	0.2732	-0.0377 (0.0071)	9.88e-08	0.0392 (0.0104)	0.0002
MUS	0.0265 (0.0078)	0.0007	0.0120 (0.0099)	0.2262	-0.0533 (0.0077)	4.40e-06	0.0186 (0.0105)	0.0755
MYA	0.0457 (0.0070)	5.63e-11	0.0128 (0.0106)	0.2279	-0.0128 (0.0065)	0.0503	0.0139 (0.0100)	0.1667
NLD	0.0088 (0.0059)	0.1372	0.0245 (0.0091)	0.0075	-0.0441 (0.0062)	1.08e-12	0.0169 (0.0098)	0.0847
NOR	0.0625 (0.0071)	0.0000	0.0509 (0.0107)	1.80e-06	-0.0201 (0.0079)	0.0113	0.0389 (0.0120)	0.0012
NZL	0.0570 (0.0082)	3.74e-12	0.0295 (0.0105)	0.0051	-0.0207 (0.0088)	0.0195	0.0336 (0.0121)	0.0055
PAN	0.0152 (0.0074)	0.0396	-0.0038 (0.0117)	0.7418	-0.0446 (0.0066)	1.35e-11	0.0174 (0.0106)	0.1009
PER	0.0122 (0.0064)	0.0544	-0.0059 (0.0109)	0.5865	-0.0388 (0.0058)	2.02e-11	0.0067 (0.0100)	0.5009
POL	0.0639 (0.0069)	0.0000	0.0313 (0.0105)	0.0028	-0.0152 (0.0077)	0.0475	0.0176 (0.0117)	0.1321
PRT	0.0406 (0.0060)	1.07e-11	0.0320 (0.0087)	0.0002	-0.0350 (0.0063)	2.95e-08	0.0256 (0.0096)	0.0077
QAT	0.0314 (0.0092)	0.0006	0.0442 (0.0077)	8.76e-09	-0.0159 (0.0094)	0.0917	0.0224 (0.0071)	0.0016
QCN	0.0304 (0.0056)	4.55e-04	0.0208 (0.0084)	0.0133	-0.0306 (0.0063)	1.16e-06	0.0296 (0.0097)	0.0022
QHP	0.0070 (0.0120)	0.5623	0.0081 (0.0182)	0.6572	-0.0564 (0.0095)	3.23e-09	0.0614 (0.0145)	0.0000
QTN	0.0174 (0.0087)	0.0442	0.0143 (0.0119)	0.2312	-0.0146 (0.0070)	0.0369	0.0061 (0.0097)	0.5294
QVE	0.0210 (0.0091)	0.0208	0.0039 (0.0144)	0.7854	-0.0297 (0.0088)	0.0007	-0.0170 (0.0141)	0.2271
ROU	0.0125 (0.0066)	0.0574	0.0138 (0.0099)	0.1656	-0.0688 (0.0066)	0.0000	0.0201 (0.0101)	0.0462
RUS	0.0655 (0.0069)	0.0000	0.0184 (0.0104)	0.0777	-0.0092 (0.0072)	0.2022	0.0080 (0.0113)	0.4767
SGP	0.0381 (0.0069)	3.49e-08	0.0043 (0.0084)	0.6442	-0.0215 (0.0075)	0.0040	0.0159 (0.0106)	0.1345
SRB	0.0399 (0.0061)	6.88e-11	0.0110 (0.0091)	0.2242	-0.0586 (0.0069)	0.0000	0.0467 (0.0102)	4.97e-06
SVK	0.0652 (0.0073)	0.0000	0.0146 (0.0103)	0.1562	-0.0337 (0.0077)	0.0000	0.0173 (0.0114)	0.1274
SVN	0.0276 (0.0060)	4.64e-06	0.0488 (0.0085)	9.34e-09	-0.0683 (0.0064)	0.0000	0.0491 (0.0094)	1.51e-07
SWE	0.0570 (0.0076)	4.37e-14	0.0406 (0.0112)	0.0003	-0.0155 (0.0081)	0.0564	0.0415 (0.0123)	0.0007
TAP	0.0610 (0.0065)	0.0000	0.0058 (0.0088)	0.5086	-0.0080 (0.0068)	0.2397	0.0050 (0.0098)	0.6139
THA	0.0602 (0.0058)	0.0000	0.0050 (0.0084)	0.5564	-0.0066 (0.0062)	0.2812	0.0109 (0.0094)	0.2439
TTO	0.0645 (0.0073)	0.0000	0.0298 (0.0113)	0.0086	-0.0177 (0.0069)	0.0101	0.0416 (0.0107)	0.0001
TUN	0.0251 (0.0068)	0.0002	0.0134 (0.0112)	0.2311	-0.0438 (0.0060)	2.96e-13	0.0211 (0.0098)	0.0316
TUR	0.0508 (0.0066)	1.11e-14	0.0098 (0.0097)	0.3130	-0.0334 (0.0066)	3.31e-07	0.0242 (0.0101)	0.0163
URY	0.0462 (0.0062)	8.08e-14	0.0205 (0.0105)	0.0502	-0.0420 (0.0066)	1.99e-10	0.0273 (0.0106)	0.0103
USA	0.0244 (0.0070)	0.0005	0.0284 (0.0100)	0.0043	-0.0376 (0.0071)	1.14e-07	0.0259 (0.0106)	0.0147

Notes: Obtained by OLS estimations of Equation (4.3). Standard errors (in parentheses) clustered at the student level.

Table A.4.3: Rotation design of the 20 PISA booklets. PISA 2009

Booklet	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Standard booklet	Easy booklet
1	Math 1	Reading 1	Reading 3A	Math 3	X	
2	Reading 1	Science 1	Reading 4A	Reading 7	X	
3	Science 1	Reading 3A	Math 2	Science 3	X	
4	Reading 3A	Reading 4A	Science 2	Reading 2	X	
5	Reading 4A	Math 2	Reading 5	Math 1	X	
6	Reading 5	Reading 6	Reading 7	Reading 3A	X	
7	Reading 6	Math 3	Science 3	Reading 4A	X	
8	Reading 2	Math 1	Science 1	Reading 6	X	X
9	Math 2	Science 2	Reading 6	Reading 1	X	X
10	Science 2	Reading 5	Math 3	Science 1	X	X
11	Math 3	Reading 7	Reading 2	Math 2	X	X
12	Reading 7	Science 3	Math 1	Science 2	X	X
13	Science 3	Reading 2	Reading 1	Reading 5	X	X
14	Math 1	Reading 1	Reading 3B	Math 3	X	
15	Reading 1	Science 1	Reading 4B	Reading 7	X	
16	Science 1	Reading 3B	Math 2	Science 3	X	
17	Reading 3B	Reading 4B	Science 2	Reading 2	X	
18	Reading 4B	Math 2	Reading 5	Math 1	X	
19	Reading 5	Reading 6	Reading 7	Reading 3B	X	
20	Reading 6	Math 3	Science 3	Reading 4B	X	

Source: OECD (2012)

Table A.4.4: Randomization test, PISA 2009

	Gender	Mother highest schooling	Father highest schooling	Self born in country	Mother born in country	Father born in country	Language at home	Possessions desk	Possessions own room	How many books at home	Age of student
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Booklet=2	0.00794 (0.72)	0.00334 (0.14)	-0.0250 (-0.05)	0.00291 (0.63)	0.00613 (0.82)	-0.00468 (-0.63)	-0.00539 (-0.67)	-0.00224 (-0.27)	-0.00317 (-0.38)	0.00143 (0.04)	0.00301 (0.47)
Booklet=3	0.00439 (0.40)	0.0172 (0.73)	0.0186 (0.75)	-0.00323 (-0.73)	-0.00140 (-0.19)	-0.00698 (-0.93)	-0.000346 (-0.04)	0.00149 (0.18)	-0.0111 (-1.39)	-0.0142 (-0.45)	-0.000698 (-0.11)
Booklet=4	0.0117 (1.07)	0.000263 (0.01)	-0.00718 (-0.30)	0.00160 (0.35)	0.00324 (0.44)	-0.00803 (-1.09)	-0.0101 (-1.25)	-0.00366 (-0.45)	0.00676 (0.82)	0.0162 (0.53)	0.00872 (1.35)
Booklet=5	0.0114 (1.03)	0.00391 (0.17)	-0.00166 (-0.07)	-0.000857 (-0.16)	0.000863 (0.11)	-0.00269 (-0.34)	-0.00484 (-0.58)	-0.00651 (-0.81)	-0.00344 (-0.42)	-0.0165 (-0.52)	0.00185 (0.29)
Booklet=6	0.0217** (1.98)	-0.0101 (-0.44)	0.000509 (0.02)	-0.000475 (-0.10)	0.00190 (0.26)	-0.000299 (-0.04)	0.00311 (0.38)	-0.00292 (-0.36)	-0.0180** (-2.27)	0.0122 (0.39)	0.00410 (0.64)
Booklet=7	-0.00131 (-0.12)	0.0199 (0.84)	0.00347 (0.14)	0.0000170 (0.00)	0.00429 (0.57)	0.00278 (0.36)	0.00220 (0.26)	0.00548 (0.65)	-0.00509 (-0.63)	-0.0189 (-0.60)	-0.000367 (-0.05)
Booklet=8	0.00208 (0.21)	-0.00763 (-0.35)	-0.00756 (-0.34)	0.00177 (0.43)	-0.00378 (-0.58)	-0.00552 (-0.83)	-0.000470 (-0.06)	0.00390 (0.51)	-0.00960 (-1.28)	-0.0312 (-1.10)	0.000895 (0.15)
Booklet=9	0.00432 (0.43)	-0.0159 (-0.75)	0.00497 (0.22)	0.00103 (0.25)	-0.00364 (0.56)	-0.00882 (-1.34)	-0.00105 (-0.14)	0.00578 (0.74)	-0.00818 (-1.09)	-0.0119 (-0.42)	0.00527 (0.88)
Booklet=10	-0.00596 (-0.60)	-0.000868 (-0.04)	0.00947 (0.42)	-0.000346 (-0.09)	-0.00139 (-0.21)	-0.00567 (-0.85)	0.00107 (0.14)	-0.00366 (-0.50)	-0.00347 (-0.46)	0.0296 (1.04)	0.00404 (0.69)
Booklet=11	0.00889 (0.89)	-0.00605 (-0.29)	0.00440 (0.20)	0.000531 (0.13)	-0.00311 (-0.48)	-0.00455 (-0.68)	-0.00606 (-0.82)	0.00546 (0.72)	-0.00311 (-0.42)	-0.0166 (-0.58)	0.00177 (0.30)
Booklet=12	0.00553 (0.56)	0.0135 (0.63)	-0.0144 (-0.64)	0.00339 (0.83)	0.00105 (0.16)	-0.000688 (-0.10)	-0.00439 (-0.59)	-0.00778 (-1.06)	-0.00332 (-0.44)	-0.00415 (-0.15)	0.00403 (0.70)
Booklet=13	-0.00356 (-0.36)	0.00992 (0.46)	0.000589 (0.03)	0.00158 (0.39)	-0.00142 (-0.22)	-0.00236 (-0.35)	0.000164 (0.02)	-0.000736 (-0.10)	-0.00618 (-0.82)	-0.0134 (-0.48)	-0.00286 (-0.49)
Booklet=14	-0.000470 (-0.04)	-0.0186 (-0.60)	0.000923 (-0.29)	0.00275 (0.69)	0.000141 (0.02)	-0.00382 (-0.60)	0.00466 (0.61)	-0.00288 (-0.28)	-0.0221** (-2.17)	-0.0260 (-0.84)	0.00442 (0.63)
Booklet=15	-0.00152 (-0.13)	-0.0244 (-0.79)	-0.0345 (-1.09)	-0.000293 (-0.08)	-0.00381 (-0.63)	-0.00464 (-0.74)	0.000911 (0.12)	-0.00628 (-0.61)	-0.00794 (-0.77)	-0.0278 (-0.91)	0.00395 (0.56)
Booklet=16	0.0100 (0.84)	-0.0191 (-0.62)	-0.0488 (-1.55)	0.00386 (0.92)	-0.00451 (-0.75)	-0.00275 (-0.43)	0.000977 (0.13)	0.00637 (0.63)	-0.00798 (-0.78)	-0.0202 (-0.66)	0.00402 (0.58)
Booklet=17	-0.000417 (-0.03)	0.0183 (0.60)	-0.00197 (-0.06)	0.000248 (0.06)	-0.00377 (-0.62)	-0.00455 (-0.72)	0.000704 (0.09)	0.00921 (0.90)	-0.0104 (-1.02)	-0.0320 (-1.06)	0.0000957 (0.01)
Booklet=18	0.00400 (0.33)	-0.0352 (-1.16)	-0.0318 (-1.02)	0.00196 (0.49)	0.00197 (0.30)	-0.000900 (-0.14)	0.00388 (0.49)	0.00915 (0.90)	-0.0129 (-1.26)	0.0303 (0.95)	0.000963 (0.14)
Booklet=19	-0.00614 (-0.51)	-0.0413 (-1.33)	-0.0261 (-0.82)	0.00291 (0.72)	-0.00258 (-0.43)	-0.00562 (-0.91)	0.000205 (0.27)	-0.0100 (-1.00)	-0.0111 (-1.10)	0.0140 (0.45)	-0.00498 (-0.72)
Booklet=20	0.00597 (0.50)	-0.0123 (-0.40)	-0.00122 (-0.04)	0.000691 (0.18)	-0.000498 (-0.08)	-0.00226 (-0.35)	-0.000666 (-0.09)	-0.00227 (-0.23)	0.00433 (0.41)	0.0218 (0.70)	-0.000613 (-0.09)
Constant	1.511*** (121.81)	2.114*** (72.09)	2.010*** (68.51)	1.013*** (245.08)	1.010*** (169.31)	1.012*** (167.36)	1.010*** (149.79)	1.074*** (131.15)	1.323*** (123.59)	2.178*** (66.12)	15.77*** (2178.87)
Observations	514865	486133	473178	506007	502761	499261	495177	504103	505341	504108	514867
F-value	0.82	0.64	0.62	0.47	0.56	0.48	0.58	0.86	1.10	1.25	0.55
p-value	0.689	0.879	0.893	0.976	0.936	0.970	0.925	0.632	0.342	0.208	0.941
Adjusted R ²	0.002	0.275	0.215	0.041	0.125	0.125	0.296	0.108	0.113	0.151	0.041

Notes: *t* statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Regressions of background characteristics upon separate dummies for every booklet and country. We directly use the coding of the answer categories from the PISA (e.g. in column (1) we have an outcome variable where 1=male and 2=female), this only affects the interpretation of the constant. The columns F-value and p-value refer to the tests for joint significance of the booklet-dummies. The PISA 2009 and PISA weights are used.

Table A.4.5: When in the test do girls close the gender gap in math and science? Data from PISA 2009

Country	Gap at beginning of test	After how long do boys and girls perform equal?
India (Himachal Pradesh)	-5.64 ⁺⁺	92 %
Malta	-3.95 ⁺⁺	72 %
Slovenia	-6.83 ⁺⁺	139 %
Jordan	-4.21 ⁺⁺	90 %
Serbia	-5.86 ⁺⁺	125 %
Ireland	-3.87 ⁺⁺	93 %
Trinidad and Tobago	-1.77 ⁺⁺	43 %
Sweden	-1.55 ⁺⁺	37 %
Montenegro	-3.77 ⁺⁺	96 %
Luxembourg	-5.41 ⁺⁺	139 %
Australia	-3.48 ⁺⁺	89 %
Norway	-2.01 ⁺⁺	52 %
Czech Republic	-5.54 ⁺⁺	156 %
New Zealand	-2.07 ⁺⁺	62 %
Latvia	-2.02 ⁺⁺	61 %
Brazil	-3.68 ⁺⁺	112 %
United Kingdom	-4.64 ⁺⁺	143 %
Denmark	-4.31 ⁺⁺	140 %
Austria	-8.41 ⁺⁺	274 %
China (Shanghai)	-3.06 ⁺⁺	103 %
Canada	-3.12 ⁺⁺	106 %
Liechtenstein	-8.33 ⁺	297 %
Azerbaijan	-1.73 ⁺⁺	63 %
Indonesia	-2.07 ⁺⁺	75 %
Uruguay	-4.20 ⁺⁺	154 %
Hungary	-6.28 ⁺⁺	235 %
Mexico	-4.99 ⁺⁺	187 %
Italy	-6.69 ⁺⁺	253 %
Germany	-6.12 ⁺⁺	232 %
United States	-3.76 ⁺⁺	145 %
Portugal	-3.50 ⁺⁺	137 %
Colombia	-5.68 ⁺⁺	230 %
Turkey	-3.34 ⁺⁺	138 %
Greece	-4.04 ⁺⁺	172 %
Lithuania	-0.55	24 %
Qatar	-1.59 ⁺⁺	71 %
Korea	-2.55 ⁺⁺	114 %
United Arab Emirates	-1.04	48 %
Tunisia	-4.38 ⁺⁺	208 %
Switzerland	-4.82 ⁺⁺	238 %
Romania	-6.88 ⁺⁺	342 %
Mauritius	-3.53 ⁺⁺	190 %
Chile	-4.96 ⁺⁺	267 %
Spain	-3.60 ⁺⁺	195 %
Israel	-3.62 ⁺⁺	205 %
Poland	-1.52 ⁺	86 %
Panama	-4.46 ⁺⁺	257 %
Slovak Republic	-3.37 ⁺	195 %
Netherlands	-4.41 ⁺⁺	261 %
Costa Rica	-5.35 ⁺	326 %
Singapore	-2.15 ⁺	135 %
Belgium	-5.97 ⁺⁺	378 %
Iceland	-1.31	84 %
Japan	-2.96 ⁺	204 %
Malaysia	-1.28 ⁺	92 %
Thailand	.066	61 %
Croatia	-5.15 ⁺	576 %
France	-5.34 ⁺	626 %
Russian Federation	-0.92	115 %
Bulgaria	-1.77 ⁺	221 %
Peru	-3.88 ⁺	578 %
China (Hong Kong)	-4.15 ⁺	622 %
India (Tamil Nadu)	-1.46 ⁺	239 %
Taiwan	-0.80	161 %

Notes: ⁺ indicates a significant gender difference at the start of the test and ⁺⁺ indicates both a significant gender difference at the start of and during the test (at the 10% level).

The gender gap at the beginning of the test can be interpreted as the percentage points difference to answer the first question correct (the definition in Section 4.7.1 and subsequently multiplied by 100). The table includes countries where boys score better at the beginning of the test and girls perform better during the test in math and science.

Table A.4.6: Overview of all the PISA background questions used the analysis of Section 4.7.3.

Question	Noncognitive skills
PISA 2012	
In the last two full weeks of school, how many times did you arrive late for school? None - five or more times.	Conscientiousness (dutifulness)
“ <i>People can count on me to keep on schedule.</i> ”	<i>Self-control, Tangney et al. (2004)</i>
Using a train timetable to work out how long it would take to get from one place to another. Not confident - very confident.	Self-concept
Understanding graphs presented in newspapers. Not confident - very confident.	Self-concept
Solving an equation like $3x+5=17$. Not confident - very confident.	Self-concept
I often worry that it will be difficult for me in mathematics classes. Strongly disagree - strongly agree.	Neuroticism (anxiety, vulnerability)
I get very tense when I have to do mathematics homework. Strongly disagree - strongly agree.	Neuroticism (anxiety, vulnerability)
I get very nervous doing mathematics problems. Strongly disagree - strongly agree.	Neuroticism (anxiety, vulnerability)
“ <i>Gets nervous easily.</i> ”	<i>Neuroticism, John and Srivastava (1999)</i>
I worry that I will get poor grades in mathematics. Strongly disagree - strongly agree.	Neuroticism (anxiety, vulnerability)
If I put in enough effort I can succeed in mathematics. Strongly disagree - strongly agree.	Locus of control
“ <i>If I try hard enough, then I will understand the course material.</i> ”	<i>MSQL (Motivation), Duncan and McKeachie (2005)</i>
Whether or not I do well in mathematics is completely up to me. Strongly disagree - strongly agree.	Locus of control
If I had different teachers, I would try harder in mathematics. Strongly disagree - strongly agree.	Locus of control
If I wanted to, I could do well in mathematics. Strongly disagree - strongly agree.	Locus of control
I do badly in mathematics whether or not I study for my exams. Strongly disagree - strongly agree.	Locus of control
I am not very good at solving mathematics problems. Not likely - very likely.	Locus of control*
My teacher did not explain the concepts well this week. Not likely - very likely.	Locus of control*
This week I made bad guesses on the quiz. Not likely - very likely.	Locus of control*
Sometimes the course material is too hard. Not likely - very likely.	Locus of control*
The teacher did not get students interested in the material. Not likely - very likely.	Locus of control*
Sometimes I am just unlucky. Not likely - very likely.	Locus of control*
“ <i>There is really no such thing as luck.</i> ”	<i>Locus of control, Rotter (1966)</i>
I avoid distractions when I am studying mathematics. Strongly disagree - strongly agree.	Conscientiousness (self-discipline)
“ <i>Pleasure and fun sometimes keep me from getting work done.</i> ”	<i>Self-control, Tangney et al. (2004)</i>
I help my friends with mathematics. Never - almost always.	Extraversion (warmth)
Most of my teachers really listen to what I have to say. Strongly disagree - strongly agree.	Agreeableness (compliance)
I make friends easily at school. Strongly disagree - strongly agree.	Extraversion (warmth)
“ <i>Is outgoing, sociable.</i> ”	<i>Extraversion, John and Srivastava (1999)</i>
If I put in enough effort, I can succeed in school. Strongly disagree - strongly agree.	Locus of control
It is completely my choice whether or not I do well at school. Strongly disagree - strongly agree.	Locus of control
If I had different teachers, I would try harder at school. Strongly disagree - strongly agree.	Locus of control
If I wanted to, I could perform well at school. Strongly disagree - strongly agree.	Locus of control
I perform poorly at school whether or not I study for my exams. Strongly disagree - strongly agree.	Locus of control

Overview of all the PISA background questions (continued)

When confronted with a problem, I give up easily. Not like me - very much like me. <i>"Setbacks do not discourage me."</i>	Conscientiousness (perseverance)** <i>Gritt, Duckworth et al. (2007)</i>
I put off difficult problems. Not like me - very much like me.	Conscientiousness (perseverance)** <i>Gritt, Duckworth et al. (2007)</i>
I remain interested in the tasks that I start. Not like me - very much like me. <i>"I have difficulty maintaining my focus on projects that take more than a few months to complete."</i>	Conscientiousness (perseverance)** <i>Gritt, Duckworth et al. (2007)</i>
I continue working on tasks until everything is perfect. Not like me - very much like me. <i>"Perseveres until the task is finished."</i>	Conscientiousness (perseverance)** <i>Conscientiousness, John and Srivastava (1999)</i>
When confronted with a problem, I do more than what is expected of me. Not like me - very much like me.	Conscientiousness (perseverance)**
I can handle a lot of information. Not like me - very much like me.	Openness (actions)
I like to participate in host culture celebrations Strongly disagree - strongly agree.	Openness (wide interests)
I like to participate in heritage culture celebrations. Strongly disagree - strongly agree.	Openness (wide interests)
Combined PISA score of the questions indicated by*	Locus of control
Combined PISA score of the questions indicated by**	Conscientiousness (perseverance)
PISA 2009	
When I study, I try to memorize everything that is covered in the text. Almost never - almost always.	Conscientiousness (self-discipline)
When I study, I start by figuring out what exactly I need to learn. Almost never - almost always. <i>"When I study the readings for this course, I outline the material to help me organize my thoughts."</i>	Conscientiousness (organized)
When I study, I try to memorize as many details as possible. Almost never - almost always.	<i>MSQL (Learning) Duncan and McKeachie (2005)</i>
When I study, I try to relate new information to prior knowledge acquired in other subjects. Almost never - almost always. <i>"Likes to reflect, play with ideas."</i>	Conscientiousness (self-discipline) Openness (fantasy)
School has been a waste of time. Strongly disagree - strongly agree.	<i>Openness, John and Srivastava (1999)</i>
I get along well with most of my teachers. Strongly disagree - strongly agree. <i>"Starts quarrels with others."</i>	Motivation towards school / learning Agreeableness (compliance) <i>Agreeableness, John and Srivastava (1999)</i>
I learn about things that are not course-related, such as sports, hobbies, people or music. Never - several times a week. <i>"Is sophisticated in art, music, or literature."</i>	Openness (actions)
Participate in online forums, virtual communities or spaces. Never - every day.	<i>Openness, John and Srivastava (1999)</i>
PISA 2006	
How informed are you about science-related careers available on the job market? Not informed at all - very well informed.	Conscientiousness (striving)
Making an effort in my subject(s) is worth it because this will help me in the work I want to do later on. Strongly disagree - strongly agree.	Motivation towards school / learning

Notes: The table displays the questions used in the analysis of Section 4.7.3 of the Supplementary Material, the related noncognitive skills and facets in parentheses, and for some questions it shows related questions of scales that have been validated by previous research.

Table A.4.7: Regression of the gender difference in performance during test on the gender difference in dynamic inputs during test

	(1)	(2)	(3)	(4)	(5)	(6)
Gender difference in time during test	0.0843*** (4.44)	0.109*** (3.28)			0.0666*** (2.88)	0.0892*** (2.81)
Gender difference in actions during test			0.00174** (2.62)	0.00398*** (3.38)	0.000986 (1.30)	0.00282** (2.12)
Gender difference in time at the start		0.0134 (0.31)				0.0251 (0.67)
Gender difference in actions at the start				0.000162 (0.44)		-0.0000534 (-0.17)
Its interaction for time		-0.258 (-0.57)				-0.176 (-0.47)
Its interaction for actions				-0.000182 (-1.52)		-0.000163 (-1.31)
Constant	0.0167*** (17.29)	0.0153*** (5.54)	0.0194*** (14.36)	0.0192*** (8.28)	0.0182*** (12.87)	0.0170*** (5.48)
<i>N</i>	58	58	58	58	58	58
Adj. <i>R</i> ²	0.203	0.183	0.132	0.150	0.227	0.205

Notes: *t* statistics in parentheses, heteroskedasticity robust standard errors

* *p* < 0.10, ** *p* < 0.05, *** *p* < 0.01

The gender difference in ability to sustain performance is β_3 of Equation (4.2). The equations estimated are as follows: $\beta_{3c} = \alpha_0 + \alpha_1 \text{gender differences in inputs during test}_c + \alpha_2 \text{gender differences in inputs start of test}_c + \alpha_3 \text{its interaction}_c + \epsilon_c$, where *c* is a subscript for country *c*.

Table A.4.8: Relationship between the gender gap in math and the length of a specific test, measured by number of questions and maximum time allowed to complete the test

	Whole sample	Exclude tests with noq \leq 10	Exclude tests with noq \leq 40	Whole sample	Recalculated gender gap	Weighted regression	Exclude five extreme long tests	Exclude tests with time \leq 5	Exclude tests with time \leq 20
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Number of questions	-0.00159** (-2.06)	-0.00347*** (-3.34)	-0.00339** (-2.09)						
Maximum time allowed				-0.000761 (-1.05)	-0.000719 (-0.99)	-0.000941 (-1.23)	-0.00214*** (-2.73)	-0.00244*** (-2.77)	-0.00281*** (-3.37)
Constant	0.200*** (4.59)	0.337*** (5.70)	0.361*** (3.02)	0.180*** (3.54)	0.174*** (3.38)	0.195*** (3.87)	0.228*** (4.27)	0.249*** (4.05)	0.282*** (4.59)
<i>N</i>	203	169	74	175	175	175	170	157	109
Adj. <i>R</i> ²	0.012	0.069	0.080	0.000	0.000	0.004	0.026	0.033	0.057

Notes: *t* statistics in parentheses, heteroskedasticity robust standard errors

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The gender gap is measured by subtracting the mean performance of girls from the mean performance of boys and dividing this by the pooled standard deviation. The equations estimated are as follows: $mathgendergap_i = \alpha_0 + \alpha_1 testlength_i + \epsilon_i$, where i is a subscript for test i .

Table A.4.9: Relationship between the gender gap in math and the number of questions in a specific test for certain subsamples

	Whole sample	Australia, Europe and Middle East	Asia	Only high-stakes tests
	(1)	(2)	(3)	(4)
Number of questions	-0.00159** (-2.06)	-0.00699*** (-3.48)	0.00183 (0.91)	-0.00557* (-1.87)
Constant	0.200*** (4.59)	0.380*** (4.36)	0.0907 (1.65)	0.192 (1.10)
<i>N</i>	203	45	20	17
Adj. <i>R</i> ²	0.012	0.303	0.005	0.151

Notes: *t* statistics in parentheses, heteroskedasticity robust standard errors

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The gender gap is measured by subtracting the mean performance of girls from the mean performance of boys and dividing this by the pooled standard deviation. The equations estimated are as follows: $mathgendergap_i = \alpha_0 + \alpha_1 numberofquestions_i + \epsilon_i$, where i is a subscript for test i .

Table A.4.10: Gender differences in performance during the test while using probit. Testing the marginal effects, no distinction per topic. PISA 2009

Country	Gender diff. in marg. effects	Welsch t-test	Country	Gender diff. in marg. effects	Welsch t-test	Country	Gender diff. in marg. effects	Welsch t-test
ALB	0.0099	1.3291	HRV	0.0285	4.3271	NZL	0.0337	5.0439
ARE	0.0340	7.3498	HUN	0.0301	4.4991	PAN	0.0092	1.1907
ARG	0.0081	1.0089	IDN	0.0159	2.3504	PER	0.0023	0.3177
AUS	0.0385	9.7755	IRL	0.0473	6.2153	POL	0.0266	3.8571
AUT	0.0372	6.3322	ISL	0.0164	2.0786	PRT	0.0308	5.2138
AZE	0.0101	1.3851	ISR	0.0236	3.3301	QAT	0.0389	7.7825
BEL	0.0237	4.7554	ITA	0.0324	11.6104	QCN	0.0245	4.2256
BGR	0.0102	1.3467	JOR	0.0562	8.8016	QHP	0.0333	2.9674
BRA	0.0380	10.7772	JPN	0.0138	2.2962	QTN	0.0138	1.8550
CAN	0.0377	12.1427	KAZ	-0.0016	-0.2429	QVE	-0.0059	-0.6024
CHE	0.0306	7.0960	KGZ	0.0041	0.5591	ROU	0.0216	3.1391
CHL	0.0279	4.3418	KOR	0.0214	3.5306	RUS	0.0143	1.9865
COL	0.0210	3.5763	LIE	0.0327	1.1886	SGP	0.0101	1.6584
CRI	0.0180	2.5292	LTU	0.0269	3.9264	SRB	0.0290	4.4584
CZE	0.0338	5.5494	LUX	0.0245	3.3317	SVK	0.0192	2.6880
DEU	0.0271	4.1157	LVA	0.0357	5.0938	SVN	0.0512	8.5132
DNK	0.0340	5.5929	MAC	-0.0034	-0.5473	SWE	0.0442	6.1275
ESP	0.0224	7.3424	MDA	0.0043	0.5842	TAP	0.0071	1.1970
EST	0.0095	1.4582	MEX	0.0200	8.0478	THA	0.0103	1.7450
FIN	0.0267	4.5598	MLT	0.0586	6.6506	TTO	0.0370	4.7927
FRA	0.0266	3.5742	MNE	0.0279	4.0536	TUN	0.0211	2.8732
GBR	0.0312	7.5104	MUS	0.0192	2.8025	TUR	0.0194	2.8759
GEO	0.0044	0.5425	MYS	0.0147	2.0660	URY	0.0248	3.4295
GRC	0.0212	2.7607	NLD	0.0215	3.4182	USA	0.0300	4.7687

Notes: See the Supplementary Material for the exact procedure of testing the statistical significance of the marginal effects.

Table A.4.11: Nonlinear Wald test on coefficients of the performance decline and initial performance, both for the whole test (all topics) and per topic (reading and math-science). PISA 2009

Country	Gender diff. in all topics	p-value	Diff. in decline for girls is statistically significant	Gender diff. in reading	p-value	Diff. in decline for girls is statistically significant	Gender diff. in math and science	p-value	Diff. in decline for girls is statistically significant
ALB	0.054	0.001	Yes	0.078	0.001	Yes	-0.001	0.963	
ARE	3.70e-14	0.067	Yes	2.85e-11	0.022	Yes	0.033	0.003	Yes
ARG	0.018	0.268	Yes	0.049	0.022	Yes	-0.034	0.161	
AUS	0.042	7.23e-24	Yes	0.039	1.58e-10	Yes	0.040	5.23e-08	Yes
AUT	0.088	5.39e-07	Yes	0.145	9.30e-06	Yes	0.056	0.034	Yes
AZE	0.025	0.045	Yes	0.020	0.424	Yes	0.048	0.009	Yes
BEL	0.027	1.09e-06	Yes	0.032	0.000	Yes	0.015	0.102	
BGR	0.034	0.053	Yes	0.043	0.055	Yes	0.011	0.623	
BRA	0.048	8.12e-27	Yes	0.060	4.96e-20	Yes	0.037	1.16e-09	Yes
CAN	0.045	8.38e-32	Yes	0.050	7.08e-21	Yes	0.032	1.96e-07	Yes
CHE	0.042	6.70e-14	Yes	0.055	6.95e-11	Yes	0.022	0.017	Yes
CHL	0.038	0.000	Yes	0.043	0.000	Yes	0.020	0.149	
COL	0.040	0.004	Yes	0.029	0.022	Yes	0.036	0.082	Yes
CRI	0.025	0.024	Yes	0.021	0.082	Yes	0.018	0.276	
CZE	0.034	5.00e-10	Yes	0.034	0.000	Yes	0.032	0.001	Yes
DEU	0.027	0.000	Yes	0.027	0.002	Yes	0.023	0.018	Yes
DNK	0.040	2.10e-08	Yes	0.043	0.000	Yes	0.033	0.007	Yes
ESP	0.028	2.19e-14	Yes	0.032	7.58e-10	Yes	0.019	0.002	Yes
EST	0.017	0.048	Yes	0.025	0.025	Yes	-0.003	0.845	
FIN	0.037	8.95e-08	Yes	0.049	2.05e-06	Yes	0.016	0.185	
FRA	0.030	0.000	Yes	0.042	0.000	Yes	0.005	0.716	
GBR	0.044	3.69e-13	Yes	0.040	1.05e-06	Yes	0.042	0.000	Yes
GEO	0.066	0.010	Yes	0.077	0.011	Yes	0.043	0.083	
GRC	0.034	0.001	Yes	0.036	0.007	Yes	0.021	0.177	
HKG	0.018	0.043	Yes	0.024	0.038	Yes	0.006	0.683	
HRV	0.050	2.70e-06	Yes	0.065	1.61e-07	Yes	0.005	0.804	
HUN	0.037	0.000	Yes	0.038	0.001	Yes	0.030	0.035	
IDN	0.035	0.003	Yes	0.021	0.160	Yes	0.049	0.005	Yes
IRL	0.064	2.27e-10	Yes	0.062	6.75e-06	Yes	0.054	0.002	Yes
ISL	0.023	0.012	Yes	0.023	0.102	Yes	0.018	0.265	
ISR	0.054	0.000	Yes	0.046	0.004	Yes	0.024	0.292	
ITA	0.034	2.53e-29	Yes	0.039	3.40e-21	Yes	0.022	1.93e-06	Yes
JOR	0.107	6.39e-14	Yes	0.092	8.33e-10	Yes	0.085	0.000	Yes
JPN	0.016	0.008	Yes	0.016	0.052	Yes	0.013	0.193	
KAZ	0.007	0.485	Yes	0.017	0.329	Yes	0.000	0.977	
KGZ	0.044	0.036	Yes	0.095	0.012	Yes	0.022	0.449	
KOR	0.033	0.000	Yes	0.027	0.017	Yes	0.030	0.045	Yes

Nonlinear Wald test (continued)

Country	Gender diff. in diff. in all topics	p-value	Diff. in decline for girls is statistically significant	Gender diff. in reading	p-value	Diff. in decline for girls is statistically significant	Gender diff. in math and science	p-value	Diff. in decline for girls is statistically significant
LIE	0.031	0.225		0.034	0.354		0.023	0.582	
LTU	0.052	1.00e-06	Yes	0.049	0.000	Yes	0.037	0.032	Yes
LUX	0.038	0.000	Yes	0.029	0.073	Yes	0.051	0.004	Yes
LVA	0.052	3.73e-09	Yes	0.065	3.14e-06	Yes	0.043	0.004	Yes
MAC	-0.003	0.717		0.003	0.714		-0.014	0.267	
MDA	0.043	0.027	Yes	0.194	0.017	Yes	0.006	0.831	
MEX	0.024	8.50e-14	Yes	0.015	0.000	Yes	0.029	3.57e-09	Yes
MET	0.066	2.48e-09	Yes	0.063	0.000	Yes	0.065	0.000	Yes
MNE	0.072	3.99e-07	Yes	0.068	0.001	Yes	0.079	0.001	Yes
MUS	0.020	0.012	Yes	0.017	0.108	Yes	0.020	0.122	
MYS	0.024	0.017	Yes	0.038	0.047	Yes	0.020	0.183	
NLD	0.021	0.000	Yes	0.024	0.004	Yes	0.014	0.143	
NOR	0.062	1.17e-12	Yes	0.069	1.43e-08	Yes	0.047	0.001	Yes
NZL	0.041	1.00e-07	Yes	0.037	0.001	Yes	0.038	0.007	Yes
PAN	0.008	0.483		-0.002	0.894		0.020	0.197	
PER	-0.012	0.568		-0.006	0.784		-0.015	0.623	
POL	0.040	2.78e-06	Yes	0.049	0.000	Yes	0.022	0.148	
PRT	0.046	1.29e-08	Yes	0.048	4.70e-06	Yes	0.032	0.018	Yes
QAT	0.062	3.77e-14	Yes	0.114	2.96e-13	Yes	0.035	0.003	Yes
QCN	0.039	3.74e-06	Yes	0.026	0.004	Yes	0.041	0.004	Yes
QHP	0.102	0.005	Yes	0.041	0.444	Yes	0.195	0.000	Yes
QTN	0.069	0.060	Yes	0.145	0.015	Yes	0.011	0.840	
QVE	-0.006	0.610		0.009	0.568		-0.027	0.136	
ROU	0.021	0.037	Yes	0.022	0.074	Yes	0.019	0.257	
RUS	0.021	0.005	Yes	0.030	0.005	Yes	0.008	0.515	
SGP	0.021	0.039	Yes	0.013	0.348	Yes	0.025	0.159	
SRB	0.106	8.54e-06	Yes	0.027	0.044	Yes	0.126	0.001	Yes
SVK	0.031	0.001	Yes	0.036	0.007	Yes	0.020	0.212	
SVN	0.061	4.86e-14	Yes	0.062	2.61e-09	Yes	0.054	6.10e-06	Yes
SWE	0.060	1.86e-10	Yes	0.066	6.89e-06	Yes	0.053	0.001	Yes
TAP	0.009	0.094	Yes	0.010	0.208	Yes	0.004	0.637	
THA	0.039	0.001	Yes	0.022	0.046	Yes	0.022	0.260	
TTO	0.231	9.66e-08	Yes	0.151	4.71e-07	Yes	0.165	0.000	Yes
TUN	0.028	0.022	Yes	0.023	0.071	Yes	0.026	0.134	
TUR	0.027	0.001	Yes	0.018	0.066	Yes	0.029	0.029	Yes
URY	0.033	0.000	Yes	0.038	0.003	Yes	0.029	0.031	Yes
USA	0.042	2.45e-06	Yes	0.041	0.001	Yes	0.032	0.030	Yes

Notes: See the Supplementary Material for the exact procedure of the nonlinear Wald test.

Chapter 5

Test Scores, Noncognitive Skills and Economic Growth

Joint work with Pau Balart and Dinand Webbink. Published in Economics of Education Review

5.1 Introduction

Many studies have found a strong association between the economic outcomes of nations and their performance on international cognitive tests such as the PISA, TIMSS or PIRLS (see, for example, Hanushek and Kimko, 2000; Hanushek and Woessmann, 2008, 2012). This association is interpreted as evidence that cognitive skills are an important determinant of productivity and economic growth. However, the performance on cognitive tests is not only the result of cognitive ability, but is also influenced by noncognitive skills. Pioneers in intelligence testing like Thorndike and Wechsler already recognized that test takers might not exert maximal effort (Wechsler, 1940). Duckworth et al. (2011) found that under low-stakes testing conditions, such as in the international cognitive tests, some individuals try harder than others. Moreover, scores of low performers can be substantially improved by offering a reward (e.g. Gneezy and Rustichini, 2000; Borghans et al., 2008b; Segal, 2012). The noncognitive skills that are important for test scores have also been found to be important for productivity and other social outcomes at the individual level (e.g. Heckman and Rubinstein, 2001; Heckman et al., 2013). This suggests that noncognitive skills might be an important omitted variable in the relationship between cognitive skills and the economic outcomes of nations. These two related issues make it unclear to what extent the strong association between the performance on international cognitive tests and economic growth should be interpreted as evidence on the importance of cognitive versus noncognitive skills. Given the differences in policy interventions required to foster cognitive

and noncognitive skills (Cunha et al., 2010), it is important to gain a better understanding of their respective roles in fostering economic growth.¹

In this paper, we explore the effects of cognitive versus noncognitive skills on economic growth. We decompose the performance on an international test (PISA) into two components: the starting performance and the decline in performance during the test. This decomposition, recently introduced by Borghans and Schils (2012), exploits the random allocation of test booklets to students, which generates exogenous variation in the position of questions in the test. This specific feature of the test allows estimation of the decline in performance during the test that is not confounded by unobserved characteristics of questions, such as the difficulty of the test items. Borghans and Schils (2012) show that differences in the decline in performance during the test are related to noncognitive skills, such as motivation and ambition. They argue that the starting performance of the test score is a measure of cognitive skills that is not confounded by the personality factors that cause the decline in performance.

Countries differ in both the starting performance and the decline in performance during the test and these differences are stable over time. We use the results of this decomposition to estimate the association between the two components and economic growth, and compare these findings with the estimated effect of test scores before the decomposition, which is the standard approach in the previous literature. For the analysis, we stay as close as possible to the seminal paper by Hanushek and Woessmann (2012), hereafter HW (2012). This study documented a strong association between test scores and economic growth, and provided convincing evidence that supports a causal interpretation of this relationship.

The exact interpretation of our findings critically depends upon which type of skills are measured by the two components. We argue that, at a minimum, our study answers the question whether noncognitive skills are partly responsible for the well-studied relationship between test scores and economic growth. To this end, we need a measure of noncognitive skills that is not confounded with cognitive skills. Most of our effort is therefore devoted to discussing the performance decline, which lends itself to two different types of interpretations. Let us call interpretation A that “*the performance decline captures noncognitive skills and does not capture cognitive skills*” while interpretation B admits the possibility that “*the performance decline captures both cognitive and noncognitive skills*”. As recognized by Borghans et al. (2008a), it is not only empirically, but also conceptually difficult to separate cognitive ability from noncognitive skills.² In fact, many aspects of personality and cognition

¹Noncognitive skills have many different names in the literature. Soft skills, personality traits, character skills, noncognitive ability and socioemotional skills are often used.

²They define cognitive skills as the “ability to understand complex ideas, to adapt effectively to the environment, to learn from experience, to engage in various forms of reasoning, to overcome obstacles by taking thought”. Noncognitive skills are referred to as patterns of thoughts, feelings and behaviors.

are closely related.³ Evidence presented in Section 5.3.1, however, supports interpretation A – that the performance decline measures only noncognitive skills.

The results from our cross-country growth regressions indicate that both components of test scores have a positive and statistically significant association with economic growth. In fact, their estimated effects are very similar in terms of magnitude. Moreover, we find that the effect of cognitive skills is approximately 40 percent smaller when we control for noncognitive skills, suggesting that noncognitive skills are important for explaining the relationship between test scores and economic growth.

Reverse causality is an obvious concern if one is interested in putting a causal interpretation on the results of macroeconomic growth regressions. In our study, we address the issue of reverse causality by applying the decomposition method to an international test administered in 1991. We find that our results are consistent with an effect of skills on growth and not vice versa.

To our knowledge, there are only two recent studies investigating the relationship between personality traits (in the form of average measures of patience per country) and productivity at the macroeconomic level (Dohmen et al., 2016; Hübner and Vannoorenbergh, 2015). Our main contribution is to test whether the well-studied relationship between economic growth and test scores is mediated by noncognitive skills and to provide estimates of the effect of noncognitive skills. The lack of works studying the relationship between noncognitive skills and economic prosperity at the aggregate level contrasts with the abundance of studies at the individual level. One reason for this might be the lack of international comparable measures for noncognitive skills. Most studies on personality traits rely on self-reports of individuals, which complicates international comparisons. Performance based measures have the advantage that they do not suffer from the typical measurement issues related to self-reports, such as reference bias (e.g. Paulhus, 1984; Kautz et al., 2014).

This study is organized as follows. Section 5.2 discusses the previous literature on the effect of cognitive skills on economic growth and the recent literature on the importance of noncognitive skills. Section 5.3 and 5.4 explain the PISA-decomposition and the estimation of the cross-country growth regressions. The data used in the analyses are described in Section 5.5. Section 5.6 shows the main estimation results. Section 5.7 investigates the robustness of the results to using stricter measures of the performance decline and addresses concerns of reverse causality. Section 5.8 concludes.

³Phelps (2006) argues that the mechanisms of emotion and cognition are intertwined from early perception to reasoning and Cunha and Heckman (2008) and Cunha et al. (2010) show noncognitive skills are important for the development of cognitive skills, but not vice versa.

5.2 Previous Studies

5.2.1 The Relationship between Cognitive Test Scores and Economic Growth

A large empirical literature has studied the impact of human capital on economic growth. One of the major challenges is to find a good proxy for human capital. As a consequence, many studies have used average educational attainment as their measure (see, for example, Barro, 1991; Krueger and Lindahl, 2001; Sala-i Martin et al., 2004; Doménech and De la Fuente, 2006; Cohen and Soto, 2007; Sunde and Vischer, 2015). However, this proxy seems quite imperfect as it assumes that a year spent in school produces the same amount of human capital across all countries. Therefore, Lee and Lee (1995) and Hanushek and Kimko (2000) introduced a new approach that uses the performance on international cognitive tests as a proxy for human capital. The main advantage of this approach is that cognitive test scores can be considered as an output measure that captures what students have learned inside and outside of school. The basic cross-country growth specification in Hanushek and Kimko (2000) regresses the average economic growth of country c (G_c) for a specific period on their measure of human capital (H_c), GDP per capita at the beginning of the period (GDP_{0c}) and control variables (Z_{nc}) such as years of schooling and population growth:

$$G_c = \beta_0 + \beta_1 H_c + \beta_2 GDP_{0c} + \sum_n \delta_n Z_{nc} + \epsilon_c \quad (5.1)$$

This approach has been extended in a series of studies, which estimate Equation (1) and have very similar results and interpretation (see Barro, 2001; Hanushek and Woessmann, 2008, 2011b,c, 2012; Hanushek, 2013; Bosworth and Collins, 2003; Jamison et al., 2007). Equation (5.1) is consistent with the endogenous growth models of Romer (1990) and Nelson and Phelps (1966). In these models, growth is attributed to the stock of human capital, which generates innovations or facilitates the adoption and imitation of new technologies. We focus on the most recent paper, HW (2012), which uses data on economic growth from 50 countries for the period 1960–2000 and cognitive test scores for the period 1964–2003. The authors find consistent evidence that cognitive test scores are strongly associated with economic growth and interpret this as indicating the importance of cognitive skills. The estimated effects of cognitive skills are large: a one standard deviation increase in test scores is associated with 1.25 to 2 percentage points higher average annual growth rate in GDP per capita across 40 years.

An important question is to what extent the association between test scores and economic growth reflects a causal effect of cognitive skills on economic performance. This is a difficult question because it is very hard to address typical identification issues like omitted variables, reverse causality

and measurement error. However, HW (2012) show that the estimated effects of cognitive test scores on economic growth are robust to alternative estimation approaches, such as instrumental variables, differences-in-differences and longitudinal analysis of changes in cognitive test scores and in growth rates. Moreover, they argue that international test scores are not driven by differences in resources across countries and note that their estimation relies upon the assumption that the average scores for a country tend to be relatively stable over time. This leads them to conclude that differences in cognitive skills lead to economically significant differences in economic growth.

Although several studies find a consistent positive relationship between cognitive test scores and economic growth, a growing literature highlights the impact of noncognitive skills on test performance, making it difficult to know how these results should be interpreted.

5.2.2 Noncognitive skills, Long-term Individual Outcomes and Cognitive Test Scores

Many studies in psychology and a more recent literature in economics have established the importance of noncognitive skills for individual socioeconomic outcomes. Noncognitive skills are defined as relatively enduring patterns of thoughts, feelings and behaviors that reflect the tendency to respond in certain ways under certain circumstances (Roberts, 2009). These studies often use the Big Five inventory as measures of noncognitive skills (Costa and McCrae, 1992; John and Srivastava, 1999) and find that these measures are as predictive as cognitive measures for important outcomes such as schooling, wages, crime, teenage pregnancy, and longevity, even after controlling for family background and cognition (see for example Mueller and Plug, 2006; Heckman et al., 2006; Almlund et al., 2011; Heckman, 2008). Intervention studies, like the Perry Pre School Program, provide evidence for a causal effect of changes in noncognitive skills on economic and social outcomes (Heckman et al., 2013). Further evidence on the importance of noncognitive skills for individual economic success can be found in Heckman and Rubinstein (2001), Borghans et al. (2008a), Heckman and Kautz (2012) and Kautz et al. (2014).

Noncognitive skills have also been related to the performance of students on cognitive tests. The possibility that test takers might not exert maximal effort has been largely recognized by researchers on intelligence testing. For instance, Wechsler (1940) noted that intelligence tests not only measure intelligence and pointed out that the tendency to try hard on low stakes intelligence tests might derive from non-intellective traits, such as competitiveness and compliance with authority. More recently, Duckworth et al. (2011) provide evidence for the role of test motivation in intelligence testing. Observer ratings of test motivation, based on the behavior of adolescent boys completing intelligence tests, explains IQ-scores and reduces the predictive validity of IQ-scores for life outcomes, particularly for nonacademic outcomes. Their findings show that under low-stakes testing conditions some

individuals try harder than others. Economists have also recognized that engaging in complex thinking is effortful and therefore motivation to exert effort affects the performance on achievement tests. For example, in Borghans et al. (2008b) subjects were given questionnaires to determine psychological traits and were asked to make trade-offs to determine relevant economic preference parameters. They found that preferences have a direct impact on cognitive test scores.⁴ Moreover, various studies have found that offering a material reward can substantially improve scores on cognitive tests (Gneezy and Rustichini, 2000; Segal, 2012).

These findings have motivated using answering patterns to obtain measures of noncognitive skills that do not rely on self-reports. In addition to Borghans and Schils (2012), Hernández and Herschaff (2014) propose using skipped items in a non-penalized test to measure noncognitive skills; Hitt (2016), Zamarro et al. (2018) and Zamarro et al. (2017) explore careless answering patterns in survey questionnaires; and Hitt et al. (2016) find that skipped questions at six nationally-representative, longitudinal surveys of American youth are a significant predictor of later-life educational attainment net of cognitive ability. The use of self-reports to measure noncognitive skills has been challenged (Duckworth et al., 2011; Duckworth and Yeager, 2015; Paulhus, 1984; West et al., 2016). By relying on Borghans and Schils (2012), we can avoid the problems associated with self-reports and simultaneously measure cognitive and noncognitive skills.

5.3 The Test Score Decomposition

Borghans and Schils (2012) developed an approach to decompose test scores into two elements: the starting performance and the decline in performance during the test. They observed that students perform worse on questions that appear later in the test. Because knowledge should be the same at the beginning and end of the test, they attribute the decline in performance during the test to motivation, which can be thought of as a noncognitive skill. One concern with this interpretation is that the performance decline might be related to unobservable characteristics, such as the difficulty of the test items.⁵ If this were the case, the performance decline would be a consequence of cognitive skills rather than noncognitive skills.

To address this important issue, Borghans and Schils (2012) exploit the variation in the question ordering of the PISA test. As shown in Table 5.1, PISA 2006 has 13 different versions of the test (booklets), all of them containing four clusters of questions (test items). A booklet contains approxi-

⁴This finding is consistent with Borghans et al. (2011) and Heckman and Kautz (2012), who find that personality variables explain roughly a third of explained variance in achievement tests.

⁵In fact, the sequencing of items from easy to difficult is used as an explicit strategy for sustaining morale (Duckworth et al., 2011).

Table 5.1: Rotation design of the 13 PISA booklets

Booklet	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1	Science 1	Science 2	Science 4	Science 7
2	Science 2	Science 3	Math 3	Reading 1
3	Science 3	Science 4	Math 4	Math 1
4	Science 4	Math 3	Science 5	Math 2
5	Science 5	Science 6	Science 7	Science 3
6	Science 6	Reading 2	Reading 1	Science 4
7	Science 7	Reading 1	Math 2	Math 4
8	Math 1	Math 2	Science 2	Science 6
9	Math 2	Science 1	Science 3	Reading 2
10	Math 3	Math 4	Science 6	Science 1
11	Math 4	Science 5	Reading 2	Science 2
12	Reading 1	Math 1	Science 1	Science 5
13	Reading 2	Science 7	Math 1	Math 3

Source: OECD (2009)

mately 60 test items. Each cluster of questions represents 30 minutes of test time, which means each student undertakes two hours of testing. Students are allowed a small break after one hour, typically shorter than 5 minutes, where they are allowed to stand up and stretch. There are 13 clusters of test items (7 science, 2 reading and 4 math) and they are distributed over the 13 different booklets according to a rotation scheme. Each cluster appears in each of the four possible positions within a booklet once (OECD, 2009). This means that one specific test item appears in four different positions of four different booklets. For instance, cluster Science 1 is included in booklets 1, 9, 12 and 10 as respectively the first, second, third and fourth cluster. This rotation scheme generates exogenous variation in the question number (position in the test) of test items because the booklets are randomly assigned to students (OECD, 2009). In other words, the random assignment of booklets ensures that the positioning of questions is unrelated to student characteristics. The results of balancing tests are consistent with random assignment. Table A.5.1 shows estimates from separate regressions of background characteristics on booklet indicators. Almost all of the coefficients of these indicators are statistically insignificant at conventional levels and the F-tests for joint significance never reject the null hypothesis.

The exogenous variation in question position can then be exploited to estimate the decline in performance during the test by using the following fixed effects model:

$$P[Y_{ij} = 1] = F(\alpha_0 + \alpha_1 Q_{ij} + \sum_{j=2}^J \mu_j) \quad (5.2)$$

where Y_{ij} is an indicator for whether student i answered question j correctly, Q_{ij} is the position of question j in the version of the test answered by student i and μ_j is a question fixed effect that takes account of unobservable characteristics of question j such as difficulty. Conditional on question fixed effects, the variation we are exploiting lies within a question across different students. As such, the identifying assumption becomes random variation in the position of a question across different students. This assumption is met due to the random allocation of booklets to students: we are comparing the performance of identical students on the same question in four different positions. The estimated parameter α_1 will not be biased by unobserved factors and can be interpreted as the decline in performance during the test. The decomposition of the test scores into the starting performance and the performance decline is based on the estimation of Equation (5.2). We estimate Equation (5.2) separately for each country by using a probit model and use the PISA weighting factors to ensure that the sample is representative.⁶ The parameter α_0 measures the starting performance of a specific country, because the question numbers have been rescaled such that the first item is numbered as 0 and the last item as 1. Both components are robust to the definition of the start of the test. For instance, including the first five questions in the starting performance does not affect the estimates of the two components. We use all test items for estimating Equation (5.2). Skipped and non-reached items were coded as incorrectly answered questions. This allows us to stay closer to the framework of HW (2012) in which uncompleted items were interpreted as incorrectly answered to compute final test scores.⁷

We have also estimated Equation (5.2) using the average performance on all test items within a cluster as the outcome variable. In this analysis the clusters have been rescaled such that the first cluster is numbered as 0 and the fourth cluster is numbered as 1. With this approach the unit of randomization exactly matches the unit of analysis. The results are very similar to the results from our main approach. We find a correlation of 0.936 for the estimates of the starting performance of the two approaches, and a correlation of 0.964 for the decline in performance.

5.3.1 Interpretation of the Two Components

The main purpose of the decomposition is to generate two components that capture both types of skills relatively well. The performance decline is a measure of noncognitive skills, where the starting

⁶Estimating Equation (5.2) with OLS gives very similar results. In fact, the correlation of the components estimated with probit and OLS equals 0.996 for the starting performance and 0.969 for the performance decline. Notice that despite using a probit model with question fixed effects, the incidental parameter problem does not apply. The number of fixed effects to be estimated (questions) remains constant when increasing the number of observations (students).

⁷Borghans and Schils (2012) note that it is unclear which type of skills determine that test items are not reached. In Section 5.7 we will investigate the sensitivity of our results to alternative ways of dealing with non-reached questions.

performance provides a measure of cognitive skills that, differently from test scores, is not confounded by the personality factors that cause the decline in performance.

Conceptually, obtaining a clean measure of cognitive skills is difficult. Cunha and Heckman (2008) and Cunha et al. (2010) show that noncognitive skills positively affect the accumulation of cognitive skills during childhood. Moreover, in our case it might be that the performance at the start of the test is influenced by ex-ante motivation as well. Although the starting performance, arguably, provides a better measure of cognitive skills than the final test scores, we cannot rule out that it is fully devoid of noncognitive skills. The consequence would be an attenuated estimate of the effect of noncognitive skills in the growth regressions.

This does not challenge, however, our investigation on whether noncognitive skills are important for the relationship between test scores and economic growth. For this purpose we only need a measure of noncognitive skills that is not contaminated with cognitive skills. Therefore it is conceptually important that cognitive skills do not affect the accumulation of noncognitive skills (Cunha and Heckman, 2008; Cunha et al., 2010). With regard to our particular measure, let us label the possibility that “*the performance decline captures noncognitive skills but does not capture cognitive skills*” interpretation A, while interpretation B admits the possibility that “*the performance decline captures both cognitive and noncognitive skills*”. Our efforts are concentrated towards obtaining a measure of noncognitive skills that fits with interpretation A. The arguments provided below are consistent with the performance decline providing such a measure.

Borghans and Schils (2012) provide four arguments in support of interpretation A. First, the performance decline differs from the students’ performance at the beginning of the test, which indicates that the two components measure different types of skills. Second, they show that the two components are stable for the years 2003 and 2006 and that there are differences between countries. This suggests that the two components are able to measure stable traits of the 15-year-old population of a country. Third, they show that the performance decline is related to specific noncognitive skills. With the data collected in the Dutch Inventaar 2010 study they find that students with higher levels of agreeableness (a Big Five personality trait), ambition and motivation towards learning have a smaller performance decline. Fourth, using data from the British Cohort Study 1970, they show that the performance decline predicts future outcomes above and beyond the pure test score.

Additional evidence comes from Balart and Oosterveen (2017). These authors noted that girls typically score better on reading tests than boys, but perform worse in science and math (Hyde and Linn, 1988; Hyde et al., 1990; Caplan et al., 1997; Kimura, 2004; Dee, 2007; Fryer Jr and Levitt, 2010; Cornwell et al., 2013; Quinn and Cooc, 2015). They argued that if the performance decline were in fact induced by cognitive skills, then we should observe girls experiencing a less pronounced decline

in reading, while boys would have a less pronounced decline when answering math and science questions. Balart and Oosterveen (2017), however, found the opposite: girls exhibited a less pronounced decline than boys in both reading and in math/science. Specifically, using the PISA 2009, they find that in 66 (62) out of the 74 countries, girls perform better (worse) in reading (math-science) than boys at the start of the test. However, there is no single country in which boys exhibit a statistically significant lower decline in performance than girls either in reading or in math-science. Girls exhibit a less pronounced decline than boys in 68 countries for reading (statistically significant for 40) and in 68 countries for mathematics and science (statistically significant for 46). A smaller performance decline independent of one's ability in a topic strongly supports the argument that the decline is not driven by cognitive skills. Moreover, it is consistent with gender differences in noncognitive skills found in previous research.⁸

Another element in support of interpretation A arises from the growing body of research that has used the decomposition strategy proposed by Borghans and Schils (2012). Using an epidemiological approach, Rodriguez-Planas and Nollenberger (2018) show that gender differences in the starting performance are related to gender equality of the country of origin of second generation immigrants while the same is not true for gender differences in the performance decline. They interpret that gender gaps in test scores are affected by social gender norms through cognitive skills rather than noncognitive skills. Zamarro et al. (2016) show that, at the country level, the performance decline is associated with other non-self reported measures of student effort such as careless answering patterns and non-response in the student background questionnaire after PISA. The non-challenging nature of filling out a questionnaire makes it unlikely that this is driven by cognitive skills (Hernández and Herschaff, 2014; Hitt, 2016; Hitt et al., 2016). As they argue, this is strengthened by the fact that careless answering patterns do not exhibit a higher correlation with test scores in reading than with non-reading ones.

Finally, following Linton (1945), Hofstede and McCrae (2004) and Benet-Martínez and Oishi (2008), we make use of the similarities between culture and noncognitive skills. Similar to noncognitive traits, culture is defined in terms of behavior and is transmitted from generation to generation.⁹ Mendez (2015) directly associates culture with differences in noncognitive skills and exploits cultural variations in second-generation immigrants to show that differences in cultural values and accompanying noncognitive skills are related to differences in PISA test scores. By contrast, cognition

⁸Girls are found to have more self-discipline (Duckworth and Seligman, 2006), have less behavioral problems (Jacob, 2002), to be less prone to overconfidence (Niederle and Vesterlund, 2007), show more developed attitudes towards learning (Cornwell et al., 2013) and report higher levels of extraversion, agreeableness, and conscientiousness (Schmitt et al., 2008).

⁹For instance, Guiso et al. (2006) or Fernandez and Fogli (2009) define culture as customary beliefs, values and actions that social groups transmit fairly unchanged from generation to generation. Intergenerational transfers of noncognitive skills is argued by Heckman (2008), he shows enhancements of family environments (socioemotional nurturing) improve child outcomes, of which personality traits are the most important channel.

Table 5.2: Regressions of Hofstede's cultural dimensions on the starting performance and performance decline

	(1) Power Distance	(2) Individualism	(3) Masculinity	(4) Uncertainty Avoidance	(5) Long term Orientation	(6) Indulgence
Starting performance	-1.626 (-0.33)	5.818* (1.81)	0.751 (0.19)	2.579 (0.48)	6.406 (1.37)	0.644 (0.16)
Performance decline	2.426 (0.89)	1.759 (0.70)	0.303 (0.11)	-7.516** (-2.10)	11.37*** (4.68)	-9.712*** (-3.37)
<i>N</i>	53	53	53	53	56	55
Adj. <i>R</i> ²	0.296	0.469	-0.056	0.165	0.255	0.225
Std.Dev. <i>Y</i>	22.23	24.19	20.58	53.28	21.97	19.86

Notes: *t* statistics in parentheses, heteroskedasticity robust standard errors

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Regressions include a constant and initial GDP per capita. The starting performance and performance decline are standardized with mean 0 and standard deviation 1.

does not have a strong link with culture. Indeed, psychologists distinguish two types of second-order factors of cognitive ability: fluid intelligence and crystallized intelligence (Cattell, 1987). Only the second is partially influenced by culture.¹⁰ We provide additional evidence on the difference between the two components by regressing the cultural dimensions of Hofstede and Hofstede (2001) upon the standardized starting performance and performance decline in Table 5.2.¹¹

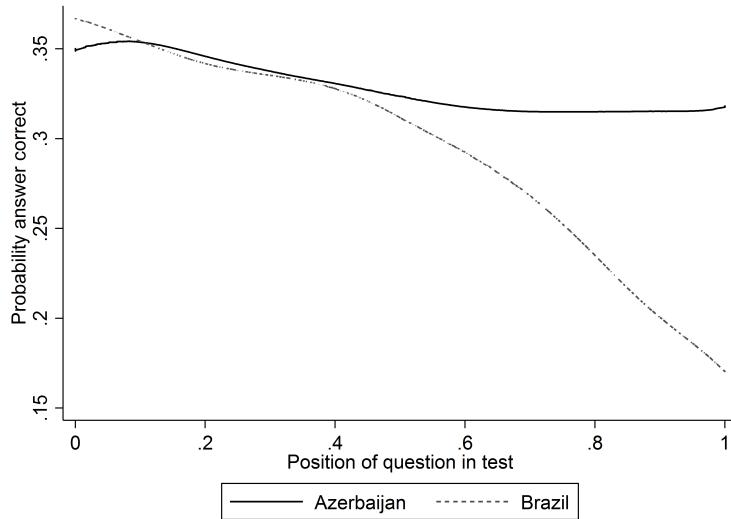
We find that the performance decline has a stronger association with the cultural values than the starting performance. Whereas the starting performance is only significantly related to higher levels of individualism at the 10%-level, the performance decline shows strong associations with measures of uncertainty avoidance, long term orientation and indulgence.¹² These results suggest that the performance decline is smaller in countries with: (i) more thriftiness, perseverance for achieving results and higher efforts in modern education (long term orientation), (ii) less preference for leisure-time, more control on the gratification of desires and stricter social norms (indulgence) and (iii) more positive preference towards uncertainty, ambiguity and curiosity (uncertainty avoidance).¹³ In sum, this can be interpreted as evidence that the performance decline is related to motivation, thriftiness, and less-preference towards leisure and certainty which are related to time- and risk-preferences.

¹⁰For instance, the Raven Progressive Matrices Test, a test commonly used to measure fluid intelligence, is referred to as “culture-free”

¹¹We use PISA 2009 to maximize the number of countries. Using the other waves of PISA does not change our results. We include initial GDP per capita in 1990 to control for economic development. Controlling for an OECD indicator or GDP per capita in 1960 gives almost identical results, but controlling for GDP per capita in 1960 sharply decreases our sample size. The components are standardized with mean 0 and standard deviation 1.

¹²Interestingly, Gorodnichenko and Roland (2016) provide evidence for a causal effect of individualism on long run growth. Gorodnichenko and Roland (2011) show that individualism is the only measure of culture, among the ones they consider, that has a robust effect on growth.

¹³See the Appendix for explanations on the six cultural dimensions of Hofstede and Hofstede (2001).

Figure 5.1: The decline in performance for Azerbaijan and Brazil

Notes: The figure is based upon PISA 2006 and uses LOWESS to visualize the relationship between Y and Q of Equation (5.2) with a bandwidth of $0.8N$, the program default. We do not use PISA weights or question fixed effects for the computation of this figure.

Remarkably, these are a large subset of the noncognitive skills that Heckman (2008) lists as being related to earnings, employment, college attendance, and other socioeconomic outcomes.

5.3.2 Differences between Countries and Years

Results of the decomposition of the PISA test of 2006 are shown in Table A.5.2. Equation (5.2) is used for computing the probability of correctly answering the first and the last question of the PISA test. Column (1) shows the average of the PISA 2006 test scores, column (2) shows the probability of correctly answering the first question, column (3) shows the probability of correctly answering the last question and column (4) shows the difference between these two probabilities. Column (2) can be interpreted as the starting performance of a country and column (4) can be interpreted as the performance decline. Countries are ranked with respect to the latter from high to low.

We observe that there are large differences between countries. Columbia and Uruguay have the largest decline in performance. That is, their probability to answer the last question correctly is 30 percentage points lower than their probability to answer the first question correctly. Within the top ten of countries with the largest decline we observe six countries from South America. Among the countries with the least pronounced performance declines, we observe Northern European and Asian countries. Moreover, Table A.5.2 indicates that these differences between countries are important

Table 5.3: Correlations between the starting performance and performance decline for PISA 2003, 2006 and 2009

	Starting performance 2003	Decline 2003	PISA 2006	Starting performance 2006	Decline 2006	Starting performance 2009	Decline 2009
Starting performance 2003	1.000						
Decline 2003	0.426	1.000					
PISA 2006	0.860	0.745	1.000				
Starting performance 2006	0.967	0.521	0.920	1.000			
Decline 2006	0.526	0.947	0.716	0.584	1.000		
Starting performance 2009	0.950	0.480	0.856	0.912	0.501	1.000	
Decline 2009	0.539	0.911	0.760	0.631	0.923	0.462	1.000

Notes: The components are estimated using Equation (5.2) with PISA weights.

for the total test score. We observe that Azerbaijan and Brazil have a very similar starting performance. However, the decline in performance for students in Brazil is much larger than for students in Azerbaijan. This translates into a difference on the PISA test of more than 19 points. This difference is shown in Figure 5.1, where we use locally weighted scatterplot smoothing to visualize the performance decline and the starting performance for the two countries. This flexible nonparametric method also suggests that, without putting assumptions on the process that generated the data, the linear specification in Q_{ij} used in Equation (5.2) seems to be a good approximation.

We have also decomposed the test scores for PISA 2003 and 2009 using the same procedure as was used for the PISA 2006. Table 5.3 shows the correlations between the different components for the three years. The correlations of the estimated starting performances (performance declines) over time are shown in bold. All of them are above 0.91. As indicated by Borghans and Schils (2012), a high correlation between the starting performances (performance declines) over the years suggests that these components capture some of the traits of the 15-year-old population of a country. Consistent with the literature on noncognitive skills, the correlation between the starting performance and the performance decline is much lower, which indicates that the two components measure different traits.

5.4 Estimation of the Relationship between Skills and Economic Growth

The starting point of our empirical analysis of the effect of skills on economic growth is the standard cross-country growth regression as shown by Equation (5.1). The main previous studies aggregate scores from all available international cognitive tests and use this as a measure for cognitive skills (see Section 5.2.1). We label the aggregate test score from HW (2012) as the HW-index. In this study we decompose the scores on an international cognitive test into the starting performance (S_c) and the performance decline during the test (PD_c). Therefore, instead of using test scores as a unidimensional proxy for human capital (H_c), we use the two components: $H_c = f(S_c, PD_c) + \nu_c$. We include these

two components into the cross-country growth regression to re-estimate Equation (5.1):

$$G_c = \beta_0 + \beta_1 S_c + \beta_2 PD_c + \beta_3 GDP_{0c} + \sum_n \delta_n Z_{nc} + \epsilon_c \quad (5.3)$$

When estimating Equation (5.3), we try to stay as close as possible to HW (2012). We use the same data on economic growth, the same growth period (1960–2000), identical covariates, estimate the same model specifications, and start with the same sample of countries. However, the decomposition method that we apply in this paper exploits a specific feature of the PISA test, namely the random allocation of the PISA booklets (see Section 5.3). Hence, we can apply the decomposition method only to one of the tests included in the HW-index. This has two implications for the estimations. First, the sample of countries that participated in the PISA test differs from the sample used in HW (2012). As a first step in our analysis, we check whether the reduction in sample size from 50 to 37 countries, due to our reliance on the PISA, changes the results obtained in HW (2012). The second implication is that we use the PISA test only for measuring skills, and not the complete set of tests used for the HW-index. However, the PISA scores are highly correlated with the HW-index ($r = 0.91$). Further we will show below that re-estimating the main models from HW (2012) with PISA scores instead of the HW-index delivers very similar results. This suggests that PISA scores are a good proxy for the HW-index and, therefore, we use the PISA scores for estimating Equation (5.1). Next, we decompose these PISA scores into the two components and we use these two components for estimating Equation (5.3). We estimate Equation (5.3) with OLS and report robust standard errors. This analysis naturally relies upon the assumption that the distribution of both the starting performance and the performance decline across countries remained relatively stable over time, which is supported by Table 5.3.

As we are using a two-step estimation approach it could be argued that the standard errors should be adjusted because the regressors are not fixed (see e.g. Murphy and Topel, 2002). However, due to the large number of observations used in the estimation of Equation (5.2), which is the number of students times the number of test items, the estimates for the starting performance and the performance decline are very precise, and can be considered as fixed (see Table A.5.3 for the standard errors of the two components and the number of students participating in PISA 2006 per country).¹⁴

¹⁴The maximum likelihood estimation of Equation (5.2) gives us consistent estimates. Since the number of observations is large, we can be confident that the ML-estimates have reached their true values. For computational tractability, standard errors in Table A.5.3 are computed using sample PISA weights but not their 80 replicates. Using the 80 replicates does not substantially increase the size of standard errors. For instance, the standard error of the starting performance and the performance decline of Iceland increase from 0.0437 to 0.0476 and from 0.0101 to 0.0132, respectively.

5.5 Data

The data used in the analysis come from various sources. The HW-index is from HW (2012). This index aggregates all available math, science and reading scores from international cognitive tests that took place in the period 1964-2003 for 50 countries.¹⁵

In addition, we use data collected by the Programme for International Student Assessment. PISA is a triennial international survey which aims to evaluate education systems worldwide by testing the skills and knowledge of 15-year-old students. The key subjects of the test are reading, science and math. The first PISA study took place in 2000. The method for decomposing test scores into a cognitive and a noncognitive component is applied for countries that participated in PISA 2003, 2006 and 2009.¹⁶ We start our analysis with PISA 2006 which allows us to include 37 countries that were included by HW (2012). We standardize the decomposed test scores and the total PISA score separately to set the mean and standard deviation equal to the HW-index, allowing us to directly compare the size of our estimates to those of the HW-index.

We follow HW (2012) for sources on the other data. Real GDP per capita comes from version 7.1 of the Penn World Tables (Aten et al., 2009).¹⁷ Data on years of schooling are taken from the most recent version of the Barro and Lee dataset (Barro and Lee, 2013, version 2.1). Further control variables used by HW (2012) are regional indicators and two proxies for the quality of economic institutions: openness of the economy and protection against expropriation. For the regional indicators we follow the classification of HW (2012). The measure of openness is the Sachs et al. (1995) index reflecting the fraction of years between 1960 and 1992 that a country was classified as having an economy open to international trade.¹⁸ For the data on protection against expropriation Acemoglu et al. (2001) is followed, the measure is an index between 0 and 10 averaged over 1985-1995. A higher score on this index means that there is more protection against expropriation. Two other controls that are used are fertility, obtained from World Development Indicators (World Bank, 2012), and tropical location measured as the proportion of a countries' area located in the tropics (Gallup et al., 1999). Table A.5.3 provides the data per country on GDP growth, the HW-index and the two components of the PISA test for the sample of 37 countries used in Section 5.6.2.

¹⁵See the Appendix of HW (2012) for further details on the computation of this measure.

¹⁶We choose not to use the two most recent PISA waves (2012 and 2015) because fewer countries participated in these waves and to mitigate concerns regarding reverse causality.

¹⁷Real GDP per capita for Tunisia was not available for 1960, so we used data from 1961 onwards.

¹⁸Because Romania was not available in Sachs et al. (1995), we used Romanian data from Sachs and Warner (1997) for the period 1965-1990.

5.6 Main Estimation Results

This section shows the main estimation results in three steps. First, we replicate the main analysis of HW (2012) for the sample of countries for which it is possible to decompose the PISA test. Second, we include the two components from the decomposition in the main estimation models.¹⁹ Third, we repeat the latter analysis and extend the sample towards 55 countries.

5.6.1 Replication of Previous Cross-Country Growth Regressions using PISA

In the first step of our analysis we check whether the estimation results obtained by HW (2012) change when we use scores of PISA 2006 instead of the HW-index. The results could, in theory, change because we are going from 50 to 37 countries, or because we use the PISA score instead of the HW-index. To show that none of these changes drive our results, we replicate the main models from HW (2012) using the sample of 37 countries. Panel A of Table 5.4 shows the results from models that use the HW-index, Panel B shows the results when using the PISA 2006 scores.

Panel A of Table 5.4 shows that the results for the growth regressions with the HW-index for the restricted sample are very similar to the results for the unrestricted sample in Table 1 of HW (2012). Column (1) shows the effect of years of schooling on economic growth. The estimated effect is statistically significant and suggests that an additional year of schooling increases the average annual growth rate in GDP per capita across 40 years with 0.2 percentage point. Column (2) shows the results from a model in which the HW-index is used as a proxy for human capital instead of years of schooling. A one standard deviation increase in cognitive test scores is associated with 2.3 percentage point increase in the annual growth rate of GDP per capita over 40 years. Similar to what HW (2012) found, replacing years of schooling with cognitive test scores increases the explained variance from one to three quarters. In column (3), we report results from a model that includes both proxies of human capital. The estimate of cognitive test scores is similar to that in column (2), but the estimated coefficient of years of schooling is no longer statistically significant. In columns (4)–(9) we report estimates from alternative specifications of the model; column (4) uses average years of schooling over the period 1960–2000 instead of the years of schooling in 1960, column (5) controls for outliers, column (6) includes eight regional indicators, column (7) includes measures for the openness of the economy and protection of property rights, column (8) adds fertility and tropical location as additional controls and column (9) controls for GDP per capita in logs instead of levels. These various estimates

¹⁹We start this analysis using test scores from the PISA 2006, but our results do not change when we use the PISA 2003, although the number of countries does decrease to 31. Using the PISA 2009, the sample increases to 40 countries and the results remain qualitatively unchanged.

Table 5.4: Growth regressions with the HW-index and PISA scores using the PISA sample

	(1)	(2)	(3)	(4) ^a	(5) ^b	(6) ^c	(7) ^d	(8) ^e	(9) ^f
Panel A: HW-index as a measure of human capital with restricted sample									
HW-index		2.256*** (9.15)	2.260*** (9.22)	2.310*** (9.14)	2.186*** (9.59)	1.144** (2.71)	1.399*** (3.74)	1.378*** (3.74)	2.213*** (9.92)
Years of schooling	0.187* (1.80)		-0.00375 (-0.05)	-0.0320 (-0.49)	-0.0661 (-1.02)	0.0420 (0.51)	0.0582 (0.85)	0.0115 (0.15)	-0.0250 (-0.30)
<i>N</i>	37	37	37	37	37	37	36	36	37
Adj. <i>R</i> ²	0.208	0.730	0.722	0.723	0.809	0.784	0.770	0.799	0.717
Panel B: PISA 2006 as a measure of human capital with restricted sample									
PISA 2006		2.282*** (7.98)	2.245*** (7.59)	2.235*** (6.84)	2.223*** (7.16)	1.181*** (3.53)	1.265*** (2.83)	1.220** (2.74)	2.299*** (9.73)
Years of Schooling	0.187* (1.80)		0.0426 (0.53)	0.0316 (0.38)	-0.0241 (-0.28)	0.0654 (0.84)	0.104 (1.46)	0.0526 (0.63)	0.0305 (0.35)
<i>N</i>	37	37	37	37	37	37	36	36	37
Adj. <i>R</i> ²	0.208	0.700	0.694	0.692	0.691	0.803	0.754	0.781	0.728

Notes: *t* statistics in parentheses, heteroskedasticity robust standard errors

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Dependent variable: average annual growth rate in GDP per capita, 1960-2000

Regressions include a constant and GDP per capita in 1960

^a Measure of years of schooling refers to the average over the period 1960-2000

^b Controlling for outliers by using *rreg* command in Stata

^c Includes indicators for the eight world regions

^d Controlled for openness of economy and protection against expropriation

^e Controls in ^d plus fertility and tropical location

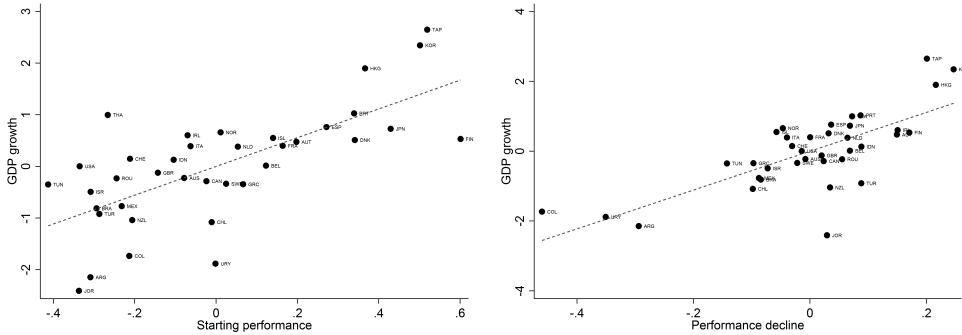
^f GDP per capita 1960 measured in logs

confirm that cognitive test scores are associated with economic growth and the results for our sample of 37 countries are very similar to the results for the full sample used by HW (2012).

In Panel B of Table 5.4 we show estimates of models that use PISA 2006 scores instead of the HW-index. We find that the estimated effects are very similar to those in Panel A. In fact, the estimates for the PISA scores are always within the 95% confidence interval of the estimates for the HW-index, which can be explained by the high correlation ($r = 0.91$) between the PISA scores and the HW-index. This indicates that PISA 2006 is a good proxy for the HW-index in models that explain differences in economic growth between countries.

In sum, we find that the results obtained by HW (2012) are robust to using the sample of countries participating in PISA 2006 and to using PISA scores instead of the HW-index, suggesting that, within the framework of HW (2012), we can use the PISA scores as a proxy for the HW-index.

Figure 5.2: The association between the conditional starting performance and the conditional decline in performance with economic growth for the period 1960-2000



5.6.2 The Relationship of the Starting Performance and the Performance Decline with Economic Growth

In this section, we present the main estimation results of models that include the two components obtained by decomposing the PISA scores. Figure 5.2 provides a first impression of the relationship between these two components and economic growth, conditional on initial GDP and years of schooling. The left panel shows a positive association between the starting test score level and GDP growth. However, the right panel shows a very similar association between the decline in performance and GDP growth, which suggests that noncognitive skills are also related to economic growth. Below, we confirm that the association is not solely driven by the three Asian countries in the upper right corner and the three Southern American countries in the lower left corner of Figure 5.2.

Table 5.5 replicates the model from Table 5.4 using the starting performance and the decline in performance as the main explanatory variables. Columns (1) and (2) show estimates of the relationships presented in Figure 5.2. The starting performance is positively and significantly associated with economic growth. The estimated effect is somewhat smaller than the previous estimate from the model that uses the PISA score in Table 5.4, suggesting that the starting performance is less confounded by personality factors than the PISA score.

The performance decline is also positively and significantly associated with economic growth. Moreover, the size of this association is quite similar to that obtained for the starting performance. A comparison of columns (1) and (2) also reveals that years of schooling is associated with economic growth only in column (2). Years of schooling is more highly correlated with the starting performance ($r = 0.64$) than with the performance decline ($r = 0.38$), perhaps indicating that the latter also

Table 5.5: Regressions of economic growth on the starting performance and performance decline

	(1)	(2)	(3)	(4) ^a	(5) ^b	(6) ^c	(7) ^d	(8) ^e	(9) ^f
Starting performance	2.143*** (4.84)		1.330*** (2.76)	1.345** (2.69)	1.049*** (3.05)	0.766* (2.00)	0.701 (1.38)	0.249 (0.49)	1.934*** (4.45)
Performance decline		1.872*** (6.36)	1.300*** (4.44)	1.257*** (4.17)	1.467*** (5.14)	0.680* (1.83)	0.917*** (3.09)	1.120*** (4.95)	0.701** (2.64)
Years of schooling	0.0415 (0.53)	0.173* (1.86)	0.0871 (1.23)	0.0651 (0.96)	0.0483 (0.64)	0.0793 (1.00)	0.131* (1.93)	0.0839 (1.14)	0.0344 (0.44)
<i>N</i>	37	37	37	37	37	37	36	36	37
Adj. <i>R</i> ²	0.579	0.623	0.726	0.721	0.746	0.787	0.754	0.798	0.716

Notes: *t* statistics in parentheses, heteroskedasticity robust standard errors

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Dependent variable: average annual growth rate in GDP per capita, 1960-2000

Regressions include a constant and GDP per capita in 1960

^{a-f} See the description of Table 5.4

captures factors that are independent of what is learned at school, which is consistent with the idea that noncognitive skills are more affected by out-of-school influences than cognitive skills.²⁰

In column (3) of Table 5.5, we report estimates from a model that includes both components on the right-hand side. We find that the starting performance and performance decline are positively and significantly associated with economic growth, but these estimates are considerably smaller than those reported in columns (1) and (2). The estimated effect of cognitive skills (the starting performance), falls approximately 40 percent as compared to the estimate in column (1), implying that noncognitive skills are partially driving the relationship between test scores and economic growth found by previous studies. These results are consistent with the models of skill formation of Cunha and Heckman (2008) and Cunha et al. (2010): noncognitive skills appear to be important for the development of cognitive skills, but not vice versa. By using the starting performance rather than test scores, we are correcting for the measurement error induced by noncognitive skills in the generation of current test scores but not for the accumulated effect of noncognitive ability on the development of cognitive skills. As a consequence, once we include a measure of noncognitive skills in the growth regression model, we are explicitly accounting for the association between growth and cognitive skills for which noncognitive skills are responsible. This explains why the coefficient of the corrected measure for cognitive skills is only marginally smaller in column (1), while it is reduced by approximately 40 percent after including noncognitive skills in the growth regressions in column (3).

In the remaining columns of Table 5.5, we report results from using the different specifications introduced in Table 5.4. In general, the results are quite robust to these sensitivity tests. Controlling for

²⁰The estimated effects become somewhat smaller but remain statistically significant at the 1% significance level if potential outliers are excluded from the analysis (Taiwan, Hong Kong, Korea, Columbia, Uruguay and Argentina).

average years of schooling over the period 1960-2000 does not change the estimates. Moreover, the results do not appear to be driven by outliers or by countries that belong to certain regions. However, when including regional indicators, noncognitive skills are only significant at the 10% level, which is consistent with the idea that cultural differences are an important determinant of noncognitive skills embedded in the performance decline. The estimated effect of the performance decline is robust to the inclusion of additional controls in columns (7) and (8). We observe that the starting performance is no longer a significant determinant of growth when controlling for the quality of economic institutions. A possible explanation for this result is that better institutions go hand in hand with better schools, capturing some of the effects of cognitive skills that the starting performance is intended to measure. Finally, controlling for the initial GDP level in logs in column (9) increases the estimated effect of starting performance and reduces the estimated effect of the performance decline. Both components, however, remain significant at conventional levels. This same pattern of results for cognitive skills was documented by HW (2012) and in Table 5.4. HW (2012) noted that specification (9) is more consistent with neoclassical growth models in which human capital affects steady-state levels of income but not growth rates.

By using PISA 2009 and the average economic growth for the period 1970-2010, we can extend our sample to 55 countries.²¹ Table 5.6 shows the results are virtually identical compared to Table 5.5.²² Not only are the estimates for both components of comparable magnitude, but the relationship between the performance decline and economic growth appears to be more robust to the inclusion of controls related to the quality of economic institutions than is the starting performance. The only difference is a marginally insignificant coefficient for the starting performance when we control for the regional dummies. Although the point estimate for the starting performance is somewhat larger than the estimated effect of the performance decline in column (6), it is less precise.

In sum, we find that both the starting performance and the performance decline are positively and significantly associated with economic growth. The estimated effects are similar in terms of magnitude, where the differences between the two components in both Tables 5.5 and 5.6 are statistically insignificant except in columns (8) and (9). The estimated effect of the performance decline is more robust to the inclusion of the quality of economic institutions than is the effect of the starting performance level. Finally, the estimate for the starting performance drops by roughly 40 percent after the

²¹PISA 2009 was the first wave where countries were allowed to include a subset of seven “easier” booklets. These seven easier booklets were identical to the standard booklets, but each had one reading cluster replaced by an easier reading cluster. We computed the performance decline on the set of 13 standard booklets that were the same for every participating country.

²²In this sample we also include China, India and Venezuela. These three countries only sample students within certain regions for PISA 2009. Results are qualitatively similar when we exclude these three countries.

Table 5.6: Regressions of economic growth on the starting performance and performance decline, maximizing our sample size

	(1)	(2)	(3)	(4) ^a	(5) ^b	(6) ^c	(7) ^d	(8) ^e	(9) ^f
Starting performance	1.674*** (3.40)		0.952** (2.61)	0.959** (2.62)	0.743* (1.98)	0.580 (1.61)	0.636 (1.28)	-0.00459 (-0.01)	1.577*** (5.02)
Performance decline		1.616*** (4.69)	1.299*** (4.32)	1.301*** (4.30)	1.287*** (4.45)	0.535* (1.73)	1.094** (2.13)	1.264** (2.55)	0.669*** (2.70)
Years of schooling	-0.00391 (-0.03)	-0.0407 (-0.43)	-0.0580 (-0.59)	-0.05298 (-0.56)	0.00447 (0.05)	0.0335 (0.34)	0.0341 (0.34)	-0.00407 (-0.04)	-0.0107 (-0.13)
<i>N</i>	55	55	55	55	55	55	47	47	55
Adj. <i>R</i> ²	0.415	0.525	0.568	0.567	0.573	0.725	0.594	0.652	0.677

Notes: *t* statistics in parentheses, heteroskedasticity robust standard errors

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Dependent variable: average annual growth rate in GDP per capita, 1970-2010

Regressions include a constant and GDP per capita in 1970

^{a-f} See the description of Table 5.4. We use average years of schooling over the period 1970-2010 and GDP per capita in 1970 in logs

inclusion of the performance decline, which implies that noncognitive skills are partly responsible for the relationship between test scores and economic growth.

5.7 Robustness Checks

In this section we perform two types of analyses. First, we focus on measurement issues and test whether our results are robust to more restrictive (and alternative) computations of the performance decline. We use our large sample because it gives us more statistical power. Second, we perform two tests of whether the observed associations reflect a relationship from skills to growth, or from growth to skills.

5.7.1 Stricter Measures of the Performance Decline

A concern with the interpretation of our previous analysis is that cognitive skills might have a direct effect on the performance decline. The correlation between the two components is 0.46, which could indicate that the performance decline is also capturing cognitive skills. We address this concern by using two stricter measures of the decline in performance. The first of these measures only exploits variation that is orthogonal to the starting performance. More precisely, we regressed the performance decline on the starting performance for all the countries participating in PISA 2009 and used the residuals of this regression as a corrected measure. As personality factors can boost the acquisition of cognition (Cunha and Heckman, 2008), the estimates obtained when using this new measure in

Table 5.7: Regressions of economic growth on components of test scores using an orthogonal measure of the performance decline

	(1)	(2)	(3)	(4) ^a	(5) ^b	(6) ^c	(7) ^d	(8) ^e	(9) ^f
Starting performance	1.674*** (3.40)		1.599*** (4.13) [4.07]	1.608*** (4.03) [3.98]	1.384*** (4.09) [2.38]	0.847** (2.52) [2.12]	1.181* (1.82) [1.82]	0.625 (0.91) [0.88]	1.910*** (5.85) [5.18]
Performance decline		1.176*** (3.55) [3.72]	1.119*** (4.32) [3.76]	1.121*** (4.30) [3.78]	1.109*** (4.45) [3.22]	0.461* (1.73) [1.52]	0.942** (2.13) [2.01]	1.088** (2.55) [2.33]	0.576*** (2.70) [2.42]
Years of schooling	-0.00391 (-0.03)	0.00206 (0.02) [0.02]	-0.0580 (-0.59) [-0.58]	-0.05298 (-0.56) [-0.55]	0.00447 (0.05) [0.04]	0.0335 (0.34) [0.30]	0.0341 (0.34) [0.33]	-0.00407 (-0.04) [-0.04]	-0.0107 (-0.13) [-0.13]
<i>N</i>	55	55	55	55	55	55	47	47	55
Adj. <i>R</i> ²	0.415	0.397	0.568	0.567	0.573	0.725	0.594	0.652	0.677

Notes: *t* statistics in parentheses, heteroskedasticity robust standard errors. Bootstrapped *z* statistics in squared brackets, based on 1000 replications

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Dependent variable: average annual growth rate in GDP per capita, 1970-2010

Regressions include a constant and GDP per capita in 1970

^{a-f} See the description of Table 5.4. We use average years of schooling over the period 1970-2010 and GDP per capita in 1970 in logs

Equation (5.3) can be thought of as a lower bound for the relationship between noncognitive skills and economic growth.

Table 5.7 shows that the results are qualitatively similar to those in Table 5.6.²³ The estimated effect of the performance decline is statistically significant in all specifications but, as a lower bound, the estimates are somewhat smaller than the corresponding estimates in Table 5.6. The effect of the starting performance is statistically significant in all but one specification and the estimated coefficients are larger than those in Table 5.6. Because the association between the starting performance and economic growth is more easily distinguished if the starting performance is uncorrelated with the performance decline, the estimates in Table 5.7 can be interpreted as an upper bound of the correlation between cognitive skills and economic growth.

Our second restrictive measure of the performance decline concerns non-reached questions. In the PISA 2009, the average number of non-reached questions per student is 1.83.²⁴ However, this number differs per country with a standard deviation of 1.65. It is unclear what factors are driving non-reached questions. They could be a consequence of cognitive and/or noncognitive skills. To take all

²³The reported *t*-statistics are based on robust standard errors, but the results do not qualitatively change if the standard errors are bootstrapped. Bootstrapped *z*-statistics are shown in square brackets. We used the bootstrap procedure for the two-step estimator as described in Cameron and Trivedi (2005). Bootstrapping is more relevant for the analysis in this section than for the analysis in Section 5.6.2 because we can only use 73 observations in the first step of the estimation, making the argument for consistency less plausible.

²⁴PISA distinguishes between non-reached and skipped test items. Non-reached questions are defined as all consecutive unanswered questions clustered at the end of test, except for the first missing answer.

Table 5.8: Regressions of economic growth on components of test scores coding non-reached questions as missing

	(1)	(2)	(3)	(4) ^a	(5) ^b	(6) ^c	(7) ^d	(8) ^e	(9) ^f
Starting performance	1.832*** (3.84)		1.528*** (3.53)	1.526*** (3.48)	1.328*** (3.11)	0.733** (2.11)	0.790 (1.26)	0.0344 (0.05)	1.912*** (6.00)
Performance decline		1.034*** (2.93)	0.616** (2.63)	0.618** (2.62)	0.548* (1.72)	0.164 (0.91)	0.484 (1.67)	0.661* (1.97)	0.306 (1.58)
Years of schooling	-0.0193 (-0.17)	0.0308 (0.30)	-0.0244 (-0.22)	-0.1792 (-0.17)	-0.000662 (-0.01)	0.0427 (0.42)	0.0760 (0.73)	0.0295 (0.26)	0.0135 (0.16)
<i>N</i>	55	55	55	55	55	55	47	47	55
Adj. <i>R</i> ²	0.454	0.354	0.486	0.486	0.417	0.717	0.553	0.593	0.669

Notes: *t* statistics in parentheses, heteroskedasticity robust standard errors

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Dependent variable: average annual growth rate in GDP per capita, 1970-2010

Regressions include a constant and GDP per capita in 1970

^{a-f} See the description of Table 5.4. We use average years of schooling over the period 1970-2010 and GDP per capita in 1970 in logs

questions into account, we coded non-reached questions in our main results as incorrectly answered. Alternatively, we can code them as missing, which has the effect of making the performance decline (α_1) weaker for all countries, whereas the starting performance (α_0) is hardly affected. The standard deviation for the performance decline also decreases by 57 percent, which makes it more difficult to find potential effects. Moreover, as noncognitive skills are potentially the cause of non-reached questions, this can be considered as another lower bound for the effect of noncognitive skills.

Recognizing this, we show results for this second stricter measure in Table 5.8. The estimated effects of the starting performance are similar in magnitude or somewhat larger than the effects shown in Table 5.6, consistent with the finding that this component is unaffected by considering non-reached questions as missing. The estimate for the performance decline is reduced in magnitude, but remains statistically significant in most specifications. The estimated coefficient of the performance decline is small and insignificant in the specification that controls for regional indicators (column (6)), which is consistent with the idea that cultural differences are important determinants of noncognitive skills. Despite the notable reduction in the estimates observed after excluding non-reached questions, this alternative measure does confirm that the relationship between test scores and economic growth is partially mediated by noncognitive skills. Table A.5.4 in the Appendix shows this conclusion does not change when combining the two restricted measures (orthogonal correction and non-reached questions as missing).

The Appendix reports results for several alternative measures of the starting performance and the decline in performance. Table A.5.5 shows the results when the two components are obtained through OLS instead of using a probit. As suggested by the correlations of 0.96 and higher of the two

components when estimated with the two different methods, the results are essentially unchanged compared to Section 5.6.2. Subsequently, in Table A.5.6, we use the probability of answering the first question correctly as the starting performance and the probability of answering the last minus the probability of answering the first question correctly to measure the performance decline. Despite being a nonlinear function of the original probit-estimates α_0 and α_1 , the results are robust to using these probabilities. Table A.5.7 reports regression results using a measure of the starting performance that incorporates performance on the first five questions, which might be considered a more accurate measure of the performance at the beginning of the test.²⁵ Results do not change when using this alternative measure of starting performance.

We have also computed the decomposition on the cluster level as explained in Section 5.3 so that the unit of analysis matches the unit of randomization and the starting performance corresponds to the first cluster of the test. Whereas the estimates reported in Table A.5.8 are qualitatively similar to those discussed in Section 5.6.2, the estimate for the starting performance (performance decline) is somewhat larger (smaller) and more (less) significant. This pattern of results can be explained by the fact that this version of the starting performance already includes the first part of the performance decline. Next, in Table A.5.9, we report estimates from using the average of the components for PISA 2003, 2006 and 2009. As both the performance decline and the starting performance are being estimated, albeit precisely, this mitigates measurement error.²⁶ We find qualitatively identical estimates to those discussed in Section 5.6.2. For a similar reason, we repeat our main analysis weighting the observations by the inverse of the standard error of the performance decline in Table A.5.10, decreasing the weight put on observations for which the performance decline is measured imprecisely. The results are qualitatively unchanged. Repeating these weighted regressions coding non-reached questions as missing confirms that the lower-bound estimate for the effect of noncognitive skills is statistically significant (see Table A.5.11). In sum, we find that the relationship between the two components and economic growth is remarkably robust and we conclude that noncognitive skills are an important mediator in the relationship between test scores and economic growth.

5.7.2 From Skills to Growth or From Growth to Skills

We have applied the decomposition strategy to the PISA test, which is mostly a post period measure. Despite the starting performance and the performance decline being stable over time and the PISA 2006 being a good proxy for the HW-index, the use of a post period measure raises obvious concerns

²⁵In particular, we set Q_{ij} of Equation (5.2) equal to zero for any item j that was ordered in any of the first five positions in the test.

²⁶We include a country in this regression if it participated in at least one of three PISA waves. Repeating this analysis with countries that participated in all three PISA waves restricts us to 31 countries, but does not change our results.

about reverse causality. In other words, our analyses may be capturing the effect of growth on the accumulation of skills. To address this issue, we test for the presence of a reversed channel from growth to skills and apply the decomposition method to an international test administered in 1991.

Most importantly, growth might provide a country with resources that are invested in human capital:

$$R_c = \eta_0 + \eta_1 G_c + \sum_n \kappa_n X_{nc} + u_c \quad (5.4)$$

$$\begin{pmatrix} S_c \\ PD_c \end{pmatrix} = \begin{pmatrix} \pi_0^S \\ \pi_0^{PD} \end{pmatrix} + \begin{pmatrix} \pi_1^S \\ \pi_1^{PD} \end{pmatrix} R_c + \begin{pmatrix} V_c^S & 0 \\ 0 & V_c^{PD} \end{pmatrix} \begin{pmatrix} \gamma^S \\ \gamma^{PD} \end{pmatrix} + \begin{pmatrix} v_c \\ \xi_c \end{pmatrix} \quad (5.5)$$

Where R_c are the (educational) resources in country c , and X_c and V_c are vectors containing control variables, where the latter could potentially be different for the starting performance versus the performance decline. Consequently, the estimate of β_1 and β_2 in Equation (5.3) could also reflect the effect of economic growth on the starting performance and the performance decline through its effect on resources. In particular, if we assume that Equations (5.3) to (5.5) only include a constant and, respectively, the starting performance, economic growth and the (educational) resources as explanatory variables, the estimate for the starting performance in Equation (5.3) equals:

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{\pi_1^S \eta_1}{1 - \pi_1^S \eta_1 \beta_1} var[\epsilon_c]}{var[S_c]}$$

However, this also shows that if π_1^S is equal to zero, it strongly reduces the concerns for reverse causality.²⁷ Previous studies failed to find consistent, strong evidence that test performance is affected by real classroom resources, financial aggregates, and other facilities such as availability of a laboratory or the size of the library (Hanushek and Kimko, 2000; Lee and Barro, 2001; Hanushek, 2002; Hanushek and Woessmann, 2011a; Woessmann, 2003). We revisit this issue by regressing the starting performance and the performance decline in the PISA 2009 on educational expenditures, which we collected from the World Development Indicators. Within our framework (i.e., if growth affects the two PISA components only through resources), consistent estimates are obtained if u is uncorrelated with v and ξ . Table 5.9 reports the results of regressing the starting performance and the performance decline on two measures of educational expenditures, average government expenditure on education as percentage of GDP for the period 1970-2009, and the average pupil-to-teacher ratio in primary

²⁷As π_1^S , η_1 , and β_1 are expected to be non negative, for a shock to die out the term $\pi_1^S \eta_1 \beta_1$ must be less than 1. Note that one can obtain a similar expression for the bias of the performance decline, where Equation (5.3) is only a function of the performance decline.

Table 5.9: Regressions of the two PISA components on measures for educational expenditures

	(1) Starting performance	(2) Performance decline	(3) Starting performance	(4) ^a Starting performance	(5) Performance decline	(6) Starting performance	(7) Performance decline
Gov. exp.	0.00214	0.0129				-0.0289	0.0102
% of GDP	(0.06)	(0.78)				(-0.84)	(0.54)
Pupil-to-teacher ratio			-0.0117* (-1.93)	0.00512 (0.59)	-0.00526 (-1.40)	-0.0118* (-1.94)	-0.00293 (-0.74)
<i>N</i>	59	59	60	60	60	53	53
Adj. <i>R</i> ²	0.480	0.227	0.438	0.327	0.204	0.460	0.204
F-test						0.1249	0.5633

Notes: *t* statistics in parentheses, heteroskedasticity robust standard errors

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Regressions include a constant and an OECD indicator

^a Controlled for regional indicators instead of an OECD indicator

We include the country in the regression if it has more than 25% nonmissing observations of the educational expenditure measure within the time period 1970-2009. As on average 75% of the variation of these two measures lies between countries (and 25% over time), we feel this criteria is strict enough in order to give a good picture of the average educational expenditures of a country. Results are robust to different criteria, also to including the country if we have just one observation within the whole time period.

school for the same period.²⁸ We control for economic development by including an OECD indicator. Although there is little evidence of an association between educational expenditures and the two components of the PISA-test, columns (3) and (6) show that the pupil-to-teacher ratio is negatively related to the starting performance (*p*-value = 0.059). In column (6), however, the two measures are jointly insignificant. Moreover, these estimates are likely to be upward biased, as favorable, but omitted, state policies tend to be positively correlated with resource-usage (Hanushek, 2002). Including regional indicators does not change our results and, in fact, column (4) shows the marginally significant estimate of the pupil-to-teacher ratio of column (3) is insignificant after controlling for regional indicators.

A more direct approach to address the possibility of reverse causality is to apply the decomposition method to an “early” international test. The main advantage of this approach is being able to explore the effects of noncognitive skills among workers in the labor force on economic growth. However, potential problems of early international tests are low country participation, bad documentation and the absence of (exogenous) variation in the ordering of the questions. Despite these potential problems, we apply the decomposition method to the Reading Literacy Study (RLS), a test administered in 1991, which is also included in the HW-index. This is the first test with relatively high country participation and we were able to retrieve the order of the questions in the test. All 14-year-old pupils

²⁸We experimented with other educational expenditure measures, such as percentage of qualified teachers in primary education, government expenditure per primary student, secondary student, tertiary student, and expenditure on education as a percentage of total government expenditure. These measures give us the same results, but are available for less countries and a shorter time period (mostly from 1998 onwards).

in the RLS were administered the same booklet, so we are unable to separate the performance decline from the difficulty of the question. To the contrary, Elley (1992) notes that the end of the RLS contains the longer and more difficult reading passages, which suggests the performance decline is contaminated by cognitive skills.

Recognizing these potential problems, we use Equation (5.2) without question fixed effects to decompose the RLS into the starting performance and performance decline.²⁹ Then, we use the average economic growth in the period 1995-2010 as our outcome in the post period.³⁰ Table 5.10 reports our results, restricting the sample to countries examined by HW (2012). Consistent with the results discussed above, we find that both components are positively related to economic growth and the estimates are statistically significant. The size of the estimates must be interpreted with care, as the performance decline is identified without variation in the order of the questions and could, therefore, be influenced by cognitive skills. This, in fact, could explain why we find somewhat larger estimates for the performance decline in Table 5.10 as compared to the results in Tables 5.5 and 5.6. Nevertheless, the overall results are very similar to those discussed in Section 5.6.2. Again, the performance decline seems to be more resilient to the inclusion of controls related to the quality of economic institutions than the starting performance, and including regional indicators reduces its coefficient.

Elley (1992) notes that the more difficult questions are concentrated at the end of the test. Therefore, we also compute the performance decline while excluding the last (two blocks of) questions from the RLS as an extra robustness check, so that the performance decline is estimated based on a more homogeneous sample of questions.³¹ Table A.5.12 in the Appendix shows that the results are insensitive to this change.

²⁹The 14-year-old pupils make two separate booklets for the RLS of 40 and 49 multiple-choice questions, we take the average of the two components to reduce measurement error. We are unsure about the length of the break in between the two tests, but results are very similar if we use the components of both booklets separately. Similarly to the previous analysis using PISA, we use the students weights provided by the RLS-dataset to ensure a representative sample. We do not use the data for the test of the 9-year-old pupils in the RLS, as this would again introduce concerns for reverse causality.

³⁰We have also considered different periods of economic growth, for example starting at 1990 or ending at 2007 to avoid an influence of the financial crisis, and results are robust. We control for initial GDP and years of schooling in 1990, as to coincide with the measurement of the starting performance and performance decline. Results are qualitatively similar controlling for the initial values in 1995, though the starting performance loses some significance.

³¹Elley (1992) does not contain specific information on when the more difficult questions are asked, so the choice of which questions to exclude is somewhat arbitrary. However, we exclude two blocks of questions that are centered around the same reading passage. These two blocks contain 9 and 13 questions, of in total 40 and 49 questions in test 1 and 2 respectively. Moreover, we did observe a somewhat sharper increase in the number of incorrectly answered questions at the start of these two blocks.

Table 5.10: Regressions of economic growth on the starting performance and performance decline using an early test (RLS 1991)

	(1)	(2)	(3)	(4) ^a	(5) ^b	(6) ^c	(7) ^d	(8) ^e	(9) ^f
Starting performance	1.961 (1.50)		2.428** (2.51)	2.347** (2.66)	2.377** (2.51)	1.693 (1.50)	0.303 (0.27)	0.00903 (0.01)	0.927 (0.61)
Performance decline		1.413* (1.93)	1.677*** (2.89)	1.750*** (2.92)	1.467** (2.49)	0.998 (1.68)	1.798** (2.77)	1.469** (2.85)	1.402*** (3.07)
Years of schooling	0.00812 (0.05)	0.140 (0.95)	0.0969 (0.67)	0.2209 (1.36)	0.143 (1.09)	0.190 (1.01)	0.254 (1.07)	0.330 (1.60)	-0.00523 (-0.05)
<i>N</i>	23	23	23	23	23	23	21	21	23
Adj. <i>R</i> ²	0.034	0.090	0.304	0.350	0.261	0.349	0.339	0.380	0.195

Notes: *t* statistics in parentheses, heteroskedasticity robust standard errors

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Dependent variable: average annual growth rate in GDP per capita, 1995-2010

Regressions include a constant and GDP per capita in 1990

^{a-f} See the description of Table 5.4. We use average years of schooling over the period 1990-2010 and GDP per capita in 1990 in logs

5.8 Conclusion

Previous studies have found a positive association between cognitive test scores and economic growth. Although this association is difficult to interpret because of the potential for reverse causality, omitted variables and measurement error, HW (2012) have found evidence consistent with a causal interpretation. The goal of the present study was to investigate whether the well-documented relationship between cognitive test scores and economic growth is, at least in part, driven by noncognitive skills. Specifically, we have applied a recently developed method for decomposing test scores into two components: the starting performance and the decline in performance during the test. Research by Borghans and Schils (2012), Balart and Oosterveen (2017) and Zamarro et al. (2016), as well as our results reported in Section 5.7.1, suggest that the performance decline provides a measure of noncognitive skills that is not confounded by cognitive skills. Consequently, it allows us to analyze whether the relationship between test scores and economic growth is, at least in part, driven by noncognitive skills.

We find that both components are associated with economic growth. The estimated effect of the performance decline is approximately equal to the estimated effect of the starting performance. Moreover, we find that the effect of the starting performance is reduced by 40 percent after controlling for the decline in performance, implying that previous estimates of cognitive skills are biased upwards and that noncognitive skills are partly responsible for the relationship between test scores and economic growth. This result is consistent with those of other recent studies that have raised concerns about the size of the effects of cognitive skills on economic growth (Atherton et al., 2013;

Breton, 2011; Levin, 2012). Our results are robust to using a variety of measures of the performance decline, to testing for the presence of a reverse channel from growth to skills, and to using a post period measurement of economic growth.

It would, of course, be ideal to use a more direct approach to test for causal effects (e.g. via the use of instrumental variables), but finding a foolproof instrument for cross-country growth regressions is extremely difficult. Table A.5.13 in the Appendix takes a first step towards this goal by exploring cultural measures as an instrument for the performance decline. Nunn (2012) and Guiso et al. (2006) argue that culture reflects customary beliefs and values inherited from previous generations and can therefore be seen as a source of exogenous variation. We exploit that Hofstede's measures are defined as stable and, specifically, that long term orientation is described as thrift and effort, something directly related to the noncognitive skills the performance decline is hypothesized to measure. When we estimate the first stage, long term orientation is strongly correlated with the decline in performance. The second-stage estimates confirm our main findings.

In this study, we have tried to stay as close as possible to the approach used in previous studies that have established a clear relationship between test scores and economic growth. It should be noted that we are not able to apply the decomposition method to the HW-index, used in the previous studies, but we have applied this method to the PISA test which is only one of the tests included in the HW-index. However, it is likely that the results are also relevant for the other tests that compose the HW-index. First, a large literature in psychology, dating back to test pioneers as Thorndike and Wechsler, and a more recent stream of studies in economics provide evidence for the importance of noncognitive skills for cognitive test scores. Second, we find a very high correlation between the HW-index and the PISA scores, and using PISA scores instead of the HW-index produces very similar results when models used by previous studies are re-estimated. Third, the components resulting from the PISA-decomposition are very stable between countries and over time. Fourth, applying the decomposition to the Reading Literacy Study 1991, an international test also included in the HW-index, gives similar results. Therefore, it seems not likely that the decomposition results found for the PISA test are relevant to this specific test only.

Given the different types of policy interventions required to foster cognitive and noncognitive skills (Cunha et al., 2010), it is important to have a good understanding of the consequences of each type of skill. This distinction has been largely studied at the microeconomic level. Our study provides a first attempt to explore the implications of distinguishing between cognitive and noncognitive skills at the macroeconomic level. Our findings imply that noncognitive skills are also important for explaining the relationship between test scores and economic growth.

5.A Appendix

Cultural Dimensions of Hofstede

- **Power Distance:** this dimension expresses the degree to which the less powerful members of a society accept and expect that power is distributed unequally. The fundamental issue here is how a society handles inequalities among people. People in societies exhibiting a large degree of Power Distance accept a hierarchical order in which everybody has a place and which needs no further justification. In societies with low Power Distance, people strive to equalise the distribution of power and demand justification for inequalities of power.
- **Individualism:** the high side of this dimension, called individualism, can be defined as a preference for a loosely-knit social framework in which individuals are expected to take care of only themselves and their immediate families. Its opposite, collectivism, represents a preference for a tightly-knit framework in society in which individuals can expect their relatives or members of a particular in-group to look after them in exchange for unquestioning loyalty. A society's position on this dimension is reflected in whether people's self-image is defined in terms of "I" or "we".
- **Masculinity:** the Masculinity side of this dimension represents a preference in society for achievement, heroism, assertiveness and material rewards for success. Society at large is more competitive. Its opposite, femininity, stands for a preference for cooperation, modesty, caring for the weak and quality of life. Society at large is more consensus-oriented. In the business context Masculinity versus Femininity is sometimes also related to as "tough versus tender" cultures.
- **Uncertainty Avoidance:** the Uncertainty Avoidance dimension expresses the degree to which the members of a society feel uncomfortable with uncertainty and ambiguity. The fundamental issue here is how a society deals with the fact that the future can never be known: should we try to control the future or just let it happen? Countries exhibiting strong UAI maintain rigid codes of belief and behavior and are intolerant of unorthodox behavior and ideas. Weak UAI societies maintain a more relaxed attitude in which practice counts more than principles.
- **Long Term Orientation:** every society has to maintain some links with its own past while dealing with the challenges of the present and the future. Societies prioritize these two existential goals differently. Societies who score low on this dimension, for example, prefer to maintain time-honoured traditions and norms while viewing societal change with suspicion. Those with

a culture which scores high, on the other hand, take a more pragmatic approach: they encourage thrift and efforts in modern education as a way to prepare for the future. In the business context this dimension is related to as “(short term) normative versus (long term) pragmatic” (PRA). In the academic environment the terminology Monumentalism versus Flexhumility is sometimes also used.

- **Indulgence:** indulgence stands for a society that allows relatively free gratification of basic and natural human drives related to enjoying life and having fun. Restraint stands for a society that suppresses gratification of needs and regulates it by means of strict social norms.

Source: <https://geert-hofstede.com/national-culture.html>

Retrieved: July 1st, 2016

Table A.5.1: Randomization test

	(1) Gender	(2) Mother highest schooling	(3) Father highest schooling	(4) Self born in country	(5) Mother born in country	(6) Father born in country	(7) Language at home	(8) Possessions desk	(9) Possessions own room	(10) How many books at home	(11) Age of student
Booklet=2	0.0219** (2.31)	0.00937 (0.36)	0.0216 (0.82)	-0.00430 (-1.12)	-0.000963 (-0.02)	-0.000727 (-0.13)	0.0110 (1.05)	0.00126 (0.15)	-0.00566 (-0.71)	-0.0385 (-1.40)	-0.00141 (-0.24)
Booklet=3	0.0138 (1.46)	0.000973 (0.04)	0.0320 (1.21)	-0.000545 (-0.13)	0.00458 (0.80)	0.00398 (0.69)	0.00969 (0.92)	0.00383 (0.48)	-0.00966 (-1.22)	-0.0178 (-0.65)	-0.00400 (-0.68)
Booklet=4	0.0105 (1.11)	0.0254 (0.99)	0.0361 (1.39)	-0.00620* (-1.65)	-0.00460 (-0.83)	0.000133 (0.02)	-0.00145 (-0.14)	-0.00271 (-0.35)	-0.00862 (-1.11)	-0.0246 (-0.91)	0.00145 (0.24)
Booklet=5	0.00380 (0.40)	0.000718 (0.03)	0.0132 (0.50)	-0.000969 (-0.23)	-0.000655 (-0.12)	0.00415 (0.72)	0.00862 (0.82)	0.00109 (0.14)	-0.00358 (-0.45)	0.00290 (0.11)	0.00161 (0.28)
Booklet=6	0.0151 (1.60)	0.00317 (0.12)	0.0463* (1.76)	-0.000802 (-0.20)	0.00137 (0.24)	0.00195 (0.34)	0.00322 (0.31)	0.0149* (1.83)	-0.00731 (-0.91)	0.0127 (0.47)	0.00453 (0.51)
Booklet=7	0.0112 (1.19)	-0.0139 (0.55)	0.0187 (0.71)	-0.000645 (-0.16)	0.00157 (0.28)	0.000849 (0.15)	0.00214 (0.21)	0.00432 (0.53)	-0.0101 (-1.28)	-0.0497* (-1.84)	0.00296 (0.50)
Booklet=8	0.0166* (1.77)	0.0213 (0.83)	0.0361 (1.37)	-0.00442 (-1.15)	-0.00145 (-0.26)	0.000277 (0.05)	0.00387 (0.38)	0.00361 (0.45)	-0.00711 (-0.88)	-0.0603** (-2.23)	-0.00148 (-0.25)
Booklet=9	0.00706 (0.75)	0.0371 (1.43)	0.0233 (0.88)	-0.00329 (-0.84)	-0.00369 (-0.66)	-0.000517 (-0.09)	0.00476 (0.47)	-0.00314 (-0.40)	0.00158 (0.19)	-0.0493* (-1.81)	0.00333 (0.56)
Booklet=10	-0.000460 (-0.05)	0.00574 (0.23)	0.0170 (0.65)	0.0000214 (0.01)	-0.000589 (-0.10)	0.00270 (0.47)	0.00300 (0.29)	-0.00194 (-0.25)	-0.00600 (-0.76)	-0.0384 (-1.41)	0.000419 (0.07)
Booklet=11	0.00842 (0.90)	-0.000262 (-0.01)	0.0146 (0.56)	-0.00526 (-1.36)	0.00428 (0.73)	0.00426 (0.73)	0.00507 (0.49)	0.00435 (0.54)	0.00357 (0.44)	0.00683 (0.25)	-0.00255 (-0.43)
Booklet=12	0.0118 (1.26)	0.0253 (0.99)	0.00712 (0.28)	-0.00374 (-0.95)	0.00228 (0.40)	0.00254 (0.44)	0.00654 (0.63)	-0.000107 (-0.01)	-0.00620 (-0.78)	-0.0125 (-0.48)	0.00600 (1.02)
Booklet=13	0.0120 (1.28)	0.0248 (0.96)	0.0273 (1.04)	-0.00200 (-0.50)	0.000499 (0.09)	0.00227 (0.40)	0.000449 (0.04)	0.00373 (0.46)	0.00191 (0.24)	-0.0213 (-0.78)	-0.00285 (-0.49)
Constant	1.485*** (223.51)	2.110*** (119.07)	2.074*** (111.35)	1.044*** (357.29)	1.092*** (271.57)	1.089*** (264.96)	1.177*** (160.00)	1.164*** (207.22)	1.240*** (218.42)	2.962*** (155.55)	15.78*** (3713.17)
Observations	397916	378276	367202	390715	389346	386517	383775	390488	391047	390014	397920
F-value	0.92	0.62	0.49	0.66	0.45	0.20	0.27	0.66	0.67	1.52	0.49
P-value	0.5267	0.8268	0.9234	0.7875	0.9448	0.9985	0.9941	0.7943	0.7816	0.1074	0.9200
Adjusted R ²	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Notes: Regressions of background characteristics upon separate indicators for every booklet. The columns ‘F-value’ and ‘P-value’ refer to the tests for joint significance of the booklet indicators. PISA 2006 and PISA weights are used.

t statistics in parentheses

* p < 0.10, ** p < 0.05, *** p < 0.01

Table A.5.2: The starting performance and decline in performance per country

Country	(1) PISA score	(2) P[$Q_0 = 1$]	(3) P[$Q_1 = 1$]	(4) Decline	Country	(1) PISA score	(2) P[$Q_0 = 1$]	(3) P[$Q_1 = 1$]	(4) Decline
Colombia	381	.59	.243	.347	Lithuania	481.3	.737	.64	.097
Uruguay	422.7	.722	.423	.299	United States	481.5	.776	.679	.097
Argentina	382	.636	.359	.277	Luxembourg	485	.807	.71	.097
Tunisia	377	.454	.229	.225	Poland	500.3	.823	.726	.097
Brazil	384.3	.547	.331	.216	China, Macao	509.3	.828	.732	.096
Kyrgyzstan	306	.393	.185	.208	Hungary	492.3	.789	.694	.095
Mexico	408.7	.594	.387	.207	Slovakia	482	.816	.725	.091
Chile	430.3	.714	.508	.207	Sweden	504	.842	.752	.09
Qatar	326.3	.525	.337	.188	Japan	517.3	.877	.791	.086
Israel	445	.686	.506	.18	Azerbaijan	403.7	.584	.499	.085
Russia	465	.832	.654	.177	Canada	529.3	.832	.748	.084
Greece	464	.786	.609	.177	Ireland	508.7	.753	.67	.083
Jordan	402.3	.51	.339	.171	Australia	520	.836	.754	.082
Romania	409.7	.6	.432	.168	Belgium	510.3	.836	.755	.081
Thailand	418.3	.533	.368	.165	Denmark	501	.866	.787	.079
Bulgaria	416.3	.667	.502	.165	Taiwan	525.7	.833	.754	.078
Indonesia	392.3	.574	.411	.163	Czech Republic	502	.845	.766	.078
Italy	468.7	.735	.582	.153	New Zealand	524.3	.814	.737	.077
Turkey	431.7	.537	.388	.149	Slovenia	505.7	.819	.742	.077
Serbia	424	.709	.578	.131	Germany	505	.826	.753	.073
Latvia	485	.761	.638	.123	Estonia	515.7	.829	.756	.072
Portugal	470.7	.779	.658	.121	Netherlands	521	.83	.76	.07
Spain	476.3	.803	.682	.121	Hong Kong	541.7	.817	.746	.07
Montenegro	401	.583	.467	.117	Korea	541.7	.823	.755	.067
France	493	.807	.693	.114	Switzerland	513.7	.85	.789	.061
United Kingdom	501.7	.762	.652	.11	Liechtenstein	519	.885	.828	.057
Norway	487	.83	.721	.109	Austria	502	.843	.789	.054
Iceland	493.7	.85	.75	.1	Finland	552.7	.898	.856	.042
Croatia	479	.759	.66	.099					

Notes: Probabilities are based on the estimates from Equation (5.2), using PISA 2006 and PISA weights.

Table A.5.3: Descriptive statistics

Country	Initial GDP (1960)	GDP growth (1960-2000)	HW- index	Starting performance (st. error)	Performance decline (st. error)	Num. of students
Argentina	6.033	1.258	3.920	.3478 (.0427)	-.710 (.0118)	4339
Australia	15.20	2.061	5.093	.9789 (.0250)	-.292 (.0059)	14170
Austria	10.54	3.173	5.089	1.007 (.0418)	-.204 (.0094)	4908
Belgium	10.16	2.975	5.041	.9766 (.0310)	-.287 (.0072)	8685
Brazil	2.469	2.709	3.637	.1183 (.0338)	-.555 (.0096)	9295
Canada	12.90	2.382	5.037	.9610 (.0284)	-.292 (.0068)	22646
Chile	3.700	2.689	4.049	.5664 (.0358)	-.546 (.0095)	5233
Colombia	2.940	1.758	4.152	.2271 (.0459)	-.924 (.0127)	4478
Denmark	11.60	2.757	4.962	1.107 (.0421)	-.310 (.0096)	4532
Finland	9.034	3.149	5.126	1.271 (.0442)	-.207 (.0094)	4714
France	10.19	2.815	5.040	.8680 (.0369)	-.362 (.0090)	4716
Greece	5.588	3.428	4.607	.7919 (.0371)	-.515 (.0093)	4873
Hong Kong	3.289	5.633	5.194	.9025 (.0413)	-.239 (.0097)	4645
Iceland	14.07	2.584	4.935	1.038 (.0437)	-.363 (.0101)	3789
Indonesia	6651	3.719	3.879	.1871 (.0343)	-.411 (.0101)	10647
Ireland	7.280	4.008	4.994	.6842 (.0363)	-.245 (.0093)	4585
Israel	6.989	3.133	4.686	.4848 (.0363)	-.470 (.0095)	4584
Italy	8.718	3.174	4.757	.6285 (.0249)	-.420 (.0064)	21773
Japan	5.594	4.521	5.310	1.160 (.0369)	-.349 (.0084)	5952
Jordan	2.721	.8659	4.263	.0257 (.0332)	-.441 (.0092)	6509
Korea	1.670	6.129	5.337	.9255 (.0365)	-.234 (.0090)	5176
Mexico	4.942	2.271	3.997	.2379 (.0278)	-.526 (.0076)	30971
Netherlands	13.43	2.606	5.114	.9557 (.0429)	-.249 (.0099)	4769
New Zealand	14.26	1.661	4.978	.8943 (.0388)	-.259 (.0093)	4823
Norway	12.50	3.286	4.830	.9542 (.0386)	-.369 (.0094)	4692
Portugal	4.181	4.134	4.563	.7694 (.0376)	-.362 (.0095)	5109
Romania	1.362	3.904	4.562	.2532 (.0478)	-.425 (.0127)	5118
Spain	6.333	3.809	4.829	.8522 (.0308)	-.379 (.0075)	19604
Sweden	14.31	1.912	5.013	1.001 (.0407)	-.320 (.0104)	4443
Switzerland	21.02	1.494	5.141	1.035 (.0339)	-.234 (.0077)	12192
Taiwan	1.858	6.459	5.451	.9650 (.0320)	-.276 (.0074)	8815
Thailand	.9620	4.713	4.564	.0835 (.0336)	-.420 (.0094)	6192
Tunisia	1.805	2.945	3.795	-.115 (.0373)	-.627 (.0105)	4640
Turkey	3.183	2.285	4.127	.0923 (.0370)	-.376 (.0103)	4942
United Kingdom	11.20	2.558	4.949	.7134 (.0304)	-.322 (.0077)	13152
United States	15.38	2.373	4.902	.7598 (.0364)	-.294 (.0098)	5611
Uruguay	5.010	1.562	4.300	.5873 (.0409)	-.782 (.0104)	4839

Notes: Descriptive statistics for the sample used in Table 5.5. GDP per capita in 1960 PPP adjusted (in 2005 international Dollars), shown in thousands. The PISA-components are related to the wave of 2006.

Table A.5.4: Regressions of economic growth on components of test scores coding non-reached questions as missing and using an orthogonal measure of the performance decline

	(1)	(2)	(3)	(4) ^a	(5) ^b	(6) ^c	(7) ^d	(8) ^e	(9) ^f
Starting performance	1.832*** (3.84)		1.778*** (4.06)	1.778*** (3.98)	1.551*** (3.90)	0.799** (2.35)	0.987 (1.48)	0.303 (0.43)	2.036*** (6.14)
Performance decline		0.656** (2.19)	0.565** (2.63)	0.567** (2.62)	0.503* (1.72)	0.150 (0.91)	0.444 (1.67)	0.606* (1.97)	0.280 (1.58)
Years of schooling	-0.0193 (-0.17)	0.0534 (0.52)	-0.0244 (-0.22)	-0.01792 (-0.17)	-0.000662 (-0.01)	0.0427 (0.42)	0.0760 (0.73)	0.0295 (0.26)	0.0135 (0.16)
<i>N</i>	55	55	55	55	55	55	47	47	55
Adj. <i>R</i> ²	0.454	0.274	0.486	0.486	0.417	0.717	0.553	0.593	0.669

Notes: *t* statistics in parentheses, heteroskedasticity robust standard errors. Bootstrapped *z* statistics in squared brackets, based on 1000 replications

* *p* < 0.10, ** *p* < 0.05, *** *p* < 0.01

Dependent variable: average annual growth rate in GDP per capita, 1970-2010

Regressions include a constant and GDP per capita in 1970

^{a-f} See the description of Table 5.4. We use average years of schooling over the period 1970-2010 and GDP per capita in 1970 in logs

Table A.5.5: Regressions of economic growth on components of test scores, where the two components are estimated via OLS instead of probit

	(1)	(2)	(3)	(4) ^a	(5) ^b	(6) ^c	(7) ^d	(8) ^e	(9) ^f
Starting performance	1.697*** (3.45)		1.164*** (3.35)	1.169*** (3.28)	1.055*** (2.87)	0.709** (2.10)	0.763 (1.43)	0.109 (0.20)	1.735*** (5.62)
Performance decline		1.466*** (4.15)	1.158*** (3.96)	1.159*** (3.94)	1.168*** (4.20)	0.392 (1.35)	0.983** (2.14)	1.123** (2.42)	0.538** (2.16)
Years of schooling	-0.0145 (-0.13)	-0.0187 (-0.19)	-0.0543 (-0.54)	-0.0464 (-0.48)	-0.00789 (-0.09)	0.0350 (0.35)	0.0422 (0.42)	0.00136 (0.01)	-0.00891 (-0.11)
<i>N</i>	55	55	55	55	55	55	47	47	55
Adj. <i>R</i> ²	0.424	0.485	0.562	0.561	0.567	0.725	0.595	0.650	0.677

Notes: *t* statistics in parentheses, heteroskedasticity robust standard errors

* *p* < 0.10, ** *p* < 0.05, *** *p* < 0.01

Dependent variable: average annual growth rate in GDP per capita, 1970-2010

Regressions include a constant and GDP per capita in 1970

^{a-f} See the description of Table 5.4. We use average years of schooling over the period 1970-2010 and GDP per capita in 1970 in logs

Table A.5.6: Regressions of economic growth on components of test scores using the probabilities related to the probit estimates

	(1)	(2)	(3)	(4) ^a	(5) ^b	(6) ^c	(7) ^d	(8) ^e	(9) ^f
Starting performance	1.600*** (3.20)		0.892** (2.62)	0.895** (2.58)	0.779** (2.13)	0.637* (1.76)	0.595 (1.21)	0.00406 (0.01)	1.534*** (5.20)
Performance decline		1.681*** (4.67)	1.380*** (4.42)	1.384*** (4.41)	1.387*** (4.67)	0.440 (1.32)	1.156** (2.15)	1.300** (2.47)	0.731*** (2.78)
Years of schooling	-0.0116 (-0.10)	-0.0179 (-0.19)	-0.0446 (-0.45)	-0.04028 (-0.42)	0.00493 (0.06)	0.0353 (0.35)	0.0474 (0.48)	0.0209 (0.20)	-0.0107 (-0.13)
<i>N</i>	55	55	55	55	55	55	47	47	55
Adj. <i>R</i> ²	0.404	0.531	0.570	0.570	0.581	0.724	0.591	0.645	0.681

Notes: *t* statistics in parentheses, heteroskedasticity robust standard errors

* *p* < 0.10, ** *p* < 0.05, *** *p* < 0.01

Dependent variable: average annual growth rate in GDP per capita, 1970-2010

Regressions include a constant and GDP per capita in 1970

^{a-f} See the description of Table 5.4. We use average years of schooling over the period 1970-2010 and GDP per capita in 1970 in logs

Table A.5.7: Regressions of economic growth on components of test scores including the first five questions of the test in the starting performance

	(1)	(2)	(3)	(4) ^a	(5) ^b	(6) ^c	(7) ^d	(8) ^e	(9) ^f
Starting performance	1.718*** (3.51)		0.957** (2.59)	0.965** (2.60)	0.751* (1.98)	0.586 (1.60)	0.644 (1.27)	-0.000494 (-0.00)	1.594*** (4.98)
Performance decline		1.626*** (4.72)	1.285*** (4.26)	1.287*** (4.25)	1.279*** (4.36)	0.529 (1.67)	1.093** (2.13)	1.277** (2.60)	0.636** (2.56)
Years of schooling	-0.00727 (-0.06)	-0.0412 (-0.44)	-0.0582 (-0.60)	-0.0529 (-0.56)	0.00388 (0.05)	0.0329 (0.33)	0.0331 (0.33)	-0.00422 (-0.04)	-0.0112 (-0.14)
<i>N</i>	55	55	55	55	55	55	47	47	55
Adj. <i>R</i> ²	0.425	0.527	0.570	0.569	0.576	0.725	0.595	0.653	0.678

Notes: *t* statistics in parentheses, heteroskedasticity robust standard errors

* *p* < 0.10, ** *p* < 0.05, *** *p* < 0.01

Dependent variable: average annual growth rate in GDP per capita, 1970-2010

Regressions include a constant and GDP per capita in 1970

^{a-f} See the description of Table 5.4. We use average years of schooling over the period 1970-2010 and GDP per capita in 1970 in logs

Table A.5.8: Regressions of economic growth on components of test scores, where the two components are estimated using the average score within a cluster as outcome variable and the position of the cluster in the test as explanatory variable

	(1)	(2)	(3)	(4) ^a	(5) ^b	(6) ^c	(7) ^d	(8) ^e	(9) ^f
Starting performance	1.960*** (4.87)		1.372*** (3.74)	1.383*** (3.67)	1.313*** (3.45)	1.023*** (2.95)	1.239** (2.26)	1.066 (1.59)	1.701*** (5.70)
Performance decline		1.497*** (4.24)	0.863*** (3.26)	0.869*** (3.29)	1.001*** (3.19)	0.301 (1.08)	0.676* (1.79)	0.763* (1.99)	0.326 (1.25)
Years of schooling	-0.0607 (-0.58)	-0.0282 (-0.29)	-0.0760 (-0.76)	-0.0734 (-0.77)	-0.0204 (-0.24)	0.00170 (0.02)	0.0107 (0.11)	0.00208 (0.02)	-0.0395 (-0.48)
<i>N</i>	55	55	55	55	55	55	47	47	55
Adj. <i>R</i> ²	0.535	0.491	0.590	0.589	0.626	0.765	0.649	0.687	0.680

Notes: *t* statistics in parentheses, heteroskedasticity robust standard errors

* *p* < 0.10, ** *p* < 0.05, *** *p* < 0.01

Dependent variable: average annual growth rate in GDP per capita, 1970-2010

Regressions include a constant and GDP per capita in 1970

^{a-f} See the description of Table 5.4. We use average years of schooling over the period 1970-2010 and GDP per capita in 1970 in logs

Table A.5.9: Regressions of economic growth on the average of both components of the test score, using PISA 2003, 2006 and 2009

	(1)	(2)	(3)	(4) ^a	(5) ^b	(6) ^c	(7) ^d	(8) ^e	(9) ^f
Starting performance	1.747*** (3.60)		1.056*** (2.86)	1.068*** (2.84)	0.886** (2.43)	0.591 (1.64)	0.792 (1.47)	0.0523 (0.09)	1.618*** (4.97)
Performance decline		1.649*** (4.67)	1.306*** (4.30)	1.315*** (4.30)	1.279*** (4.48)	0.655** (2.06)	1.118** (2.10)	1.353** (2.58)	0.685*** (2.73)
Years of schooling	-0.0324 (-0.29)	-0.0375 (-0.41)	-0.0739 (-0.77)	-0.07441 (-0.79)	-0.0225 (-0.27)	0.0236 (0.24)	0.0126 (0.12)	-0.0157 (-0.14)	-0.0377 (-0.47)
<i>N</i>	55	55	55	55	55	55	47	47	55
Adj. <i>R</i> ²	0.432	0.528	0.585	0.584	0.589	0.729	0.599	0.660	0.688

Notes: *t* statistics in parentheses, heteroskedasticity robust standard errors

* *p* < 0.10, ** *p* < 0.05, *** *p* < 0.01

Dependent variable: average annual growth rate in GDP per capita, 1970-2010

Regressions include a constant and GDP per capita in 1970

^{a-f} See the description of Table 5.4. We use average years of schooling over the period 1970-2010 and GDP per capita in 1970 in logs

Table A.5.10: Regressions of economic growth on components of test scores using the standard error of performance decline as weights

	(1)	(2)	(3)	(4) ^a	(5) ^b	(6) ^c	(7) ^d	(8) ^e	(9) ^f
Starting performance	1.815*** (3.63)		0.908** (2.08)	0.939** (2.20)	0.743* (1.98)	0.470 (1.35)	0.598 (1.22)	0.133 (0.29)	1.633*** (4.55)
Performance decline		1.764*** (4.84)	1.394*** (4.05)	1.385*** (4.03)	1.287*** (4.45)	0.608 (1.58)	1.367** (2.44)	1.420** (2.53)	0.731*** (2.72)
Years of schooling	-0.0391 (-0.36)	-0.102 (-1.03)	-0.101 (-0.98)	-0.0897 (-0.85)	0.00447 (0.05)	-0.0358 (-0.34)	0.0280 (0.25)	0.00234 (0.02)	-0.0632 (-0.81)
<i>N</i>	55	55	55	55	55	55	47	47	55
Adj. <i>R</i> ²	0.487	0.578	0.604	0.600	0.573	0.744	0.639	0.670	0.715

Notes: *t* statistics in parentheses, heteroskedasticity robust standard errors

* *p* < 0.10, ** *p* < 0.05, *** *p* < 0.01

Dependent variable: average annual growth rate in GDP per capita, 1970-2010

Regressions include a constant and GDP per capita in 1970

^{a-f} See the description of Table 5.4. We use average years of schooling over the period 1970-2010 and GDP per capita in 1970 in logs

Table A.5.11: Regressions of economic growth on components of test scores coding non-reached questions as missing and using the standard error of performance decline as weights

	(1)	(2)	(3)	(4) ^a	(5) ^b	(6) ^c	(7) ^d	(8) ^e	(9) ^f
Starting performance	1.942*** (3.94)		1.549*** (3.46)	1.561*** (3.47)	1.328*** (3.11)	0.654* (1.74)	0.916 (1.39)	0.313 (0.48)	1.998*** (6.00)
Performance decline		1.095*** (3.01)	0.646** (2.58)	0.646** (2.57)	0.548* (1.72)	0.214 (0.88)	0.646** (2.10)	0.741* (1.98)	0.329 (1.51)
Years of schooling	-0.0533 (-0.49)	-0.0383 (-0.38)	-0.0655 (-0.61)	-0.0546 (-0.51)	-0.000662 (-0.01)	-0.0298 (-0.29)	0.0852 (0.80)	0.0566 (0.50)	-0.0434 (-0.55)
<i>N</i>	55	55	55	55	55	55	47	47	55
Adj. <i>R</i> ²	0.518	0.449	0.548	0.546	0.417	0.738	0.592	0.611	0.711

Notes: *t* statistics in parentheses, heteroskedasticity robust standard errors

* *p* < 0.10, ** *p* < 0.05, *** *p* < 0.01

Dependent variable: average annual growth rate in GDP per capita, 1970-2010

Regressions include a constant and GDP per capita in 1970

^{a-f} See the description of Table 5.4. We use average years of schooling over the period 1970-2010 and GDP per capita in 1970 in logs

Table A.5.12: Regressions of economic growth on the starting performance and performance decline using an early test (RLS 1991) and excluding the last two blocks of questions

	(1)	(2)	(3)	(4) ^a	(5) ^b	(6) ^c	(7) ^d	(8) ^e	(9) ^f
Starting performance	2.133 (1.63)		1.936* (1.85)	1.913* (1.84)	0.707 (0.78)	0.973 (0.89)	0.579 (0.64)	0.0590 (0.06)	0.237 (0.18)
Performance decline		1.781** (2.68)	1.697** (2.63)	1.680** (2.51)	0.910* (1.96)	1.185* (2.12)	1.727** (2.45)	1.412** (2.34)	1.490** (2.67)
Years of schooling	0.00556 (0.03)	0.164 (1.29)	0.113 (0.90)	0.120 (1.35)	0.144 (1.40)	0.222 (1.28)	0.332 (1.60)	0.391* (2.12)	0.0105 (0.09)
<i>N</i>	23	23	23	23	22	23	21	21	23
Adj. <i>R</i> ²	0.061	0.232	0.359	0.388	0.154	0.437	0.378	0.399	0.262

Notes: *t* statistics in parentheses, heteroskedasticity robust standard errors

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Dependent variable: average annual growth rate in GDP per capita, 1995-2010

Regressions include a constant and GDP per capita in 1990

^{a-f} See the description of Table 5.4. We use average years of schooling over the period 1990-2010 and GDP per capita in 1990 in logs

Instrumental Variable Analysis

Despite the usual set of controls in the cross-country growth regressions, it is desirable to use a more direct approach to identify causal effects. To this end we use an instrumental variable approach. This small-sample analysis has to be interpreted with caution, IV estimators can have a finite-sample distribution that differ from the asymptotic distribution. As in Guiso et al. (2006), we use cultural measures as an instrument, in our case for the performance decline. They argue that culture reflects customary beliefs and values that are inherited by an individual from previous generations, rather than voluntarily accumulated. Because of the difficulty of changing culture, it is largely given to individuals throughout their lifetime and can therefore be seen as a source of exogenous variation. Similarly, Nunn (2012) conceptualizes culture through decision making heuristics or rules of thumb that have evolved given our need to make decisions in complex and uncertain environments. He argues these are typically slow-moving.

The growing body of research investigating the effects of culture upon economic growth raises questions regarding the validity of using culture as an instrument (Tabellini, 2010; Gorodnichenko and Roland, 2011, 2016). For example, Gorodnichenko and Roland (2016) use genes as an instrument to document a direct effect of individualism on growth. To the best of our knowledge, however, the cultural measures used below have not been well studied (also not by Gorodnichenko and Roland (2011)). Moreover, Guiso et al. (2006), arguably, define cultural measures as potential instruments, whereas this IV-analysis might also reduce the risk of reverse causality as we call upon slow-moving variation in the performance decline.

Table A.5.13 shows our results, where the upper and lower panel display the first- and second stage respectively. First, we exploit the Weber-hypothesis which states that the emergence of the spirit of capitalism, accumulation of wealth, and virtues of hard work can be attributed to Protestant work ethic (Nunn, 2012). We use the share of Protestantism in 2000 from Barro (2003) as an instrument for the performance decline (column (1)). We find a positive first-stage relationship, but the F-statistic reveals we are dealing with a weak instrument. The second stage shows an IV-estimate that is close to OLS, but it is very imprecisely estimated which can be explained by the amount of noise introduced in the first stage. Moreover, problems related to finite-sample bias and potential endogeneity of the instrument are magnified if the instrument is weakly correlated with the endogenous variable.

Next we exploit Hofstede's long term orientation cultural component as an alternative instrument for the performance decline. Hofstede explains his components as follows: "These relative scores have been proven to be quite stable over time. The forces that cause cultures to shift tend to be global or continent-wide. This means that they affect many countries at the same time, so if their cultures

Table A.5.13: Growth regressions using instrumental variables for the performance decline and starting performance

	(1) ^a	(2) ^a	(3) ^a	(4) ^a	(5) ^{b,c}	(6) ^{b,c}	(7) ^{b,d}	(8) ^{b,d}
First stage								
Protestant share in 2000	0.342 (1.52)							
Long-term orientation		0.0132*** (5.19)	0.00798** (2.72)	0.0114*** (4.15)				
Private enrollment share					0.0046** (2.58)	0.0022 (1.38)		
Catholic share in 1900							0.766 (1.31)	0.358 (0.73)
Years of schooling					0.0433 (1.10)	0.0495 (1.65)	0.035 (0.82)	0.0624 (1.34)
Additional controls	No	No	Yes	No	No	Yes	No	Yes
F-test	2.31	26.94	7.40	17.22	3.74	3.05	1.12	1.12
Second stage								
Performance decline	1.095 (0.79)	2.604*** (4.39)	2.998*** (3.17)	2.375*** (3.67)				
Starting performance					0.457 (0.94)	2.472* (1.93)	3.474* (1.95)	1.50 (0.83)
Years of schooling	0.006 (0.07)	-0.104 (-1.01)	-0.081 (-0.65)	-0.108 (-1.08)				
<i>N</i>	53	51	45	51	21	21	42	41
Adj. <i>R</i> ²	0.495	0.415	0.504	0.461	0.395	0.152	0.617	0.666

Notes: *t* and *z* statistics in parentheses, heteroskedasticity robust standard errors

* *p* < 0.10, ** *p* < 0.05, *** *p* < 0.01

Dependent variable: average annual growth rate in GDP per capita, 1970-2010

Regressions include a constant and GDP per capita in 1970. Additional controls include: openness of the economy and protection against expropriation (column (6) and (8)), plus fertility and tropical location (column (3))

^a Dependent variable in first stage is the performance decline

^b Dependent variable in first stage is the starting performance. We treat initial years of schooling (in 1970) as an extra instrumental variable

^c Sample restricted to OECD countries

^d Controlled for the share of Catholics in 1970 in both the first and second-stage

shift, they shift together and their relative positions remain the same".³² Using long term orientation as an instrument for the performance decline gives a strong first-stage relationship with an F-statistic of 26.94 (column (2)). The second-stage estimate for the performance decline is statistically significant, the OLS-estimate falls within the 95% confidence interval of the IV-estimate. Regarding the validity of long term orientation as an instrument, the exclusion restriction is violated if culture affects economic growth through other channels than the performance decline. In particular, Nunn

³²<https://geert-hofstede.com/national-culture.html>, retrieved July 1st, 2016.

(2012) argues historical shocks can have a persistent effect upon culture (only) if formal institutions change with it. Column (3) and (4) respectively show our results are robust to the full set of controls, including the quality of economic institutions, and to controlling for the potentially endogenous starting performance, which addresses concerns on this particular violation of the exclusion restriction. Moreover, long term orientation is described as thrift and effort which are directly related to the noncognitive skills the performance decline is hypothesized to measure.³³

Ideally we would instrument both of the potentially endogenous components within one model. To this end, column (5) until (8) investigate whether two of the instruments used by HW (2012) can be used for the starting performance. For comparability we perform this analysis on the same sample as HW (2012) and also use initial years of schooling as an additional instrument.³⁴ The potential instruments are private enrollment share in 1985 and the catholic share in 1900, which use the idea of private competition being beneficial for student achievement. HW (2012) state one can plausibly assume this institutional feature to be exogenous.³⁵ For both instruments the first stage shows weak F-statistics, there is only a significant relationship in column (5) at the 5%-level. This reduces the interest in the second stage, which show positive estimates that are either borderline significant or insignificant. We will refrain from an analysis with both components instrumented within one model.

³³For the significant IV-estimates (columns (2), (3) and (4)) we tested for exogeneity of the performance decline using the Durbin-Wu and Hausman test. Under the assumption that the instrument is valid, we reject exogeneity at the 5%-level for all three specifications. While IV has the property of being consistent, keep in mind we are working with a small sample.

³⁴As we have more instruments than endogenous variables, we can use the Sargan test to test whether the moment conditions are valid. For columns (5) until (8) we cannot reject the null-hypothesis of exogeneity for the instruments at conventional significance levels.

³⁵See HW (2012) for details, but in particular they argue that many educational institutions are slow-moving and reflect long-standing policies that are not the outcome of economic growth. The data for the private enrollment share refers to the private enrollment as a percentage of total enrollment in general secondary education in 1985 and come from UNESCO (1998). The Catholic shares in 1900 and 1970 are obtained from Barro (2003).

Nederlandse Samenvatting

(Summary in Dutch)

Onderwijs is een investering in menselijk kapitaal en heeft een grote invloed op private en maatschappelijke uitkomsten. Beleidsmakers hebben dan ook een enorme verantwoordelijkheid wanneer zij schaarse middelen gebruiken om het onderwijsstelsel vorm te geven. Een weloverwogen beslissing van een beleidsmaker vereist kennis over welk beleid het meest effectief is om voorgestelde doelen te bereiken. Dit proefschrift beoogt aan deze kennis toe te voegen door middel van vier hoofdstukken over de impact van ontwerkenmerken in het onderwijs.

Hoofdstuk 2 behandelt peer-effecten in het klaslokaal; hebben klasgenoten een effect op elkaar gedrag en uitkomsten? Zo ja, dan kan een alternatieve toewijzing van leerlingen over klassen schoolprestaties verbeteren, zonder dat hierbij extra geld wordt uitgegeven. In dit hoofdstuk bestuderen wij peer-effecten op een grote Europese universiteit, waar studenten willekeurig worden geplaatst in een werkgroep en in één van twee subgroepen binnen een werkgroep. De universiteit stimuleert sociale interactie en peer-binding binnen, en niet tussen, de subgroepen via informele bijeenkomsten. Elke student kan haar peers binnen de werkgroep dus indelen in *nabije* en *verre* peers. Ons onderzoek toont aan dat er positieve peer-effecten zijn op de prestaties van studenten, welke uitsluitend afkomstig zijn van nabije peers. Onze interpretatie is dat peer-effecten worden gegenereerd door sociale interactie tussen studenten en niet door effecten die spelen op klassikaal niveau. De peer-effecten zijn heterogeen; studenten van hoge en lage bekwaamheid presteren beter (slechter) wanneer zij nabije peers hebben die hoog (laag) bekwaam zijn. Een alternatieve toewijzing van studenten - zoals het toewijzen van hoog bekwame studenten aan dezelfde werkgroepen - kan de prestaties van studenten gemiddeld verbeteren. De baten zijn echter geconcentreerd bij studenten van hoge bekwaamheid.

In hoofdstuk 3 onderzoeken wij de impact van een beleidsmaatregel die de aanwezigheid van studenten op een grote Europese universiteit verplicht. Deze beleidsmaatregel verplicht studenten om naar ten minste 70 procent van de tweedejaars werkgroepen te gaan als zij gemiddeld lager dan een 7 hebben gescoord in het eerste jaar. Door studenten met een gemiddeld cijfer rond de 7 te vergelijken

kunnen wij de causale impact van verplichte en regelmatige aanwezigheid in kaart brengen. Recent onderzoek beargumenteert dat het aanbrengen van een dergelijke structuur tot betere prestaties van studenten leidt. Wij vinden echter dat de gedwongen studenten lagere cijfers en slagingspercentages hebben. Wij stellen dat de beleidsmaatregel studenten dwingt om een substantieel aantal uren te spenderen aan werkgroepen waardoor zij minder tijd hebben voor andere belangrijke studieactiviteiten. Prestaties van studenten dalen doordat de negatieve impact van dit tijdsverlies groter is dan de potentiële positieve impact van meer aanwezigheid bij de werkgroepen. Een analyse naar het totaal aantal studie-uren suggereert dat studenten ook minder tijd besteden aan activiteiten buiten de universiteit. Oftewel, studenten hebben minder vrije tijd. Deze beleidsmaatregel zorgt er dus voor dat gedwongen studenten slechter af zijn.

Hoofdstuk 4 bestudeert genderverschillen in de testuitslagen van 15- en 16-jarigen in de PISA test (Programme for International Student Assessment). In dit hoofdstuk vergelijken wij het antwoord van jongens en meisjes op dezelfde vraag die zij op willekeurig verschillende posities beantwoordden in de test. Wij vinden dat meisjes hun prestatie beter kunnen vasthouden gedurende de test dan jongens. Dit resultaat geldt voor de meerderheid van de landen die meedoen met de PISA test en wordt niet beïnvloed door het onderwerp van de test. Dit biedt nieuwe inzichten in de genderverschillen van testuitslagen. Aan het begin van de test presteren jongens beter in wiskunde en natuurkunde en presteren meisjes beter in lezen. Aan het eind van de test is er geen genderverschil meer in wiskunde en natuurkunde, of is het juist omgekeerd in eenderde van de participerende landen. In meer dan de helft van de landen hebben meisjes het initiële genderverschil met ten minste 50 procent gereduceerd aan het eind van de test. Dit suggereert dat het ontwerp van testen een rol zou kunnen spelen bij het verder bevorderen van gendergelijkheid in STEM-vakken.

In hoofdstuk 5 analyseren wij of non-cognitieve vaardigheden (gedeeltelijk) verantwoordelijk zijn voor de, in eerder onderzoek aangetoonde, positieve relatie tussen cognitieve test scores en economische groei. Om te beginnen gebruiken wij een soortgelijke methode als in hoofdstuk 4 om de PISA test score te ontbinden in twee componenten: het startniveau en de prestatiedaling gedurende de test. De prestatiedaling is gerelateerd aan non-cognitieve vaardigheden, terwijl het startniveau een geschoonde meting is voor cognitieve vaardigheden. Studenten uit verschillende landen verschillen in zowel hun prestatie aan het begin van de test als in de daling van hun prestatie gedurende de test. Ons onderzoek toont aan dat beide componenten positief gerelateerd zijn aan economische groei. De relatie tussen beide componenten en economische groei is ongeveer even groot en vrij robuust. Deze resultaten suggereren dat non-cognitieve vaardigheden ook belangrijk zijn voor de relatie tussen test scores en economische groei.

Bibliography

Acemoglu, D., Johnson, S., Robinson, J.A., 2001. The colonial origins of comparative development: An empirical investigation. *The American Economic Review* 91, pp. 1369–1401.

Akyol, P., Key, J., Krishna, K., 2016. Hit or miss? test taking behavior in multiple choice exams. NBER Working Paper: 22401 .

Almlund, M., Duckworth, A.L., Heckman, J., Kautz, T., 2011. Chapter 1 - personality psychology and economics1, in: Eric A. Hanushek, S.M., Woessmann, L. (Eds.), *Handbook of The Economics of Education*. Elsevier. volume 4, pp. 1 – 181.

Angrist, J.D., 2014. The perils of peer effects. *Labour Economics* 30, 98–108.

Angrist, J.D., Krueger, A.B., 1991. Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics* 106, 979–1014.

Angrist, J.D., Oreopoulos, P., Williams, T., 2014. When opportunity knocks, who answers? new evidence on college achievement awards. *Journal of Human Resources* 1, 1–29.

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

Angrist, J.D., Pischke, J.S., 2010. The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of economic perspectives* 24, 3–30.

Arcidiacono, P., Nicholson, S., 2005. Peer effects in medical school. *Journal of public Economics* 89, 327–350.

Aten, B., Heston, A., Summers, R., 2009. Penn world table version 7.1. Center for International Comparisons of Production, Income, and Prices at the University of Pennsylvania .

Atherton, P., Appleton, S., Bleaney, M., 2013. International school test scores and economic growth. *Bulletin of Economic Research* 65, 82–90.

Athey, S., Imbens, G.W., 2017. The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives* 31, 3–32.

Azmat, G., Calsamiglia, C., Iribarri, N., 2016. Gender differences in response to big stakes. *Journal of the European Economic Association* 14, 1372–1400. doi:10.1111/jeea.12180.

Balart, P., Oosterveen, M., 2017. Wait and see: Gender differences in performance during cognitive tests. *JOLE Working Paper* 17679 .

Baldiga, K., 2013. Gender differences in willingness to guess. *Management Science* 60, 434–448.

Barro, R., 2003. Religion adherence data. <http://scholar.harvard.edu/barro/publications/religion-adherence-data>.

Barro, R.J., 1991. Economic growth in a cross section of countries. *The quarterly journal of economics* 106, 407–443.

Barro, R.J., 2001. Human capital and growth. *The American Economic Review* 91, 12–17.

Barro, R.J., Lee, J.W., 2013. A new data set of educational attainment in the world, 1950–2010. *Journal of development economics* 104, 184–198.

Benet-Martínez, V., Oishi, S., 2008. *Culture and Personality*. New York: Guilford.

Booij, A.S., Leuven, E., Oosterbeek, H., 2017. Ability peer effects in university: Evidence from a randomized experiment. *The Review of Economic Studies* 84, 547–578.

Borghans, L., Duckworth, A.L., Heckman, J.J., Ter Weel, B., 2008a. The economics and psychology of personality traits. *Journal of human Resources* 43, 972–1059.

Borghans, L., Golsteyn, B.H., Heckman, J., Humphries, J.E., 2011. Identification problems in personality psychology. *Personality and Individual Differences* 51, 315–320.

Borghans, L., Golsteyn, B.H., Heckman, J.J., Humphries, J.E., 2016. What grades and achievement tests measure. *Proceedings of the National Academy of Sciences* 113, 13354–13359.

Borghans, L., Meijers, H., Ter Weel, B., 2008b. The role of noncognitive skills in explaining cognitive test scores. *Economic Inquiry* 46, 2–12.

Borghans, L., Schils, T., 2012. The leaning tower of pisa: Decomposing achievement test scores into cognitive and noncognitive components. *JOLE Working Paper* 13260 .

Bosworth, B., Collins, S.M., 2003. The empirics of growth: An update. *Brookings Papers on Economic Activity* 2003, 113–206.

Breton, T.R., 2011. The quality vs. the quantity of schooling: What drives economic growth? *Economics of Education Review* 30, 765–773.

Brown, R.P., Josephs, R.A., 1999. A burden of proof: Stereotype relevance and gender differences in math performance. *Journal of personality and social psychology* 76, 246.

Brunello, G., De Paola, M., Scoppa, V., 2010. Peer effects in higher education: Does the field of study matter? *Economic Inquiry* 48, 621–634.

Burke, M.A., Sass, T.R., 2013. Classroom peer effects and student achievement. *Journal of Labor Economics* 31, 51–82.

Buser, T., Yuan, H., 2016. Do women give up competing more easily? evidence from the lab and the dutch math olympiad. *Tinbergen Institute Discussion Paper* 16-096/I .

Calonico, S., Cattaneo, M.D., Farrell, M.H., Titiunik, R., 2016. rdrobust: Software for regression discontinuity designs. Unpublished manuscript available at: http://faculty.chicagobooth.edu/max.farrell/research/Calonico-Cattaneo-Farrell-Titiunik_2016_Stata.pdf .

Cameron, A.C., Trivedi, P.K., 2005. *Microeometrics: methods and applications*. Cambridge university press.

Caplan, P.J., Crawford, M., Hyde, J.S., Richardson, J.T., 1997. Gender Differences in Human Cognition. *Counterpoints: Cognition, Memory, and Language Series*. ERIC.

Carrell, S.E., Fullerton, R.L., West, J.E., 2009. Does your cohort matter? measuring peer effects in college achievement. *Journal of Labor Economics* 27, 439–464.

Carrell, S.E., Sacerdote, B.I., West, J.E., 2013. From natural variation to optimal policy? the importance of endogenous peer group formation. *Econometrica* 81, 855–882.

Castleman, B.L., 2014. Prompts, personalization, and pay-offs: Strategies to improve the design of college and financial aid information. The George Washington University Graduate School of Education and Human Development .

Cattaneo, M.D., Idrobo, N., Titiunik, R., 2018. *A practical introduction to regression discontinuity designs: Part ii. Preparation for Cambridge Elements: Quantitative and Computational Methods for Social Science*, Cambridge University Press .

Cattell, R.B., 1987. *Intelligence: Its structure, growth and action*. volume 35. Elsevier.

Cerulli, G., Dong, Y., Lewbel, A., Poulsen, A., 2017. Testing stability of regression discontinuity models, in: *Regression Discontinuity Designs: Theory and Applications*. Emerald Publishing Limited, pp. 317–339.

Chapman, B.P., Duberstein, P.R., Sørensen, S., Lyness, J.M., 2007. Gender differences in five factor model personality traits in an elderly cohort. *Personality and individual differences* 43, 1594–1603.

Chen, J., Lin, T.F., 2008. Class attendance and exam performance: A randomized experiment. *The Journal of Economic Education* 39, 213–227.

Cheng, A., Zamarro, G., Orriens, B., 2018. Personality as a predictor of unit nonresponse in an internet panel. (in press). *Sociological Methods & Research* .

Cohen, D., Soto, M., 2007. Growth and human capital: good data, good results. *Journal of economic growth* 12, 51–76.

Cohodes, S., Goodman, J., 2014. Merit aid, college quality and college completion: Massachusetts' adams scholarship as an in-kind subsidy. *American Economic Journal: Applied Economics* 6, 251–285.

Cornwell, C., Mustard, D.B., Van Parys, J., 2013. Noncognitive skills and the gender disparities in test scores and teacher assessments: Evidence from primary school. *Journal of Human Resources* 48, 236–264.

Correspondence, 1994. Correspondence: Should class attendance be mandatory. *Journal of Economic Perspectives* 8, 205–216.

Costa, P.T., McCrae, R.R., 1992. Four ways five factors are basic. *Personality and individual differences* 13, 653–665.

Croson, R., Gneezy, U., 2009. Gender differences in preferences. *Journal of Economic literature* , 448–474.

Cunha, F., Heckman, J.J., 2008. Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation. *Journal of Human Resources* 43, 738–782.

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

Cunha, F., Karahan, F., Soares, I., 2011. Returns to skills and the college premium. *Journal of Money, Credit and Banking* 43, 39–86.

De Giorgi, G., Pellizzari, M., Redaelli, S., 2010. Identification of social interactions through partially overlapping peer groups. *American Economic Journal: Applied Economics* 2, 241–75.

De Paola, M., Scoppa, V., Nistico, R., 2012. Monetary incentives and student achievement in a depressed labor market: Results from a randomized experiment. *Journal of Human Capital* 6, 56–85.

Dee, T.S., 2007. Teachers and the gender gaps in student achievement. *Journal of Human Resources* 42, 528–554.

Desalegn, A.A., Berhan, A., Berhan, Y., 2014. Absenteeism among medical and health science undergraduate students at hawassa university, ethiopia. *BMC medical education* 14, 81.

Dohmen, T., Enke, B., Falk, A., Huffman, D., Sunde, U., 2016. Patience and the wealth of nations. Mimeo .

Doménech, R., De la Fuente, A., 2006. Human capital in growth regressions: how much difference does data quality make? *Journal of the European Economic Association* 4, 1–36.

Duckworth, A.L., Peterson, C., Matthews, M.D., Kelly, D.R., 2007. Grit: perseverance and passion for long-term goals. *Journal of personality and social psychology* 92, 1087.

Duckworth, A.L., Quinn, P.D., Lynam, D., Loeber, R., Stouthamer-Loeber, M., Moffit, T.E., Caspi, A., 2010. What intelligence tests test: Individual differences in test motivation and iq. Unpublished manuscript, University of Pennsylvania .

Duckworth, A.L., Quinn, P.D., Lynam, D.R., Loeber, R., Stouthamer-Loeber, M., 2011. Role of test motivation in intelligence testing. *Proceedings of the National Academy of Sciences* 108, 7716–7720.

Duckworth, A.L., Seligman, M.E., 2006. Self-discipline gives girls the edge: Gender in self-discipline, grades, and achievement test scores. *Journal of educational psychology* 98, 198.

Duckworth, A.L., Yeager, D.S., 2015. Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher* 44, 237–251.

Duflo, E., Dupas, P., Kremer, M., 2011. Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in kenya. *American Economic Review* 101, 1739–74.

Duncan, T.G., McKeachie, W.J., 2005. The making of the motivated strategies for learning questionnaire. *Educational psychologist* 40, 117–128.

DUO, 2014. Examenmonitor VO 2014. Technical Report. Dienst Uitvoering Onderwijs.

Durden, G.C., Ellis, L.V., 1995. The effects of attendance on student learning in principles of economics. *The American Economic Review* 85, 343–346.

Dynarski, S., 2008. Building the stock of college-educated labor. *Journal of Human Resources* 43, 924–937.

Eastwood, J.D., Frischen, A., Fenske, M.J., Smilek, D., 2012. The unengaged mind: Defining boredom in terms of attention. *Perspectives on Psychological Science* 7, 482–495.

Eklöf, H., 2007. Test-taking motivation and mathematics performance in timss 2003. *International Journal of Testing* 7, 311–326. doi:10.1080/15305050701438074.

Elley, W.B., 1992. How in the world do students read? ie a study of reading literacy. *International Association for the Evaluation of Educational Achievement* .

Espinosa, M.P., Gardeazabal, J., 2013. Do students behave rationally in multiple choice tests? evidence from a field experiment. *Journal of Economics and Management* 9, 107–135.

Feld, J., Salamanca, N., Zölitz, U., 2018. Are professors worth it? the value-added and costs of tutorial instructors. University of Zurich, Department of Economics, Working Paper No. 293 .

Feld, J., Zölitz, U., 2017. Understanding peer effects: on the nature, estimation, and channels of peer effects. *Journal of Labor Economics* 35, 387–428.

Fernandez, R., Fogli, A., 2009. Culture: An empirical investigation of beliefs, work, and fertility. *American Economic Journal: Macroeconomics* 1, 146–177.

Fisher, C.D., 1993. Boredom at work: A neglected concept. *Human Relations* 46, 395–417.

Foster, G., 2006. It's not your peers, and it's not your friends: Some progress toward understanding the educational peer effect mechanism. *Journal of public Economics* 90, 1455–1475.

Fredriksson, P., Öckert, B., Oosterbeek, H., 2013. Long-term effects of class size. *The Quarterly Journal of Economics* 128, 249–285.

Fryer Jr, R.G., Levitt, S.D., 2010. An empirical analysis of the gender gap in mathematics. *American Economic Journal: Applied Economics* 2, 210–40.

Gallup, J.L., Sachs, J.D., Mellinger, A.D., 1999. Geography and economic development. *International regional science review* 22, 179–232.

Garlick, R., 2018. Academic peer effects with different group assignment policies: Residential tracking versus random assignment. *American Economic Journal: Applied Economics* (forthcoming)

Gennaioli, N., La Porta, R., Lopez-de Silanes, F., Shleifer, A., 2012. Human capital and regional development. *The Quarterly journal of economics* 128, 105–164.

Gneezy, U., List, J., Livingston, J., Qin, X., Xu, Y., 2017. Measuring success in education: The role of effort on the test itself. *NBER Working Paper*: 24004 .

Gneezy, U., Meier, S., Rey-Biel, P., 2011. When and why incentives (don't) work to modify behavior. *The Journal of Economic Perspectives* 25, 191–209.

Gneezy, U., Niederle, M., Rustichini, A., et al., 2003. Performance in competitive environments: Gender differences. *Quarterly Journal of Economics* 118, 1049–1074.

Gneezy, U., Rustichini, A., 2000. Pay enough or don't pay at all. *Quarterly journal of economics* , 791–810.

Gneezy, U., Rustichini, A., 2004. Gender and competition at a young age. *American Economic Review* 94, 377–381. doi:10.1257/0002828041301821.

Goldin, C., Katz, L.F., Kuziemko, I., 2006. The homecoming of american college women: The reversal of the college gender gap. *The Journal of Economic Perspectives* 20, 133–133.

Gorodnichenko, Y., Roland, G., 2011. Which dimensions of culture matter for long-run growth? *The American Economic Review* 101, 492–98.

Gorodnichenko, Y., Roland, G., 2016. Culture, institutions and the wealth of nations. *Review of Economics and Statistics* 99, 402–416.

Guiso, L., Sapienza, P., Zingales, L., 2006. Does culture affect economic outcomes? *Journal of Economic Perspectives* 20, 23–48.

Guryan, J., Kroft, K., Notowidigdo, M.J., 2009. Peer effects in the workplace: Evidence from random groupings in professional golf tournaments. *American Economic Journal: Applied Economics* 1, 34–68.

Hanushek, E.A., 2002. Publicly provided education. *Handbook of Public Economics* 4, 2045–2141.

Hanushek, E.A., 2013. Economic growth in developing countries: The role of human capital. *Economics of Education Review* 37, 204–212.

Hanushek, E.A., Kimko, D.D., 2000. Schooling, labor-force quality, and the growth of nations. *American economic review* , 1184–1208.

Hanushek, E.A., Woessmann, L., 2008. The role of cognitive skills in economic development. *Journal of economic literature* , 607–668.

Hanushek, E.A., Woessmann, L., 2011a. Chapter 2 - the economics of international differences in educational achievement 3, 89 – 200. doi:<https://doi.org/10.1016/B978-0-444-53429-3.00002-8>.

Hanushek, E.A., Woessmann, L., 2011b. How much do educational outcomes matter in oecd countries? *Economic Policy* 26, 427–491.

Hanushek, E.A., Woessmann, L., 2011c. Sample selectivity and the validity of international student achievement tests in economic research. *Economics Letters* 110, 79–82.

Hanushek, E.A., Woessmann, L., 2012. Do better schools lead to more growth? cognitive skills, economic outcomes, and causation. *Journal of Economic Growth* 17, 267–321.

Harmon, C., Oosterbeek, H., Walker, I., 2003. The returns to education: Microeconomics. *Journal of economic surveys* 17, 115–156.

Harris, M.B., 2000. Correlates and characteristics of boredom proneness and boredom. *Journal of Applied Social Psychology* 30, 576–598.

Heckman, J., Pinto, R., Savelyev, P., 2013. Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review* 103, 2052–86. doi:10.1257/aer.103.6.2052.

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

Heckman, J.J., 2008. Schools, skills, and synapses. *Economic inquiry* 46, 289–324.

Heckman, J.J., Kautz, T., 2012. Hard evidence on soft skills. *Labour economics* 19, 451–464.

Heckman, J.J., Rubinstein, Y., 2001. The importance of noncognitive skills: Lessons from the ged testing program. *The American Economic Review* , 145–149.

Hernández, M., Hershaff, J., 2014. Skipping questions in school exams: The role of socio-emotional skills on educational outcomes. Mimeo. Draft version: March 18, 2014 .

Hitt, C., 2016. Just filling in the bubbles: Using careless answers patterns on surveys as a proxy measure of non cognitive skills. EDRE Working Paper 2015-06 .

Hitt, C., Trivitt, J., Cheng, A., 2016. When you say nothing at all: The predictive power of student effort on surveys. *Economics of Education Review* 52, 105–119.

Hoffmann, F., Oreopoulos, P., 2009. A professor like me the influence of instructor gender on college achievement. *Journal of Human Resources* 44, 479–494.

Hofstede, G., McCrae, R.R., 2004. Personality and culture revisited: Linking traits and dimensions of culture. *Cross-Cultural Research: The Journal of Comparative Social Science* .

Hofstede, G.H., Hofstede, G., 2001. Culture's consequences: Comparing values, behaviors, institutions and organizations across nations. SAGE Publications.

Hong, S.C., Lee, J., 2017. Who is sitting next to you? peer effects inside the classroom. *Quantitative Economics* 8, 239–275.

Hoxby, C., 2000. Peer effects in the classroom: Learning from gender and race variation. NBER Working Paper No. 7867 .

Hoxby, C.M., Weingarth, G., 2005. Taking race out of the equation: School reassignment and the structure of peer effects. Mimeo .

Hübner, M., Vannoorenberghe, G., 2015. Patience and long-run growth. *Economics Letters* 137, 163 –167.

Hyde, J.S., Fennema, E., Lamon, S.J., 1990. Gender differences in mathematics performance: a meta-analysis. *Psychological bulletin* 107, 139.

Hyde, J.S., Linn, M.C., 1988. Gender differences in verbal ability: A meta-analysis. *Psychological bulletin* 104, 53.

Hyde, J.S., Mertz, J.E., 2009. Gender, culture, and mathematics performance. *Proceedings of the National Academy of Sciences* 106, 8801–8807.

Imbens, G.W., 2010. Better late than nothing: Some comments on deaton (2009) and heckman and urzua (2009). *Journal of Economic literature* 48, 399–423.

Imbens, G.W., Lemieux, T., 2008. Regression discontinuity designs: A guide to practice. *Journal of econometrics* 142, 615–635.

Iribarri, N., Rey-Biel, P., 2018. Competitive pressure widens the gender gap in performance: Evidence from a two-stage competition in mathematics. *The Economic Journal*. Accepted Author Manuscript

Jacob, B.A., 2002. Where the boys aren't: Non-cognitive skills, returns to school and the gender gap in higher education. *Economics of Education review* 21, 589–598.

Jamison, E.A., Jamison, D.T., Hanushek, E.A., 2007. The effects of education quality on income growth and mortality decline. *Economics of Education Review* 26, 771–788.

Joensen, J.S., Nielsen, H.S., 2009. Is there a causal effect of high school math on labor market outcomes? *Journal of Human Resources* 44, 171–198.

John, O.P., Srivastava, S., 1999. The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research* 2, 102–138.

Johns, M., Schmader, T., Martens, A., 2005. Knowing is half the battle: Teaching stereotype threat as a means of improving women's math performance. *Psychological Science* 16, 175–179.

Jones, E.B., Jackson, J.D., 1990. College grades and labor market rewards. *The Journal of Human Resources* 25, 253–266.

Kass, S.J., Beede, K.E., Vodanovich, S.J., 2010. Self-report measures of distractibility as correlates of simulated driving performance. *Accident Analysis & Prevention* 42, 874 – 880. doi:<https://doi.org/10.1016/j.aap.2009.04.012>. assessing Safety with Driving Simulators.

Kautz, T., Heckman, J.J., Diris, R., ter Weel, B., Borghans, L., 2014. Fostering and measuring skills: Improving cognitive and non-cognitive skills to promote lifetime success. *NBER Working Paper* No. 20749 .

Kimbrough, E.O., McGee, A.D., Shigeoka, H., 2017. How do peers impact learning? an experimental investigation of peer-to-peer teaching and ability tracking. *NBER Working Paper* No. 23439 .

Kimura, D., 2004. Human sex differences in cognition, fact, not predicament. *Sexualities, Evolution & Gender* 6, 45–53.

Kirby, A., McElroy, B., 2003. The effect of attendance on grade for first year economics students in university college cork. *The Economic and Social Review* 34, 311–326.

Van der Klaauw, W., 2002. Estimating the effect of financial aid offers on college enrollment: A regression–discontinuity approach. *International Economic Review* 43, 1249–1287.

Kottasz, R., et al., 2005. Reasons for student non-attendance at lectures and tutorials: An analysis. *Investigations in university teaching and learning* 2, 5–16.

Krueger, A.B., 1999. Experimental estimates of education production functions. *The quarterly journal of economics* 114, 497–532.

Krueger, A.B., Lindahl, M., 2001. Education for growth: Why and for whom? *Journal of Economic Literature* 39, 1101–1136.

Latif, E., Miles, S., 2013. Class attendance and academic performance: a panel data analysis. *Economic Papers: A journal of applied economics and policy* 32, 470–476.

Lavecchia, A.M., Liu, H., Oreopoulos, P., 2014. Behavioral economics of education: Progress and possibilities. *NBER Working Paper No. 20609* .

Lavy, V., Paserman, M.D., Schlosser, A., 2012a. Inside the black box of ability peer effects: Evidence from variation in the proportion of low achievers in the classroom. *The Economic Journal* 122, 208–237.

Lavy, V., Schlosser, A., 2011. Mechanisms and impacts of gender peer effects at school. *American Economic Journal: Applied Economics* 3, 1–33.

Lavy, V., Silva, O., Weinhardt, F., 2012b. The good, the bad, and the average: Evidence on ability peer effects in schools. *Journal of Labor Economics* 30, 367–414.

Lazear, E.P., 2001. Educational production. *The Quarterly Journal of Economics* 116, 777–803.

Lee, D.S., 2008. Randomized experiments from non-random selection in us house elections. *Journal of Econometrics* 142, 675–697.

Lee, D.S., Lemieux, T., 2010. Regression discontinuity designs in economics. *Journal of economic literature* 48, 281–355.

Lee, D.W., Lee, T.H., 1995. Human capital and economic growth tests based on the international evaluation of educational achievement. *Economics Letters* 47, 219–225.

Lee, J.W., Barro, R.J., 2001. Schooling quality in a cross-section of countries. *Economica* 68, 465–488.

Leuven, E., Oosterbeek, H., van der Klaauw, B., 2010. The effect of financial rewards on students' achievement: Evidence from a randomized experiment. *Journal of the European Economic Association* 8, 1243–1265.

Levin, H.M., 2012. More than just test scores. *Prospects* 42, 269–284.

Lin, T.F., Chen, J., 2006. Cumulative class attendance and exam performance. *Applied Economics Letters* 13, 937–942.

Lindberg, S.M., Hyde, J.S., Petersen, J.L., Linn, M.C., 2010. New trends in gender and mathematics performance: a meta-analysis. *Psychological bulletin* 136, 1123.

Linton, R., 1945. The cultural background of personality. New York: Appleton-Century.

Lu, F., Anderson, M.L., 2014. Peer effects in microenvironments: The benefits of homogeneous classroom groups. *Journal of Labor Economics* 33, 91–122.

Lyle, D.S., 2009. The effects of peer group heterogeneity on the production of human capital at west point. *American Economic Journal: Applied Economics* 1, 69–84.

Machin, S., Pekkarinen, T., 2008. Global sex differences in test score variability. *Science* .

Malkovsky, E., Merrifield, C., Goldberg, Y., Danckert, J., 2012. Exploring the relationship between boredom and sustained attention. *Experimental Brain Research* 221, 59–67.

Manski, C.F., 1993. Identification of endogenous social effects: The reflection problem. *The review of economic studies* 60, 531–542.

Marburger, D.R., 2001. Absenteeism and undergraduate exam performance. *The Journal of Economic Education* 32, 99–109.

Marburger, D.R., 2006. Does mandatory attendance improve student performance? *The Journal of Economic Education* 37, 148–155.

Marie, O., Zölitz, U., 2017. “high” achievers? cannabis access and academic performance. *The Review of Economic Studies* 84, 1210–1237.

Marmaros, D., Sacerdote, B., 2006. How do friendships form? *The Quarterly Journal of Economics* 121, 79–119.

Sala-i Martin, X., Doppelhofer, G., Miller, R.I., 2004. Determinants of long-term growth: A bayesian averaging of classical estimates (bace) approach. *American Economic Review* 94, 813–835. doi:10.1257/0002828042002570.

Martins, P.S., Walker, I., 2006. Student achievement and university classes: Effects of attendance, size, peers, and teachers. *IZA Discussion Paper No. 2490* .

McCrory, J., 2008. Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of econometrics* 142, 698–714.

Mendez, I., 2015. The effect of the intergenerational transmission of noncognitive skills on student performance. *Economics of Education Review* 46, 78 – 97.

Misra, R., McKean, M., 2000. College students' academic stress and its relation to their anxiety, time management, and leisure satisfaction. *American Journal of Health Studies* 16, 41.

Mueller, G., Plug, E., 2006. Estimating the effect of personality on male and female earnings. *Industrial & Labor Relations Review* 60, 3–22.

Murphy, K.M., Topel, R.H., 2002. Estimation and inference in two-step econometric models. *Journal of Business & Economic Statistics* 20, 88–97.

Murphy, L.O., Ross, S.M., 1990. Protagonist gender as a design variable in adapting mathematics story problems to learner interests. *Educational Technology Research and Development* 38, 27–37.

Nelson, R.R., Phelps, E.S., 1966. Investment in humans, technological diffusion, and economic growth. *The American Economic Review* 56, 69–75.

Niederle, M., Vesterlund, L., 2007. Do women shy away from competition? do men compete too much? *The Quarterly Journal of Economics* , 1067–1101.

Nollenberger, N., Rodríguez-Planas, N., Sevilla, A., 2016. The math gender gap: The role of culture. *American Economic Review* 106, 257–61. doi:10.1257/aer.p20161121.

Nunn, N., 2012. Culture and the historical process. *Economic History of Developing Regions* 27, 108–126.

OECD, 2009. PISA 2006 Technical Report. OECD Publishing.

OECD, 2012. PISA 2009 Technical Report. OECD Publishing.

OECD, 2014. PISA 2012 Technical Report. OECD Publishing.

OECD, 2015. PISA 2015 Technical Report. OECD Publishing.

Oosterbeek, H., Van Ewijk, R., 2014. Gender peer effects in university: Evidence from a randomized experiment. *Economics of Education Review* 38, 51–63.

Oreopoulos, P., 2007. Do dropouts drop out too soon? wealth, health and happiness from compulsory schooling. *Journal of Public Economics* 91, 2213–2229.

Oreopoulos, P., Petronijevic, U., 2013. Making college worth it: A review of research on the returns to higher education. *NBER Working Paper No. 19053* .

Ors, E., Palomino, F., Peyrache, E., 2013. Performance gender gap: does competition matter? *Journal of Labor Economics* 31, 443–499.

Paulhus, D.L., 1984. Two-component models of socially desirable responding. *Journal of personality and social psychology* 46, 598.

Pekkarinen, T., 2015. Gender differences in behaviour under competitive pressure: Evidence on omission patterns in university entrance examinations. *Journal of Economic Behavior & Organization* 115, 94–110.

Phelps, E.A., 2006. Emotion and cognition: insights from studies of the human amygdala. *Annual Review of Psycholgy* 57, 27–53.

Quinn, D.M., Cooc, N., 2015. Science achievement gaps by gender and race/ethnicity in elementary and middle school trends and predictors. *Educational Researcher* 44, 336–346.

Roberts, B.W., 2009. Back to the future: Personality and assessment and personality development. *Journal of Research in Personality* 43, 137–145.

Rodriguez-Planas, N., Nollenberger, N., 2018. Let the girls learn! it is not only about math ... it's about gender social norms. *Economics of Education Review* 62, 230 – 253. doi:<https://doi.org/10.1016/j.econedurev.2017.11.006>.

Romer, D., 1993. Do students go to class? should they? *Journal of Economic Perspectives* 7, 167–174. doi:[10.1257/jep.7.3.167](https://doi.org/10.1257/jep.7.3.167).

Romer, P.M., 1990. Endogenous technological change. *Journal of Political Economy* 98, pp. S71–S102.

Rotter, J.B., 1966. Generalized expectancies for internal versus external control of reinforcement. *Psychological monographs: General and applied* 80, 1.

Rubin, D.B., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* 66, 688.

Sacerdote, B., 2001. Peer effects with random assignment: Results for dartmouth roommates. *The Quarterly journal of economics* 116, 681–704.

Sacerdote, B., 2011. Peer effects in education: How might they work, how big are they and how much do we know thus far?, in: *Handbook of the Economics of Education*. Elsevier. volume 3, pp. 249–277.

Sacerdote, B., 2014. Experimental and quasi-experimental analysis of peer effects: two steps forward? *Annu. Rev. Econ.* 6, 253–272.

Sachs, J.D., Warner, A., Åslund, A., Fischer, S., 1995. Economic reform and the process of global integration. *Brookings papers on economic activity 1995*, 1–118.

Sachs, J.D., Warner, A.M., 1997. Fundamental sources of long-run growth. *The American Economic Review* 87, 184–188.

Sapienza, P., Guiso, L., Monte, F., Zingales, L., 2010. Culture, gender and math. *Science* , 1164–1165.

Schmitt, D.P., Realo, A., Voracek, M., Allik, J., 2008. Why can't a man be more like a woman? sex differences in big five personality traits across 55 cultures. *Journal of personality and social psychology* 94, 168.

Scott-Clayton, J., 2011. The shapeless river: Does a lack of structure inhibit students' progress at community colleges. *CCRC Working Paper No. 25* .

Segal, C., 2012. Working when no one is watching: Motivation, test scores, and economic success. *Management Science* 58, 1438–1457.

Sieversten, H.H., Gino, F., Piovesan, M., 2016. Cognitive fatigue influences students' performance on standardized tests. *Proceedings of the National Academy of Sciences* 113, 2621–2624.

Sjøberg, S., 2007. PISA and "Real Life Challenges": Mission Impossible? Technical Report. Contribution to Hopman (Ed).

Snyder, J.L., Lee-Partridge, J.E., Jarmoszko, A.T., Petkova, O., D'Onofrio, M.J., 2014. What is the influence of a compulsory attendance policy on absenteeism and performance? *Journal of Education for Business* 89, 433–440.

Stanca, L., 2006. The effects of attendance on academic performance: Panel data evidence for introductory microeconomics. *The Journal of Economic Education* 37, 251–266.

Steele, C.M., 1997. A threat in the air: How stereotypes shape intellectual identity and performance. *American psychologist* 52, 613.

Stinebrickner, R., Stinebrickner, T.R., 2006. What can be learned about peer effects using college roommates? evidence from new survey data and students from disadvantaged backgrounds. *Journal of public Economics* 90, 1435–1454.

Stinebrickner, R., Stinebrickner, T.R., 2008. The causal effect of studying on academic performance. *B.E. Journal of Economic Analysis and Policy* 8.

Sunde, U., Vischer, T., 2015. Human capital and growth: Specification matters. *Economica* 82, 368–390.

Tabellini, G., 2010. Culture and institutions: economic development in the regions of europe. *Journal of the European Economic Association* 8, 677–716.

Tangney, J.P., Baumeister, R.F., Boone, A.L., 2004. High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of personality* 72, 271–324.

Tannenbaum, D.I., 2012. Do gender differences in risk aversion explain the gender gap in sat scores? uncovering risk attitudes and the test score gap. Unpublished paper, University of Chicago, Chicago

Terrier, C., 2016. Boys lag behind: How teachers' gender biases affect student achievement. IZA Discussion Paper No. 10343 .

Thiemann, P., 2017. The persistent effects of short-term peer groups in higher education. IZA Discussion Paper No. 11024 .

Torija, P., 2012. Straightening pisa: When students do not want to answer standardized tests. Mimeo

Trueman, M., Hartley, J., 1996. A comparison between the time-management skills and academic performance of mature and traditional-entry university students. *Higher education* 32, 199–215.

UNESCO, 1998. World education report: Teachers and teaching in a changing world. Paris: UNESCO .

Vodanovich, S.J., Kass, S.J., 1990a. Age and gender differences in boredom proneness. *Journal of Social Behavior and Personality* 5, 297.

Vodanovich, S.J., Kass, S.J., 1990b. A factor analytic study of the boredom proneness scale. *Journal of Personality Assessment* 55, 115–123.

Vodanovich, S.J., Wallace, J.C., Kass, S.J., 2005. A confirmatory approach to the factor structure of the boredom proneness scale: Evidence for a two-factor short form. *Journal of Personality Assessment* 85, 295–303.

Wang, M.T., Degol, J.L., 2017. Gender gap in science, technology, engineering, and mathematics (stem): current knowledge, implications for practice, policy, and future directions. *Educational psychology review* 29, 119–140.

Webbink, D., Koning, P., Vujić, S., Martin, N.G., 2012. Why are criminals less educated than non-criminals? evidence from a cohort of young australian twins. *The Journal of Law, Economics, & Organization* 29, 115–144.

Wechsler, D., 1940. Nonintellectual factors in general intelligence , 444–445.

West, M.R., Kraft, M.A., Finn, A.S., Martin, R.E., Duckworth, A.L., Gabrieli, C.F., Gabrieli, J.D., 2016. Promise and paradox: Measuring students' non-cognitive skills and the impact of schooling. *Educational Evaluation and Policy Analysis* 38, 148–170.

Willingham, W.W., Cole, N.S., 2013. *Gender and fair assessment*. Routledge.

Woessmann, L., 2003. Schooling resources, educational institutions and student performance: the international evidence. *Oxford Bulletin of Economics and Statistics* 65, 117–170.

WorldBank, 2002. *World Development Indicators 2002*. World Bank Publications.

Ye, J., Han, S., Hu, Y., Coskun, B., Liu, M., Qin, H., Skiena, S., 2017. Nationality classification using name embeddings, in: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, ACM. pp. 1897–1906.

Zamarro, G., Cheng, A., Shakeel, M.D., Hitt, C., 2018. Comparing and validating measures of non-cognitive traits: Performance task measures and self-reports from a nationally representative internet panel. *Journal of Behavioral and Experimental Economics* 72, 51 – 60.

Zamarro, G., Hitt, C., Mendez, I., 2016. When students don't care: Reexamining international differences in achievement and non-cognitive skills. *EDRE Working Paper No. 2016-18* .

Zamarro, G., Nichols, M., Duckworth, A., D'Mello, S., 2017. Further validation of survey-effort measures of conscientiousness: results from a sample of high school students. *The Character Assesment Initiative Working Paper* .

Zimmerman, D.J., 2003. Peer effects in academic outcomes: Evidence from a natural experiment. *Review of Economics and statistics* 85, 9–23.

Zuckerman, M., Eysenck, S.B., Eysenck, H.J., 1978. Sensation seeking in england and america: Cross-cultural, age, and sex comparisons. *Journal of consulting and clinical psychology* 46, 139.

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:

684. U. TURMUNKH, *Ambiguity in Social Dilemmas*
685. U. KESKIN, *Essays on Decision Making: Intertemporal Choice and Uncertainty*
686. M. LAMMERS, *Financial Incentives and Job Choice*
687. Z. ZHANG, *Topics in Forecasting Macroeconomic Time Series*
688. X. XIAO, *Options and Higher Order Risk Premiums*
689. D.C. SMERDON, 'Everybody's doing it': *Essays on Trust, Norms and Integration*
690. S. SINGH, *Three Essays on the Insurance of Income Risk and Monetary Policy*
691. E. SILDE, *The Econometrics of Financial Comovement*
692. G. DE OLIVEIRA, *Coercion and Integration*
693. S. CHAN, *Wake Me up before you CoCo: Implications of Contingent Convertible Capital for Financial Regulation*
694. P. GAL, *Essays on the role of frictions for firms, sectors and the macroeconomy*
695. Z. FAN, *Essays on International Portfolio Choice and Asset Pricing under Financial Contagion*
696. H. ZHANG, *Dealing with Health and Health Care System Challenges in China: Assessing Health Determinants and Health Care Reforms*
697. M. VAN LENT, *Essays on Intrinsic Motivation of Students and Workers*
698. R.W. POLDERMANS, *Accuracy of Method of Moments Based Inference*
699. J.E. LUSTENHOUWER, *Monetary and Fiscal Policy under Bounded Rationality and Heterogeneous Expectations*
700. W. HUANG, *Trading and Clearing in Modern Times*
701. N. DE GROOT, *Evaluating Labor Market Policy in the Netherlands*

702. R.E.F. VAN MAURIK, *The Economics of Pension Reforms*

703. I. AYDOGAN, *Decisions from Experience and from Description: Beliefs and Probability Weighting*

704. T.B. CHILD, *Political Economy of Development, Conflict, and Business Networks*

705. O. HERLEM, *Three Stories on Influence*

706. J.D. ZHENG, *Social Identity and Social Preferences: An Empirical Exploration*

707. B.A. LOERAKKER, *On the Role of Bonding, Emotional Leadership, and Partner Choice in Games of Cooperation and Conflict*

708. L. ZIEGLER, *Social Networks, Marital Sorting and Job Matching. Three Essays in Labor Economics*

709. M.O. HOYER, *Social Preferences and Emotions in Repeated Interactions*

710. N. GHEBRIHWET, *Multinational Firms, Technology Transfer, and FDI Policy*

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. CIURILA, *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. THUMMEL, *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*



Education is an investment in human capital which has large positive impacts upon individual and social outcomes. This implies that policymakers have an enormous responsibility when using scarce resources as to design the education system. Policymakers like to make informed decisions, but this requires knowledge about the consequences of intended policies. This thesis aims at contributing to this knowledge by means of four self-contained chapters on the impact of design features in education. It contains, among others, a chapter on ability peer effects in the classroom. If peer effects exist, reorganizing students across classes could increase aggregate student performance without spending additional resources.

Matthijs Oosterveen (1991) obtained both his Bachelor's and Master's (Cum Laude) degree in Economics from the Erasmus University Rotterdam. In September 2014 he started his PhD at the Tinbergen Institute and the Department of Economics in Rotterdam. His research interests include Economics of Education, Labor Economics, and extend to the fields of Policy Evaluation and Applied Microeconomics.