

THREE ESSAYS ON CAUSAL EFFECTS: IDENTIFICATION IN A NOVEL SETTING AND TWO EMPIRICAL STUDIES

by

János K. Divényi

Submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy at
Central European University

Supervisors: Gábor Kézdi, Róbert Lieli

Budapest, Hungary

© Copyright by János K. Divényi, 2020.
All rights reserved.

10.14754/CEU.2020.12

CENTRAL EUROPEAN UNIVERSITY
DEPARTMENT OF ECONOMICS AND BUSINESS

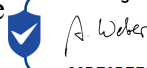
The undersigned hereby certify that they have read and recommend to the Department of Economics and Business for acceptance a thesis entitled **“Three Essays on Causal Effects: Identification in a Novel Setting and Two Empirical Studies”** by Janos Karoly Divenyi.

Dated: October 26, 2020

I certify that I have read this dissertation and in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Chair of the Thesis Committee

DocuSigned by:



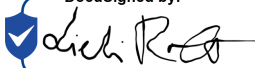
86DE0BEF29874CD...

Andrea Weber

I certify that I have read this dissertation and in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Advisor:

DocuSigned by:



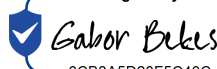
243BCFE3F32841C...

Robert Lieli

I certify that I have read this dissertation and in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Internal Examiner:

DocuSigned by:



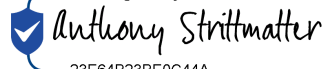
6CB3A5D20F5C40C...

Gabor Bekes

I certify that I have read this dissertation and in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

External Examiner:

DocuSigned by:



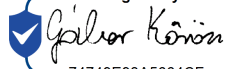
23E64B23BE0C44A...

Anthony Strittmatter

I certify that I have read this dissertation and in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

External Member:

DocuSigned by:



74749E93A5664CF...

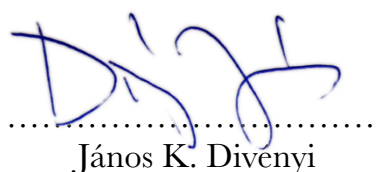
Gabor Korosi

CENTRAL EUROPEAN UNIVERSITY
DEPARTMENT OF ECONOMICS AND BUSINESS

Author: János K. Divényi
Title: Three Essays on Causal Effects:
Identification in a Novel Setting and Two Empirical Studies
Degree: Ph.D.
Dated: October 26, 2020

Hereby I testify that this thesis contains no material accepted for any other degree in any other institution and that it contains no material previously written and/or published by another person except where appropriate acknowledgement is made.

Signature of the author:



.....
János K. Divényi

Disclosure of coauthor contribution

Do Elite Schools Benefit the Average Student

Co-author: Sándor Sóvágó

Sándor Sóvágó came up with the idea to apply non-parametric bound estimation for analyzing the effect of elite schools in Hungary. Both authors contributed substantially to developing the precise theoretical arguments. The empirical analysis was largely implemented by János Divényi, while the paper was mainly written by Sándor Sóvágó.

Abstract

The thesis consists of three chapters on causal effects: in each of them, I aim to deepen our understanding of causal relationships. In the first chapter, I investigate how we can identify causal effects from adaptively collected data which is becoming more and more widespread in the modern on-line world. Running Monte-Carlo simulations I illustrate how adaptivity biases the standard treatment effect estimators and I recommend a new strategy that balances between the welfare and estimation goals of the decision-maker. The remaining two chapters are empirical studies aimed at estimating causal effects in two policy-relevant settings. In the second chapter, I examine the effect of retirement on cognitive performance, taking a unifying approach by replicating previous studies and tracing back the differences in their results to various identification problems. I propose a new identification strategy that suggests that the effect is close to zero. In the third chapter (co-authored with Sándor Sóvágó) we study the effect of elite schools on competency scores using non-parametric bound identification. Our results show that the effect is heterogeneous being more expressed for high-ability students; this result has important implications for the validity of previous studies that mainly estimate the effect around the admission cutoffs exploiting the discontinuity.

Chapter 1: Eliminating Bias in Treatment Effect Estimation Arising from Adaptively Collected Data

It is well understood that bandit algorithms that collect data adaptively - balancing between exploration and exploitation - can achieve higher average outcomes than the "experiment first, exploit later" approach of the traditional treatment choice literature. However, there has been much less work on how data arising from such algorithms can be used to estimate treatment effects. This paper contributes to this growing literature in three ways. First, a systematic simulation exercise characterizes the behavior of the standard average treatment effect estimator on adaptively collected data: I show that treatment effect estimation suffers from amplification bias and illustrate that this bias increases in noise and adaptivity. I also show that the traditional correction method of inverse propensity score weighting (IPW) can even exacerbate this bias. Second, I suggest an easy-to-implement bias correction method: limiting the adaptivity of the data collection by requiring sampling from all arms results in an unbiased IPW estimate. Lastly, I demonstrate a trade-off between two natural goals: maximizing expected welfare and having a good estimate of the treatment effect. I show that my correction method extends the set of choices regarding this trade-off, yielding higher expected welfare while allowing for an unbiased and relatively precise estimate.

Chapter 2: Examining the Effect of Retirement on Cognitive Performance – A Unifying Approach

Several recent works investigate the effect of retirement on cognitive performance, arriving at different conclusions. The key ingredient of the various approaches is how they handle the endogeneity of the retirement decision. In order to examine this issue more deeply, I replicate the results of previous works using three waves from the Survey of Health, Ageing and Retirement in Europe (SHARE). I draw attention to potential biases inherent in the standard instrumental variable identification strategies and assess their magnitudes. Based on the lessons learned, I propose a new instrument that utilizes the panel structure of the data, enabling the comparison of individual cognitive paths. I show that if retirement has any adverse effect on cognitive performance it must be really small in magnitude.

Chapter 3: Do Elite Schools Benefit their Students?

with Sándor Sóvágó

This paper studies the effects of enrollment in an elite school on elite-school students' academic achievement in Hungary. Enrollment in a Hungarian elite school entails having academically stronger peers and early switching to a secondary school. We examine effects for elite-school students throughout the outcome distribution using a mild stochastic dominance assumption. We find that enrollment in an elite school decreases female and low-ability students' mathematics test scores two years after enrollment. However, these negative effects are short-lived, and we obtain estimates that are consistent with substantial positive effects four years after enrollment. School value-added estimates lie within our non-parametric bounds, and confirm the positive effects on the medium run.

Acknowledgements

Writing the Acknowledgements section of someone's thesis is one of the most satisfying tasks. It comes after the real job is done. However, it is far from easy, especially if the person writing it spent so much time doing the real job as I did. I am going to try my best to list everyone who helped me to finally reach this important milestone.

First I have to thank my teachers prior to my graduate studies who gave me strong fundamentals and inspired me to pursue a deeper understanding in the field of econometrics (a non-exhaustive list): Zsolt Vízhányó, Béla Futó, László Görbe, Balázs Varga, Zsombor Ligeti, Ádám Reiff, Róbert Lieli and Gábor Kézdi.

It would have been impossible for me to write this thesis without the support of my advisors. Gábor Kézdi taught me to seek an intuitive understanding of complicated formulas. He opened a lot of valuable opportunities for me to try myself in academia, both as a teacher and as a researcher. Róbert Lieli admitted me among his students when it was already clear that I will not follow an academic career. In spite of that, he relentlessly pushed me forward and looked at everything I sent to him with unprecedented care.

Sándor Sóvágó, my co-author and friend, stood always next to me. Working with him has been not only inspiring but also fun. His constant support and encouragement were essential for me to finish up my theses.

I am indebted to the examiners of my thesis, Gábor Békés and Anthony Strittmatter, who gave constructive comments that helped me a lot to improve upon the quality of my papers. I also appreciate the work of the other members of the Committee, Andrea Weber (chair) and Gábor Kőrosi (external member) who made my defense possible.

I learnt a lot from discussions of my fellow PhD students, Márta Bisztray, Gergely Hajdu, Kinga Marczell, Bálint Menyhért, Ágoston Reguly, István Szabó, and Péter Zsohár. I would like to particularly mention Jenő Pál whose friendship and continuous support have been fundamental assets.

Veronika Orosz, Corinne Freiburger, Márta Jombach, and Katalin Szimler, staff members at the Department of Economics and Business, have always been very helpful and freed me of any administrative burden.

In the last 5 years, I worked on my theses beside my day-to-day job as a data scientist at Emarsys. My manager, Levente Otti, supported me in every possible way. Without his backing, it would have been much harder to complete my thesis.

I would like to express my utmost gratitude to my family. All of them helped me in their own way, my father being always my number one supporter. The feeling that I

can make him proud provided the energy and motivation I crucially needed over the years.

The most important companion in this journey has been my wife. She unsuspectingly said yes to my proposal just before I started my Ph.D. She endured with endless patience and understanding how I suffered from fruitless research, worked through weekends before important presentations, and regularly consumed my days off for working on the thesis instead of spending time with our kids. I cannot be grateful enough for her unconditional love and support.

All of these people's help, and all the hard work I put in would have been useless had the loving God not blessed my efforts. As Saint Paul writes in his Letter to the Romans: *"Non est volentis, neque currentis, sed miserentis Dei"*.

Contents

1	<i>Eliminating Bias in Treatment Effect Estimation Arising from Adaptively Collected Data</i>	1
1.1	<i>Introduction</i>	1
1.2	<i>Setup</i>	5
1.3	<i>Demonstration of welfare and estimation properties</i>	7
1.3.1	<i>Parametrization</i>	7
1.3.2	<i>Welfare</i>	7
1.3.3	<i>Estimation</i>	9
1.3.4	<i>Welfare-Estimation Trade-off</i>	11
1.4	<i>Bias correction</i>	12
1.4.1	<i>Inverse Propensity Weighting (IPW)</i>	12
1.4.2	<i>Using the first batch only</i>	14
1.4.3	<i>Limiting the propensity scores</i>	15
1.5	<i>Monte Carlo Simulation</i>	19
1.5.1	<i>Uncertainty</i>	19
1.5.2	<i>Horizon</i>	25
1.5.3	<i>Non-Gaussian Potential Outcomes</i>	26
1.6	<i>Data-driven simulations</i>	29
1.7	<i>Concluding remarks</i>	32

2	<i>Examining the Effect of Retirement on Cognitive Performance - A Unifying Approach</i>	34
2.1	<i>Introduction</i>	34
2.2	<i>Model</i>	37
2.3	<i>Data</i>	38
2.4	<i>Replications</i>	39
2.4.1	<i>Rohwedder and Willis (2010)</i>	40
2.4.2	<i>Mazzonna and Peracchi (2012)</i>	41
2.4.3	<i>Bonsang et al. (2012)</i>	46
2.5	<i>My strategy</i>	47
2.6	<i>Concluding remarks</i>	52
3	<i>Do Elite Schools Benefit the Average Student</i>	54
3.1	<i>Introduction</i>	54
3.2	<i>Context: Elite schools in Hungary</i>	57
3.3	<i>Data and summary statistics</i>	58
3.3.1	<i>Data</i>	58
3.3.2	<i>Summary statistics</i>	59
3.4	<i>Empirical strategies</i>	63
3.4.1	<i>Non-parametric bounds: conditional MTS throughout the outcome distribution</i>	63
3.4.2	<i>Validity check</i>	65
3.4.3	<i>School value-added</i>	67
3.5	<i>Results</i>	68
3.5.1	<i>Short-run academic achievement</i>	69

3.5.2	<i>Medium-run academic achievement</i>	71
3.5.3	<i>School value-added</i>	72
3.6	<i>Conclusions</i>	77
A	<i>Appendix for Chapter 1</i>	84
A.1	<i>Simulation distributions</i>	84
A.2	<i>Detailed simulation results</i>	89
B	<i>Appendix for Chapter 2</i>	101
B.1	<i>Comparison of methodologies in the literature</i>	101
B.2	<i>Additional tables for the replication exercises</i>	101
B.3	<i>Detailed estimation tables for my strategy</i>	105
C	<i>Appendix for Chapter 3</i>	114
C.1	<i>Additional Tables</i>	114
C.1.1	<i>Data</i>	114
C.1.2	<i>Validity check</i>	116
C.1.3	<i>Results</i>	118
C.2	<i>The relative effects of elite-school enrollment</i>	121
C.3	<i>Data (for online publication)</i>	128
C.3.1	<i>Sample restrictions</i>	128
C.3.2	<i>Variable description</i>	128
C.3.3	<i>Imputation</i>	129

List of Tables

1.1	Comparison of $\hat{\tau}_{IPW}$ by number of batches with control assignment	14
1.2	Descriptive statistics of JTPA experiment	30
2.1	Comparing the methodology of Rohwedder and Willis (2010) by two versions of the instrumental variable: 2SLS estimation	41
2.2	Moving from the strategy of Rohwedder and Willis (2010) to that of Mazzonna and Peracchi (2012)	43
2.3	Moving to the strategy of Mazzonna and Peracchi (2012) - the effect of gender control for different measures of cognitive performance	44
2.4	Estimating separately by gender, closest to Mazzonna and Peracchi (2012) . . .	45
2.5	Replication of Bonsang et al. (2012)	47
2.6	Summary statistics	49
2.7	Panel estimation: change in total word recall score between wave 1 and 4 . . .	51
3.1	The overview of the education system in Hungary	57
3.2	Summary statistics	61
A.1	Expected welfare for different strategies ($n = 10,000$)	89
A.2	Bias for different strategies ($n = 10,000$)	91
A.3	MSE for different strategies ($n = 10,000$)	93
A.4	Expected welfare for different strategies ($\sigma = 10$)	96
A.5	Bias for different strategies ($\sigma = 10$)	97

A.6	<i>MSE for different strategies ($\sigma = 10$)</i>	98
B.1	<i>Comparing the methodologies of the literature</i>	101
B.2	<i>Comparing the methodology of Rohwedder and Willis (2010) by two versions of the instrumental variable: first stage</i>	102
B.3	<i>Moving from the strategy of Rohwedder and Willis (2010) to that of Mazzonna and Peracchi (2012): first stage</i>	102
B.4	<i>First stages, FE-IV estimation mimicing Bonsang et al. (2012)</i>	103
B.5	<i>Replication of Bonsang et al. (2012) on various subsamples</i>	103
B.6	<i>First stages, replication of Bonsang et al. (2012) on various subsamples</i>	104
B.7	<i>Panel estimation: change in total word recall score between wave 1 and 4: first stage</i>	105
B.8	<i>Panel estimation: change in total word recall score between wave 1 and 2</i>	106
B.9	<i>Panel estimation: change in total word recall score between wave 1 and 2: first stage</i>	106
B.10	<i>Panel estimation: change in total word recall score between wave 2 and 4</i>	107
B.11	<i>Panel estimation: change in total word recall score between wave 2 and 4: first stage</i>	107
B.12	<i>Panel estimation: change in numeracy score between wave 1 and 4</i>	108
B.13	<i>Panel estimation: change in numeracy score between wave 1 and 4: first stage</i>	108
B.14	<i>Panel estimation: change in numeracy score between wave 1 and 2</i>	109
B.15	<i>Panel estimation: change in numeracy score between wave 1 and 2: first stage</i>	109
B.16	<i>Panel estimation: change in numeracy score between wave 2 and 4</i>	110
B.17	<i>Panel estimation: change in numeracy score between wave 2 and 4: first stage</i>	110
B.18	<i>Panel estimation: change in fluency score between wave 1 and 4</i>	111
B.19	<i>Panel estimation: change in fluency score between wave 1 and 4: first stage</i>	111
B.20	<i>Panel estimation: change in fluency score between wave 1 and 2</i>	112

<i>B.21 Panel estimation: change in fluency score between wave 1 and 2: first stage . .</i>	<i>112</i>
<i>B.22 Panel estimation: change in fluency score between wave 2 and 4</i>	<i>113</i>
<i>B.23 Panel estimation: change in fluency score between wave 2 and 4: first stage . .</i>	<i>113</i>
<i>C.1 Additional summary statistics</i>	<i>115</i>
<i>C1 Evolution of the sample size</i>	<i>128</i>

List of Figures

1.1	<i>Expected welfare by batch size</i>	8
1.2	<i>Evolution of bandit algorithms</i>	9
1.3	<i>Bias in group mean estimates by batch size</i>	10
1.4	<i>Density of mean estimates, using the first batch versus the whole sample</i>	11
1.5	<i>Performance of the bandit assignment rule in the welfare-estimation space . . .</i>	12
1.6	<i>Batch average for the control mean across batches ($n_B = 1000$)</i>	14
1.7	<i>Performance of different strategies in the welfare-estimation space</i>	16
1.8	<i>Welfare and estimation performance of the LTS-IPW strategy</i>	17
1.9	<i>Performance of different strategies in the welfare-estimation space</i>	18
1.10	<i>Expected total welfare and bias</i>	20
1.11	<i>Average treated share in the second batch</i>	21
1.12	<i>Performance of different strategies in the welfare-estimation space</i>	22
1.13	<i>Expected welfare of different combinations of n_B and L</i>	23
1.14	<i>MSE of different combinations of n_B and L</i>	24
1.15	<i>Welfare performance of bandit algorithm with various levels of adaptivity across different horizons</i>	26
1.16	<i>Performance of different strategies in the welfare-estimation space, for different horizons</i>	27
1.17	<i>Welfare performance by different strategies compared by the distribution of the potential outcome</i>	28

1.18	<i>Estimation performance by different strategies compared by the distribution of the potential outcome</i>	29
1.19	<i>Welfare-estimation trade-off for the JTPA experiment</i>	31
2.1	<i>Pattern of cognitive scores across waves by working history</i>	50
2.2	<i>Comparison of the retirement effect estimate by gender</i>	52
3.1	<i>Peer quality and elite-school enrollment</i>	62
3.2	<i>Validity check: Elite-school enrollment and student characteristics</i>	66
3.3	<i>Validity-check: The distribution of students' 6th-grade standardized test scores by elite-school enrollment</i>	66
3.4	<i>Validity-check: The p-values of the Kolgomorov-Smirnov test</i>	67
3.5	<i>The effect of elite-school enrollment on the distribution of elite-school students' 8th-grade mathematics grade</i>	70
3.6	<i>The effect of elite-school enrollment on the distribution of elite-school students' 8th-grade standardized test scores</i>	71
3.7	<i>The effect of elite-school enrollment on the distribution of elite-school students' 8th-grade standardized test scores: Comprehensive schools</i>	72
3.8	<i>The effect of elite-school enrollment on the distribution of elite-school students' 10th-grade standardized test scores</i>	73
3.9	<i>The effect of elite-school enrollment on the distribution of elite-school students' 10th-grade standardized test scores: Elite-school subsample</i>	74
3.10	<i>The effect of elite-school enrollment on the distribution of elite-school students' 8th-grade test scores: School VA</i>	75
3.11	<i>The effect of enrollment in an elite school on the distribution of elite-school students' 10th-grade test scores: School VA</i>	76
A.1	<i>Distribution of welfare by batch size</i>	85
A.2	<i>Distribution of $\hat{\tau}_0$ by batch size</i>	86
A.3	<i>Distribution of $\hat{\tau}_{IPW}$ by batch size</i>	87

A.4	<i>Distribution of $\hat{\tau}_{FB}$ by batch size</i>	88
C.1	<i>Validity check: Elite-school enrollment and student characteristics – 10th-grade sample</i>	116
C.2	<i>Validity check: The p-values of the Kolgomorov-Smirnov test – 10th-grade sample</i>	117
C.3	<i>The effect of elite-school enrollment on the distribution of elite-school students' 8th-grade standardized test scores: Non-comprehensive schools</i>	118
C.4	<i>The effect of elite-school enrollment on the distribution of elite-school students' 8th-grade standardized test scores: Bounds and school VA</i>	119
C.5	<i>The effect of elite-school enrollment on the distribution of elite-school students' 10th-grade standardized test scores: Bounds and school VA</i>	120
B1	<i>The relative effect of elite-school enrollment on the distribution of elite-school students' 8th-grade mathematics grade</i>	121
B2	<i>The relative effect of elite-school enrollment on the distribution of elite-school students' 8th-grade standardized test scores</i>	122
B3	<i>The relative effect of elite-school enrollment on the distribution of elite-school students' 8th-grade standardized test scores: Comprehensive schools</i>	123
B4	<i>The relative effect of elite-school enrollment on the distribution of elite-school students' 10th-grade standardized test scores</i>	124
B5	<i>The relative effect of elite-school enrollment on the distribution of elite-school students' 10th-grade standardized test scores: Elite secondary grammar schools</i>	125
B6	<i>The relative effect of elite-school enrollment on the distribution of elite-school students' 8th-grade standardized test scores: School VA</i>	126
B7	<i>The relative effect of elite-school enrollment on the distribution of elite-school students' 10th-grade standardized test scores: School VA</i>	127

Chapter 1

Eliminating Bias in Treatment Effect Estimation Arising from Adaptively Collected Data

1.1 Introduction

We are often interested in whether an innovative treatment should be introduced and applied for individuals arriving in succession. Suppose an online shop wants to change its pricing scheme. They can experiment with a new scheme introducing it to part of their daily visitors, with the ultimate goal of applying the better scheme as soon as possible to maximize their profit. Once they change to the new scheme, they also want to know how much value they can hope from it for their next year's budget, i.e. they also want to measure the treatment effect.

This problem is ubiquitous today. Innovation is crucial to survival. We want to apply the procedure that yields the best expected outcome according to our current knowledge (status quo) but we also want to experiment with new ideas that might yield even higher outcome (exploitation versus exploration, earning versus learning). We are also interested in learning what to expect from introducing an innovation.

The standard procedure in economics to decide on the introduction of a new pricing scheme is to first learn its effect, and then to introduce it if the effect is positive. The traditional treatment choice literature (e.g. Manski, 2004; Dehejia, 2005; Hirano and Porter, 2009; Kitagawa and Tetenov, 2018; Athey and Wager, 2019) assumes that an experimental sample with randomized assignment exists and derives the welfare-maximizing policy rule given the information that can be learnt on the previously collected data. The welfare of the experimental subjects is disregarded. However, in practice, exploration and exploitation do not naturally separate. The

decision-maker always decides (sometimes unconsciously) whether it is worth experimenting or simply applying the best practice.

Multi-armed bandit algorithms (for comprehensive reviews see, e.g. Lattimore and Szepesvári, 2019; Slivkins, 2019) seek to optimize the exploration-exploitation trade-off suggesting heuristic rules that "learn and earn" in parallel. Instead of aiming for a one-off decision, they involve a sequence of decisions where each decision balances between experimenting and exploiting. As such, it is suitable for situations where the feedback is quick (as in our pricing scheme example). The goal is to maximize the expected welfare during the whole process, including the experimentation phase. Bandit algorithms continuously balance between choosing the treatment arm with the highest expected payoff (exploitation) and choosing treatment arms that are not yet known well (exploration) – the result of each decision contributes to later decisions. There is a quickly evolving literature (in the field of computer science) that investigates different algorithms in different setups and prove their optimality by various criteria. As algorithms aim to find the arm with the highest expected reward (or finding the better pricing scheme), measuring the exact effect of the various arms relative to a baseline is not part of the problem considered.

My paper is at the intersection of the traditional treatment choice literature of econometrics and the growing literature on multi-armed bandits of machine learning. I consider situations similar to the online shop example above, where the decision-maker assigns individuals to different treatments with two goals in mind: (1) maximizing profit (or welfare) and (2) estimating the treatment effect. There are two treatments (status quo and innovation, control and treatment) and individuals arriving in groups or batches should be assigned to one of them. The individual-level treatment effect is fixed but its magnitude (relative to the variation in the potential outcomes) is ex ante unknown. The length of the process (total number of arriving individuals, also called as "horizon") is finite but also unknown. The size of the batches, ie. the frequency of allocation decisions is controlled by the decision-maker.

I run Monte Carlo simulations to understand the welfare and estimation behavior of different strategies in this setup. I study a well-known multi-armed bandit heuristic, Thompson sampling, suggested by Thompson (1933). I chose this method because it is one of the most well-known algorithms, it is widely used in the industry (see e.g. Graepel et al., 2010; Scott, 2010) and it is a probabilistic rule that has some appealing features I am going to rely on later. However, the focus is not on the specific heuristic, but on the basic features of adaptively collected data when used for statistical inference. All of my results should extend to other popular heuristics that are deterministic, such as the Upper Confidence Bound algorithm (see e.g. Lai and Robbins, 1985).

What we know so far *The welfare performance of bandit algorithms in a stochastic context are measured by their expected reward (total welfare) relative to the reward gained by the best*

possible assignment policy (which is usually infeasible). The difference between these two measures is the expected regret. Each bandit can be characterized by their worst-case regret (within a given set of environments formed by the distribution of rewards and the length of the horizon). The seminal paper of [Lai and Robbins \(1985\)](#) derived an asymptotic lower bound on regret that any bandit algorithm should suffer.

Recent papers ([Agrawal and Goyal, 2012, 2013](#); [Korda et al., 2013](#)) prove that Thompson sampling is asymptotically optimal in terms of regret in various settings. [Perchet et al. \(2016\)](#) extends their result to batched bandits, where individuals arrive in groups (or batches) instead of one-by-one. The traditional solution in econometrics to experiment first and form an appropriate assignment rule later is welfare-suboptimal (see e.g. [Lattimore and Szepesvári, 2019](#)).

There are much less result that considers estimation after bandits. [Nie et al. \(2018\)](#) prove in theory that the estimated means of the treatment arms suffer from negative bias. They suggest a complex modification of the data collection process that can eliminate the bias.

[Villar et al. \(2015\)](#) compare various bandit algorithms in terms of outcome and also estimation performance in a simulated clinical trial. They show biased treatment effect estimations simulating many different multi-armed bandit algorithms.

My contribution To my knowledge, this is the first paper that considers welfare and estimation goals parallel and compares different strategies in the welfare-estimation space. I have three main contributions to the literature:

First, I characterize the welfare and estimation behavior of Thompson sampling and the traditional treatment effect estimator on adaptively collected data. I show that, generally, smaller batch size (ie. deciding more often) increases the expected welfare. However, if adaptivity is too quick adaptivity (the batch size is below a certain cutoff) the welfare cost of higher volatility outweighs the gains from smaller opportunity cost. Quicker adaptivity also increases the negative bias in means (for which I provide an intuitive explanation) that results in a larger amplification bias in the treatment effect estimate. These results highlight an important trade-off: strategies that achieve high welfare (adaptive algorithms) lead to highly biased treatment effect estimates - whereas running a randomized controlled trial on the whole sample (the gold standard for measuring the effect) suffers from a huge opportunity cost (resulting from assigning too many individuals to the inferior treatment).

Second, I prove that inverse propensity weighting (IPW) – traditionally used for bias correction – is equivalent to taking the simple averages of the batch averages (if the propensity weights are estimated). I show that in this setup, IPW does not work – in fact, it can even exacerbate the bias.

Finally, I suggest an easy-to-implement bias correction method: limiting the propensity scores away from the extremes that practically moderates the adaptivity of the data collection by requiring sampling from both arms in each batch. This assignment rule allows for unbiased inverse-propensity-weighted treatment effect estimate, whereas it preserves almost all of the welfare gain stemming from adaptivity. I show that limiting extends the set of choices regarding the welfare-estimation trade-off relative to some established strategies (such as the standard "explore first, exploit later" or explore-then-commit strategy).

Related recent literature A recent paper of *Hadad et al. (2019)* deals with a similar problem: they suggest data-adaptive weighting schemes to correct the standard treatment effect estimator on adaptively collected data, also ensuring asymptotic normality to make statistical inference possible. They deal only with estimation, and do not consider welfare.

Dimakopoulou et al. (2018) look at so called contextual bandits that include observable variables in the algorithms to capture heterogeneity in the treatment effect. They focus on bias in treatment effect originating from imbalances in the observables. In contrast, I focus on the general characteristics of the standard treatment effect estimator that are apparent even if the effect itself is constant.

A new line of research focuses on optimal experimentation design where the goal is to learn the treatment effect (see *Kasy (2016)* for one-off experiments, and *Hahn et al. (2011)* for adaptive experiments). Another deals with adaptive treatment assignment where the goal is to choose among a set of policies for large-scale implementation (*Kasy and Sautmann, 2019*). The latter's setup is especially close to mine but there is a major difference: these works assume away the welfare of the experimental subjects and only focus on learning. I consider both welfare and estimation under adaptive treatment assignment.

This paper The paper is structured as follows. Section 1.2 gives a formal setup for the problem. Section 1.3 characterizes the basic welfare and estimation properties of the bandit assignment rule using the standard treatment effect estimate and shows the welfare-estimation trade-off. Section 1.4 discusses different methods for correcting the bias: inverse-propensity weighting, first batch treatment effect and propensity score limiting. Section 1.5 demonstrates the results of the systematic Monte Carlo simulation which illustrate the behavior of the previously discussed strategies in different scenarios. Section 1.6 assesses the simulation results in a practically relevant setting using data from the well-known National Job Training Partnership Act (JTPA) study. Section 1.7 concludes.

1.2 Setup

There is a set of n individuals indexed by $i \in \{1, \dots, n\}$ whose outcome Y is of interest. There is a binary treatment $W_i \in \{0, 1\}$ where $W_i = 0$ stands for the no-treatment case, i.e. the status quo. $\{Y_i(1), Y_i(0)\}$ are potential outcomes that would have been observed for individual i with or without the treatment (potential outcomes might include the cost of the corresponding treatment). The actual (observed) outcome is $Y_i = Y_i(1)W_i + Y_i(0)(1 - W_i)$. Let us denote the expected value of the potential outcomes by $\mu_w = E[Y_i(w)]$, for $w \in \{0, 1\}$. The individual-level treatment effect is fixed, i.e. $Y_i(1) = Y_i(0) + \tau$ for each i where τ denotes the treatment effect. Therefore, the population is characterized by $\{Y_i(0)\}_{i=1}^n$. For simplicity, I assume $Y(0)$ is Gaussian with known variance (I show in Section 1.5.3 that the Gaussian assumption is only technical, the main results stand for skewed and fat-tailed distributions as well as long as they have finite means).

Individuals arrive randomly in equal-sized batches denoted by B and indexed by $j \in \{1, \dots, m\}$. The batch size is under the control of the decision-maker¹ and is denoted by n_B so $mn_B = n$. Arrival is sequential and the outcome is observed right after the assignment. The process can be described as follows:

1. A group of individuals $i \in B_j$ arrive, and are assigned to either treatment or control.
2. Outcomes $\{Y_i\}_{i \in B_j}$ are observed.
3. A next group of individuals $i \in B_{j+1}$ arrive and the first two steps are repeated.

Let us denote the observed history (assignments and outcomes) up until the k th batch by $H^{(k)} = \{Y_i, W_i\}_{i \in \bigcup_{j=1}^k B_j}$. Therefore, the whole history of n individuals is $H^{(m)}$.

The decision-maker has two goals: she wants to maximize profit (or welfare) based on outcomes, and she also wants to estimate the treatment effect τ with an unbiased, precise estimator. She decides about two things in parallel:

1. **assignment rule** A function that maps the history to a probability that expresses the share of the next batch assigned to the treatment: $\pi(H^{(k)}) = \mathbb{P}(W_i = 1 | i \in B_{k+1}) = p_{k+1}$. The choice of assignment rule incorporates the choice of batch size as well: $n_B = |B_k|$.
2. **estimation method** A function that maps the whole history (observed data of the population) to a number that expresses the treatment effect: $\hat{\tau}(H^{(m)})$.

¹It is natural to assume that the decision-maker has some control over the batch size. Even if the arrival of individuals is dictated by an external process, one can still increase the batch size by collapsing original batches. How frequently the decision-maker decides about allocation is a decision itself.

I will call a combination of an assignment rule and an estimation method a **strategy**. The decision-maker chooses a strategy to pursue both of her goals. Throughout this paper I use two simple objective functions to measure these goals:

1. **welfare goal** $\max \sum_{i=1}^n Y_i$ ²
2. **estimation goal** $\min E [(\hat{\tau} - \tau)^2]$ subject to $E[\hat{\tau}] = \tau$ ³

To illustrate adaptive assignment rules that blend exploitation with exploration I use an old heuristic, the Thompson Sampling (Thompson, 1933). It suggests to assign each individual to treatment by the probability that corresponds to your actual beliefs that the treatment outcome is the highest⁴. I implement this rule as follows (for a chosen batch size):

Thompson Sampling (TS)

1. Split the first batch equally between treatment and control.
2. Form beliefs about the treatment and control means by deriving posterior distributions using normal density with calculated averages (recall the known-variance assumption)^a:

$$\mathcal{N} \left(\hat{\mu}_1^{(k)}, \frac{\sigma^2}{n_1^{(k)}} \right) \text{ for treatment, and } \mathcal{N} \left(\hat{\mu}_0^{(k)}, \frac{\sigma^2}{n_0^{(k)}} \right) \text{ for control,}$$

where

$$n_1^{(k)} = \sum_{i \in \cup_{j=1}^k B_j} W_i, \quad n_0^{(k)} = \sum_{i \in \cup_{j=1}^k B_j} (1 - W_i).$$

3. Calculate the probability that the treatment mean is higher than the control mean (let us denote it with $r^{(k)}$). Technically, this can be achieved by sampling from the corresponding distributions.
4. Split the next batch according to this probability: $p_{k+1} = r^{(k)}$
5. Repeat from step (2) until assigning the last batch.

^aThis is equivalent to the posterior of mean of a normal variable with known variance using non-informative Jeffreys prior

²Assuming the outcome contains the cost of treatment, it is the profit of a firm. Assuming a utilitarian social welfare function, it is the total welfare.

³Recall the bias-variance decomposition: $E [(\hat{\tau} - \tau)^2] = (E[\hat{\tau}] - \tau)^2 + E[\hat{\tau}^2] - E^2[\hat{\tau}]$ where the last two terms give the variance of the estimator. So minimizing the mean-squared error is just minimizing the variance if the estimator is unbiased.

⁴For more detail, see Russo et al. (2017)

Intuitively, we will choose the treatment more likely (for a larger fraction of individuals in the batch) if (1) we are uncertain about its expected outcome (exploration), or (2) we are certain that its expected outcome is high (exploitation).

1.3 Demonstration of welfare and estimation properties

1.3.1 Parametrization

I assume – without loss of generality – a positive average treatment effect with unit value ($\tau = 1$). The population consists of $n = 10,000$ individuals, the potential outcomes are Gaussian with $\sigma = 10$. The noise-to-signal ratio is high to make the treatment effect hard to measure, and thus, the problem interesting. The potential outcomes are constructed such that $\mu_1 = 1$ and $\mu_0 = 0$ within the population. The minimum batch size is 10 (where $m = 1000$), and I simulate the following choices for the decision-maker: $n_B \in \{10, 20, 50, 100, 200, 500, 1000, 2000, 5000, 10000\}$. The maximum value corresponds to a simple random split on the whole sample.

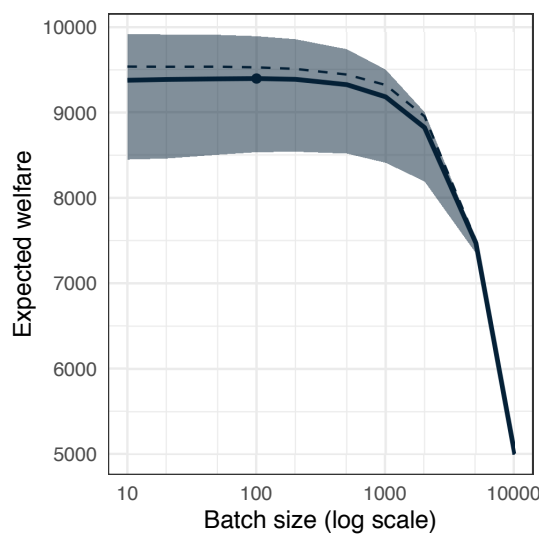
In this setup, the (infeasible) optimal treatment rule is to treat everyone ($\pi = 1$) that would achieve a total welfare of 10,000. Due to the fact that the treatment effect is normalized and is fixed for everyone, the sum of outcomes equals to the sum of individuals assigned to the treated, so both measures express the total welfare.

I run 20,000 simulations for each assignment rule. The runs differ only in the sequence of how the individuals arrive; they all use the same population of 10,000 with the average of potential outcomes equaling to 0 and 1, respectively.

1.3.2 Welfare

One would expect that smaller batch size (more batches, quicker adaptivity) leads to higher welfare, as it extends the possibilities of the policy maker. Also, as the first batch is a simple random split, the maximum welfare an adaptive rule could achieve in the best case is $10,000 - \frac{n_b}{2}$. Smaller batch sizes give the chance of reacting more quickly to a positive treatment effect, hence, suffering less opportunity cost.

However, the simulation results only partially justify this expectation. Figure 1.1 shows the expected welfare by batch size: generally, smaller batch size leads to higher expected welfare, but focusing on the small batch size region (left panel) reveals that being too “quick” can also do harm; the optimum is around $n_B = 100$. The reason for this is that being more adaptive

Figure 1.1: Expected welfare by batch size

Notes: The figure shows the expected welfare by batch size using a logarithmic scale to focus on the interesting region. The shaded area shows the 90% confidence interval, the dashed line depicts the median, the point highlights the batch size with maximum expected welfare. Smaller batches (quicker adaptivity) generally lead to higher welfare, but only until a certain point: really small batch size can harm. Number of simulations = 20,000.

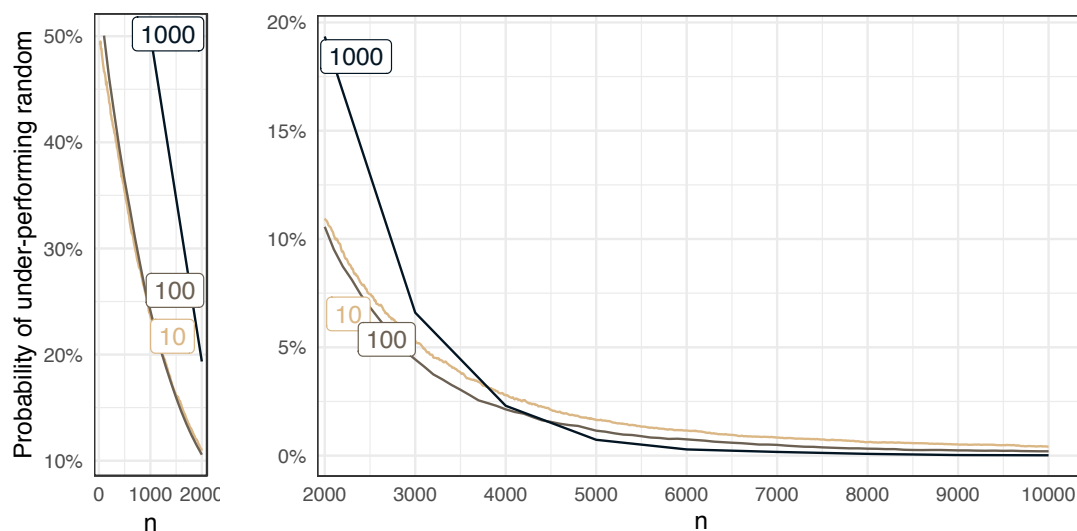
means deciding based on more volatile estimates that increases the probability of adapting to the wrong pattern (in this case, "learning" a negative treatment effect)⁵. Under a certain threshold of batch size, the loss on volatility seems to outweigh the gain on opportunity cost⁶.

Figure 1.2 illustrates this phenomenon by showing the probability of under-performing a simple random split in terms of welfare at each point of the process, for different batch sizes. At the beginning, quicker adaptivity allows for smaller opportunity cost as smaller batch sizes mean that the algorithm can allocate less people to the inferior treatment (recall that the first batch size is a random split). However, quicker adaptivity also means making decisions based on more volatile measures due to smaller sample sizes. These decisions turn out more likely to be false, therefore, the probability of under-performing remains relatively high at the later stages of the process. The welfare result of Figure 1.1 originates from these two contradicting processes.

The fact that for this given setup a constrained algorithm works better than a less constrained one does not contradict to the literature. The Thompson Sampling algorithm is a general solution, working well in different setups whose parameters (mainly τ and n) are ex-ante unknown.

⁵Figure A.1 in the Appendix shows the whole distribution of welfare for each batch size: the achieved welfare (that is equivalent to the number of individuals assigned to the treatment) is much more volatile for smaller batch sizes

⁶The behavior of the batch size parameter lets us raise an interesting analogy from the machine learning literature: regularization (see e.g. Hastie et al., 2001) is a technique that discourages learning a too complex or flexible model (e.g. by shrinking coefficients). Regularization leads to higher bias to gain on variance, increasing predictive accuracy. In our case, larger batch size means more regularization: it constrains the set of choices and loses on opportunity cost at the beginning, but wins on generalization in the longer term – especially if the noise is high.

Figure 1.2: Evolution of bandit algorithms

Notes: Each point depicts the probability that the bandit algorithm under-performs a simple random split after the first n arriving individuals (evaluated across the simulation runs). Quicker adaptivity results in smaller opportunity cost at the beginning (left panel), but leads to higher probability of getting wrong at later stages (right panel). Number of simulations = 20,000.

As we are going to see later, avoiding too small batches helps only if the noise is high, or equivalently, if the treatment effect is small.

1.3.3 Estimation

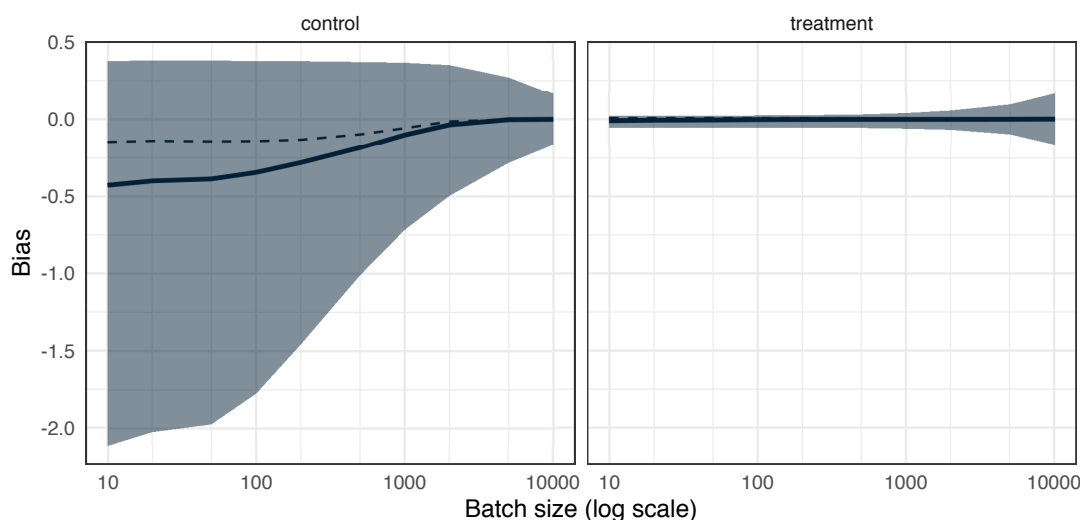
The standard method to estimate the treatment effect is to compare the observed averages of the individuals in both groups:

$$\hat{\tau}_0 = \frac{\sum_{i=1}^n Y_i W_i}{\sum_{i=1}^n W_i} - \frac{\sum_{i=1}^n Y_i (1 - W_i)}{\sum_{i=1}^n (1 - W_i)} \quad (1.1)$$

According to the theoretical results of Nie et al. (2018) the averages are negatively biased estimator for the true expected values of the outcomes. Figure 1.3 characterizes the bias for different choices of batch size. It confirms the negative bias result and shows two additional interesting result: (1) quicker adaptivity leads to a more volatile estimate with larger bias and (2) the control mean contains a larger (negative) bias that is more volatile than the treatment mean. The latter result follows from the fact that the treatment effect is positive so we end up with much more treatment observations (recall that the expected welfare equals to the number of individuals assigned to the treatment). As a result, the treatment effect estimator suffers from amplification bias but because of partial compensation, the bias is smaller than the bias in the control

mean (Figure A.2 in Appendix shows the distribution of τ_0 for different batch sizes).

Figure 1.3: Bias in group mean estimates by batch size



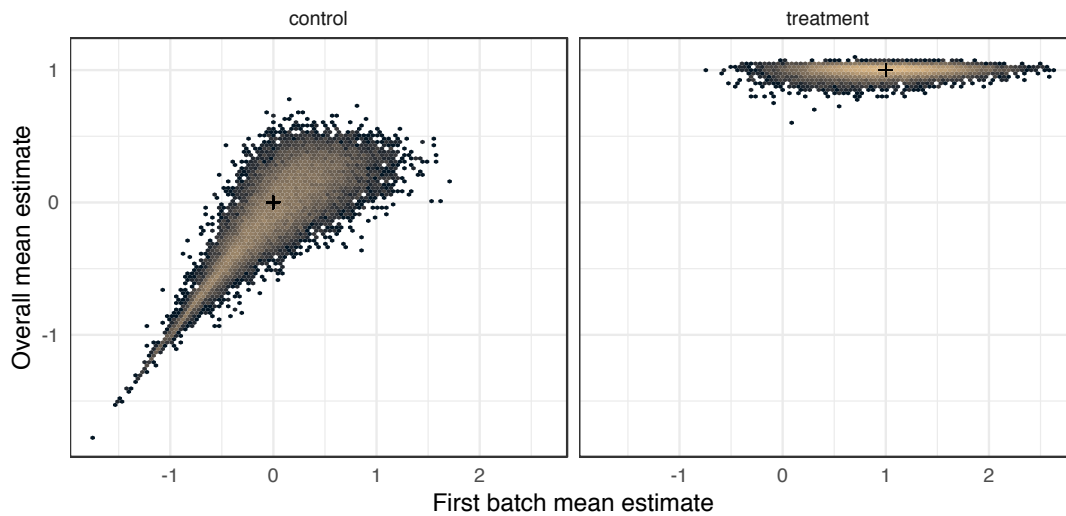
Notes: The figure shows the bias in the group mean estimates by batch size using a logarithmic scale to focus on the interesting region. The shaded area shows the 90% confidence interval, the dashed line depicts the median. Quicker adaptivity results in larger negative bias that is much more expressed for the control group (as we end up with more treatment observations). Number of simulations = 20,000.

The negative bias in group means results from an asymmetry in sampling that is an inherent feature of the adaptive data collection. For the sake of an intuitive understanding of this process, let us focus only on the control estimate where the bias is larger. As the first batch is a simple random split, the first batch average is an unbiased estimate for the control mean: $E[\hat{\mu}_0^{(1)}] = \mu_0$. However, the actual estimate contains some estimation error: $\hat{\mu}_0^{(1)} = \mu_0 + \varepsilon_0^{(1)}$. If this error is negative $-\varepsilon_0^{(1)} < 0$ – there will be a positive error in the treatment effect estimate. As a result, the bandit's belief will be distorted towards the treatment being effective, so more individuals will be assigned to the treatment and only a few to the control. Few new observations in the control group cannot compensate for the original error in the control estimate. However, if the error in the first batch is positive $-\varepsilon_0^{(1)} > 0$ – the belief will be distorted towards the treatment being ineffective, so more individuals will be assigned to control, and these new observations can outweigh the original error in the control estimate.

Figure 1.4 provides a visual illustration for this mechanism. If the first batch results in a negative control estimate, this error is more likely to remain there also in the overall estimate of the experiment, than in the case when the first batch results in a positive control estimate.

Note that this asymmetry by the estimation error is not restricted to the first versus later batches but is present throughout the whole process. It is only most visible after the first batch as the first round of assignment does not depend on previous observations.

The asymmetry can be highlighted using a simple decomposition of $\hat{\tau}_0$: the treatment and control averages can be calculated as weighted averages of the batch group averages where the

Figure 1.4: Density of mean estimates, using the first batch versus the whole sample

Notes: First batch mean estimate is evaluated on individuals arriving in the first batch ($n_B = 1000$). Darker regions mean higher density. The importance of the first batch estimate is clear, especially for the control outcome: an underestimated group mean from the first batch remains uncompensated in the overall estimate. Number of simulations = 20,000.

weights are the shares of the given batch within the total size of the given group (see Equation 1.2). The batch group estimates are unbiased as they arise from simple random splits of batches (only the way how the split is done changes but it does not matter regarding unbiasedness). The bias in the overall averages results only from compositional effect: as a negative error in the estimate of a given batch leads to under-sampling in the following batches, it means lower weights for these batches, thus, a relatively higher weight to the given erroneous batch. In contrast, a positive error leads to over-sampling in the following batches, which gives a relatively lower weight for the erroneous batch. Also, over-sampling in the next batch quickly leads to the correction of the error, thus the over-sampling itself remains only a temporary issue.

$$\hat{\tau}_0 = \sum_{j=1}^m \underbrace{\frac{\sum_{i \in B_j} Y_i W_i}{\sum_{i \in B_j} W_i}}_{\text{batch treated average}} \underbrace{\frac{\sum_{i \in B_j} W_i}{\sum_{i=1}^n W_i}}_{\text{share of batch within all treated}} - \sum_{j=1}^m \underbrace{\frac{\sum_{i \in B_j} Y_i (1 - W_i)}{\sum_{i \in B_j} (1 - W_i)}}_{\text{batch control average}} \underbrace{\frac{\sum_{i \in B_j} (1 - W_i)}{\sum_{i=1}^n (1 - W_i)}}_{\text{share of batch within all control}} \quad (1.2)$$

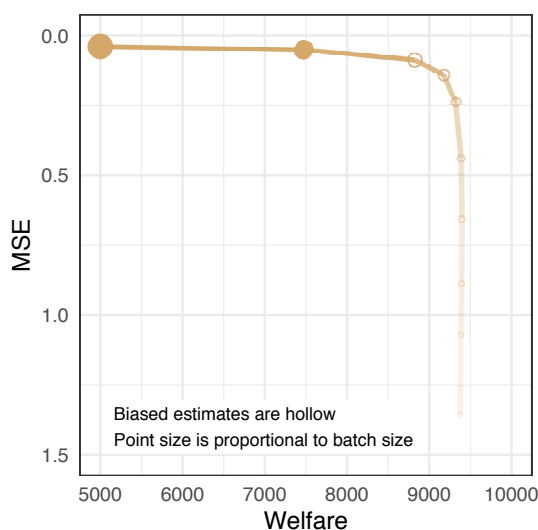
1.3.4 Welfare-Estimation Trade-off

My previous results suggest an interesting trade-off: quicker adaptivity generally results in higher expected outcome (welfare goal) but leaves us with a more biased and more volatile treatment effect estimate (estimation goal). Using the maximum batch size of 10,000 is equivalent to running a randomized controlled trial (RCT) on the whole sample: being the gold standard for

measuring an effect it results in a reasonable estimate, but also a much lower expected welfare.

To compare the performance of different strategies in this space I plot the expected welfare (x axis) against the mean squared error of the estimator (reversed y axis), the two objective functions of the decision-maker (see Figure 1.5). To highlight the decision-maker's constraint of unbiasedness, biased estimates are shown with hollow circles whose transparency is proportional to the size of bias. The best strategy would be a strong point at the top right corner: with a total welfare of 10,000 and an unbiased treatment effect estimate with zero MSE. Obviously, such a strategy does not exist.

Figure 1.5: Performance of the bandit assignment rule in the welfare-estimation space



Notes: Each dot shows the achieved welfare and the mean squared error of the standard treatment effect estimator of the bandit assignment rule with a given batch size. Smaller batch size (quicker adaptivity) leads to higher welfare but also larger bias and larger MSE. Number of simulations = 20,000.

Each strategy on the figure combines the adaptive allocation rule with $\hat{\tau}_0$, the only difference is the choice of n_B . A decision-maker who only cares about the estimation goal would choose the top left point of full RCT. Moving towards more adaptive rules brings significant welfare gains for a slow increase in the variance of the estimator. However, the bias needs to be corrected.

1.4 Bias correction

1.4.1 Inverse Propensity Weighting (IPW)

A standard technique to correct bias in the treatment effect estimator is inverse propensity weighting (also mentioned by Nie et al., 2018; Dimakopoulou et al., 2018). I prove in Equation

1.3 that using IPW with estimated⁷ propensity score (the actual share of a batch assigned to the treatment) is equivalent to using simple average of the batch averages (without weighting as in $\hat{\tau}_0$). Following from the fact that each group average is an unbiased estimate for the corresponding group mean, this method takes the averages of multiple unbiased estimates and thus gets rid of the compositional effect and takes the averages of multiple unbiased estimates. As individuals arrive in batches, individual propensity scores depend only on the individual's batch: $p_i = \mathbb{P}(W_i = 1) = p_j$ for $i \in B_j$.

$$\begin{aligned}
 \hat{\tau}_{IPW} &= \frac{1}{n} \left(\sum_{i=1}^n \frac{Y_i W_i}{p_i} - \sum_{i=1}^n \frac{Y_i (1 - W_i)}{1 - p_i} \right) \\
 &= \frac{1}{n} \sum_{j=1}^m \left(\sum_{i \in B_j} \frac{Y_i W_i}{p_j} - \sum_{i \in B_j} \frac{Y_i (1 - W_i)}{1 - p_j} \right) \\
 &= \frac{1}{n} \sum_{j=1}^m \left(\sum_{i \in B_j} \frac{Y_i W_i n_B}{\sum_{i \in B_j} W_i} - \sum_{i \in B_j} \frac{Y_i (1 - W_i) n_B}{\sum_{i \in B_j} (1 - W_i)} \right) \\
 &= \frac{1}{m} \sum_{j=1}^m \underbrace{\frac{\sum_{i \in B_j} Y_i W_i}{\sum_{i \in B_j} W_i}}_{\text{batch treated average}} - \frac{1}{m} \sum_{j=1}^m \underbrace{\frac{\sum_{i \in B_j} Y_i (1 - W_i)}{\sum_{i \in B_j} (1 - W_i)}}_{\text{batch control average}} \tag{1.3}
 \end{aligned}$$

However, IPW does not seem to be effective: instead of eliminating the bias, it can even exacerbate the problem (Figure A.3 in Appendix shows the distributions of $\hat{\tau}_{IPW}$ for different batch sizes). The volatility of the estimator is also much higher.

The reason for this lies again in the asymmetry of sampling. Taking the average of averages as explained above should work but only if there are averages available to average on. However, in some cases the bandit might assign everyone to the treatment leaving no control assignees to use for calculating the control batch average. These cases are exactly the ones where the treatment effect is estimated with the highest positive error (hence the extreme assignment share of the treated). I illustrate this process for $n_B = 1000$. Table 1.1 summarizes the expected value of the estimator by how many batches contained any control assignee: the more batch is without controls (everyone is assigned to the treatment) the more over-estimated is the effect. As the natural consequence of this selection, runs with controls in every batch (the majority) result in an under-estimated treatment effect.

Figure 1.6 provides a visual illustration for this phenomenon on the control group. The left panel shows that each batch average in itself is an unbiased estimate for the corresponding

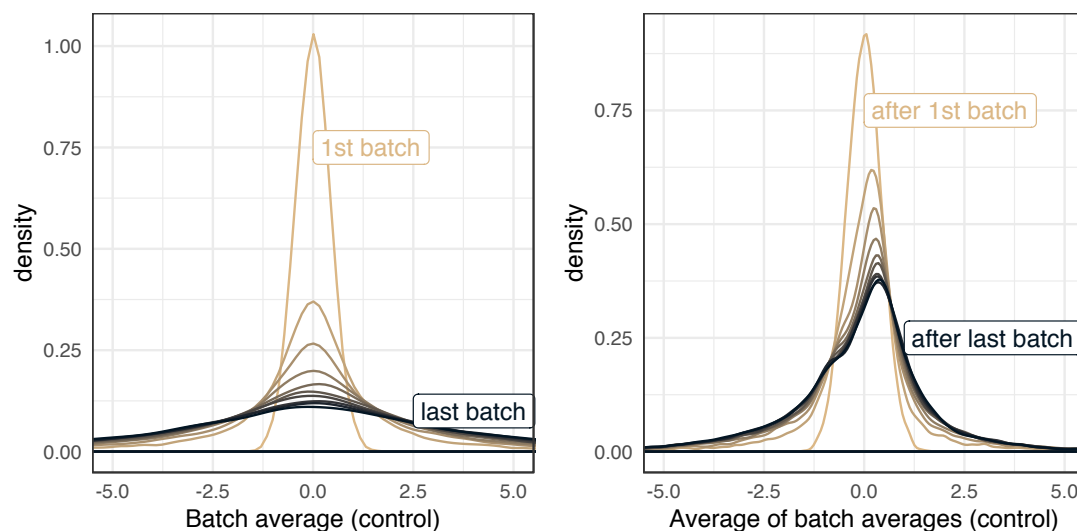
⁷Other works, such as Hadad et al. (2019), use true propensity scores instead. This requires that one stores the allocation probabilities as well. For me, $\{Y_i, W_i\}$ suffice.

Table 1.1: Comparison of $\hat{\tau}_{IPW}$ by number of batches with control assignment

# of batches with controls	1	2	3	4	5	6	7	8	9	10
$E[\hat{\tau}_{IPW}]$	1.99	1.85	1.76	1.79	1.71	1.46	1.64	1.32	1.22	0.80
Probability	2.0%	3.2%	3.5%	3.5%	3.8%	4.4%	5.2%	6.8%	11.4%	56.3%

Notes: Selection bias: Runs with controls in every batch ($n_B = 1000$) underestimate the treatment effect while runs with batches without controls overestimate the treatment effect, using the average of averages ($\hat{\tau}_{IPW}$) for estimator. Number of simulations = 20,000.

control mean. As we tend to sample less and less control in later batches, the estimate is more and more volatile. The right panel shows how the average of averages evolve through batches. If the average of averages after a given batch is small, we tend to sample either less control in the following batch so we update the average with a more volatile average, or no control at all so we do not update the average. This process results in the negatively biased, negatively skewed distribution plotted with the darkest color in the chart.

Figure 1.6: Batch average for the control mean across batches ($n_B = 1000$)

Notes: Each batch in itself is unbiased. Average of batch averages is getting biased due to selection. Number of simulations = 20,000.

1.4.2 Using the first batch only

One can overcome the problem with inverse propensity weighting by using only the data collected in the first batch. I call this as First Batch Estimator ($\hat{\tau}_{FB}$):

$$\hat{\tau}_{FB} = \frac{\sum_{i \in B_1} Y_i W_i}{\sum_{i \in B_1} W_i} - \frac{\sum_{i \in B_1} Y_i (1 - W_i)}{\sum_{i \in B_1} (1 - W_i)} \quad (1.4)$$

This estimator is unbiased, so the strategy of Thompson sampling assignment rule combined with the first batch estimation method (TS-FB) works. However, it loses on efficiency as it drops a large fraction of observations, especially for small batch sizes (Figure A.4 in Appendix shows the distributions of $\hat{\tau}_{FB}$ for different batch sizes).

To better understand the efficiency cost relative to the welfare gain of this strategy, I visualize its performance on the welfare-estimation plot (Figure 1.7). As a benchmark, I add the traditional strategy in economics where the assignment rule is not adaptive: first, concentrate on the estimation goal and run an RCT on an experimental sample, and then, focus on the outcome and form a deterministic rule based on the result that can be applied from then on (subject of the classic treatment choice literature). This process can be translated to my case as the rule of Explore-then-commit (ETC):

Explore-then-commit (ETC)

1. Split the first batch equally between treatment and control^a.
2. Estimate the average treatment effect by comparing the treatment and control averages calculated on the collected data^b:

$$\hat{\tau}^{(1)} = \hat{\mu}_1^{(1)} - \hat{\mu}_0^{(1)} = \frac{\sum_{i \in B_1} Y_i W_i}{\sum_{i \in B_1} W_i} - \frac{\sum_{i \in B_1} Y_i (1 - W_i)}{\sum_{i \in B_1} (1 - W_i)}$$

3. Apply the assignment with the higher mean to everyone onwards:

$$p_k = \arg \max_w \left\{ \hat{\mu}_w^{(1)} \right\} \text{ for } k \geq 2$$

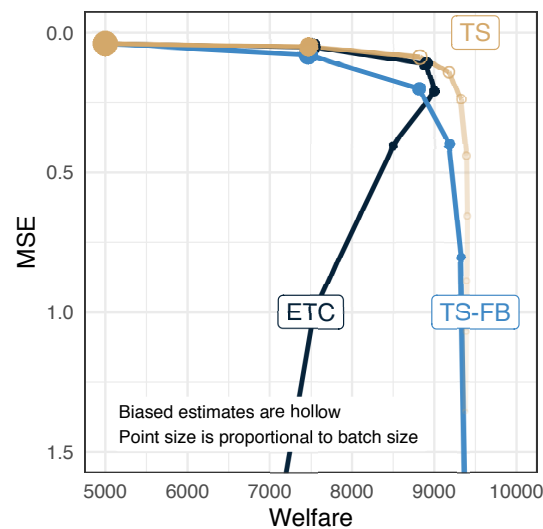
^aTypically, the size of the batch is calculated by assuming a minimum size for the treatment effect and deriving a required sample size that yields enough power given a predetermined false positive rate (or significance level).

^bComparing the averages corresponds to the Conditional Empirical Success Rule of Manski (2004).

Adaptive data collection using $\hat{\tau}_{FB}$ clearly dominates the Explore-then-Commit (ETC) strategy (using $\hat{\tau}_0$) for decision-makers valuing welfare more, but it loses when MSE is more important. The closest choices to the optimal top right point are $n_B \in \{1000, 2000\}$ for both strategies.

1.4.3 Limiting the propensity scores

With a slight modification of the assignment rule the efficiency problem of the TS-IPW strategy can be improved (while preserving the bias-corrected estimate). As I showed in section 1.4.1, the reason why τ_{IPW} is biased after adaptive data collection is that the algorithm does not assign to

Figure 1.7: Performance of different strategies in the welfare-estimation space

Notes: Each dot shows the achieved welfare and the mean squared error of the corresponding treatment effect estimator for a given strategy with a given batch size. Generally, quicker adaptivity leads to higher welfare but also larger MSE. ETC with moderate batch size works well, but smaller batch size harms not only MSE but also welfare. TS-FB approximates the standard TS strategy with higher MSE but ensuring an unbiased estimate. Number of simulations = 20,000.

both groups in each batch, and this unanimous assignment asymmetrically depends on previous observations. A simple solution for this issue is to ensure that people are assigned to both groups in each batch, that is to limit the (realized) propensity score away from the extremes of zero and one. Although this method needs the modification of the data collection process, in the digital world this is typically not very costly. Also, this solution is easy-to-implement.

Limited Thompson Sampling (LTS)

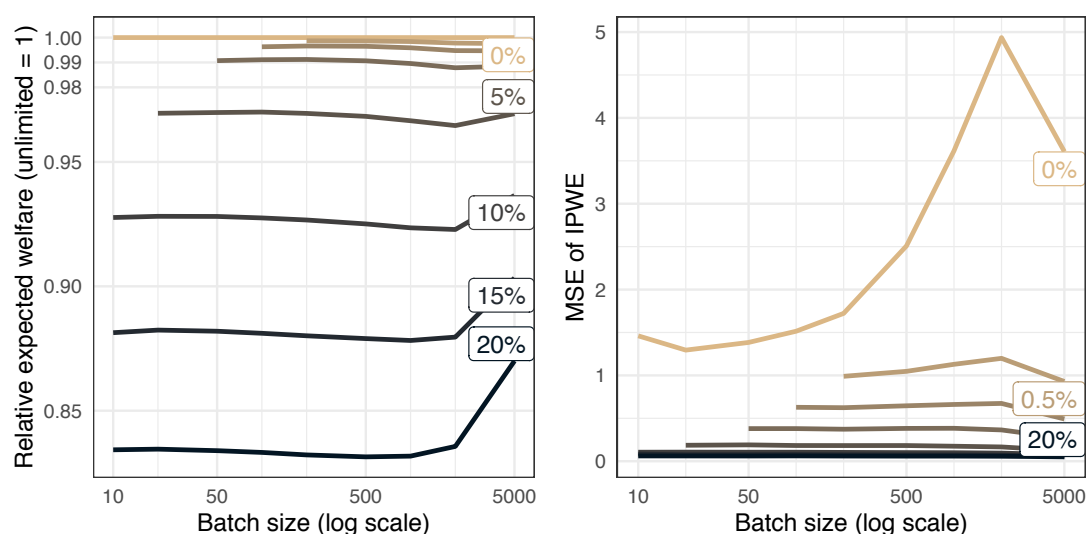
The difference to the native Thompson Sampling is highlighted in bold.

1. Split the first batch equally between treatment and control.
2. Form beliefs about the treatment and control means by deriving posterior distributions using normal density with calculated averages (assuming that standard deviation is known).
3. Assign individuals to the treatment in the next batch by the probability that the treatment mean is higher than the control mean. **If this probability is too extreme, use a limited probability instead. Denoting the amount of limitation by L , and the probability after the k th batch by $p^{(k)}$, the assigning probability is $\tilde{p}^{(k)} = \max\left(\min\left(p^{(k)}, 1 - L\right), L\right)$.**
4. Repeat from step (2) until assigning the last batch.

The smallest possible limitation (e.g. 1% for the batch size of 100) would yield an unbiased $\hat{\tau}_{IPW}$ estimate. The amount of limitation incorporates the welfare-estimation trade-off. Limiting to higher extent requires higher opportunity cost, but also allows for more robust estimates. It forms a smooth transition between two endpoints: the unlimited bandit (0% limit, previously used in TS and TS-FB strategies) and a random split of the full sample (50% limit, ETC with $n_B = 10000$, full RCT).

Figure 1.8 shows the effect of limitation on welfare and estimation goals simulating 8 different limit levels⁸. As expected, higher limit means lower welfare and more precise $\hat{\tau}_{IPW}$ estimate⁹.

Figure 1.8: Welfare and estimation performance of the LTS-IPW strategy



Notes: The left panel shows the relative welfare achieved by the limited bandit rule compared to the unlimited one for various limit choices, by batch size. The right panel compares the MSE of the inverse-propensity-weighted estimators on the resulting data. Higher limits incur higher welfare cost but bring more precision. The loss and gain by the amount of limit are disproportionate. Number of simulations = 20,000.

The loss in welfare and the gain in precision is disproportionate: while the loss is linear in the amount of limitation, the gain is not: using a 1% limit, MSE drops dramatically for each batch size (by as much as 80% for $n_B = 2000$ - see right panel) while it costs no more than 1% of welfare (left panel).

It is interesting to note that limitation affects differently the different batch sizes. Small and large batch sizes induce lower cost than the middle range for a given limit. This is the result of two factors: First, limitation acts as a regularization tool, similarly to what we have seen with larger batch sizes. Limitation decreases the probability of over-fitting, and can thus improve welfare for some runs. Second, limitation obviously does not affect the simple random split of

⁸0%, 0.5%, 1%, 2%, 5%, 10%, 15% and 20%.

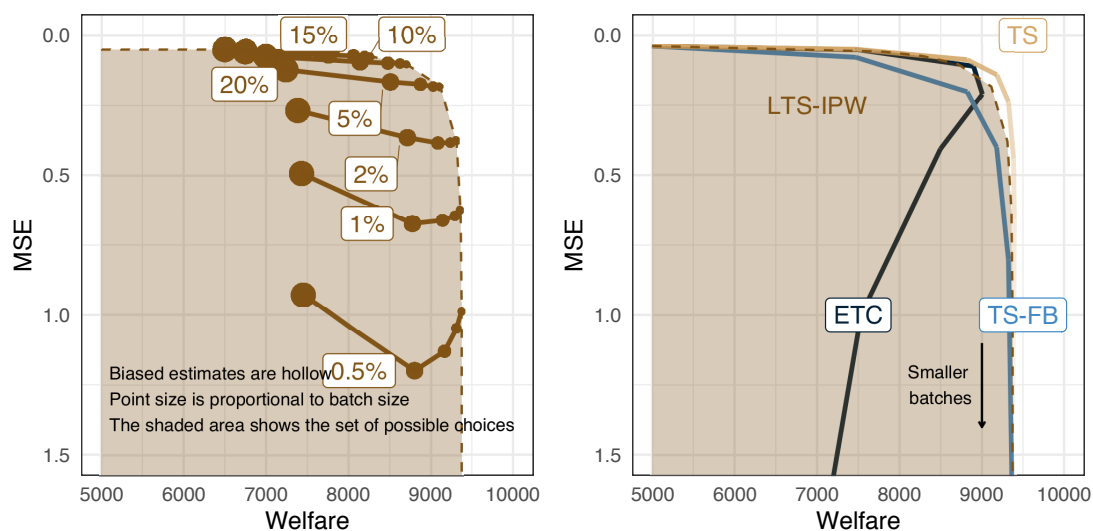
⁹Limitation also decreases the bias of the $\hat{\tau}_0$, but due to the inherent weighting in Equation 1.2, some bias remains until the limit reaches the level of the simple random split.

the first batch. For larger batch sizes, the share of the first batch is higher, thus, the limitation cost is relatively lower.

On the other hand, the improvement on the estimation precision is about stable by batch size. This result follows from the fact that limitation is defined as share of the batch, so it means closely the same for each batch size. Higher limitation - in line with approaching the simple random split strategy - also improves the skewness of the estimator and the variance of the reached welfare.

As the estimation improvement does not depend on the batch size, strategies with quicker adaptivity should fare better in the welfare-estimation space. The left panel of Figure 1.9 shows the performance of LTS-IPW with different limits. Lower limitation can achieve higher welfare with an appropriate batch size, but only for a growing cost on MSE. The lines are close to horizontal, showing that smaller batch sizes can achieve higher expected welfare for practically no estimation cost. Different points of this chart depict different parametrizations (n_B, L) of LTS-IPW strategy; some of them dominate each other (e.g. large batch sizes with low limitation are clearly worse than smaller batch sizes with higher limitation). Connecting the best parametrizations give us the Performance Frontier of this strategy in the welfare-estimation space. Any of these point could be achieved by choosing an appropriate batch size (n_B) and amount of limitation (L) - not necessarily simulated in this exercise.

Figure 1.9: Performance of different strategies in the welfare-estimation space



Notes: The right panel shows the achieved welfare and the MSE of the inverse-propensity-weighted estimator of the limited bandit rule, by various limits and batch sizes. The dashed line connects the best available choices (Performance Frontier). The left panel shows only this frontier compared to the previous strategies: LTS-IPW extends the possibilities by approximating the TS strategy while also ensuring an unbiased estimate. Number of simulations = 20,000.

The right panel of figure shows only the frontier for the LTS-IPW strategy, along with our previous strategies. Limitation with inverse propensity weighting clearly extends the possibilities

of the decision-maker: It gets the closest to the TS strategy but also allows for an unbiased estimate, and dominates TS-FB and also ETC for $n_B < 2000$. If the decision-maker cares about welfare as well, collecting data adaptively with some limitation and estimating the treatment effect with inverse propensity weighting is the best strategy.

1.5 Monte Carlo Simulation

1.5.1 Uncertainty

Parametrization I investigate the behavior and performance of different strategies with different levels of uncertainty (σ) holding the treatment effect constant at unit value, so σ expresses the noise-to-signal ratio. As the important measure in this problem is the relative effect size τ/σ , it does not matter which one is fixed. Fixing τ allows me to directly compare the welfare and estimation performance of the strategies. I investigate 8 different values for σ with $n = 10,000$ ¹⁰. Each setup is simulated with 10 values of batch size and 8 values of limit¹¹, 10 – 50 thousand runs for each¹².

Welfare Figure 1.10 summarizes the results of the expected total welfare and the bias in $\hat{\tau}_0$ by batch size for each σ . Less uncertainty (smaller variation in the potential outcomes) increases the expected gain and decreases the bias. Both of these results are intuitive.

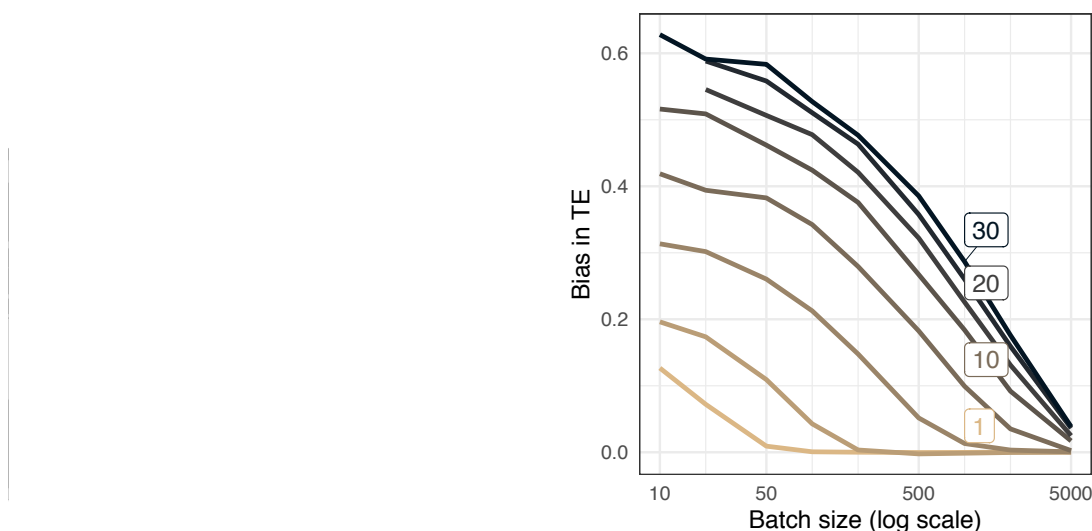
Unlike in the setup of the previous section ($\sigma = 10$), the quickest adaptivity results in the highest expected welfare for low levels of noise ($\sigma < 5$). For these setups, the danger of overfitting is low, so regularizing by increasing the batch size does not help, only incurs a higher opportunity cost.

There is another interesting pattern to note: For welfare, each line approaches the one with the smallest σ as batch size increases, some also reach it. This means that less uncertainty does not lead to higher outcome under a certain value of σ if batches are large enough. The reason for this is that for each batch size there is a maximum of outcome that cannot be exceeded: when the positive treatment effect is learnt immediately in the first batch and all subsequent batches are assigned to the treatment. It is possible if the noise in the outcomes are small relative to the batch size. This maximum possible welfare is depicted by the dashed line on the chart - if the standard deviation in potential outcomes is not larger than the treatment effect, practically

¹⁰ $\sigma \in \{1, 2, 5, 10, 15, 20, 25, 30\}$

¹¹As small batch sizes do not work with low limits, it means 63 parametrizations for each setup.

¹²The number of runs depends on the level of noise: for setups with larger noise I run more simulations to get robust results: 10,000 for σ below 10, 20,000 for σ at least 10 but below 20 and 50,000 for larger values of σ .

Figure 1.10: Expected total welfare and bias

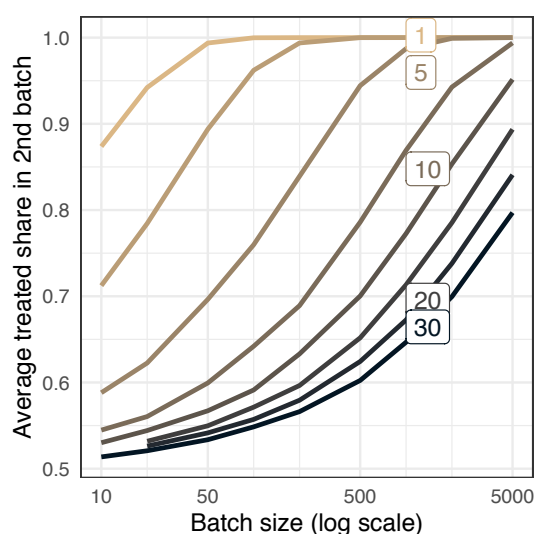
Notes: The left panel shows the expected welfare achieved by the (unlimited) bandit rule with different batch sizes (along the x axis) by different levels of noise (labelled). The dashed line highlights the maximum welfare that each strategy could achieve, and the points depict the batch sizes with the maximum welfare for a given σ . The right panel compares the bias in the standard treatment effect estimators. Larger noise results in lower welfare and larger bias. Number of simulations = 10-50,000.

each batch size achieves this maximum. Table A.1 in the Appendix contains the results for each scenario.

Estimation A similar pattern is visible in the bias (right panel) as well: if the noise is sufficiently low and the batch size is large enough, there is no bias. Obviously, if the treatment effect is perfectly learnt in the first batch, the asymmetric sampling that causes the bias does not kick in. Figure 1.11 shows the average share of treated in the second batch across batch sizes for each setup. It confirms that full learning in first batch can explain the observed patterns in welfare and bias. Table A.2 and A.3 in the Appendix contain the expected bias and MSE values for each scenario.

Welfare-Estimation Trade-off The previous results are in line with the main message of this paper: welfare and estimation goals are working against each other. Mainly, quicker adaptivity leads to higher outcome but also higher bias, for each level of σ . This observation works differently only for two special regions: (1) for high levels of noise, extreme adaptivity hurts both goals, whereas (2) for low levels of noise, adaptivity can be increased until a certain point gathering the welfare gain but without introducing any bias.

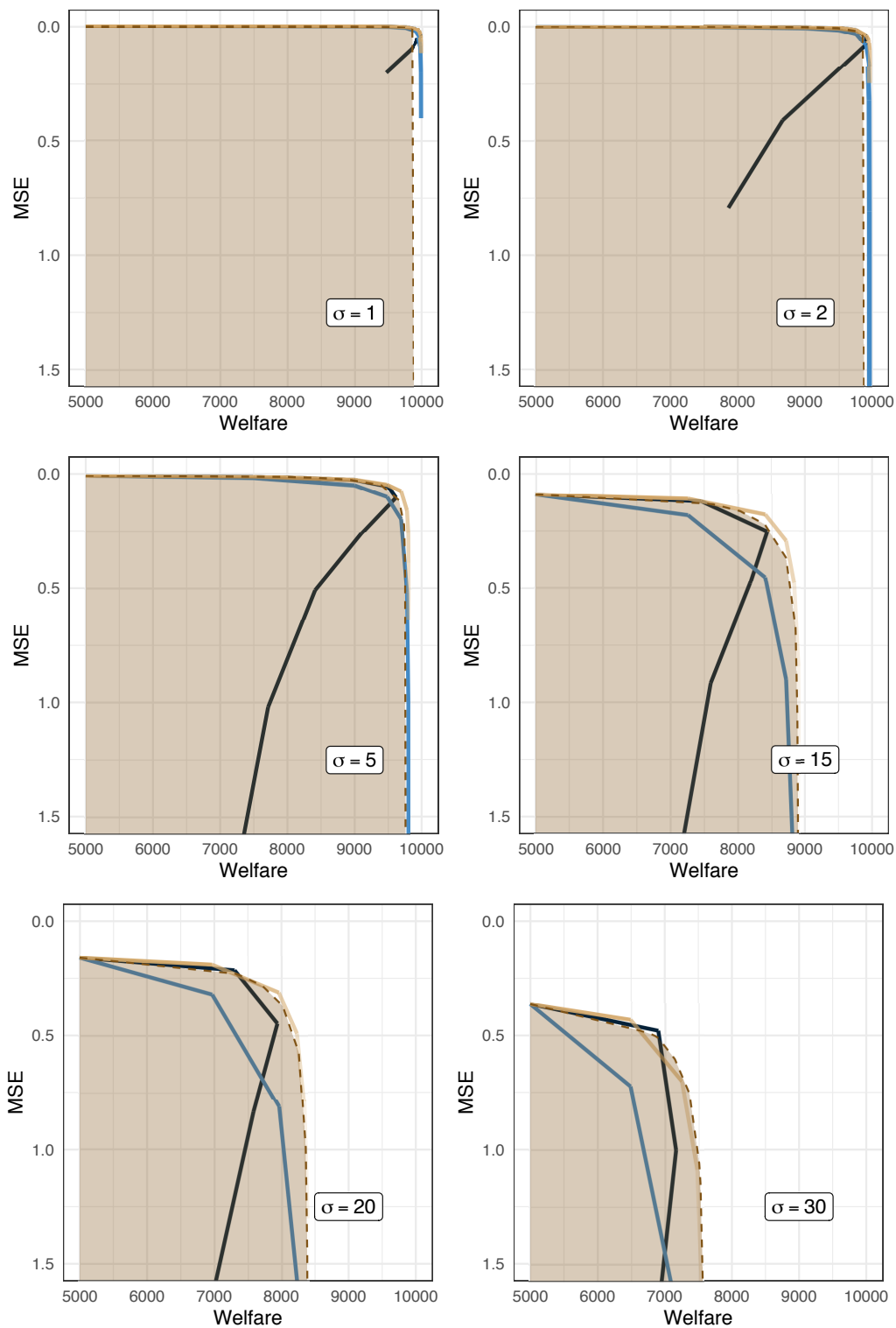
I suggested limiting as a working method for bias correction in section 1.4.3. I showed that small amounts of limitation result in unbiased treatment effect estimates with highly improved MSE

Figure 1.11: Average treated share in the second batch

Notes: The figure shows the expected share of individuals assigned to the treatment in the second batch for various batch sizes, under different noise levels. If the noise is small and the adaptivity is slow enough, full learning occurs. These situations do not cause any bias, and they end up with the highest possible welfare (see the left panel of Figure 1.10). Number of simulations = 10-50,000.

for only a low price in achieved welfare, and this disproportionality allows for the extension of the set of available choices for the decision-maker in the welfare-estimation space.

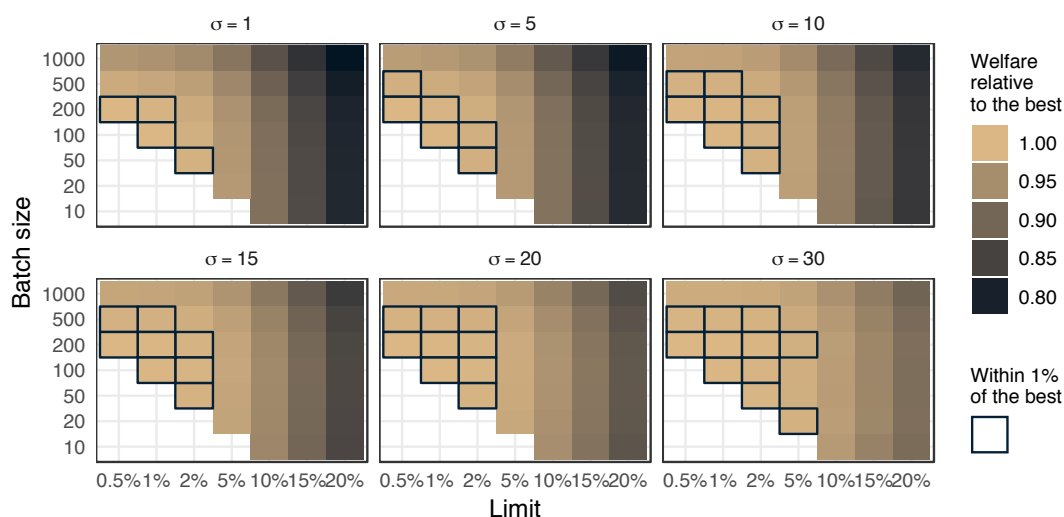
Figure 1.12 shows the performance of the different strategies in the welfare-estimation space for each setup. Similarly to Figure 1.7, it only shows the frontier for the TS-IPW strategy that is formed by the best combinations of batch size and limit. Obviously, as the problem gets harder (as the uncertainty grows), each strategy performs worse (are farther away from the top right corner). My previous result is strengthened: adaptivity with limitation almost always extends the feasible set of welfare-MSE pairs. For high noise, my suggested strategy even extends upon the unlimited TS that were excluded because the estimate is biased. Only in low-noise setups is this extension ambiguous. However, in these setups the problem to solve is easy, and the whole question is of less importance. The treatment effect can be learnt perfectly right in the first batch, so an unlimited bandit could deliver an unbiased estimate next to near-optimal welfare (see Figure 1.10).

Figure 1.12: Performance of different strategies in the welfare-estimation space

Notes: Each panel is a replication of the left panel of Figure 1.9 for different levels of noise. The TS-IPW strategy always extends the set of choices, especially if the problem is hard (the noise is large). Number of simulations = 10-50,000.

In practice it is important to know which combinations form the frontier that extends the possibilities. For welfare, it is obvious, that smaller limits are expected to fare better. However, a small limit excludes small batch sizes as we need control assignees in every batch to ensure unbiasedness. So, it is not straightforward how to choose the best strategy. Figure 1.13 shows the expected welfare for all batch size - limit combinations, for different levels of uncertainty. There are three interesting results to note:

Figure 1.13: Expected welfare of different combinations of n_B and L



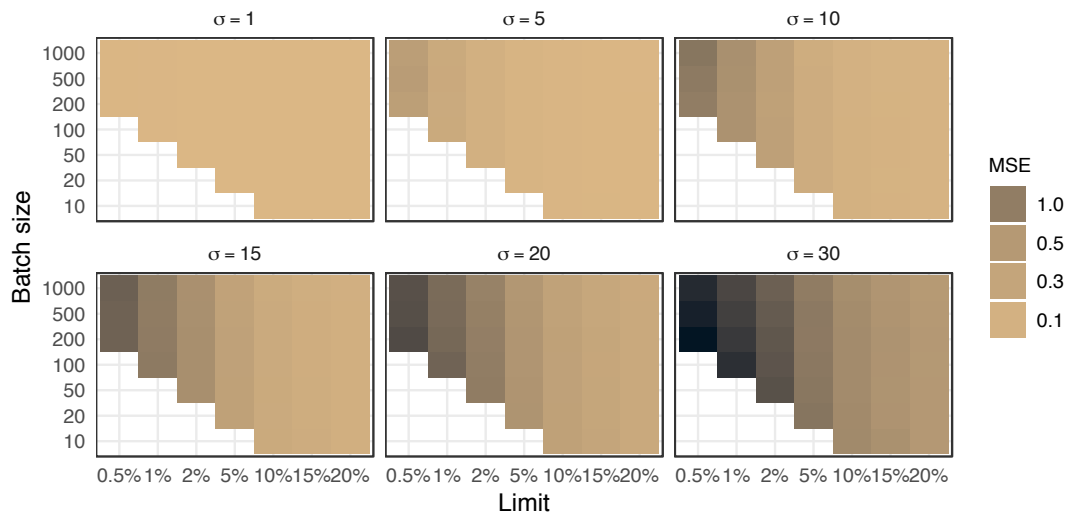
Notes: Each panel shows the expected welfare relative to the best strategy for each batch size and limit combinations, for different levels of noise. The best strategies are highlighted within each scenario. Number of simulations = 10-50,000.

1. Quicker adaptivity is generally better, but not beyond $n_B = 50$. Too small batch size requires too large limit to preserve unbiasedness that adversely affects welfare. Also, the opportunity cost they could possibly win is no more than the size of the batch which is obviously small for small batches.
2. One can increase limit and decrease batch size to achieve about the same welfare. For large noise cases, many combinations result in the same level of welfare. Note, however, that this level is smaller in absolute value than in low-noise scenarios (recall Figure 1.1).
3. Limiting does not eliminate the problem of over-fitting: too quick adaptivity has a detrimental effect on expected welfare if the noise is high (e.g. for $\sigma = 20$ the achieved welfare is smaller with $n_B = 10$ than with $n_B = 50$ even with larger limits).

Figure 1.14 shows the same chart for the estimation goal, plotting the MSE of different combinations. As in this case, the important comparison is the estimated treatment effect itself, I use the levels of MSE: a value above 1 means an error that is larger than what is measured.

1. Intuitively, larger noise means larger MSE, across each combinations.
2. Smaller adaptivity and larger limits improve MSE. More interestingly, limiting matters more than batch size: in terms of estimation precision, increasing the limit is more effective than increasing the batch size.
3. The combination that results in the smallest MSE while still achieving the maximal welfare is: $\{n_B = 50, L = 2\%\}$ for $\sigma = 1$ while $\{n_B = 200, L = 5\%\}$ for $\sigma \geq 5$.

Figure 1.14: MSE of different combinations of n_B and L



Notes: Each panel shows the MSE of the inverse-propensity-weighted treatment effect estimator of the limited bandit for each batch size and limit combinations, for different levels of noise. Recall, MSE above the unit level means an error that is larger than what is measured. Number of simulations = 10-50,000.

To better understand the behavior of different strategies, it is worth considering the limiting cases of uncertainty:

1. **no-noise scenario** $\sigma \rightarrow 0$ For low-noise cases, smaller limits reach higher welfare while the MSE remains stable, so as $\sigma \rightarrow 0$ it is reasonable to $L \rightarrow 0$. Also recall that the standard treatment effect estimator on unlimited bandit data is unbiased for sufficiently large batches (see Figure 1.10), where the sufficiently large batch size decreases in noise. Last, smaller batch size reaches higher welfare, and for low noise levels we should not worry about over-fitting either. All of these suggest that we should run an unlimited bandit with the smallest batch size ($n_B = 2$) for the no-noise scenario (the estimation method does not matter as $\sigma = 0 \Rightarrow \hat{\tau}_0 = \hat{\tau}_{IPW}$). This strategy is equivalent to the intuitive solution of this problem: assign one observation to both groups and then assign everyone based on the comparison of these outcomes.
2. **no-treatment-effect scenario** $\sigma \rightarrow \infty$ For high-noise cases, high limits are needed to keep MSE at moderate level. As noise increases, so decreases the achievable welfare (see

Figure 1.10) and expands the set of batch size and limit choices that result in about the same welfare as the best combination. These suggest to use the maximum limit of 0.5 for the limiting no-treatment-effect scenario which strategy is equivalent to the simple random split. Again, this is an intuitive solution as zero treatment effect means there is no potential welfare to gain from being adaptive so it would only incur losses on the estimation goal.

Generally, we can conclude to following practical recommendations: choose the limit based on the welfare-MSE trade-off and then use the smallest possible batch size. This choice of the batch size gets less relevant as the noise increases.

1.5.2 Horizon

I also consider different lengths for the horizon¹³. Note that this is similar to changing the noise and batch size appropriately: e.g. a 4 times larger sample size is equivalent to a setup with 2 times larger σ with 4 times larger batches (i.e. holding the number of batches fixed). Simulating the illustrative case ($\sigma = 10$) for different lengths makes the comparison easier.

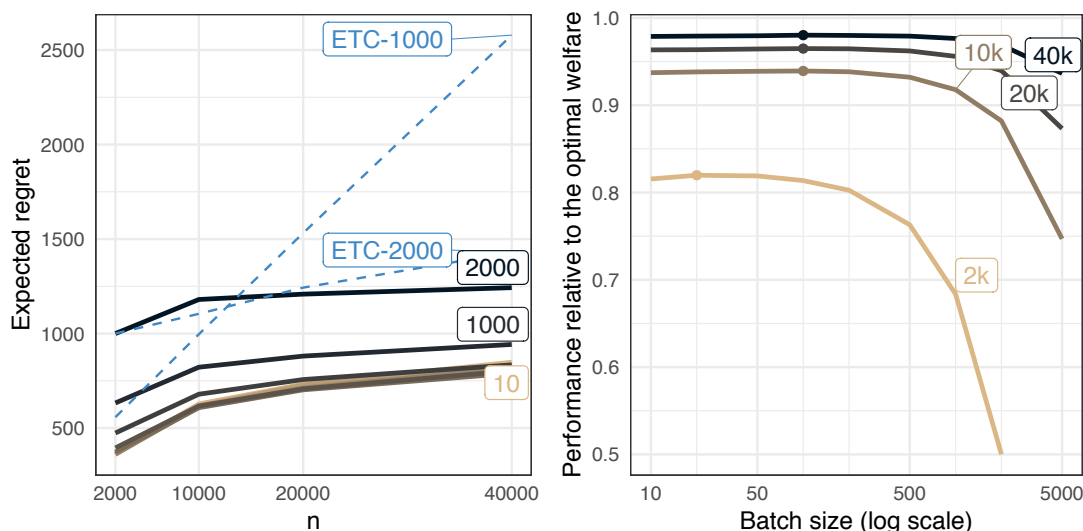
The right panel of Figure 1.15 validates the theoretical result, that the regret of Thompson sampling with any batch size grows slower than the regret of the exploit-then-commit (ETC) rule typical in the treatment choice literature.

The left panel of the chart focuses on the choice of batch size by different horizons. If the horizon is shorter, smaller batch sizes are better: quicker adaptivity means less opportunity cost at the beginning. Extreme adaptivity can still lead to over-fitting and thus, lower welfare. As the horizon gets longer, larger batch sizes fare better. This result might be explained by the fact that in the longer run, one has more time to invest in learning as there will be more time to gather the interests. Note also, that for shorter horizon, smaller batch size means the same number of batches. E.g. for $n = 2000$, the best batch size of 20 means 100 batches, the same, as the optimal batch size of 100 for the $n = 10,000$ case. The most allocation decisions should be made in the longest horizon setup (400 batches deliver the best result for $n = 40,000$). It is also worth noting, that the importance of the batch size gets less important as the horizon grow: smaller batch sizes reach about the same level of expected welfare.

Figure 1.16 depicts the performance of different strategies in the welfare-estimation space. The limited IPWE strategy extends the available set of choices, especially if the horizon is shorter. Note that decreasing the horizon is making the learning problem harder, similarly to increasing the noise. Therefore, it is not surprising that the chart for the longest horizon resemble more

¹³The simulated values are the followings: 2000, 10,000, 20,000, and 40,000.

Figure 1.15: Welfare performance of bandit algorithm with various levels of adaptivity across different horizons



Notes: The left panel shows the expected regret of different strategies by the horizon: the regret of Thompson sampling grows slower with n than for the explore-then-commit rule common in the econometric practice. The right panel shows the expected welfare achieved by different strategies relative to the (infeasible) optimal welfare (treatment-only scenario). Longer horizons lessen the importance of the choice of batch size. Number of simulations = 10,000.

for the small noise setups of Figure 1.12. Table A.4, A.5 and A.6 in the Appendix contain the expected welfare, bias and MSE values for each scenario.

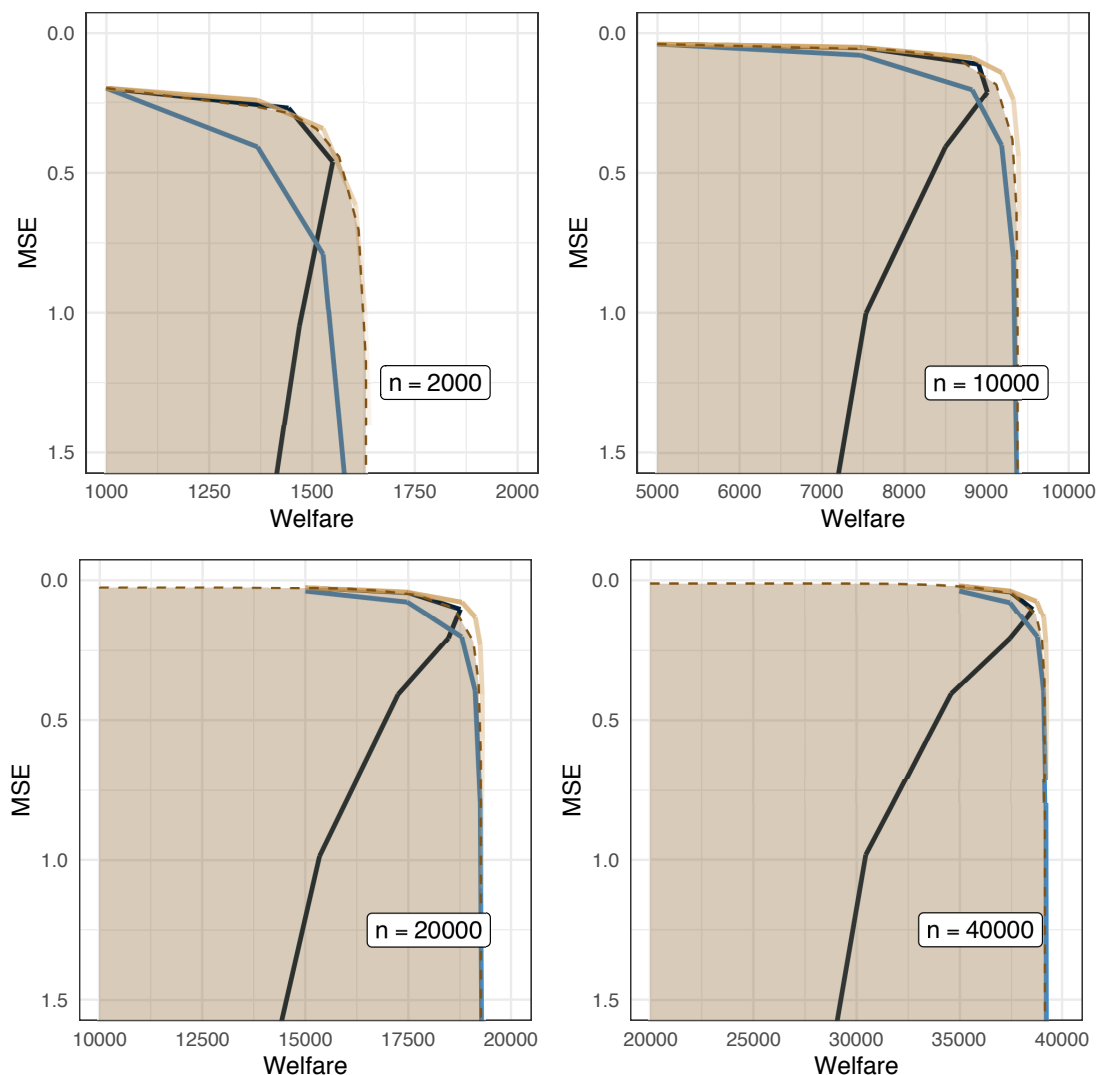
1.5.3 Non-Gaussian Potential Outcomes

All the previous results were built on the Gaussian assumption for the potential outcomes. In this subsection I show how relevant this assumption is by considering less well-behaved distributions as well. I focus on two common behavior: fat tails and skewness. I compare the behavior of the strategies by simulating untreated potential outcomes by four distributions:

1. Normal distribution
2. Student's t -distribution with 4 degrees of freedom (fat tails)
3. χ^2 distribution with 5 degrees of freedom (positive skewness)
4. negative χ^2 distribution with 5 degrees of freedom (negative skewness)

All of the simulated outcomes are standardized to have $\mu_0 = 0$ and $\sigma = 10$ in the population (as in the original setup, see Section 1.3.1) to allow for a strict comparison by the shape of the distribution.

Figure 1.16: Performance of different strategies in the welfare-estimation space, for different horizons



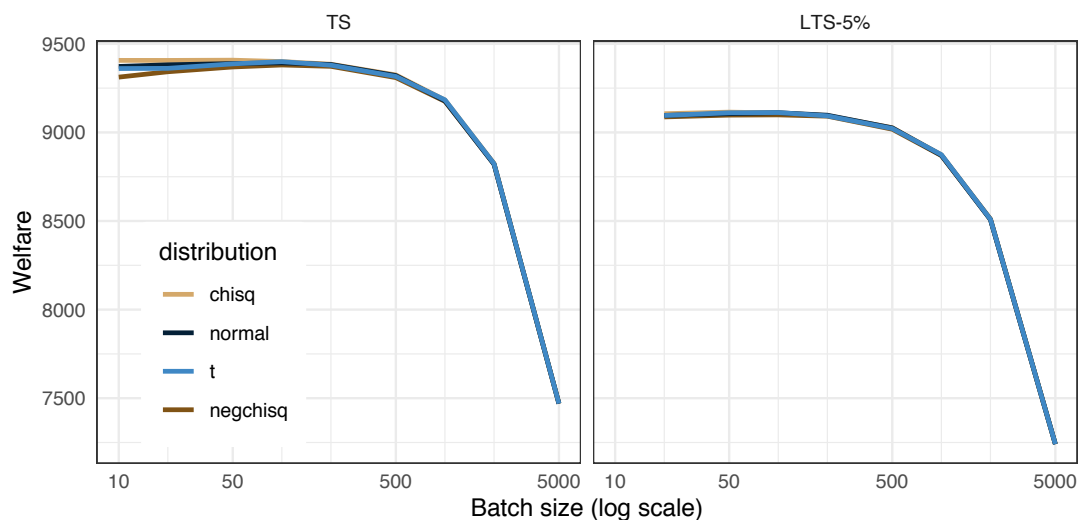
Notes: Each panel is a replication of the left panel of Figure 1.9 for different horizons. LTS-IPW always extends the set of possible choices, especially if the problem is hard (n is small). Number of simulations = 10,000.

For the comparison I choose two strategies: Thompson sampling with the standard treatment effect estimator ($\hat{\tau}_0$) and the limited Thompson sampling with 5% limit using the inverse-propensity-weighted estimator ($\hat{\tau}_{IPW}$)¹⁴.

Figure 1.17 compares welfare performance of the strategies by distribution. The only difference can be detected in the TS strategy with quick adaptivity: the fat-tailed and the negatively-skewed distribution fare worse (but this difference is relatively small). The difference disappears with the limited strategy.

¹⁴Note that I do not change the assignment mechanism, so the posterior beliefs about the group means are still formed using normal distributions. This better approximates a real-world situation where the exact distribution of the outcomes are not known.

Figure 1.17: Welfare performance by different strategies compared by the distribution of the potential outcome



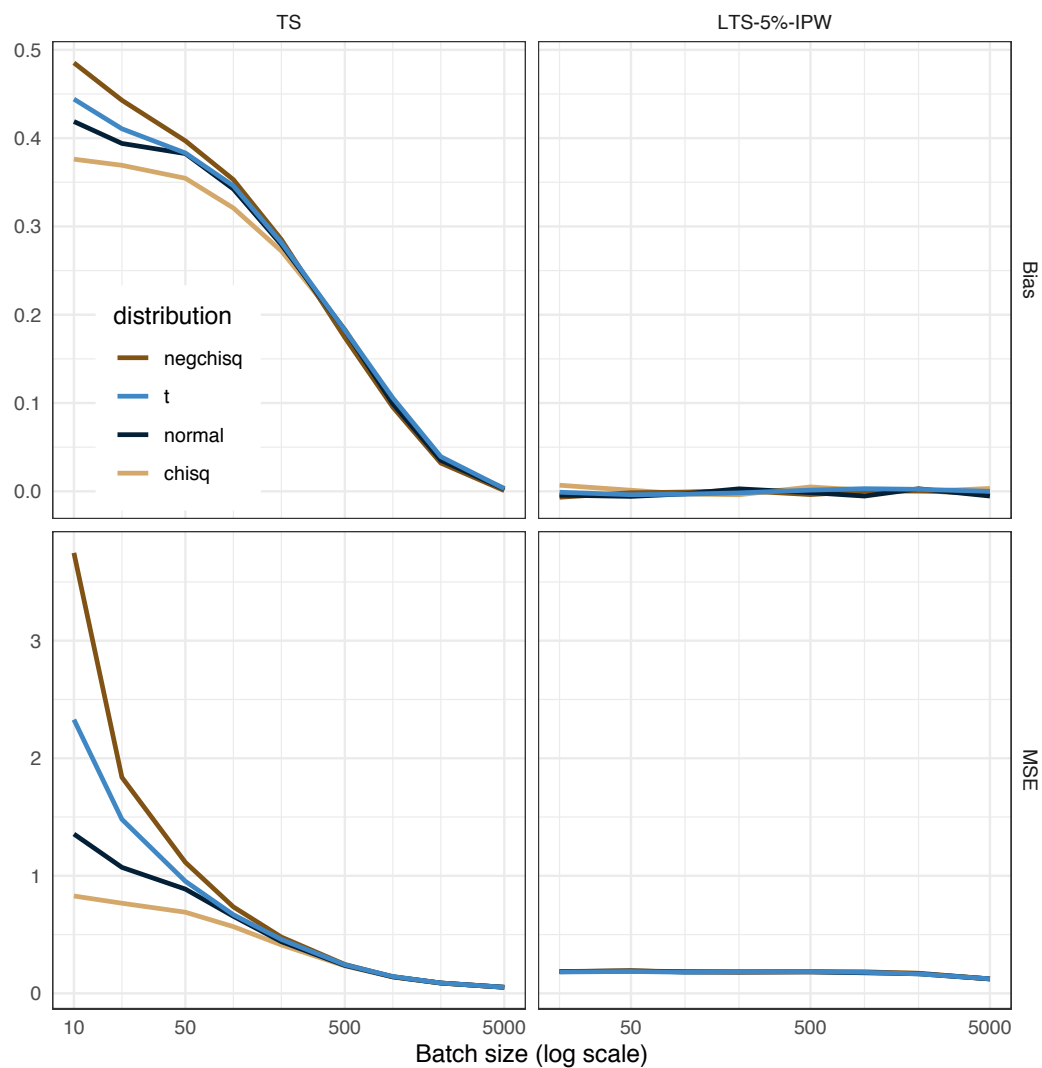
Notes: The figure shows the expected welfare of different strategies by batch size for each outcome distribution. TS: Thompson sampling, LTS-5%: limited Thompson sampling with 5% limit. There is no difference in the expected welfare by the distribution of the potential outcomes for the LTS strategy. Number of simulations = 10,000.

The expected welfare solely depends on each strategy's ability to assign as much individual to the best group (here: to the treatment) as possible. In adaptive allocation rules this ability is determined by how the estimated means compare to each other. As the assignment rule compares averages of the observed outcomes, for large enough sample size the central limit theorem kicks in and this makes the underlying distribution less relevant. In small sample cases, certain shapes of the underlying distribution makes the true means harder to estimate: if it has fat tails or a negative skewness. Interestingly, positive skewness seems to help.

Figure 1.18 shows the estimation performance of the same strategies using the standard treatment effect estimator on the unlimited bandit data and the inverse-propensity-weighted estimator on the limited bandit data. The general patterns are very similar to that of welfare. Practically, one could detect differences only for the TS strategy with small enough batch sizes. Fat-tailed and negatively skewed distributions of potential outcomes are worse, a positively skewed distribution is better than the standard normal one. However, all of these differences disappear when we apply limiting even if we want to preserve adaptivity.

The result that fat tails make our problem harder is intuitive. The differential result in skewness needs some explanation. As I discussed in Section 1.3.3, the bias originates from the belief about the control mean getting stuck a very low region. For this to happen we should draw from the low end of the distribution. If the distribution of the potential outcomes is such that drawing an observation negatively far from the mean has a higher probability, our problem gets harder. Negative skewness means that the mode is below the mean so the probability of drawing negative

Figure 1.18: Estimation performance by different strategies compared by the distribution of the potential outcome



Notes: The figure shows the expected bias and MSE of the estimator of different strategies by batch size for each outcome distribution. TS: Thompson sampling using $\hat{\tau}_0$, LTS-5%-IPW: limited Thompson sampling with 5% limit using $\hat{\tau}_{IPW}$. Number of simulations = 10,000.

outliers is higher. In contrast, positive skewness brings more positive outliers that – due to the asymmetric sampling – only makes our problem easier. Obviously, if the treatment effect is negative, positive skewness is better.

1.6 Data-driven simulations

To assess the behavior of different strategies and the welfare-estimation trade-off in a practical setting, I will run data-driven Monte Carlo simulations using the well known National Job Training Partnership Act (JTPA) study (Bloom et al., 1997). I take the experimental sample

that was used by the influential paper of *Abadie et al. (2002)*. This sample has been used many times for illustrative purposes in the treatment choice literature (see among others *Kitagawa and Tetenov, 2017*). Participants of the JTPA study assigned to the treatment group were offered job training. The outcome of interest is the earnings of the participants in the next 30-months period.

Table 1.2 shows the main numbers of the experiment. The program seems to be effective. The average earnings of the treatment group is \$1,159 higher, even though only 64% of them actually got the training. This shows a positive intention-to-treat effect (ITT), that is my main interest here focusing on treatment assignment rules. The positive ITT more than compensates for the actual cost of the treatment, resulting in a net intention-to-treat effect of \$674.

Table 1.2: Descriptive statistics of JTPA experiment

	Assignment		All
	Treatment	Control	
Number of participants	7,487	3,717	11,204
Share of trainees	64.2%	1.5%	
Mean outcome	\$16,200	\$15,041	\$15,815
ITT			\$1,159
Mean net outcome	\$15,703	\$15,029	\$15,480
net ITT			\$674

Notes: Mean outcome is calculated as the 30-month earnings of the participants. Mean net outcome accounts for the occasional cost of training (\$774, borrowed from *Bloom et al., 1997*).

JTPA was a one-off experiment lasting for more than a year. For the sake of illustration, I will assume participants could have arrived in batches to simulate how adaptive assignment rules would have behaved with the JTPA-participants. Considering that such programs last over years this might be a relevant thought experiment: one can regard batches as yearly participants of such programs where each year's policy depends on available observations until that point¹⁵.

For data-driven simulation, I relax the distributional assumption and the homogeneous treatment effect assumption. Instead, I simulate potential outcomes by bootstrapping from the available data. This way, I can simulate arbitrary assignments using the original data - as a result, it will not be true any more that only the arrival is random: each simulation run will consist of a different population (bootstrapped from the same original population). Besides this difference, the JTPA data could be translated to my setup by scaling the number of individuals and the

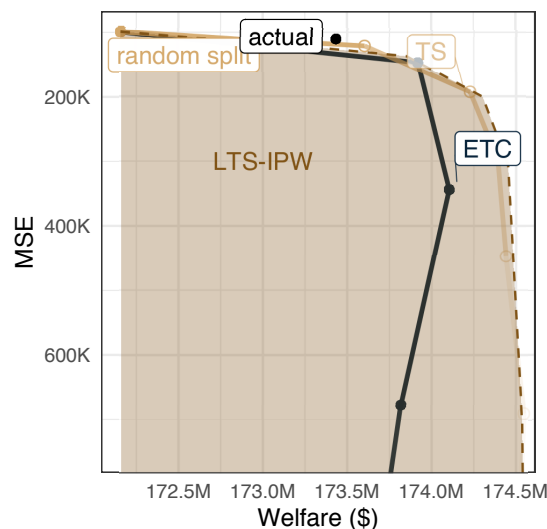
¹⁵In such setup, the independent and identically distributed arrival is a very strong assumption: unemployed people in different years are likely to behave differently. *Dimakopoulou et al. (2018)* investigate the estimation problem in the exploration-exploitation framework in settings where the outcome is heterogeneous by the arrival. They suggest a balancing method for contextual bandits to eliminate bias. In this paper I maintain the IID assumption to focus on the bias that arises even without any heterogeneity.

treatment effect (average net ITT) corresponding to a scenario of $\sigma = 13.7$ ¹⁶.

Figure 1.19 compares the performance of different strategies in the welfare-estimation space. As in the actual study 67% of the participants was assigned to the treatment and the net intention-to-treatment effect happened to be positive, the actual strategy fares very well¹⁷. However, the adaptive rules can win an additional welfare of as much as \$1M while still providing an unbiased estimate (in exchange for higher variance in the estimate). Furthermore, had the treatment effect happened to be negative, the actual strategy would have suffered a huge loss whereas the adaptive strategy could adapt to that scenario as well. Compared to a neutral 50-50% random split, the welfare gain of the adaptive strategy is much bigger and a large share of it could be realized without much loss on MSE.

The patterns of the figure are really similar to that of the $\sigma = 15$ case in Figure 1.12, only the uncertainty seems to be larger (limiting beats the unlimited strategy in welfare). This could result from the fact that the treatment effect is no longer constant and there is also variability in the population due to the bootstrap.

Figure 1.19: Welfare-estimation trade-off for the JTPA experiment



Notes: Each dot shows the achieved welfare and the mean squared error of the standard treatment effect estimator for a given strategy. The shaded area shows the available choices for the limited bandit strategy with inverse-propensity-weighted estimator for an appropriate batch size and limit. The dashed line connects the best possible combinations (Performance Frontier). The LTS-IPW strategy extends the available set of choices. Number of simulations = 10,000.

¹⁶First scale the outcome to have $\mu_0 = 0$ and $\mu_1 = 1$. Then scale the standard deviation of this scaled outcome by $\sqrt{10000/n}$.

¹⁷To assess the MSE of the actual assignment, I simulated a random split with a treatment share of 67%.

1.7 Concluding remarks

In our digital world, collecting data and base our decisions on them are getting technologically feasible. Therefore, online experimentation is getting more and more popular. In this paper, I dealt with this problem from a new perspective. Instead of focusing either on welfare maximization or estimation, I take a more practical viewpoint by considering both goals together. I borrow ideas from program evaluation and apply them on multi-armed bandits to improve upon the established methods valued by both welfare and estimation metrics.

Running a systematic Monte Carlo study, I highlight an important trade-off between welfare and estimation: experimentation strategies that result in good estimators (such as randomized controlled trial) suffer from huge opportunity cost, whereas the bandit algorithm that optimizes for welfare leads to biased treatment effect estimate. Some straight-forward strategies (e.g. explore-then-commit, bandit with estimation on randomized subsample) form transitions between the two extremes, so they provide good choices for decision-makers who have both welfare and estimation goals.

My contribution is threefold: First, I characterize the behavior of a well-known bandit heuristic, the Thompson sampling, across different setups. The standard treatment effect estimator on adaptively collected data suffers from amplification bias, and this bias increases in the relative size of the treatment effect and in the speed of adaptivity of the algorithm (smaller batches). The traditional bias correction method of inverse propensity weighting (IPW) does not work, it can even exacerbate the bias. Second, I highlight the welfare-estimation trade-off for established solutions. Finally, I suggest an easy-to-implement trick to correct the bias: limiting the adaptivity of the data collection by requiring sampling from all arms. Using inverse propensity weighting on data that arise from limited adaptivity results in an unbiased treatment effect estimate, whereas it preserves almost all of the welfare gain stemming from adaptivity.

If you face an easy problem where the relative size of the treatment effect is large, quick adaptivity along with small (or even no) limiting is the best choice to reach both high welfare and a reasonable estimator. If the noise is larger, choosing a higher batch size (skipping some decisions) is a better idea, as it could improve the expected outcome (similarly to how regularization improves prediction accuracy if the noise is large). Limiting more has small welfare cost while it can highly improve the precision of the estimator.

Running a bandit algorithm with limiting has a major advantage over the explore-then-commit strategy. While the latter could beat the frontier defined by the best batch size and limit combinations in certain setups, one should choose the sample for exploration optimally to realize this result. However, this sample should be chosen in advance where we do not know the relative treatment effect, nor the horizon. In contrast, when running an adaptive experiment, one can change the batch size and limiting parameters throughout the whole process, and adjust them

according to the actual knowledge about the environment – without risking unbiasedness.

My simulation considered only a very simple setup. Real world scenarios often include fat tail distributions, or much more than just one treatment. I stick to the simple setup to concentrate on the basic mechanisms of adaptive data collection. The main result of the welfare-estimation trade-off should hold for a much broader set of environments. I suppose that regularization with higher limits and larger batch sizes gets more important for fat tail distributions. However, this question should be answered by future research.

I expect that adaptive experiments are becoming more popular in every field, including economics. Understanding its mechanisms is essential to be able to use this tool correctly. This paper hopefully could contribute to this purpose.

Chapter 2

Examining the Effect of Retirement on Cognitive Performance - A Unifying Approach

2.1 Introduction

In developed countries, increased life expectancy, together with the parallel decline in the average retirement age, has increased the average spell of retirement in the last decades (the expected number of years in retirement for OECD countries increased from 10.6 years in 1970 to 18.2 years in 2015 for men, and from 14.6 to 22.7 years for women¹). Even if eligibility ages have been raised recently, people often spend 15-20 years of their lives as pensioners, which makes this phase of their life more and more relevant. Beside the individual level, the period of retirement is also of growing importance at the social level as well, because the proportion of retirees is increasing in the ageing population. As a natural consequence, various fields of research began to deal with the quality of the life of retirees. In this agenda, a particular aspect – namely the cognitive performance of old age individuals – has captured the attention of economists as it highly influences the decisions they make forming their consumption or saving behavior which affects the work of the economy to an increasing extent. Therefore, the age profile of cognitive abilities at the later stages of life is fundamental for many fields from marketing to pension and health policy.

It has been widely documented that individual cognitive performance tends to decline in older ages. According to Schaie (1989) cognitive abilities are relatively stable until the age of 50 but begin to decline afterwards. However, there is large heterogeneity in the progress of cognitive decay, raising the natural question of what are the driving forces behind and whether there is

¹OECD, <https://stats.oecd.org/index.aspx?queryid=54758>

a way to decelerate it in order to maintain cognitive abilities as long as possible. A popular hypothesis, which is often called as use-it-or-lose-it hypothesis (see for example Rohwedder and Willis, 2010), suggests that the natural decay of cognitive abilities in older ages can be mitigated by intellectually engaging activities. Thus, retirement which goes together with the cease of cognitively demanding tasks at work, might accelerate the natural declining process, having a negative causal effect on cognition. In this respect, the notion of retirement simply refers to not working, and thus incorporates a broader definition than usual (for example, people on disability benefit or who are unemployed could also be regarded as retirees).

Many papers have been investigating recently the effect of retirement on cognitive abilities in developed countries (e.g. Rohwedder and Willis, 2010; Mazzonna and Peracchi, 2012; Bonsang et al., 2012), yet the results they have delivered are ambiguous. The inconclusive outcome is most likely due to the difficulty of identification and the resulting variety in the identification strategies.

In this paper I investigate the effect of retirement and cognition by two methods: First, I replicate the estimations of previous papers uncovering the factors behind the differences. I show that they fail to disentangle the true effect highlighting how sensitive their results are to minor modifications of the set of controls. Second, I apply a novel identification strategy which aims to handle the problems which the current literature suffers from. Applying a difference-in-differences approach I can account for all time-invariant individual heterogeneity, and get a smaller result than any other previous work. This result is robust to choosing different time-periods or including more controls.

My analysis is based on the first, second and the fourth waves of the Survey of Health, Ageing and Retirement in Europe (SHARE)² which collects rich multidisciplinary data about the socio-economic status, health (including cognitive functioning), and other relevant characteristics (like social networks) of people aged 50 or over across 10 developed European countries. The survey is harmonized not only across European countries but also with other surveys such as the Health and Retirement Study (HRS), that serves similar purposes in the United States and is used by some of the replicated papers (e.g. Bonsang et al., 2012). To my knowledge, this paper is the first which makes use of a large longitudinal cross-country sample to go after the effect of retirement on cognition.

²This paper uses data from SHARE wave 4 release 1.1.1, as of March 28th 2013 (DOI: 10.6103/SHARE.w4.111) or SHARE wave 1 and 2 release 2.6.0, as of November 29 2013 (DOI: 10.6103/SHARE.w1.260 and 10.6103/SHARE.w2.260) or SHARELIFE release 1, as of November 24th 2010 (DOI: 10.6103/SHARE.w3.100). The SHARE data collection has been primarily funded by the European Commission through the 5th Framework Programme (project QLK6-CT- 2001-00360 in the thematic programme Quality of Life), through the 6th Framework Programme (projects SHARE- I3, RII-CT-2006-062193, COMPARE, CIT5- CT-2005-028857, and SHARELIFE, CIT4-CT-2006-028812) and through the 7th Framework Programme (SHARE-PREP, No 211909, SHARE-LEAP, No 227822 and SHARE M4, No 261982). Additional funding from the U.S. National Institute on Aging (U01 AG09740-13S2, P01 AG005842, P01 AG08291, P30 AG12815, R21 AG025169, Y1-AG-4553-01, IAG BSR06-11 and OGHA 04-064) and the German Ministry of Education and Research as well as from various national sources is gratefully acknowledged (see www.share-project.org for a full list of funding institutions).

The main challenge in uncovering the true effect is the endogeneity of retirement: a simple comparison of cognitive abilities of retirees and employees is likely to lead to biased estimates. Ideally, we would like to compare individuals from the same cohort and country, with the same age and education, one of them randomly assigned to be retired for a period of time while the other is working. As retirement is mainly an individual choice, this comparison is clearly impossible. For example, one can conveniently argue that the decay of cognitive abilities may induce the individual to retire, that is there is reverse causality going from cognition to retirement. This may result in overestimating the retirement effect on cognition in a simple comparison, even if we control for age. The standard solution is employing instrumental variables. Public policy rules (like official retirement age) seem to be good candidates for being relevant and exogenous instruments, as they clearly affect whether an individual retires but they generally refer to everyone irrespective of their actual cognitive performance.

While replicating previous cross-sectional analyses I show that applying public policy rule instruments in such settings can easily violate the exogeneity assumption: eligibility rules typically vary by country and gender, so not controlling for them properly can wrongly attribute general differences to the effect of retirement. This explains the sensitivity of cross-sectional results to alternative sets of controls.

Bonsang et al. (2012) use panel data from the HRS estimating a fixed-effect specification that naturally controls for such general differences. However, their results cannot be replicated on my sample. Also, their strategy uses an unbalanced panel over 10 years that mixes together the effect of retirement with other factors: cohort, attrition and learning effects could bias their results in an ambiguous way.

My strategy directly compares individual cognitive paths of employed and retired people. This way I also control for all time-invariant individual characteristics that allows for safely applying the standard public policy rules as instruments. Comparing cognitive paths implies a balanced panel specification, making it easier to assess panel concerns. I estimate the same relationship on different samples (comparing different waves) that contain different cohorts and also might suffer from different attrition rates, and my results are robust to these alternative specifications. Contrary to previous findings, my results suggest that retirement does not seem to cause serious harm for cognition: only around 0.02 standard deviation yearly.

This paper is structured as follows. Section 2.2 gives a formal setup for the problem and discusses the main challenges in the identification. Section 2.3 describes the data, detailing the various cognitive measures. Section 2.4 contains the replications of the previous results backed up by additional estimates that put them into context. Section 2.5 exhibits the results of my strategy. Section 2.6 concludes.

2.2 Model

A general way to model parametrically the relationship between cognitive performance and retirement is the following:

$$C_i = \alpha + f(R_i; \beta) + u_i \quad (2.1)$$

$$u_i = \mathbf{X}_i' \gamma + \varepsilon_i \quad (2.2)$$

where C_i denotes the cognitive performance of individual i , R_i is the number of years the individual has spent in retirement (i.e. not working). I allow for the cognitive performance to depend upon these years through an arbitrary function f with parameter β . The term u_i contains all factors associated with C_i except for R_i , for example: age. Equation (2.2) makes this dependency explicit where \mathbf{X}_i is the vector containing these factors.

Clearly, $E[C_i|R_i] = \alpha + f(R_i; \beta) + E[u_i|R_i]$. Assuming that we know f and have a good measure for C_i , the parameter of interest (β) can be consistently estimated if $E[u_i|R_i] = 0$. However, this is hardly the case. There are two sources which make the exogeneity assumption dubious: omitted variable bias and reverse causality.

Omitted variable bias There are lots of factors which are associated with the cognitive performance and also the years spent in retirement. These are factors in \mathbf{X}_i which are correlated with R_i . The most obvious candidate is age: older individuals are expected to have spent more years in retirement and they also have worse cognitive skills due to age-related decline. Education is also incorporated in \mathbf{X}_i : worse educated individuals retire earlier and they also have worse cognition. One should take care of these factors when estimating the effect of retirement on cognitive performance. The main challenge here is that we do not know exactly what factors are in \mathbf{X}_i .

Reverse causality One can conveniently argue that the decay of cognitive abilities may induce the individual to retire, so there is reverse causality going from cognition to retirement. That may result in overestimating the retirement effect on cognition in a simple comparison, even if we control for all factors in \mathbf{X}_i .

Most attempts trying to uncover β apply instrumental variables, as they might be able to eliminate both problems. Good instrumental variables (let us denote them by the vector \mathbf{Z}_i) satisfy two requirements: first, they are correlated with the possibly endogenous retirement variable ($\text{Cov}(\mathbf{Z}_i, R_i) \neq 0$), and second, they are related to the cognitive performance only through years of retirement ($E[u_i|\mathbf{Z}_i] = 0$). If these two assumptions hold, both omitted variable bias and reverse causality are resolved.

2.3 Data

Most papers which are after the effect of interest use the same sources of data provided by three large longitudinal surveys: the Health and Retirement Study (HRS), the English Longitudinal Survey of Ageing (ELSA) and the Survey of Health, Ageing and Retirement (SHARE).

Aiming to provide a multidisciplinary data about ageing, the United States of America launched the Health and Retirement Study (HRS) in 1992, and since then the study has collected detailed information about socio-economic status, health (including cognitive functioning), and other relevant characteristics (like social networks) of people aged 50 or over. Respondents of the survey are visited biannually and put through in-depth interviews to collect rich panel micro data about ageing population. The English Longitudinal Survey of Ageing (ELSA) was designed according to the HRS with its first wave launched in 2002. 2 years later Continental Europe also decided to set up an ageing database by establishing the Survey of Health, Ageing and Retirement in Europe (SHARE), a cross-nationally comparable panel database of micro data. SHARE started with 12 countries (Austria, Belgium, Denmark, France, Germany, Greece, Israel, Italy, the Netherlands, Spain, Sweden and Switzerland) in 2004 with wave 1, three countries (the Czech Republic, Ireland and Poland) joined in wave 2, and another four countries (Estonia, Hungary, Portugal and Slovenia), joined in wave 4. The three surveys (HRS, ELSA and SHARE) are carefully harmonized, and thus provide an excellent basis for cross- country investigation of ageing population in developed countries.

What makes the surveys appropriate for this particular analysis is that they include a battery of tests about cognitive abilities (memory, verbal fluency and numeracy). The test of memory is done as follows: 10 simple words are read out by the interviewer and the respondent should recall them once immediately after hearing and then at the end of the cognitive functioning module. As a result, both immediate recall and delayed recall scores range from 0 to 10. Often, the two variables are merged to a composite one by adding them up, which is called total word recall. Verbal fluency is tested by asking the respondent to name as many distinct animals as she can within one minute. The length of this list provides a measure for verbal fluency. SHARE also consists of several questions about individual numeracy skills. Respondents who answer the first one correctly get a more difficult one, while those who failed get an easier one. The last question requires the respondent to calculate compound interest. The number of correct answers to these questions provides an objective measure of numeracy ranging from 0 to 4. Finally, there is a test of orientation of four questions which examines whether the respondent is aware of the date of the interview (day, month, year) and the day of the week. This test may be used to detect individuals with serious cognitive problems or progressed dementia.

Various measures of cognitive skills might grab its different aspects as argued in [Mazzonna and Peracchi \(2012\)](#). As most of the papers use the results on memory tests I also focus on that measure for comparison purposes. To have a common unit I use standardized scores to express

scales in standard deviation.

Throughout the paper I make use of the first, second and fourth waves of SHARE. The third wave of data collection (SHARELIFE) is omitted, as it is of different nature: it focuses on people's life histories instead of current characteristics.

2.4 Replications

In this section I replicate the main results of the literature, specifically that of [Rohwedder and Willis \(2010\)](#), [Mazzonna and Peracchi \(2012\)](#), and [Bonsang et al. \(2012\)](#). I put all of these results in my unified framework and show that their differing conclusions actually fit in the broader picture. The ambiguity of their results stems from the differences in their identification strategies that implies that their estimated "effects" of retirement on cognitive performance measure different kinds of things.

The papers differ in three crucial aspect: first, what is their assumption about how retirement should affect cognitive performance (i.e. what is their assumption for f), second, how they handle omitted variable bias (i.e. which factors they are controlling for from \mathbf{X}_i), and third, what is their choice for instrumental variable to overcome endogeneity. Besides the methodology, they also differ in the data they use for estimation. However, considering the goal of uncovering a general relationship this fact should be of secondary importance as far as the measurements are comparable across the datasets.

The structural equation the papers try to estimate could be summarized as follows:

$$S_i = \alpha + f(R_i; \beta) + \mathbf{X}_i^{*'} \gamma^* + \tilde{u}_i \quad (2.3)$$

$$\tilde{u}_i = \tilde{\mathbf{X}}_i' \tilde{\gamma} + \varepsilon_i \quad (2.4)$$

where S_i is a cognitive score, a measurement of cognitive performance. This formulation helps to differentiate between factors which are controlled for (\mathbf{X}_i^*) versus factors which remain in the error term ($\tilde{\mathbf{X}}_i$). To get a clear causal effect equation (2.3) is estimated by a 2SLS procedure where the first stage is

$$R_i = \mathbf{Z}_i' \pi + \mathbf{X}_i^{*'} \rho + v_i \quad (2.5)$$

From now on let us assume that the cognitive measurements detailed in the previous section describe well the actual cognitive skills. To be more precise, I assume that $C_i = S_i + e_i$ where e_i is a classical measurement error in the dependent variable, i.e. $\text{Cov}(e_i, S_i) = \text{Cov}(e_i, R_i) = 0$. In this case our estimators remain consistent although less precise.

All of the papers use various public policy rules to instrument retirement (such as pension eligibility rules). Such rules are good candidates for instrument as they vary across country and gender and are strongly correlated with employment status. The crucial question is whether it also satisfies the exogeneity assumption. Formally, the exogeneity assumption can be expressed as $E[\tilde{u}_i | \mathbf{Z}_i] = 0$. It essentially says that there is no systematic difference in the cognitive performance of an eligible and a non-eligible individual in the sample (after controlled for some other features).

2.4.1 Rohwedder and Willis (2010)

The first serious attempt to uncover the causal relationship between retirement and individual cognitive performance (Rohwedder and Willis, 2010) uses a simple setup: they only include a dummy for not working on the right hand side on a restricted sample of people aged between 60 and 64. This is equivalent to estimating the average effect of retirement on cognition conditional on the average duration of retirement the sample, that is assuming that $f(R_i; \beta) = \tilde{\beta} \mathbf{1}(R_i > 0)$ where $\tilde{\beta} = \beta \bar{R}_i$. Beside restricting the sample on a narrow age-range they do not include anything in \mathbf{X}_i^* . To handle endogeneity they use public pension eligibility rules as instruments: whether the individual is eligible for early or full benefits. See Table B.1 for a summary of the methodologies.

Rohwedder and Willis (2010) estimate their model on the 2004 waves of SHARE, ELSA and HRS, and find that retirement has a large adverse effect on cognition among 60-64 years old, amounting to one-and-a-half standard deviation. Unfortunately, they do not report the average duration of retirement in their sample which makes it hard to convert this number to yearly average.

Using only the first wave of SHARE (and thus having a much smaller sample than theirs, 4464 versus 8828 observations) I was able to replicate their main findings (see the first column of Table 2.1). The pattern is the same: retirement seems to decrease cognitive performance. However, my estimation is somewhat smaller, amounting to only 1 standard deviation. Considering that the average duration of retirement in my sample is 6.6 years, it could be translated to an average yearly decline of 0.15 standard deviation (if I use the same number for conversion, the estimate of Rohwedder and Willis (2010) corresponds to 0.23 standard deviation yearly decline).

In order to be able to interpret the previous result as causal effect it should be true that $E[\tilde{u}_i | \mathbf{Z}_i] = 0$. Clearly, eligibility rules are not related to unobserved individual idiosyncrasies in cognition, as they generally refer to everyone. So using the instrument indeed helps with the problems. However, there are other factors left in \tilde{u}_i which are likely to be correlated with the instrument. For example, in most countries eligibility rules differ for males and females: women tend to become eligible earlier. Women also have higher memory scores than men in

Table 2.1: Comparing the methodology of Rohwedder and Willis (2010) by two versions of the instrumental variable: 2SLS estimation

	(1) Rohwedder and Willis (2010)	(2) Mazzonna and Peracchi (2012)
Retired	−1.010*** (0.14)	−0.500*** (0.13)
Constant	0.736*** (0.10)	0.365*** (0.097)
Observations	4,464	4,464
Weak IV F statistic	154.45	156.12

Notes: Both results are from the second stage estimation of $S_i = \alpha + \beta 1(R_i > 0) + u_i$ where the retirement dummy is instrumented by early and normal eligibility dummies. The coefficient of interest in Rohwedder and Willis (2010) is -4.66^{***} on a sample of 8,828 observations which amounts to 1.5 standard deviation. The corresponding first stage regressions are summarized in Table B.2.

Weak IV F statistic is calculated according to Angrist and Pischke (2008). Stock et al. (2002) suggest that an F below 10 should make us worry about the potential bias in the IV estimation.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

the same age, even before retirement (for people below 55 the mean difference amounts to 0.19 standard deviation in my data). Not controlling for gender is likely to lead to underestimated effects as women with better scores are overrepresented in the eligible population. Moreover, people from different countries might differ in their average education as well (e.g. because of different compulsory schooling laws affecting today's pensioners). As different countries also have different eligibility rules, ignoring schooling is also likely to undermine the exogeneity of the instruments. Bingley and Martinello (2013) show that countries with higher eligibility ages also tend to have better educated old age people, and thus the effect of Rohwedder and Willis (2010) is overestimated. The violation of the exogeneity assumption makes the causal interpretation of the results in the first column of Table 2.1 questionable.

2.4.2 Mazzonna and Peracchi (2012)

The paper of Mazzonna and Peracchi (2012) improves upon Rohwedder and Willis (2010) along all of the three aspects: they allow for a yearly retirement effect instead of including just a retired dummy, they control for a set of features (age, gender and country), and they use a modified instrumental variable that has some variation within the country-gender cells. They end up with an estimated yearly decline of around 0.04 standard deviation, an order of magnitude less than the first estimate. I replicate their strategy by implementing their improvements one by one, to shed some light on what causes the reasonable drop in the effect.

I start with the modified instrumental variable: as opposed to the eligibility rules that were in

effect at the time when the interviews were conducted, [Mazzonna and Peracchi \(2012\)](#) consider also the changes that the rules might have had during the times. For each individual they apply the eligibility rules that were in effect for the individual's cohort. This way they have some variation in the rules within country-gender cells. Both instrumental variables reach the same level of relevance (see the first stage regression results in [Table B.2](#) in the appendix). However, using the refined IV results in a reasonable drop in the coefficient of interest (see the second column of [Table 2.1](#)) even with the original specification. Introducing within-country-gender variation into the instrumental variable leads to halving of the effect, to a decline of only 0.075 standard deviation per year.

The methodology of [Mazzonna and Peracchi \(2012\)](#) differs from that of [Rohwedder and Willis \(2010\)](#) not only in respect of the instrumental variable. They also assume a different functional form, and control for a different set of features. Instead of using just a retirement dummy (and thus estimating the effect conditional on the average duration of retirement) they enter the number of years spent in retirement linearly in the equation (i.e. they assume that $f(R_i; \beta) = \beta R_i$). To adapt to the different endogenous variable, they also modify the instrument accordingly: instead of using eligibility dummies, they calculate the years lived after reaching the eligibility age (i.e. $\max(0, \text{age} - \text{age}_{\text{eligibility}})$). They control for age in a different manner: instead of restricting the sample to 60-64 years old they estimate a linear age coefficient on a sample of people aged 50-70. They also control for country dummies and estimate the equation separately for men and women.

[Table 2.2](#) summarizes the results of moving from the strategy of [Rohwedder and Willis \(2010\)](#) to that of [Mazzonna and Peracchi \(2012\)](#) step by step. ([Table B.3](#) in Appendix shows the corresponding first stage regression results.) This simple exercise illustrates how sensitive the estimates are to various specifications. In the followings, I comment on each specification, explaining what could be the reason behind the change (each point discusses the estimated specification with the corresponding number):

- (1) The estimated effect of 0.05 standard deviation yearly (first column) is comparable to the effect estimated with the retirement dummy (see second column [Table 2.1](#) and considering the average retirement duration of 6.6 years: $0.5 / 6.6 = 0.075$).
- (2) Extending the age range does not really matter.
- (3) Restricting to those with labor market history makes the effect a bit larger. This is practically due to excluding some outliers with 50+ years spent in retirement.
- (4) Controlling for age delivers weird results. The effect doubles and the coefficient on age is positive: age seems to improve cognitive performance until retirement, whereas it deteriorates it by around 0.14 standard deviation after that. This could be explained by country differences: as [Bingley and Martinello \(2013\)](#) draws the attention to, eligibility age and

Table 2.2: Moving from the strategy of Rohwedder and Willis (2010) to that of Mazzonna and Peracchi (2012)

	(1) aged 60-64	(2) aged 50-70	(3) + worked at 50	(4) + age	(5) + country
Years in retirement	-0.051*** (0.0072)	-0.053*** (0.0021)	-0.083*** (0.0029)	-0.169*** (0.015)	0.158*** (0.032)
Age				0.044*** (0.0075)	-0.112*** (0.015)
Constant	0.376*** (0.051)	0.359*** (0.015)	0.226*** (0.012)	-2.157*** (0.41)	6.090*** (0.77)
Country dummies	No	No	No	No	Yes
Observations	4,052	17,448	14,052	14,052	14,052
Weak IV F statistic	118.05	1614.48	5779.83	246.16	60.57

Notes: The results are from the second stage estimation of $S_i = \alpha + \beta R_i + X_i^* \gamma + u_i$ with different samples and different X^* where years of retirement is instrumented by early and normal eligibility dummies. The corresponding first stage regressions are summarized in Table B.3.

Weak IV F statistic is calculated according to Angrist and Pischke (2008). Stock et al. (2002) suggest that an F below 10 should make us worry about the potential bias in the IV estimation.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

schooling is positively correlated (in my sample the correlation is 0.21 and 0.14 for the early and normal eligibility age, respectively). Therefore, comparing two individuals with the same age but differing years after eligibility likely means comparing two individuals from different countries with the older one being from the better educated country. This reasoning justifies the positive age coefficient and underlines the importance of controlling for both age and country.

- (5) Controlling for country indeed solves the puzzle of positive age coefficient, changing its sign to what is expected. However, now the coefficient of interest changes sign and gets positive. The unexpected sign results again from omitted variable bias: gender is not controlled for. As mentioned previously, women perform significantly better on memory tests (controlling for age) than men (for this sample, they are by 0.28 standard deviation better. Thus, when we control for both age and country, we mainly identify the retirement effect from gender variation. To see that this is really the case, check the results in the first two columns of Table 2.3 where I also included a control for gender. The positive sign of the coefficient of interest reverses back to what is expected. The next two columns of the table shows the same result when the numeracy score is used to measure the cognitive skills. Women perform on average by -0.28 standard deviation worse on the numeracy test and correspondingly, we see larger negative effect of years in retirement on numeracy when not controlling for gender. For fluency, there is no notable difference in the performance of men and women.

There is one more puzzle in Table 2.3. Why is the coefficient on age is positive for numeracy

Table 2.3: Moving to the strategy of [Mazzonna and Peracchi \(2012\)](#) - the effect of gender control for different measures of cognitive performance

	(1) TWR	(2) TWR	(3) numeracy	(4) numeracy	(5) fluency	(6) fluency
Years in retirement	0.158*** (0.032)	-0.176*** (0.036)	-0.360*** (0.042)	-0.013 (0.032)	-0.038 (0.026)	-0.048 (0.031)
Age	-0.112*** (0.015)	0.049*** (0.017)	0.151*** (0.020)	-0.016 (0.016)	-0.007 (0.013)	-0.002 (0.015)
Female		0.290*** (0.022)		-0.302*** (0.019)		0.010 (0.019)
Constant	6.090*** (0.77)	-2.270** (0.90)	-7.318*** (1.03)	1.392* (0.80)	0.951 (0.64)	0.689 (0.76)
Country dummies	Yes	Yes	Yes	Yes	Yes	Yes
Observations	14,052	14,052	14,145	14,145	14,004	14,004
Weak IV F statistic	60.57	43.41	61.11	44.23	62.38	45.15

Notes: The results are from the second stage estimation of $S_i = \alpha + \beta R_i + X_i' \gamma^* + u_i$ where years in retirement is instrumented by early and normal eligibility dummies. Column (1) is equivalent to column (5) of Table 2.2, it is included to ease the comparison.

Weak IV F statistic is calculated according to [Angrist and Pischke \(2008\)](#). [Stock et al. \(2002\)](#) suggest that an F below 10 should make us worry about the potential bias in the IV estimation.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

when controlling for country but not for gender? (The same coefficient is negative for TWR.) There is a possible explanation for that: as the sample ages so decreases the share of women (interestingly, as mortality rates would predict the opposite). The coefficient on age is mainly identified on non-eligible population (as for eligible population the age effect is actually the sum of the coefficients on age and years in retirement). As women are better in memory tests, and their share is smaller in older cohorts, the composition effect implies a negative coefficient for age. By contrast, the opposite is true for numeracy (women perform worse), so the composition effect implies a positive coefficient for age. Controlling for gender eliminates the level differences in cognitive scores. However, the rate of cognitive decline due to retirement might still be different by gender (i.e. heterogeneous retirement effect for men and women) that could further complicate the results and make the direction of possible bias hard to assess.

To allow for heterogeneous retirement effect by gender, [Mazzonna and Peracchi \(2012\)](#) estimate the equation separately for men and women in their preferred specification. Table 2.4 show my replication for their strategy for total word recall, numeracy and fluency. According to my results, the rate of decline is indeed different: the relatively better performing gender suffers a larger decline. These estimations differ from that of [Mazzonna and Peracchi \(2012\)](#) only in how cognitive performance is measured: they adjust the cognitive scores by the time spent on answering them whereas I do not do. Nevertheless, the replicated numbers are comparable to theirs, although less precise, showing a yearly decline of about 0.02-0.05 standard deviation.

Table 2.4: Estimating separately by gender, closest to [Mazzonna and Peracchi \(2012\)](#)

	(1) TWR men	(2) TWR women	(3) numeracy men	(4) numeracy women	(5) fluency men	(6) fluency women
Years in retirement	0.015 (0.037)	−0.041* (0.024)	−0.058 (0.039)	−0.035 (0.024)	−0.025 (0.037)	−0.029 (0.022)
Age	−0.041** (0.017)	−0.017 (0.012)	0.008 (0.018)	−0.009 (0.012)	−0.011 (0.017)	−0.015 (0.011)
Constant	2.403*** (0.90)	1.309** (0.61)	0.053 (0.94)	0.832 (0.61)	1.083 (0.89)	1.427** (0.56)
Country dummies	Yes	Yes	Yes	Yes	Yes	Yes
Observations	8,058	5,994	8,126	6,019	8,028	5,976
Weak IV <i>F</i> statistic	33.19	79.78	32.76	79.66	33.48	83.38

Notes: The results are from the second stage estimation of $S_i = \alpha + \beta R_i + X_i' \gamma^* + u_i$ where years in retirement is instrumented. [Mazzonna and Peracchi \(2012\)](#) estimate different equations for immediate and delayed word recall that I use as an aggregate total word recall. Their estimates are −0.018** and 0.15* (men) and −0.051*** and −0.025*** (women) for IWR and DWR respectively, −0.029*** (men) and −0.041*** for numeracy, and −0.006 (men) and −0.023** (women) for fluency. See Column 2B of their Table 7.

Weak IV *F* statistic is calculated according to [Angrist and Pischke \(2008\)](#). [Stock et al. \(2002\)](#) suggest that an *F* below 10 should make us worry about the potential bias in the IV estimation.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Although this strategy is more robust than the original one, there is a potential issue that could contaminate the results. As already mentioned, education matters in old age cognitive skills even if gained in early stages of life ([Banks and Mazzonna, 2012](#)). Today's pensioners are highly affected by the expansion of average schooling: the 50 years old cohort spent on average 2.7 years more in school than the cohort of 70. [Mazzonna and Peracchi \(2012\)](#) try to control for education by including a low-education dummy (they also interact this dummy with the effect) and show that education indeed plays a significant role in explaining the heterogeneity in the levels of cognitive skills (and to a smaller extent in their age-related decline). However, there is also evidence (see for example the PISA surveys) that countries are different in how effectively they improve cognitive abilities in childhood. The first PISA survey was performed in 2000, the mean scores of countries in math, science and reading are positively correlated with the eligibility ages ([OECD, 2001](#)). If there is some persistence in the quality of education systems from the time when today's pensioners went to school and that of today, this might introduce a new type of bias in the estimations even if the number of years spent in education is controlled for.

[Mazzonna and Peracchi \(2012\)](#) improve a lot on the first estimate, getting a much smaller effect, but the sensibility of the results to different control sets show how hard it is to ensure the exogeneity of the instrument in a cross-sectional setting. As public policy rules only vary across countries and gender (mostly), and these are related to a lot of factors that also affect individuals' cognitive performance, controlling for all of them seems nearly impossible.

2.4.3 Bonsang et al. (2012)

The first paper that uses panel data to estimate the causal effect of retirement on cognitive performance (Bonsang et al., 2012) looks only at the US: they extract information from 6 waves (1998-2008) of the HRS.

In their main specification they follow the approach of Rohwedder and Willis (2010) by estimating the effect of retirement through a simple dummy, that is not taking into account the length of retirement. The only condition is that they restrict their attention to those who have been retired for at least one year. Formally, they assume that $f(R_i; \beta) = \tilde{\beta} \mathbf{1}(R_i \geq 1)$.

Using unbalanced panel data they could control for individual heterogeneity by estimating a fixed effects specification. Thus, they could include a_i in X_i^* that means a control for all time-invariant factors, like (mostly) gender and country. Additionally, they also control for age in a quadratic form.

To handle the endogeneity of the retirement decision they also use eligibility ages as instruments: first, the age of 62 which is the eligibility age for social security in the US and second, the normal retirement age which varies by cohort. They measure the cognitive performance of the individual with the total word recall score. They find that being retired has an effect of 1 less word recalled, that could be translated into a yearly drop of 0.05 standard deviation, a slightly larger effect than what was found by Mazzonna and Peracchi (2012).

In the replication of their analysis, I use the same eligibility dummies for early and normal retirement benefits as before - these should correspond to the dummies which Bonsang et al. (2012) apply for the case of the US. Table 2.5 summarizes the results of the estimation which follows the main specification of Bonsang et al. (2012) but for a different sample (using SHARE instead of HRS). The age range (50-75) is the same, but the time period is shorter due to data limitation (spanning over only 6 years of 3 waves). The corresponding first stages can be found in the Appendix (Table B.4).

I could not replicate their results. Both the unrestricted retirement dummy that was used by Rohwedder and Willis (2010) and the dummy requiring a retirement spell of at least one year lead to positive coefficient estimates. Nonetheless, none of them are significant.

The difference might originate from at least two sources: First, the US might be different from the European countries. Either because eligibility rules have a stronger influence on the time of retirement which is suggested by my weaker first stage, or because retiring might mean something else in terms of cognition. This is a topic worth investigating further. Second, my time period is shorter.

Estimating the fixed-effects model on an unbalanced panel makes the results hard to interpret. Mazzonna and Peracchi (2012) argue that panel data is inappropriate for identifying the causal

Table 2.5: Replication of Bonsang et al. (2012)

	(1) Retirement duration > 0	(2) Retirement duration ≥ 1
Retired	0.0955 (0.14)	0.173 (0.14)
Age	0.185*** (0.016)	0.189*** (0.017)
Age (sq.)	−0.001*** (0.0001)	−0.001*** (0.0001)
Observations	41,476	37,374
Weak IV F statistic	136.82	131.89

Notes: The results are from the second stage estimation of $S_i = \alpha + a_i + \beta g(R_i) + u_i$ where $g(R_i) = 1(R_i > 0)$ or $1(R_i \geq 1)$ and these dummies are instrumented by early and normal eligibility dummies. The coefficient of interest in Bonsang et al. (2012) is -0.942^{***} on a sample of 54,377 observations. This amounts to 0.27 standard deviation, or (considering the average duration of retirement in their sample) a yearly drop of 0.05 standard deviation. The corresponding first stage estimates are summarized in Table B.4.

Weak IV F statistic is calculated according to Angrist and Pischke (2008). Stock et al. (2002) suggest that an F below 10 should make us worry about the potential bias in the IV estimation.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

effect of retirement on cognition because of two reasons: First, as in each wave the exact same cognitive exercises are performed, the participants may remember their answers from previous waves. Second, there is considerable attrition in the sample which might also bias the result. The unbalance in the data means that people who are present throughout the whole period get more weight. This makes it ambiguous how learning and attrition could bias the estimates. Also, pooling waves that extend over 10 years mixes cohorts that are differently affected by the extension of education. This could lead to compositional effect with unclear consequences. To assess the magnitude of these problems, I estimating the same relationship on different time periods. The resulting point estimates vary a lot, but all remain positive and insignificant (see Table B.5 in the Appendix). We need a cleaner identification strategy to reliably estimate the causal effect.

2.5 My strategy

I demonstrated in the previous section that getting a clear causal effect in cross-section is really challenging. Some straightforward improvements upon the strategy of Rohwedder and Willis (2010) resulted in a drop of the estimated effect from 0.23 standard deviation per year to a magnitude less. But still, there are plenty of factors left in u_i^ which are likely to bias our results (to a yet unknown extent). Moving to panel identification solves a lot of issues. It enables us to compare the cognitive scores of the exact same individual instead of assuming that another one*

is a good subject for the comparison. The fixed-effects specification on unbalanced panel data resulted in a negative estimated yearly decline with comparable magnitude to what we get with the improved cross-sectional method, but this does not let itself be replicated on my data and might be subject to other biases resulting from attrition, learning and cohort effects.

I suggest a novel identification strategy that also exploits the panel nature of the data but aims for more clarity in the mechanisms. Instead of pooling periods, I look at changes between two periods to identify whether the cognitive paths differ for retired and non-retired individuals. Formally, I estimate the following equation

$$\Delta S_i = \alpha^* + \beta \Delta R_i + \theta' W_i + \Delta \tilde{u}_i \quad (2.6)$$

where ΔS_i is the change in cognitive performance between two waves (measured by TWR, numeracy or fluency score) and ΔR_i is the number of years spent in retirement during this period. Note that this specification implies a balanced panel. It also controls for all time-invariant individual heterogeneity like gender, amount and quality of education, country, etc.³ Variables W_i control for different trends (instead of just level differences). This way I can handle issues like different rates of cognitive decline by gender, as discussed in the replication of Mazzonna and Peracchi (2012). The use of years spent in retirement instead of a simple retirement dummy captures better the actual treatment. To handle endogeneity, I use the same instruments as Mazzonna and Peracchi (2012): years after early and normal eligibility ages.

Following from the nature of the data collection, the time elapsed between interviews of two waves is not the same for all individual (e.g. it ranges from 11 to 40 months between the first two waves). Therefore, the amount of ageing between two waves is not the same across the sample so I need to control for this by including the years elapsed variable in W .

This strategy makes evaluating the panel concerns easier. The learning effect is captured by the constant term: if people are indeed better at their second and third interviews because of the repetition, $\hat{\alpha}^*$ should be positive. However, learning effect shall not bias the coefficient of interest due to the balanced panel. It only matters if learning is different for employed and retired persons, but in this case, this is part of the retirement effect so it should be estimated within the treatment effect – and this is exactly what is going to happen in this specification.

Using a balanced panel also resolves some of the attrition concerns. To see whether it still causes any a problem, I will estimate the same relationship on different samples (comparing different waves) which are likely to suffer from different attrition rates and check whether they are different.

³For two time periods, regression on differences gives the same coefficient estimates as the fixed effect specification on a balanced panel.

I have three waves of SHARE, so I can estimate the differences in three ways: between wave 1 and 4 (the longest period), between wave 1 and 2, and between wave 2 and 4. I restrict my sample to those who were aged between 50 and 70 and were either employed or retired in the first period, and has worked at age 50. I exclude those who returned to the labor market during this period (around 5% of the sample). Table 2.6 summarizes the data for the different comparison periods.

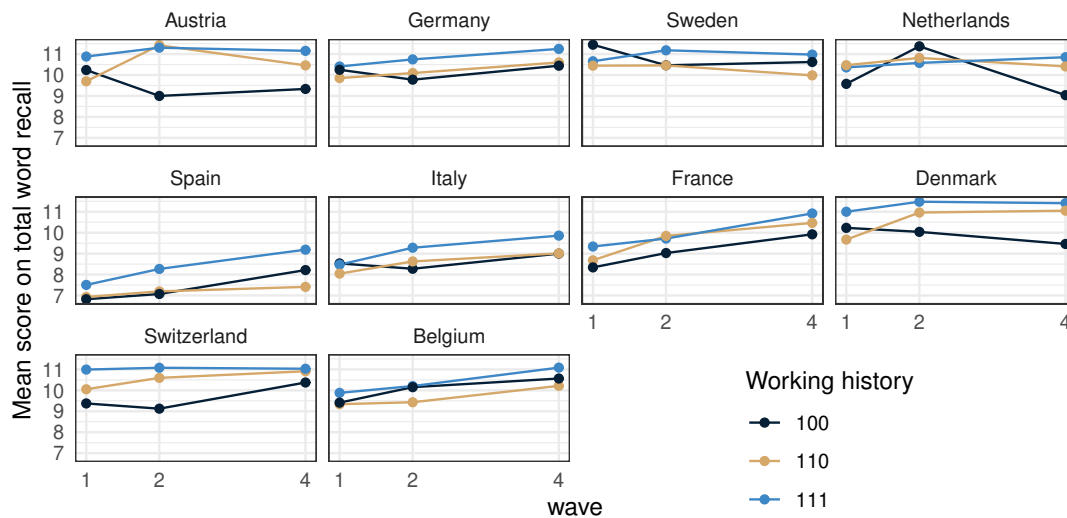
Table 2.6: Summary statistics

Sample		Men				Women			
		wave 1	wave 2	wave 4	N	wave 1	wave 2	wave 4	N
wave 1-4	age	59.7 (5.9)		66.3 (6.0)	3,551	58.9 (5.8)		65.5 (5.9)	2,877
	TWR	8.73 (3.17)		9.22 (3.27)	3,528	9.82 (3.30)		10.49 (3.37)	2,866
	numeracy	3.74 (1.03)		3.82 (1.02)	3,547	3.48 (1.02)		3.56 (1.00)	2,873
	fluency	21.34 (7.08)		20.41 (6.94)	3,510	22.00 (7.04)		21.25 (7.24)	2,858
	retired	1,537		2,244	3,551	1,137		1,761	2,877
wave 1-2	age	59.7 (5.9)	62.1 (6.0)		5,332	58.9 (5.9)	61.2 (5.9)		4,026
	TWR	8.71 (3.13)	9.01 (3.17)		5,258	9.78 (3.26)	10.13 (3.23)		3,998
	numeracy	3.76 (1.01)	3.80 (1.04)		5,295	3.48 (1.04)	3.54 (1.03)		4,000
	fluency	20.48 (7.13)	20.56 (7.10)		5,214	21.31 (7.11)	21.43 (7.35)		3,986
	retired	2,346	2,702		5,332	1,651	1,933		4,026
wave 2-4	age		60.3 (5.8)	64.5 (5.8)	4,359		59.2 (5.8)	63.5 (5.9)	3,783
	TWR		9.23 (3.10)	9.53 (3.20)	4,330		10.47 (3.23)	10.81 (3.30)	3,765
	numeracy		3.86 (1.03)	3.87 (1.01)	4,345		3.58 (1.02)	3.60 (1.00)	3,765
	fluency		21.69 (7.17)	20.87 (6.94)	4,313		21.31 (7.14)	21.43 (7.07)	3,757
	retired		1,712	2,142	4,359		1,319	1,711	3,783

Notes: The table presents the means and standard deviations (in parentheses) of the individuals' characteristics for each sample, by gender and wave.

Figure 2.1 shows the average paths of cognitive score of the individuals who were present in all waves, by working history for each country. For example, working history of 110 indicates a person who worked during the first two waves but left between wave 2 and 4. The first thing to note is that no clear pattern arises regarding the effect of retirement. There are some signs of selection and learning, but it does not seem like retirement would have any clear-cut effect on cognitive scores.

The first three columns of Table 2.7 shows the 2SLS results for changes in total word recall score between wave 1 and 4 with different controls (again, years in retirement is instrumented

Figure 2.1: Pattern of cognitive scores across waves by working history

Notes: Working history code corresponds to the three waves used in the analysis, each digit displaying 1 if the individual worked in the given wave (100: worked only in first wave, 110: worked in first two waves, 111: worked in each wave).

by years after the eligibility ages). It seems that one more year in retirement decreases the cognitive score by 0.03-0.04 standard deviation point. My preferred specification is in column (3) where I allow for different trends in cognitive decline by gender and country. According to the estimates, women tend to lose their abilities slower. The positive constant term includes the effect of learning. The magnitude of the effect is comparable to that of Table 2.4, the closest replication of *Mazzonna and Peracchi (2012)*. Note that my estimates are much more stable to including additional controls (e.g. allowing for different trends by country) which I interpret as a sign that the panel specification in itself eliminates many biases inherent in a cross-sectional analysis.

However, there is a reasonable scenario we have not considered yet. If natural age-related cognitive decline is concave instead of being linear then controlling only for a linear age trend might attribute the larger decline in older ages to the effect of retirement. Unfortunately, we are unable to allow for heterogeneous age effect and still use our instruments (there is not enough variability within country, gender and age). It still makes sense to run simple OLS regressions to see the difference.

If we compare the third and fourth columns of the tables we can see that OLS results are slightly smaller in absolute value than those of 2SLS. Originally, we were afraid of a negative selection bias. This difference could be possibly explained by two facts: First, we eliminated all time-invariant individual heterogeneity so selection only matters if the rate of decline is different as well. Second, the 2SLS estimate the Local Average Treatment Effect (LATE), the effect of retirement on those who retired because reaching the eligibility age. These people might have been exposed a bigger change in their lifestyle than those who were retired anyway, and thus,

Table 2.7: Panel estimation: change in total word recall score between wave 1 and 4

	(1) 2SLS	(2) 2SLS	(3) 2SLS	(4) OLS	(5) OLS
Years in retirement	-0.025*** (0.0058)	-0.026*** (0.0058)	-0.040*** (0.0060)	-0.028*** (0.0044)	-0.016** (0.0065)
Years elapsed	-0.263*** (0.033)	-0.263*** (0.033)	-0.084* (0.050)	-0.093* (0.050)	-0.093* (0.050)
Female		0.042 (0.026)	0.053** (0.026)	0.055** (0.026)	0.049* (0.026)
Age at first wave					-0.009*** (0.0032)
Constant	1.819*** (0.22)	1.808*** (0.22)	0.568* (0.34)	0.571* (0.34)	1.086*** (0.39)
Country dummies	No	No	Yes	Yes	Yes
Observations	6,394	6,394	6,394	6,394	6,394
Weak IV <i>F</i> statistic	3729.06	3787.19	3665.10		

Notes: All results are from the estimation of $\Delta S_i = \alpha^* + \beta \Delta R_i + W_i^{*'} \nu + \Delta \tilde{u}_i$ where $\Delta M_i = M_{i,w=4} - M_{i,w=1}$ and $S_i = \text{Total word recall}_i$. For 2SLS, the second stage regressions are reported where ΔR_i is instrumented by the distance from early and normal retirement age, and W include years elapsed, female dummy and country dummies. For OLS, W additionally include age at first wave. The corresponding first stage estimates are summarized in Table B.7.

Weak IV *F* statistic is calculated according to Angrist and Pischke (2008). Stock et al. (2002) suggest that an *F* below 10 should make us worry about the potential bias in the IV estimation.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

might have been affected by a more expressed cognitive shock at retirement. Including age control (column 5) decreases the OLS estimate further, to less than 0.02 standard deviation yearly. This decline suggests that the 2SLS estimate in column (3) without age control slightly overestimates the true effect.

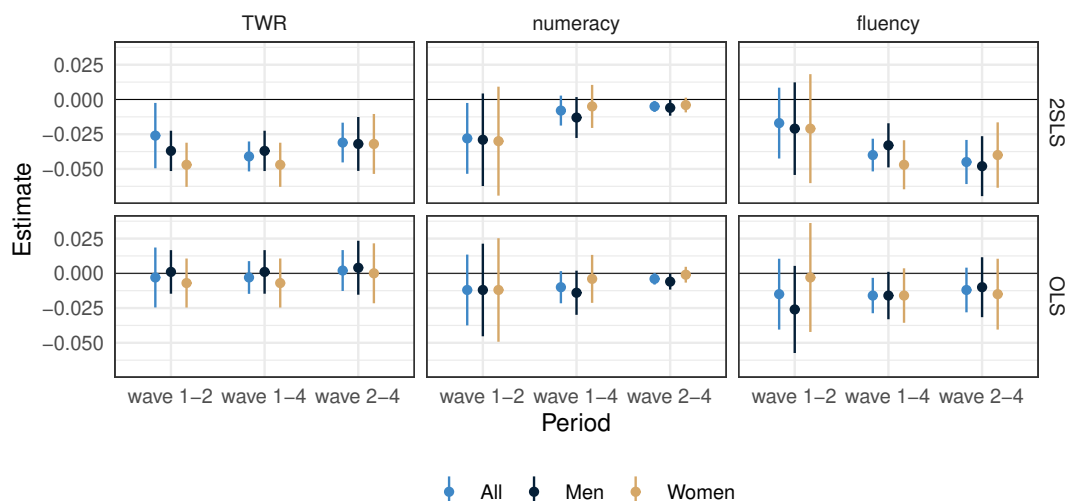
The same results for the other periods and other cognitive measures along with the corresponding first stages can be found in Appendix. The results are mainly consistent: 2SLS estimate a yearly decline of 0.01-0.05 standard deviation. The OLS estimates are smaller in absolute value, ranging between 0 and 0.02 standard deviation. There is no clear pattern by periods and measures which serves as a robustness check for the results⁴. My results are even smaller than the previous ones, like Mazzonna and Peracchi (2012) and Bonsang et al. (2012), especially if we take into account that the 2SLS estimates might still overestimate the true effect.

As an additional robustness check, we can estimate the preferred specifications separately for men and women. We saw before that it was crucial in the cross-sectional setting to pin down the estimates. In contrast, in the panel setting it is less of concern. Figure 2.2 compares the estimates by gender along with the pooled estimates. The estimated coefficients lie close to each

⁴The other coefficients are less consistent. Sometimes, the constant term and the years elapsed estimate switch sign – it can be a result of the high noise and the difficulty to separate the effects as there is a little variation in the years elapsed variable. The female coefficient is mostly positive.

other in every setting. The picture they draw is consistent: the effect of retirement on cognition cannot exceed a yearly 0.05 standard deviation decline (that is close in magnitude to what Mazzonna and Peracchi (2012) and Bonsang et al. (2012) get with different methods), and it might still be an overestimate of the true effect (see the near-zero OLS-estimates).

Figure 2.2: Comparison of the retirement effect estimate by gender



Notes: The dots represent $\hat{\beta}$, the lines show the 95% confidence intervals from the estimation of $\Delta S_i = \alpha^* + \beta \Delta R_i + W_i^{*'} \nu + \Delta \tilde{u}_i$, pooled and separately by gender. For 2SLS, the second stage regressions are reported where ΔR_i is instrumented by the distance from early and normal retirement age, and W include years elapsed, and country dummies. For OLS, W additionally include age at first wave.

2.6 Concluding remarks

Several recent works investigate the effect of retirement on cognitive performance, arriving to very different conclusions. To better understand the mechanisms that cause the excursive outcomes, I replicated the methods of the literature using data from the first three general waves of Survey of Health, Ageing and Retirement in Europe (SHARE). This exercise showed that identifying a clear causal effect is far from straightforward. Ideally, we would need to compare individuals from the same cohort and country, with the same age and education, one of them randomly assigned to be retired for a period of time while the other being working. Such a comparison is clearly impossible. I traced back the major difference in the estimations to omitted variable bias: the large estimated retirement effect can be reduced a lot controlling for relevant variables such as gender and country. However, even after the inclusion of controls, the estimate is likely to be biased.

I came up with a novel identification strategy to solve the issues. I use the longitudinal feature of the data and compare the cognitive paths of employed and retired individuals. My estimated

effect is even smaller than the smallest estimate in the literature, amounting to at most 0.05 standard deviation point decline yearly (that is still likely to be overestimated).

I reckon that retirement in itself has essentially no effect. What matters is the accompanying change in the lifestyle – that could be very heterogeneous across agents: some may retire from a stressful job to get more time to relax and care for grandchildren, others may lose the meaning of their lives. There are some works in the psychology literature that point to this direction. [Stine-Morrow \(2007\)](#) investigates the so called Dumbledore hypothesis, that individual choice plays a crucial role in old age cognitive decline. "it is our choices... that show what we truly are, far more than our abilities" ([Rowling, 1999](#)). [Aichberger et al. \(2010\)](#) find that physical inactivity goes together with worse cognitive performance. [Hertzog et al. \(2009\)](#) summarize previous results about what older adults can do to preserve their functional capacity. Future empirical work in this topic should build in these considerations and try to uncover the heterogeneous effect of retirement on cognitive performance, differentiating by the nature of the retirement, that is how it affected the individual's lifestyle.

Apart from the actual question I see my paper contributing to the literature in a more general way as well: it shows that replicating previous works and synthesizing their findings can lead to better understanding of the problem. Amidst of the replication crisis ([Ioannidis, 2018](#)) I consider this as a valuable lesson and hope to see more efforts towards this direction in the future.

Chapter 3

Do Elite Schools Benefit the Average Student

joint with Sándor Sóvágó

3.1 Introduction

Most school districts feature elite schools in which admission is merit-based (i.e., depends on a qualifying priority score) and the student body consists of affluent, high-achieving students. Admission to these elite schools is highly selective and preparation for the admission requires substantial pecuniary and non-pecuniary resources from the parents of prospective students. Therefore, it is important for these parents to understand the extent to which their child may benefit from enrollment in an elite school.

A large literature studies the effectiveness of elite schools (for a recent summary, see [Beuermann and Jackson, Forthcoming](#)). This literature exploits discontinuities in admission chances that are created by the merit-based priority-score cutoffs. Thereby, these studies provide estimates for students who are on the margin of admission. However, if the effect of elite schools is heterogeneous, existing estimates may not be informative for parents, who would be interested in the effect for the average elite-school student with similar characteristics to theirs.

This paper studies the effects of elite schools on academic achievement in Hungary. Merit-based admission combined with high demand implies that elite-school students have high-achieving peers. Elite schools offer a more advanced, higher-paced curriculum, and teachers of high-qualification. Moreover, elite-school enrollment entails early switching to a secondary grammar school. Using administrative data, our study examines how enrollment in an elite school

affects elite-school students' short- and medium-run academic achievement. Motivated by the potential heterogeneity in the effectiveness of elite schools, we conduct our analysis by gender and baseline ability (Oosterbeek et al., 2020).

Our main empirical strategy identifies the effects of enrollment in an elite school using non-parametric bounds. Our non-parametric bounds approach builds on a weak stochastic dominance assumption. We assume that conditional on observable student characteristics more able students are more likely to enroll in an elite school (conditional Monotone Treatment Selection, MTS; see Manski and Pepper, 2000; de Haan, 2011). The conditional MTS assumption yields an upper bound on the average treatment effect on the treated (ATET). We estimate the effect of elite-school enrollment throughout the outcome distribution, thus, we can study whether the effects differ between the top and the bottom end of the outcome distribution (de Haan and Leuven, 2020).

The non-parametric bounds approach offers several advantages. First, our approach identifies the effect of elite-school enrollment for elite-school students, and not only for students on the margin of admission. Second, the conditional MTS assumption is consistent with elite schools' admission policy and yields testable implications.

We find that enrollment in an elite school has a negative effect on female and low-ability elite-school students' mathematics test scores two years after enrollment. Elite-school enrollment reduces the probability of scoring above the median in mathematics by more than 7.5 percentage points, corresponding to 30 percent, for low-ability female students, and by more than 1 percentage points, corresponding to 10 percent, for low-ability male students. For high-ability female students, we find that elite-school enrollment reduces the probability of scoring in the top 10 percent in mathematics by more than 2 percentage points, corresponding to 10 percent. We argue that our findings for female and low-ability students are not the consequence of grading policies (e.g., ceiling effects, grading on a curve), but they reflect negative effects on skill formation. By contrast, we cannot rule out positive effects for high-ability male students at the upper half of the outcome distribution.

We next investigate whether the negative short-run effects persist four years after enrollment. We find that the upper bounds are positive and relatively large in magnitude for each ability and gender group. Thus, our non-parametric bounds strategy is uninformative about the sign of the effect on the medium run.

We further examine the medium-run effects of elite-school enrollment using a complementary empirical strategy. Using the selection on observables assumption, we estimate school value-added models to identify the average treatment effect on the treated. Our school value-added models control for students' lagged test scores and well as a rich set of student and school characteristics. We find that elite-school enrollment has a positive effect elite-school students' mathematics and reading test scores on the medium run. Even if the school value-added and

non-parametric bounds empirical strategies build on different identifying assumptions, reassuringly, the school value-added estimates are always consistent with our non-parametric bounds on the effect of elite schools.

Our paper contributes to the understanding of the effectiveness of elite schools in a number of ways. First, numerous studies examine the effects of elite schools (or attending a better school) using regression discontinuity design (RDD) in settings where admission is merit-based (Jackson, 2010; Clark, 2010; Pop-Eleches and Urquiola, 2013; Bui et al., 2014; Lucas and Mbiti, 2014; Abdulkadiroglu et al., 2014; Dobbie and Fryer, 2014; Anderson et al., 2016; Barrow et al., Forthcoming).¹ These RDD estimates are informative for students who are on the margin of admission. Instead, our study focuses on the effect for the average student, and thus complements this literature. Moreover, these RDD studies have limited ability to identify heterogeneous effects by student ability, since student ability varies little at the margin of admission (Oosterbeek et al., 2020). By contrast, our identification strategy allows us to estimate the effect for different ability groups. When we split the sample by baseline ability, our estimates suggest that the benefits of elite schools are concentrated on high-ability students. When looking at the outcome distribution, our estimates indicate that the benefits of elite schools materialize at the top of the outcome distribution.

Second, our paper studies the effect of elite-school enrollment on outcomes that are measured in different points in time. Documenting how the effects of elite school enrollment change over time is important, since behavioral responses of students, teachers, or parents may materialize on different time horizons (Pop-Eleches and Urquiola, 2013). We show that elite-school enrollment has a negative effect on female and low-achieving students' mathematics score on the short run. This finding suggests that it more costly for certain groups of students to adjust to the elite-school environment.

Finally, we also contribute to the studies evaluating the performance of Hungarian elite schools. In the same context as ours, (Horn, 2013) finds positive, albeit, imprecisely estimated effects of selective secondary-school attendance on short-run academic achievement. Using less stringent identifying assumptions, our study considerably narrows the upper bound on these estimates.

The remainder of the paper is organized as follows. Section 3.2 provides background information about Hungarian education and the organization of elite schools. Section 3.3 describes our data and provides summary statistics. Section 3.4 discusses our empirical strategies and presents the results of our validity checks. Section 3.5 presents our results. Section 3.6 concludes.

¹A very much related literature studies elite-school effectiveness (or the effects of attending a better school) in settings where a lottery-based admission system is in place (e.g., Cullen et al., 2006; Deming, 2011; Dobbie and Fryer, 2011; Oosterbeek et al., 2020). Our study is similar to these papers in a sense that it identifies the effect of elite-school enrollment away from the priority-score cutoff. A key difference is that admission is merit-based in our setting, which implies that elite-school students are more selected than those who are admitted in a lottery-based admission system.

3.2 Context: Elite schools in Hungary

This section describes the institutional details of elite schools in Hungary. We describe admission to elite schools and detail the treatment.

Students begin primary education at age 6 in Hungary. After eight years in primary schools, they transition to secondary education. Secondary education is tracked in Hungary: students may choose a secondary grammar school, which has a more academic orientation and prepares students for higher education, or students may choose a vocational school, which has a less academic orientation and gives a vocational degree (see Table 3.1). Primary and secondary education are organized together in some schools. That is, in these comprehensive schools, students do not switch school when they transition from primary to secondary education. Even in these comprehensive schools, students are required to do the admission exam to proceed on the secondary track.

Primary-school students, at the end of grade 6, may decide to enroll in the 6-years long academic track of a secondary grammar school. We label these 6-years long academic tracks as elite schools. These elite schools are typically separate classes in a secondary grammar school, which have regular tracks as well. Thus, an elite school is essentially an “elite track” in a secondary grammar school (cf. Pop-Eleches and Urquiola, 2013).²

Table 3.1: The overview of the education system in Hungary

Grade	1–4	5–6	7–8	9–12
Age	6–9	10–11	12–13	14–17
Path	Primary school	Regular track in a secondary grammar school		
		Vocational school		
		6-years long academic track in a secondary grammar school (elite school)		
		8-years long academic track in a secondary grammar school (excluded)		

Notes: The table provides an overview of the education system in Hungary. The regular and the 6-years long academic tracks take place in secondary grammar schools. We refer to the 6-years long academic track in a secondary grammar school as an elite school. We exclude the 8-years long academic track from our analysis.

Admission to elite schools is merit-based. Elite schools organize admission exams where students have to solve numeracy and literacy tests. While all applicants solve the same test in the country, the priority-score formula used for the actual rankings is school-specific: it is a combination of the students’ primary-school grades, the result of their written admission test, and the result of their oral admission exam. Since elite schools are highly oversubscribed, they are highly selective, and only high-achieving students are admitted.

Enrollment in an elite school is a composite treatment. First, elite-school students have academically stronger peers, which follows from the combination of merit-based admission and the

²Some of the secondary grammar schools offer a 8-years long academic track. Due to data limitations we do not study these 8-years long academic tracks, and drop students who enrolled in these 8-years long academic tracks from all of our samples.

high demand. Second, elite-school students enter the secondary education environment sooner than non-elite-school students. Teachers in secondary schools are typically more qualified than primary school teachers, and they are more experienced with more mature students. Third, elite schools offer a more advanced, higher-paced curriculum. Finally, elite-school students switch school at age 12. By contrast, students who do not enroll in an elite school attend their primary school for an additional 2 years, and switch only at age 14.

3.3 Data and summary statistics

This section describes the data we use to estimate the effect of enrollment in an elite school on elite-school students' academic achievement. We begin, in Section 3.3.1, by describing the data and discussing the construction of our samples. In Section 3.3.2, we present summary statistics showing that elite-school students are positively selected based on socioeconomic status and academic achievement, and that the peer quality of elite-school students is better than those of non-elite-school students.

3.3.1 Data

For the analysis, we use administrative data from the National Assessment of Basic Competencies (NABC). Our data are longitudinal and cover every student in grades 6, 8, and 10 in the period of 2008–2014.

The backbone of our administrative data are the standardized test scores of the NABC. The NABC is similar to OECD's Programme for International Student Assessment (PISA), but it extends to every student in grades 6, 8, and 10. The NABC (and the corresponding survey on students' background) is conducted annually in the last week of May. The NABC measures students' mathematics and reading skills. Students' 6th-grade test scores are measured on a scale, which has a mean of 1,500 and standard deviation of 200. In grades 8 and 10, the test scores are standardized in a way, such that the test scores are comparable over time (e.g., a student's 6th-grade and 8th-grade test scores are comparable) and across cohorts (e.g., the average test scores in grade 8 are comparable across cohorts). The standardized test scores differ from students' school-grades in many aspects: the standardized test scores are of low-stakes, they are graded blindly and externally, and they are not top-coded.³

Our data also include information on students' demographics (gender), socioeconomic status (the number of books at home, disadvantaged status, and parental education), and schools

³Figure 3.3 presents the cumulative distribution function of students' 6th-grade mathematics and reading test scores.

(school identifiers in grades 6, 8, and 10, school type, the county of the school, and the type of the settlement where the school is located). We also have rich information on students' academic achievement. Our data also include students' GPA in grade 5, i.e., one year prior the students' first NABC taking place. Finally, our data have information on students' 8th-grade mathematics grade.⁴ These grades are given by students' own teachers and they are measured on a scale of 1 to 5.

We study two (overlapping) samples in order to maximize the power of our analysis. When studying short-run outcomes (i.e., outcomes measured 2 years after elite-school enrollment), we focus on students who we observe in grades 6 and 8. We refer to this sample as the 8th-grade sample. The 8th-grade sample consists of five cohorts, whose 8th-grade outcomes are measured in the period of 2010–2014. When studying medium-run outcomes (i.e., outcomes measured 4 years after elite-school enrollment), we focus on students who we observe in grades 6, 8, and 10. We refer to this sample as the 10th-grade sample. The 10th-grade sample consists of three cohorts, whose 10th-grade outcomes are measured in the period of 2012–2014.

In our analysis we make a number of sample restrictions to focus on students for whom enrolling in an elite school is a viable option. First, we focus on students' with a complete academic path with no missing information. Second, our samples include students who are enrolled either in an elite or in a primary school in grade 8. Third, we focus on students whose propensity to enroll in an elite school is non-negligible. Therefore, our samples only include students (1) whose 5th-grade GPA is at least 4, (2) whose 6th-grade mathematics grade is at least 3, and (3) who attend a primary school in grade 6 from which at least one student enrolled in an elite school in our sample period. Appendix Table C1 provides information on the sample size reduction for each sample restriction. After the sample restrictions, we end up with about 25,000 students in each cohort (approximately 25% in each cohort).

In our analysis, we standardize students' test scores on our restricted samples. We conduct our analysis for four student groups, defined by (baseline) ability and gender, separately. We refer to students whose 6th-grade standardized test score is below the sample median as low-ability students, and whose 6th-grade standardized test score is above the sample median as high-ability students.

3.3.2 Summary statistics

Table 3.2 presents summary statistics of student characteristics by elite-school enrollment for each of our samples. Panel A focuses on students' pre-treatment characteristics. About 55 percent of the students are female. Approximately 37 percent of the students has maximum 150

⁴Information on students' socioeconomic status (number of books at home and parental education) and their school grades (5th-grade GPA and 8th-grade mathematics grade) is self-reported.

books at home, 37 percent has between 150 and 600 books at home, and 27 percent has more than 600 books at home. On average, students' 5th-grade GPA is 4.6, on a scale of 1 to 5.⁵ The composition of the 8th-grade and 10th-grade samples are almost identical.

Out of the 126,196 students in the 8th-grade sample, 16,702 students (13.2 percent) enrolled in an elite school. Students who enrolled in an elite school have higher socioeconomic status: about 45 percent of elite-school students has more than 600 books at home compared to 24 percent of those who did not enroll in an elite school. Elite-school students' 6th-grade mathematics (reading) test score is higher than the sample average by 53 (50) percent of a standard deviation. By contrast, students who did not enroll in an elite school have a 6th-grade test score that is by 6 percent of a standard deviation lower than the sample average in both subjects. These patterns indicate that elite-school students are positively selected based on their socioeconomic status and (baseline) academic achievement.

Panel B of Table 3.2 reports summary statistics on students' outcomes. On average, students' 8th-grade mathematics grade is 4.0. Elite-school and non-elite-school students' average 8th-grade mathematics grades are similar. By contrast, elite-school students' 8th-grade standardized test scores are considerably higher than those of non-elite-school students. For example, elite-school students' average 8th-grade mathematics test score is 36 percent of a standard deviation higher the sample average. Non-elite-school students' average 8th-grade mathematics test score is 6 percent of a standard deviation lower than the sample average. Summary statistics of the 10th-grade sample indicate that elite-school students' average 10th-grade mathematics test score is 58 percent of a standard deviation higher the sample average. By contrast, non-elite-school students' average 10th-grade mathematics test score is 8 percent of a standard deviation lower than the sample average. These patterns indicate that the differences in standardized scores between elite-school and non-elite-school students persist up until 4 years after elite-school enrollment.

⁵The relatively high average 5th-grade GPA follows from the fact that our samples include students whose 5th-grade GPA is at least 4; see Section 3.3.1.

Table 3.2: Summary statistics

	8th-grade sample			10th-grade sample		
	Elite-school students (1)	Non-elite-school students (2)	Total (3)	Elite-school students (4)	Non-elite-school students (5)	Total (6)
<i>A. Pre-treatment characteristics</i>						
Female	0.54 (0.50)	0.56 (0.50)	0.55 (0.50)	0.54 (0.50)	0.56 (0.50)	0.55 (0.50)
Max. 150 books at home	0.19 (0.39)	0.39 (0.49)	0.37 (0.48)	0.17 (0.37)	0.36 (0.48)	0.33 (0.47)
B/w 150 and 600 books at home	0.36 (0.48)	0.37 (0.48)	0.37 (0.48)	0.34 (0.47)	0.37 (0.48)	0.37 (0.48)
More than 600 books at home	0.45 (0.50)	0.24 (0.43)	0.27 (0.44)	0.49 (0.50)	0.27 (0.45)	0.30 (0.46)
5th-grade GPA (1–5)	4.77 (0.28)	4.55 (0.34)	4.58 (0.34)	4.78 (0.28)	4.57 (0.34)	4.59 (0.34)
6th-grade mathematics test score (std.)	0.53 (1.01)	-0.08 (0.97)	0.00 (1.00)	0.56 (1.02)	-0.08 (0.97)	0.00 (1.00)
6th-grade reading test score (std.)	0.50 (0.96)	-0.08 (0.98)	0.00 (1.00)	0.51 (0.96)	-0.07 (0.99)	-0.00 (1.00)
<i>B. Outcomes</i>						
8th-grade mathematics grade (1–5)	4.04 (0.87)	3.98 (0.88)	3.99 (0.88)	· (·)	· (·)	· (·)
8th-grade mathematics test score (std.)	0.36 (1.01)	-0.06 (0.99)	0.00 (1.00)	· (·)	· (·)	· (·)
8th-grade reading test score (std.)	0.42 (0.95)	-0.06 (0.99)	0.00 (1.00)	· (·)	· (·)	· (·)
10th-grade mathematics test score (std.)	· (·)	· (·)	· (·)	0.58 (0.98)	-0.08 (0.98)	0.00 (1.00)
10th-grade reading test score (std.)	· (·)	· (·)	· (·)	0.53 (0.91)	-0.07 (0.99)	-0.00 (1.00)
Number of students	16,702	109,494	126,196	8,850	63,112	71,962

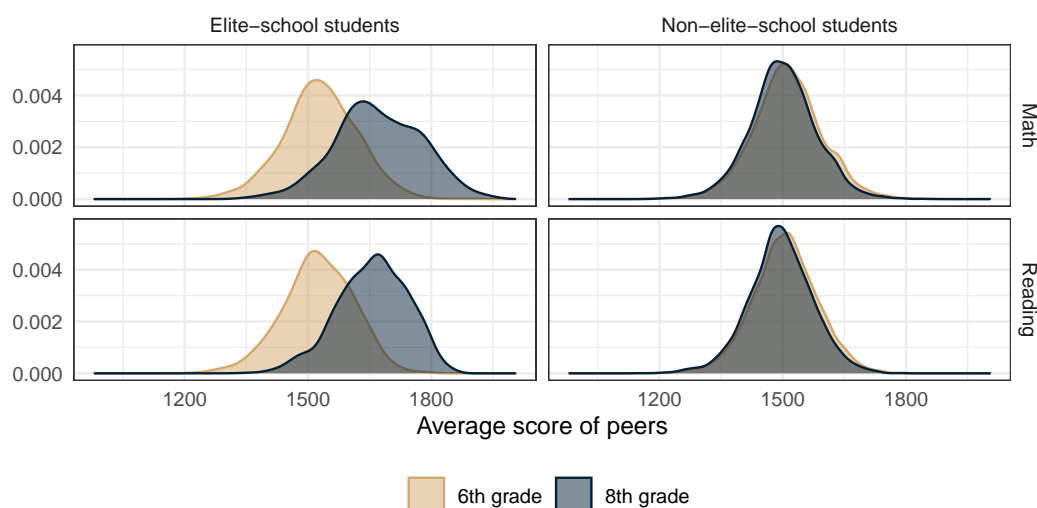
Notes: The table presents the means and standard deviations of student characteristics for each sample. Columns (1) and (4) focus on students who did not enroll in an elite school, columns (2) and (5) focus on students who enrolled in an elite school, and columns (3) and (6) focus on the entire sample.

Figure 3.1 presents the distribution of students' peer quality, which is the leave-out mean of peers' 6th-grade standardized test scores within a school. The figure displays the distribution of peer quality before and after elite-school enrollment, i.e., in grades 6 and 8. We present these distributions for elite-school and non-elite-school students separately, and for students' 6th-grade mathematics and reading test scores.⁶

The left panels of Figure 3.1 show that the peer quality distribution of elite-school students shifts to the right between grades 6 and 8. The average 6th-grade mathematics test score of elite-school students' peers in grade 6 is 1,530. By contrast, the average 6th-grade mathematics test score of elite-school students' peers in the elite school is 1,671. Thus, elite-school students' peer quality substantially improves after they enroll in an elite school.

The right panels of Figure 3.1 show that the peer quality distribution of non-elite-school students largely remain unchanged between grades 6 and 8. The average 6th-grade mathematics test score of non-elite-school students' peers in grade 6 is 1,512 and it is 1,501 in grade 8. This small reduction in the peer quality of non-elite school students is consistent with positive selection into elite schools.⁷ We also note that elite-school students' peer quality considerably exceeds the peer quality of non-elite-school students in grade 8. Thus, enrollment in an elite school entails having academically stronger peers.

Figure 3.1: Peer quality and elite-school enrollment



Notes: The figure shows the distributions of students' peer quality in grades 6 and 8. Peer quality is the leave-out mean of peers' 6th-grade standardized test scores (within school). The left (right) panels focus on elite-school (non-elite-school) students. The top (bottom) panels focus on peers' 6th-grade mathematics (reading) test scores.

⁶When we compute students' peer quality, we include all students in the calculations. Therefore, in Figure 3.1, we preserve the original scale of students' 6th-grade standardized test scores; see Section 3.3.1.

⁷In the same context, (Schiltz et al., 2019) show that the departure of smart peers to elite schools has a small, negative effect on the academic achievement of students who are left behind.

3.4 Empirical strategies

This section discusses our empirical strategies to identify the effect of enrollment in an elite school on the outcome distribution. We begin, in Section 3.4.1, by deriving a non-parametric bound on the effect of enrollment in an elite school using a weak stochastic dominance assumption (conditional Monotone Treatment Selection). In Section 3.4.2, we present evidence supporting the validity of our identifying assumption. Finally, in Section 3.4.3, we discuss a complementary empirical strategy that builds on the selection on observables assumption.

3.4.1 Non-parametric bounds: conditional MTS throughout the outcome distribution

We are interested in the effect of enrollment in an elite school on elite-school students' outcomes, that is, we focus on the the average treatment effect on the treated (ATET). We study the entire distribution of potential outcomes. Thus, the causal effect of interest, denoted by $\tau(\gamma)$, is the effect of elite-school enrollment on the probability of obtaining an outcome greater than γ for students who enrolled in an elite school:

$$\tau(\gamma) = \mathbb{P}[Y(1) \geq \gamma | D = 1] - \mathbb{P}[Y(0) \geq \gamma | D = 1],$$

where we denote student i 's potential outcome by $Y_i(D)$ and D takes a value of one if the student enrolled in an elite school and zero otherwise.

The causal effect of interest focuses on students who enrolled in an elite school. This means that we identify the effect not only for students on the margin of admission, as studies that exploit priority-score cutoffs, but also for inframarginal students. Thus, we address whether elite schools academically benefit students who enrolled in them, and not whether the expansion of elite-school capacities benefits marginally admitted students.

The identification problem is that the potential untreated outcome distribution of students who enrolled in an elite school is unobservable. Therefore, we make an assumption on the selection of students into elite schools. Specifically, we impose a weak stochastic dominance condition on the potential untreated outcome distribution of students who enrolled in an elite school.

Assumption 1 (Monotone Treatment Selection, MTS). The distribution of potential untreated outcomes of students who enrolled in an elite school weakly dominates that of those who did not enroll in an elite school:⁸

$$\mathbb{P}[Y(0) \geq \gamma | D = 1] \geq \mathbb{P}[Y(0) \geq \gamma | D = 0], \quad \forall \gamma. \quad (\text{MTS})$$

⁸Our MTS assumption requires stochastic dominance of the potential outcome distributions, and thus is stronger than the MTS assumption originally proposed by (Manski and Pepper, 2000).

The MTS assumption implies that the effect of enrollment in an elite school does not exceed the difference between the (observed) outcome distributions of students who enrolled in an elite school and of those who did not enroll:

$$\begin{aligned}\tau(\gamma) &= \mathbb{P}[Y(1) \geq \gamma | D = 1] - \mathbb{P}[Y(0) \geq \gamma | D = 1] \\ &\leq \mathbb{P}[Y(1) \geq \gamma | D = 1] - \mathbb{P}[Y(0) \geq \gamma | D = 0].\end{aligned}$$

We further sharpen the upper bound on the effect of enrollment in an elite school by assuming that the MTS assumption holds for certain subgroups of students.

Assumption 2 (conditional MTS). The distribution of potential untreated outcomes of students who enrolled in an elite school weakly dominates that of those who did not enroll in an elite school conditional on each values of the variable of X :

$$\mathbb{P}[Y(0) \geq \gamma | D = 1, X] \geq \mathbb{P}[Y(0) \geq \gamma | D = 0, X], \quad \forall \gamma. \quad (\text{conditional MTS})$$

In the analysis, for a given value of γ , we use the conditional MTS assumption to derive bounds on $\mathbb{P}[Y(0) \leq \gamma | D = 1, X]$ for each values of X . We then combine these bounds to obtain an upper bound on the causal effect of interest:⁹

$$\tau(\gamma) \leq \sum_{x \in X} (\mathbb{P}[Y(0) \leq \gamma | D = 0, X = x] - \mathbb{P}[Y(1) \leq \gamma | D = 1, X = x]) \mathbb{P}[D = 1 | X = x] \quad \forall \gamma.$$

The upper bound on the effect of enrollment in an elite school equals to the corresponding exact matching estimator (Rubin, 1973). The exact matching estimator identifies the effect of enrollment in an elite school on the treated under the Conditional Independence Assumption (CIA). Instead of maintaining the CIA, we assume that selection into an elite school is positive conditional on X , and thus we identify an upper bound on the causal effect of interest.

We estimate the non-parametric bound on $\tau(\gamma)$ by the corresponding sample means and empirical probabilities. The 95% confidence intervals on the causal effect of interest are based on 1,000 bootstrap replications using the methodology derived by Imbens and Manski (2004).

We combine two pre-treatment variables, students' 5th-grade GPA and the number of books at home, to sharpen the bounds on the effect of enrollment in an elite school. We assume that conditional on students' 5th-grade GPA and the number of books at home the potential untreated outcome distribution of students who enrolled in an elite school weakly dominates of those who did not enroll in an elite school. Since admission to elite schools is merit-based, elite-school

⁹Since we investigate the effect of enrollment in an elite school on the entire outcome distribution, we could derive a *no-assumption* lower bound on $\tau(\gamma)$. This lower bound is, by construction, never positive, thus we do not report it.

students are positively selected. Consistent with the merit-based admission procedure, we assume that positive selection is present even conditional on students' 5th-grade GPA, which is a pre-treatment proxy of students' academic ability, and the number of books at home, which is a proxy of socioeconomic status. The next section presents evidence using pre-treatment outcomes to support the validity of our identifying assumptions.¹⁰

3.4.2 Validity check

We next assess the validity of the conditional MTS assumption for the interaction of students' 5th-grade GPA and the number of books at home. A prerequisite of the conditional MTS assumption is a sufficient overlap between elite-school and non-elite-school students for each value of the interaction of students' 5th-grade GPA and the number of books at home (cf. Common Support Assumption). Moreover, if the conditional MTS assumption is met, then the distribution of elite-school students' pre-treatment outcomes should weakly dominate those of non-elite-school students' pre-treatment outcomes. We use students' 6th-grade standardized test score, which is realized prior to elite-school enrollment, to test this implication.

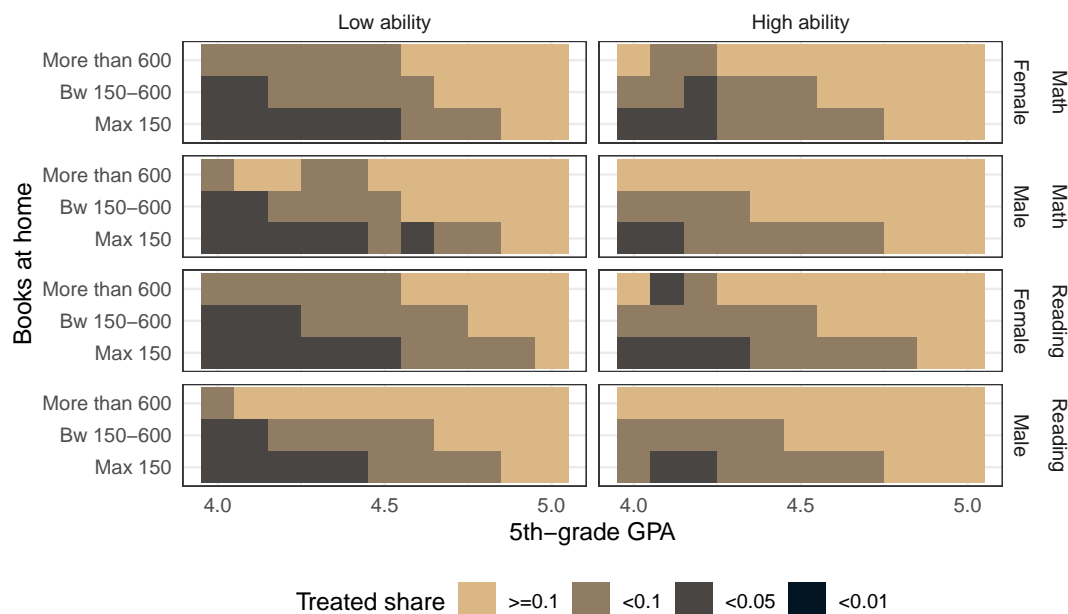
We first provide evidence supporting the overlap between elite-school and non-elite-school students' characteristics. Figure 3.2 presents the share of students who enrolled in an elite school for each value of the combination of 5th-grade GPA and the number of books at home. The figure splits the sample by (pre-treatment) ability and gender. There is no combination of 5th-grade GPA and the number of books at home such that the share of elite-school students is below 1 percent.¹¹

We next provide graphical evidence supporting the MTS assumption. Figure 3.3 displays the distribution of students' 6th-grade standardized test scores by elite-school enrollment and gender. The figure shows that the 6th-grade test score distribution of students who enrolled in an elite school weakly dominates of those who did not enroll for each gender and test type (mathematics and reading).

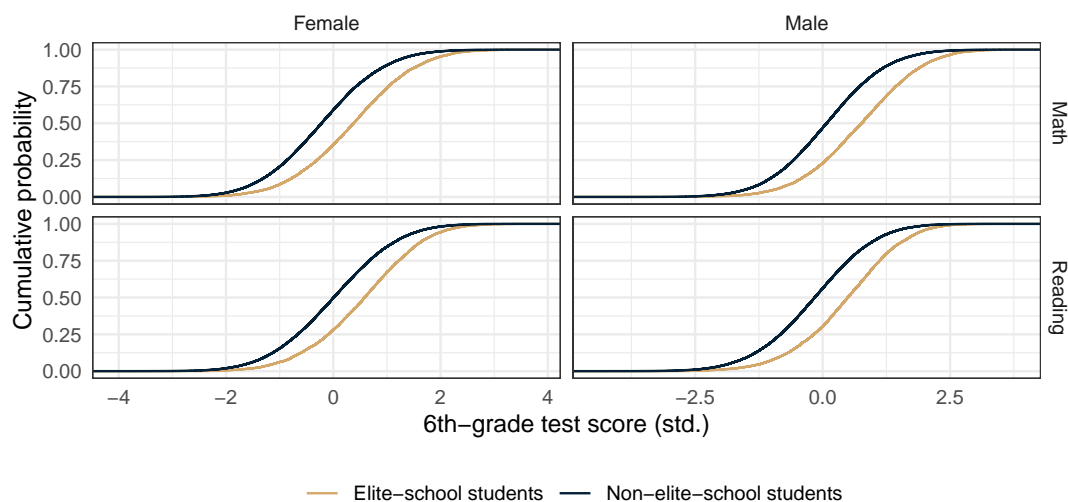
Figure 3.4 presents a formal test of the validity of the conditional MTS assumption. For each value of the combination of students' 5th-grade GPA and the number of books at home, we perform a one-sided Kolmogorov-Smirnov test. Figure 3.4 presents the corresponding *p*-values of

¹⁰Students' 5th-grade GPA and the number of books at home are likely to be positively related to students' potential outcome distribution, and thus are valid monotone instrumental variables (MIVs) (Manski and Pepper, 2000; de Haan, 2011). We find empirical support for the combination of students' 5th-grade GPA and the number of books at home being valid MIVs. However, consistent with the argument of Richey (2016), the combination of the conditional MTS and MIV assumptions does not tighten our bounds.

¹¹Appendix Figure C.1 provides the same information for the 10th-grade sample, with the same conclusion.

Figure 3.2: Validity check: Elite-school enrollment and student characteristics

Notes: The figure presents the share of students who enrolled in an elite school by student characteristics. Each cell shows the share of elite-school students for a combination of 5th-grade GPA and the number of books at home. In the top panels (bottom) high-ability/low-ability is defined as having 6th-grade mathematics (reading) test score above/below the median. Sample: 8th-grade sample, N = 126,196.

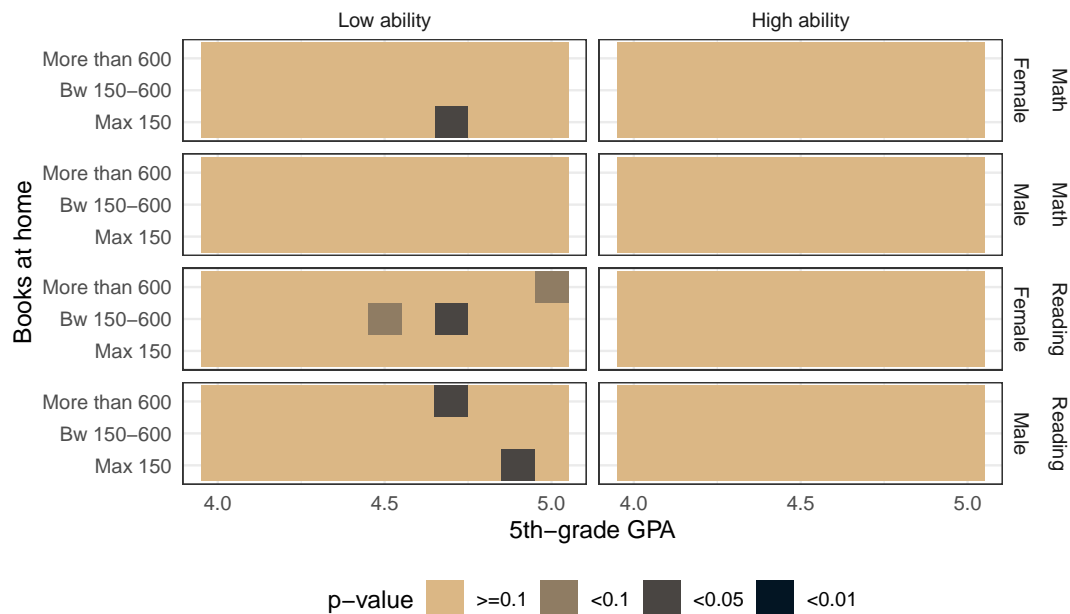
Figure 3.3: Validity-check: The distribution of students' 6th-grade standardized test scores by elite-school enrollment

Notes: The figure displays the cumulative distribution function of students' 6th-grade standardized test score by elite-school enrollment and gender. The left (right) panels present the distributions for female (male) students. The top (bottom) panels present the distribution of the 6th-grade mathematics (reading) test scores.

the Kolmogorov-Smirnov test by gender and pre-treatment ability for each value of the combination of 5th-grade GPA and number of books at home. Out of the 264 tests we perform, we

reject the null hypothesis only 6 times at a 10 percent significance level. We interpret these results as strong evidence supporting the validity of the conditional MTS assumption.¹²

Figure 3.4: Validity-check: The p-values of the Kolgomorov-Smirnov test



Notes: The figure displays the p-values of the one-sided Kolgomorov-Smirnov test. The Kolgomorov-Smirnov test tests the equality of the distributions of elite-school and non-elite-school students' 6th-grade standardized test scores. Each cell shows the p-value for a combination of 5th-grade GPA and the number of books at home. In the top panels (bottom) high-ability/low-ability is defined as having 6th-grade mathematics (reading) test score above/below the median. Sample: 8th-grade sample, N = 126,196.

3.4.3 School value-added

The non-parametric bound approach offers several advantages (e.g., mild identifying assumptions with testable implications), however, its ability to recover informative estimates may be limited. Therefore, we consider a complementary empirical strategy, which builds on selection on observables—a more demanding assumption.

Assumption 3 (Selection on observables).

$$\mathbb{P}[Y(d) \geq \gamma | D = d, Z] = \alpha_d^\gamma + \beta_d^\gamma Z, \quad d = 0, 1; \forall \gamma,$$

where Z is a vector of variables.

¹²Appendix Figure C.2 provides evidence supporting the validity of the conditional MTS assumption for the 10th-grade sample.

The selection on observables assumption provides point identification for the causal effect of interest:

$$\tau(\gamma) = \sum_{z \in Z} (\alpha_1^\gamma - \alpha_0^\gamma + (\beta_1^\gamma - \beta_0^\gamma)z) \mathbb{P}[D = 1|Z = z].$$

The selection on observables assumption ensures that an ordinary least squares (OLS) regression of Y_i on an indicator of elite-school enrollment interacted with Z_i recovers unbiased estimates of α_d^γ and β_d^γ . We consider two specifications. First, the vector Z includes students' 6th-grade standardized test scores and cohort fixed effects (simple model). Second, the vector Z includes additional covariates, such as 5th-grade GPA, number of books at home, education of the mother, education of the father, being disadvantaged, the county of the school, and the type of the residence where the school is located (full model).¹³ Both specifications resemble the commonly used "school value-added" approach to evaluate the effectiveness of schools (Koedel et al., 2015).

The plausibility of the selection on observables assumption is debated in the context of school value-added (school VA) estimates (see e.g., Chetty et al., 2014a,b; Guarino et al., 2015; Rothstein, 2010, 2017). For elite schools in Amsterdam, (Oosterbeek et al., 2020) finds that school value-added estimates are severely biased when they are compared to admission lottery-based estimates. In the context of Hungarian elite schools, we view the school value-added estimates complementary to our non-parametric bound approach. We note that when the selection on observables assumption is violated, the direction of the bias of the school value-added estimate is unclear. Therefore, the school value-added estimates might be less informative about our causal effect of interest than the non-parametric bounds.

3.5 Results

This section presents our results. We begin, in Section 3.5.1, by showing that enrollment in an elite school has a negative effect on the short-run academic achievement of female and low-ability students. In Section 3.5.2, we show that our non-parametric bounds strategy cannot rule out moderately positive effects on academic achievement 4 years after enrollment. Finally, in Section 3.5.3, using value-added models, we show that elite schools improve students' academic achievement 4 years after enrollment.

¹³Appendix C.3 describes the construction of our variables, and Appendix Table C.1 presents summary statistics of the additional covariates of the school value-added specifications.

3.5.1 Short-run academic achievement

We begin by studying the effect of enrollment in an elite school on elite-school students' 8th-grade mathematics grade. Figure 3.5 displays the upper bounds on the causal effect of interest, split by ability and gender.¹⁴ The figure also displays the raw (unconditional) differences between the outcomes of elite-school and non-elite school students for each group.

Figure 3.5 shows that the upper bounds are negative throughout the outcome distribution for each gender–ability group. For example, enrollment in an elite school decreases the probability of having an 8th-grade mathematics grade larger than 4.5 by at least 15 percentage points for high-ability male students. The upper bound of 15 percentage points corresponds to a 21 percent relative decrease (Appendix Figure B1). For high-ability students, the upper bounds are the lowest at the top of the distribution. The upper bounds of the causal effect are considerably lower than the unconditional differences, which is consistent with positive selection into elite schools.

The negative upper-bound estimates on students' 8th-grade mathematics grade do not necessarily reflect negative effects on students' skill formation. Alternative explanations are ceiling effects, grading on a curve (Calsamiglia and Loviglio, 2019), or more demanding study requirements in elite schools. To rule out these alternative explanations, we next examine the effect of elite schools on students' 8th-grade standardized test scores. This test score is not top-coded, blindly graded, and standardized nationwide, therefore, it is not susceptible to the above mentioned alternative explanations.

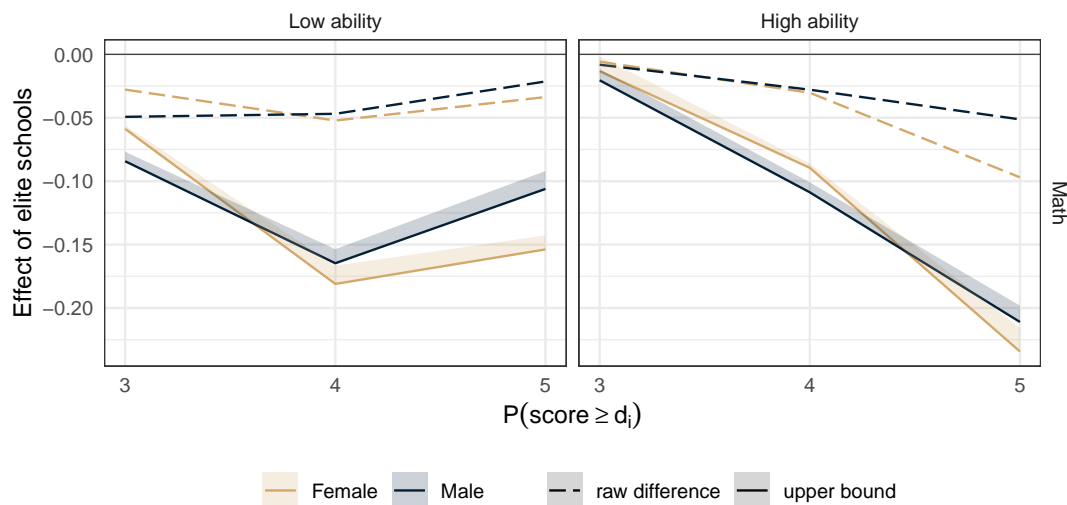
Figure 3.6 presents the upper bounds on the effect of enrollment in an elite school on the distribution of 8th-grade standardized test scores. The figure displays the upper bounds on the deciles of the distribution of students' mathematics and reading test scores.¹⁵

The figure shows that enrollment in an elite school has a negative effect on female and low-ability elite-school students' mathematics test scores two years after enrollment. Elite-school enrollment reduces the probability of scoring above the median in mathematics by more than 7.5 percentage points, corresponding to 30 percent, for low-ability female students, and by more than 1 percentage points, corresponding to 10 percent, for low-ability male students. For high-ability female students, we find that elite-school enrollment reduces the probability of scoring in the top 10 percent in mathematics by more than 2 percentage points, corresponding to 10 percent (Appendix Figure B2). The figure also shows that enrollment in an elite school does not meaningfully increase the 8th-grade mathematics test scores of low-ability male students. The

¹⁴The figure also displays the 95% confidence intervals on the causal effect of interest. Since we present the upper bound estimates of the causal effect exclusively, we mark the area between the upper confidence band and the estimate itself (shaded area).

¹⁵To simplify the exposition, Figure 3.6 displays the bounds only on the bottom- and top-6 deciles of the outcome distribution of low- and high-ability students, respectively.

Figure 3.5: The effect of elite-school enrollment on the distribution of elite-school students' 8th-grade mathematics grade



Notes: The figure presents our upper-bound estimates of the effect of elite-school enrollment on the distribution of elite-school students' 8th-grade mathematics grade (solid lines). The dashed line denotes the raw difference between the outcomes of elite-school and non-elite-school students. We report the estimates separately by gender and low- and high-ability students. Low- and high-ability students are defined by whether the students' 6th-grade mathematics test score is below or above the median. Students' 6th-grade mathematics grade is measured on the scale of 1–5. The shaded area represents the area between the upper confidence band (95%) and the upper bound estimate itself. The 95% confidence intervals are based on 1,000 bootstrap draws. Sample: 8th-grade sample, $N = 126,196$.

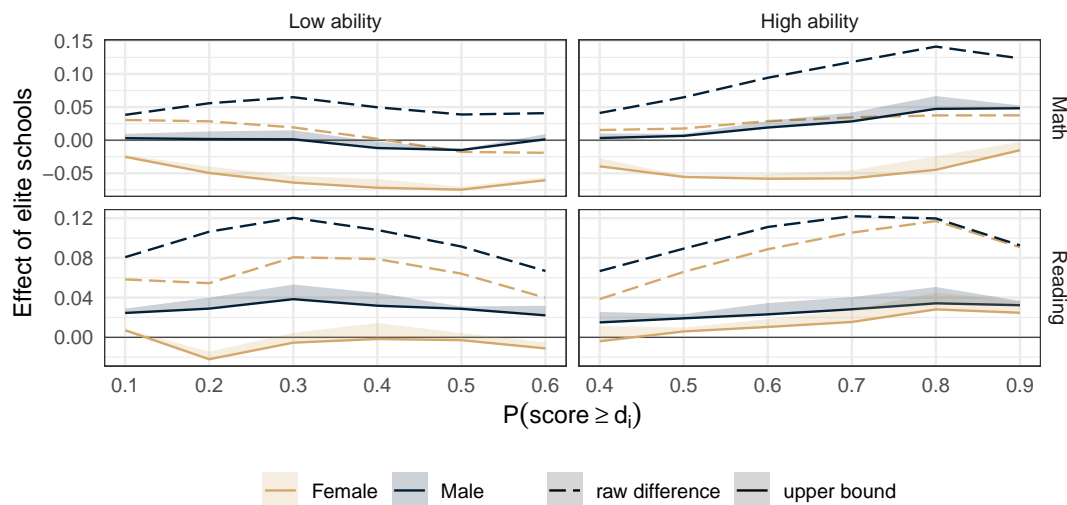
upper bounds are actually negative for the fourth and fifth deciles of the outcome distribution. If anything, elite-school enrollment may only raise high-achieving male students' mathematics test score at the top of the outcome distribution, by at most 2–5 percentage points.

The bottom panels of the figure show the effect of enrollment in an elite school on students' 8th-grade reading test scores. We find that the effect is negative for low-ability female students. The bounds do not exclude positive effects for male and high-achieving female students.

A potential mechanism underlying the negative short-run effects is that elite-school students switch schools, and it may take time for them to adapt to the new environment. We test this hypothesis by conducting a heterogeneity analysis. To this end, we focus on elite-school students who attend a comprehensive school, and thus enrollment in an elite school does not involve switching schools. Figure 3.7 presents the estimates for this subsample of students. The upper bounds are negative for both male and female students (irrespective of ability) in mathematics. These results suggest that the fact that students in the comparison group do not switch school does not drive our short-run results.¹⁶

¹⁶Appendix Figure C.3 shows the estimates for elite-school students who did switch to an elite school. The estimates are largely similar to those presented in Figure 3.6.

Figure 3.6: The effect of elite-school enrollment on the distribution of elite-school students' 8th-grade standardized test scores



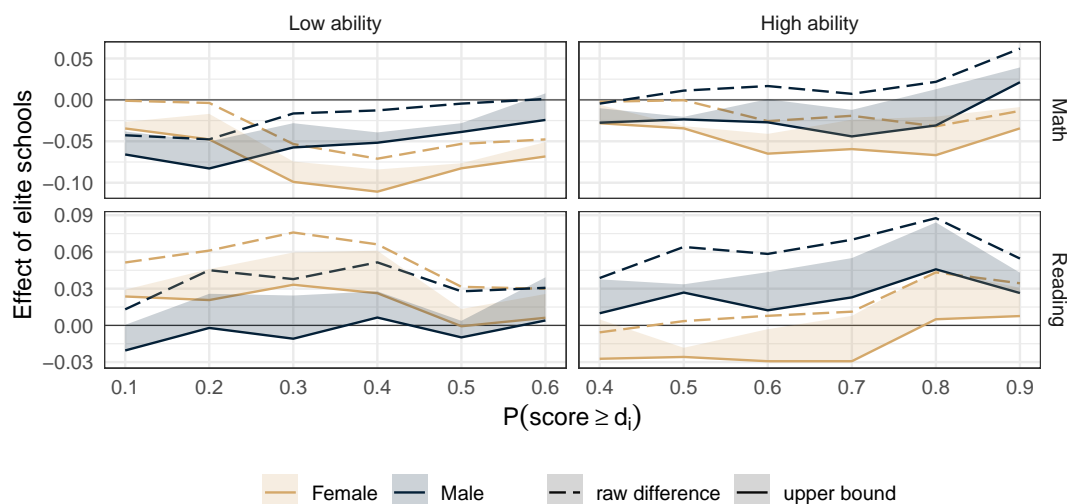
Notes: The figure presents our upper-bound estimates of the effect of elite-school enrollment on the distribution of elite-school students' 8th-grade mathematics grades (solid lines). The dashed line denotes the raw difference between the outcomes of elite-school and non-elite-school students. We report the estimates separately by gender and low- and high-ability students. Low- and high-ability students are defined by whether the students' 6th-grade standardized test score is below or above the median. The shaded area represents the area between the upper confidence band (95%) and the upper bound estimate itself. The 95% confidence intervals are based on 1,000 bootstrap draws. Sample: 8th-grade sample, $N = 126,196$.

3.5.2 Medium-run academic achievement

Having found that enrollment in an elite school has a negative effect on female and low-ability students' 8th-grade mathematics test score, we next investigate whether these negative effects persist in grade 10. Figure 3.8 presents the upper bounds on the effect of enrollment in an elite school on the distribution of 10th-grade test scores. We find that the upper bounds are positive, and relatively large in magnitude, for each subgroup and test (mathematics and reading). These findings are consistent with negative as well as substantial positive effects, and therefore they are inconclusive about the question whether the short-run negative effects persist.

A potential explanation for the high upper bounds is that students who did not enroll in an elite school attend low value-added schools. This concern, for example, is particularly pronounced for vocational schools, whose curriculum has less of an academic orientation. We thus focus on the subset of secondary grammar schools that have elite school tracks and regular tracks as well (see Table 3.1). Figure 3.9 presents the upper bounds on this subset of secondary grammar schools. We find that the upper bounds are positive, but considerably lower than the ones presented in Figure 3.8. For example, the upper bound on the effect of having a mathematics test score in the top decile for high-ability male students drops from 12.5 to 6 percentage points. These findings indicate that low value-added secondary-schools (e.g., vocational schools) do not

Figure 3.7: The effect of elite-school enrollment on the distribution of elite-school students' 8th-grade standardized test scores: Comprehensive schools



Notes: The figure presents our upper-bound estimates of the effect of elite-school enrollment on the distribution of elite-school students' 8th-grade standardized test scores (solid lines). The dashed line denotes the raw difference between the outcomes of elite-school and non-elite-school students. We report the estimates separately by gender and low- and high-ability students. Low- and high-ability students are defined by whether the students' 6th-grade standardized test score is below or above the median. The shaded area represents the area between the upper confidence band (95%) and the upper bound estimate itself. The 95% confidence intervals are based on 1,000 bootstrap draws. Sample: 8th-grade sample – comprehensive schools, $N = 111,501$.

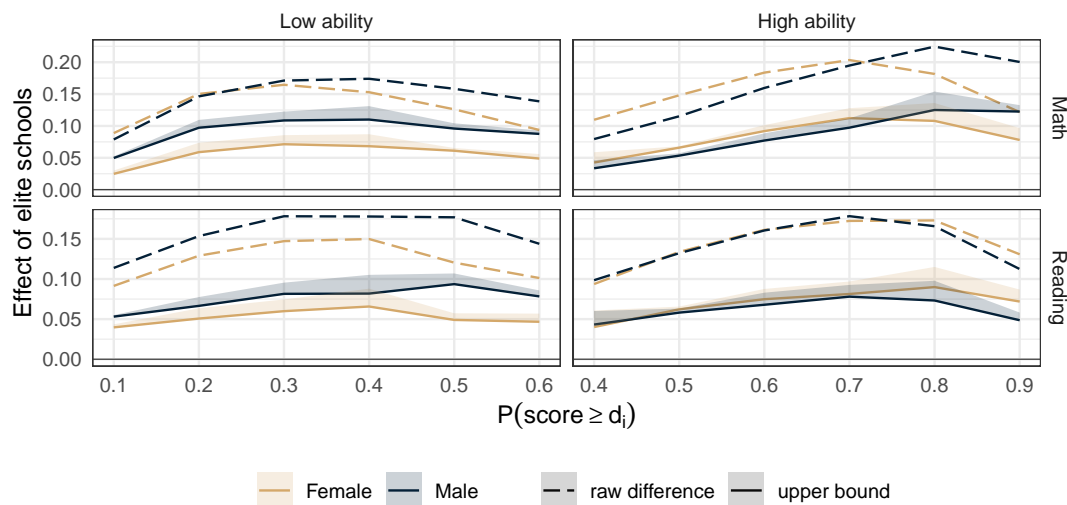
drive the positive upper bounds. Yet, these bounds do not reveal whether the negative effects of elite schools persist.

3.5.3 School value-added

As our non-parametric bounds strategy is inconclusive about the sign of the effects of elite-school enrollment in the medium run, we turn to the results of our school value-added models. Figure 3.10 presents the school value-added estimates for students' 8th-grade standardized test scores, in each gender–ability group using two alternative specifications. The simple model controls for students' 6th-grade standardized test score and cohort fixed effects. The full model adds additional controls for student and school characteristics (see Section 3.4.3).

The figure corroborates that enrollment in an elite school has a negative effect on female elite-school students' 8th-grade mathematics test scores throughout the outcome distribution. The simple model shows that enrollment in an elite school decreases the probability of having a mathematics test score above the median by 5 percentage points for low-ability, and 3 percentage points for high-ability female students. The full model with additional covariates suggests that the effect is about minus 7.5 percentage points for low-ability, and minus 8 percentage points

Figure 3.8: The effect of elite-school enrollment on the distribution of elite-school students' 10th-grade standardized test scores



Notes: The figure presents our upper-bound estimates of the effect of elite-school enrollment on the distribution of elite-school students' 10th-grade standardized test scores (solid lines). The dashed line denotes the raw difference between the outcomes of elite-school and non-elite-school students. We report the estimates separately by gender and low- and high-ability students. Low- and high-ability students are defined by whether the students' 6th-grade standardized test score is below or above the median. The shaded area represents the area between the upper confidence band (95%) and the upper bound estimate itself. The 95% confidence intervals are based on 1,000 bootstrap draws. Sample: 10th-grade sample, $N = 71,962$.

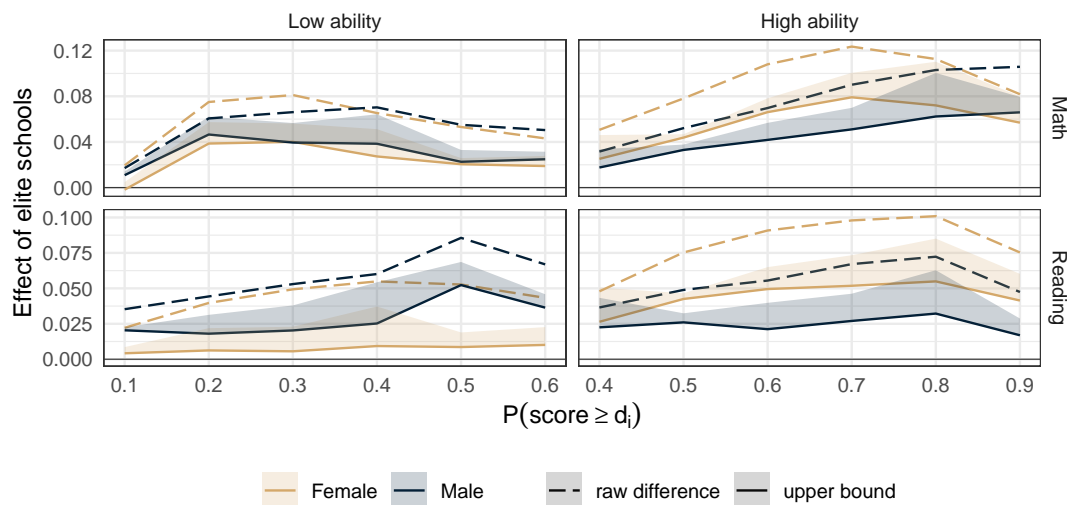
for high-ability female students. The negative relative effect of elite-school enrollment is larger for low-ability female students and at the top of the outcome distribution (Appendix Figure B6).

Figure 3.10 also reiterates that elite-school enrollment does not meaningfully increase the 8th-grade mathematics test scores of low-ability male students. The estimates of the simple school VA model are close to zero for both low- and high ability male students throughout the entire outcome distribution. The full model with additional covariates indicates that the effect is negative. Similar to female students, we find that the size of the relative effects are larger for low-ability male students relative to high-ability male students, and at the top of the outcome distribution (Appendix Figure B6).

The school value-added estimates are less robust when we study the effect of elite-school enrollment on elite-school students' 8th-grade reading scores. The simple model yields positive estimates for each ability-gender group throughout the outcome distribution. By contrast, the full model results in negative or insignificant estimates. We find that the full model's results are consistent with our non-parametric bounds (Appendix Figure C.4).

Figure 3.11 presents the school VA estimates for elite-school students' 10th-grade standardized test scores. We find that elite-school enrollment has a positive effect on elite-school students' 10th-grade test scores for both ability groups. The full model suggests that elite-school en-

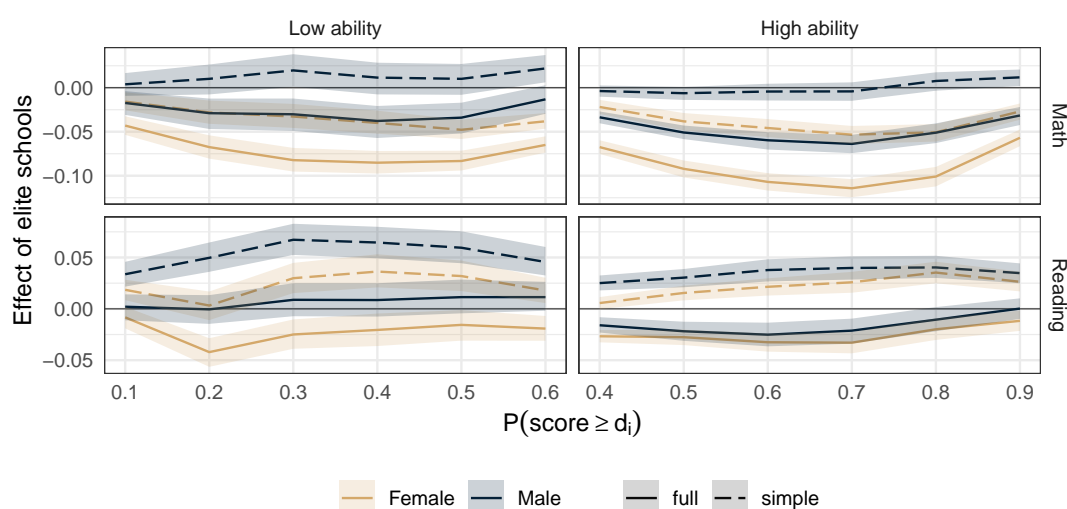
Figure 3.9: The effect of elite-school enrollment on the distribution of elite-school students' 10th-grade standardized test scores: Elite-school subsample



Notes: The figure presents our upper-bound estimates of the effect of elite-school enrollment on the distribution of elite-school students' 10th-grade standardized test scores (solid lines). The dashed line denotes the raw difference between the outcomes of elite-school and non-elite-school students. We report the estimates separately by gender and low- and high-ability students. Low- and high-ability students are defined by whether the students' 6th-grade standardized test score is below or above the median. The shaded area represents the area between the upper confidence band (95%) and the upper bound estimate itself. The 95% confidence intervals are based on 1,000 bootstrap draws. Sample: 10th-grade sample – elite-school subsample, $N = 21,384$.

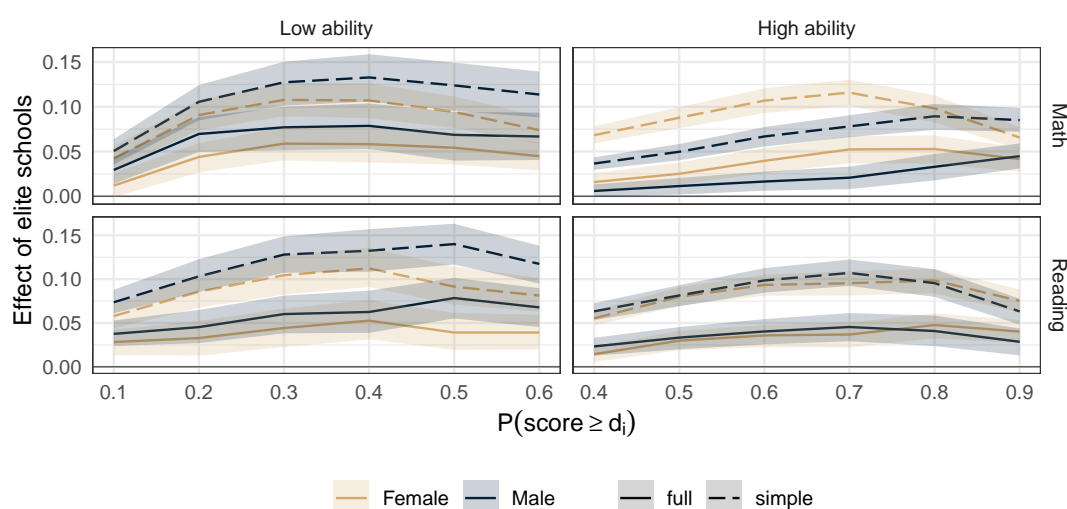
rollment increases the probability of having a mathematics test score above the median by 10 percentage points for low-ability female students, and 12.5 percentage points for low-ability male students. For high-ability students, we find that the effect on the probability of having a mathematics test score above the 9th decile is 7 percentage points for girls, and 8 percentage points for boys. The relative effect of elite-school enrollment on the mathematics test score is heterogeneous for high-ability students. We find that the relative effect for high-ability female students is higher than those of high-ability male students. By contrast, the relative effects do not differ between low-ability female and male students. We also find that the relative effect is higher at the top of the outcome distribution (Appendix Table B7).

Figure 3.10: The effect of elite-school enrollment on the distribution of elite-school students' 8th-grade test scores: School VA



Notes: The figure presents the school value-added estimates of the effect of elite-school enrollment on the distribution of elite-school students' 8th-grade standardized test scores. The figure presents school VA estimates for the deciles of the outcome distribution. The top (bottom) panels focus on mathematics (reading). The left (right) panels focus on students whose 6th-grade test score is above (below) the median. The dashed lines refer to the estimates of the simple school VA model (6th-grade standardized test score, cohort fixed effects) and the solid lines refer to the full school VA model (6th-grade standardized test score, cohort fixed effects, 5th-grade GPA, number of books at home, parental education, disadvantaged status, county of the school, type of the settlement where the school is located). The shaded area represents the 95% confidence intervals around the school VA estimates. The confidence intervals are based on 1,000 bootstrap draws. Sample: 8th-grade sample, $N = 126,196$.

Figure 3.11: The effect of enrollment in an elite school on the distribution of elite-school students' 10th-grade test scores: School VA



Notes: The figure presents the school value-added estimates of the effect of elite-school enrollment on the distribution of elite-school students' 10th-grade standardized test scores. The figure presents school VA estimates for the deciles of the outcome distribution. The top (bottom) panels focus on mathematics (reading). The left (right) panels focus on students whose 6th-grade test score is below (above) the median. The dashed lines refer to the estimates of the simple school VA model (6th-grade standardized test score, cohort fixed effects) and the solid lines refer to the full school VA model (6th-grade standardized test score, cohort fixed effects, 5th-grade GPA, number of books at home, parental education, disadvantaged status, county of the school, type of the settlement where the school is located). The shaded area represents the 95% confidence intervals around the school VA estimates. The confidence intervals are based on 1,000 bootstrap draws. Sample: 10th-grade sample, $N = 71,962$.

3.6 Conclusions

We have studied the effect on enrollment in an elite school on elite-school students' short- and medium-run academic achievement. Motivated by the fact that local effects identified at the margin of admission may not generalize to the average elite-school student, the causal effect of interest has been the average treatment effect on the treated. To identify the causal effects of interest, we have used non-parametric bounds and school value-added models. To further understand the heterogeneous effects of elite schools, we have conducted our analysis by gender-ability groups, and we have studied the effects throughout the outcome distribution.

Our main finding is that enrollment in an elite school has a negative effect on female and low-ability students' mathematics test scores on the short run. By contrast, our non-parametric bounds strategy does not allow us to exclude small, positive effects for high-ability male students. We argue that the short-run negative effects are not the consequence of grading policies (e.g., grading on a curve, ceiling effects), but they reflect real effects in skill formation. As non-parametric bounds are uninformative about the sign of the medium-run effects, we have used school value-added models to test whether the negative short-run effects persist. These school value-added estimates suggest that elite-school enrollment has a positive effect on academic achievement for each gender-ability group on the medium run.

The most salient difference between an average elite-school student and a student on the margin of admission is her (baseline) academic ability. Therefore, our analysis has focused on effect heterogeneity along the ability distribution. Splitting the sample by baseline ability suggests that high-ability students benefit more from elite-school enrollment. By studying the effect of elite-school enrollment throughout the outcome distribution, our findings indicate that the benefits of elite schools are concentrated at the top of the test score distribution.

Our study design does not allow us to identify the exact mechanisms driving our results. Nonetheless, heterogeneity analysis along school characteristics enables us to exclude some important mechanisms. First, the negative upper bounds on the short-run effects in comprehensive schools suggest that school switching does not explain the negative short-run effects of elite schools. Second, the positive upper bounds on medium-run outcomes are not solely driven by less selective schools in the counterfactual.

We have focused on how elite-school enrollment affect academic achievement. However, parents may also value schools along other dimensions, such as non-cognitive skills, field of study, college enrollment, or labor market outcomes (see, e.g., [Beuermann and Jackson, Forthcoming](#); [Beuermann et al., 2018](#)).¹⁷ We view the understanding of how elite schools affect these outcomes as fruitful directions for future research.

¹⁷The strong positive association between Hungarian students' standardized test scores and their labor market outcomes (such as employment and wages) suggests that the medium-run benefits in academic achievement may turn into long-run gains in the labor market ([Hermann et al., 2019](#)).

Bibliography

Abadie, Alberto, Joshua Angrist, and Guido Imbens, “Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings,” *Econometrica*, 2002, 70 (1), 91–117.

Abdulkadiroglu, Atila, Joshua Angrist, and Parag Pathak, “The Elite Illusion: Achievement Effects at Boston and New York Exam Schools,” *Econometrica*, 2014, 82 (1), 137–196.

Agrawal, Shipra and Navin Goyal, “Analysis of thompson sampling for the multi-armed bandit problem,” in “*Journal of Machine Learning Research*,” Vol. 23 Microtome Publishing 2012.

— **and** —, “Further optimal regret bounds for thompson sampling,” in “*Journal of Machine Learning Research*,” Vol. 31 Microtome Publishing 2013, pp. 99–107.

Aichberger, M.C., M.a. Busch, F.M. Reischies, A. Ströhle, A. Heinz, and M.a. Rapp, “Effect of Physical Inactivity on Cognitive Performance after 2.5 Years of Follow-Up,” *GeroPsych: The Journal of Gerontopsychology and Geriatric Psychiatry*, 2010, 23 (1), 7–15.

Anderson, Kathryn, Xue Gong, Kai Hong, and Xi Zhang, “Do Selective High Schools Improve Student Achievement? Effects of Exam Schools in China,” *China Economic Review*, 2016, 40, 121–134.

Angrist, Joshua D and Jörn-Steffen Pischke, *Mostly harmless econometrics: An empiricist’s companion*, Princeton: Princeton University Press, 2008.

Athey, Susan and Stefan Wager, “*Efficient Policy Learning*,” 2019.

Banks, James and Fabrizio Mazzonna, “The Effect of Education on Old Age Cognitive Abilities: Evidence from a Regression Discontinuity Design,” *The Economic Journal*, 2012, 122 (560), 418–448.

Barrow, Lisa, Lauren Sartain, and M. De la Torre, “Increasing Access to Selective High Schools through Place-based Affirmative Action: Unintended Consequences,” *American Economic Journal: Applied Economics*, Forthcoming.

- Beuermann, Diether W. and C. Kirabo Jackson**, “The Short and Long-Run Effects of Attending the Schools that Parents Prefer,” *Journal of Human Resources*, Forthcoming.
- , —, **Laia Navarro-Sola, and Francisco Pardo**, “What is a Good School, and Can Parents Tell? Evidence on the Multidimensionality of School Output,” Working Paper 25342, National Bureau of Economic Research December 2018.
- Bingley, Paul and Alessandro Martinello**, “Mental retirement and schooling,” *European Economic Review*, 2013, 63, 292–298.
- Bloom, Howard S., Larry L. Orr, Stephen H. Bell, George Cave, Fred Doolittle, Winston Lin, and Johannes M. Bos**, “The Benefits and Costs of JTPA Title II-A Programs: Key Findings from the National Job Training Partnership Act Study,” *The Journal of Human Resources*, 1997, 32 (3), 549–576.
- Bonsang, Eric, Stéphane Adam, and Sergio Perelman**, “Does retirement affect cognitive functioning?,” *Journal of Health Economics*, 2012, 31 (3), 490–501.
- Bui, Sa A., Steven G. Craig, and Scott A. Imberman**, “Is Gifted Education a Bright Idea? Assessing the Impact of Gifted and Talented Programs on Students,” *American Economic Journal: Economic Policy*, 2014, 6 (3), 30–62.
- Calsamiglia, Caterina and Annalisa Loviglio**, “Grading on a Curve: When Having Good Peers is not Good,” *Economics of Education Review*, 2019, 73 (101916), 1–21.
- Chetty, Raj, John N. Friedman, and Jonah E. Rockoff**, “Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates,” *American Economic Review*, 2014, 104 (9), 2593–2632.
- , —, and —, “Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood,” *American Economic Review*, 2014, 104 (9), 2633–2679.
- Clark, Damon**, “Selective Schools and Academic Achievement,” *The B.E. Journal of Economic Analysis & Policy*, February 2010, 10 (1), 1–40.
- Cullen, Julie Berry, Brian A. Jacob, and Steven Levitt**, “The Effect of School Choice on Participants: Evidence from Randomized Lotteries,” *Econometrica*, 2006, 74 (5), 1191–1230.
- de Haan, Monique**, “The Effect of Parents’ Schooling on Child’s Schooling: A Nonparametric Bounds Analysis,” *Journal of Labor Economics*, 2011, 29 (4), 859–892.
- and **Edwin Leuven**, “Head Start and the Distribution of Long-Term Education and Labor Market Outcomes,” *Journal of Labor Economics*, 2020, 38 (3), 727–765.
- Dehejia, Rajeev H.**, “Program evaluation as a decision problem,” *Journal of Econometrics*, 2005, 125 (1-2 SPEC. ISS.), 141–173.

Deming, David J., “Better Schools, Less Crime?,” *Quarterly Journal of Economics*, 2011, 126 (4), 2063–2115.

Dimakopoulou, Maria, Zhengyuan Zhou, Susan Athey, and Guido Imbens, “Estimation Considerations in Contextual Bandits,” 2018.

Dobbie, Will and Jr. Fryer Roland G., “Are High-Quality Schools Enough to Increase Achievement Among the Poor? Evidence from the Harlem Children’s Zone,” *American Economic Journal: Applied Economics*, 2011, 3 (3), 158–187.

— **and** —, “The Impact of Attending a School with High-Achieving Peers: Evidence from the New York City Exam Schools,” *American Economic Journal: Applied Economics*, July 2014, 6 (3), 58–75.

Graepel, Thore, Joaquin Quinonero Candela, Thomas Borchert, and Ralf Herbrich, “Web-Scale Bayesian Click-Through Rate Prediction for Sponsored Search Advertising in Microsoft’s Bing Search Engine,” in “Proceedings of the 27th International Conference on Machine Learning (ICML)” 2010, pp. 13–20.

Guarino, Cassandra M., Mark D. Reckase, and Jeffrey M. Wooldridge, “Can Value-Added Measures of Teacher Performance Be Trusted?,” *Education Finance and Policy*, 2015, 10 (1), 117–156.

Hadad, Vitor, David A. Hirshberg, Ruohan Zhan, Stefan Wager, and Susan Athey, “Confidence Intervals for Policy Evaluation in Adaptive Experiments,” 2019.

Hahn, Jinyong, Keisuke Hirano, and Dean Karlan, “Adaptive experimental design using the propensity score,” *Journal of Business and Economic Statistics*, jan 2011, 29 (1), 96–108.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning Springer Series in Statistics*, New York, NY, USA: Springer New York Inc., 2001.

Hermann, Zoltán, Dániel Horn, János Köllő, Anna Sebők, András Semjén, and Júlia Varga, “Szövegértési és matematikai kompetencia hatása a keresetre és a foglalkoztatási esélyekre,” in Károly Fazekas, Márton Csillag, Zoltán Hermann, and Ágota Scharle, eds., *Munkaerőpiaci Tükör 2018*, Budapest: Közgazdaság- és Regionális Tudományi Kutatóközpont, 2019, pp. 45–52.

Hertzog, Christopher, Arthur F Kramer, Robert S Wilson, and Ulman Lindenberger, “Enrichment Effects on Adult Cognitive Development,” *Psychological Science*, 2009, 9 (1), 1–65.

Hirano, Keisuke and Jack R. Porter, “Asymptotics for Statistical Treatment Rules,” *Econometrica*, 2009, 77 (5), 1683–1701.

- Horn, Dániel**, “Diverging Performances: The Detrimental Effects of Early Educational Selection on Equality of Opportunity in Hungary,” *Research in Social Stratification and Mobility*, 2013, 32 (June), 25–43.
- Imbens, Guido W. and Charles F. Manski**, “Confidence Intervals for Partially Identified Parameters,” *Econometrica*, 2004, 72 (6), 1845–1857.
- Ioannidis, John P.A.**, “Why most published research findings are false,” *PLoS Medicine*, 2018, 2 (8), 2–8.
- Jackson, Kirabo C.**, “Do Students Benefit from Attending Better Schools? Evidence from Rule-Based Student Assignments in Trinidad and Tobago,” *The Economic Journal*, 11 2010, 120 (549), 1399–1429.
- Kasy, Maximilian**, “Why experimenters might not always want to randomize, and what they could do instead,” *Political Analysis*, 2016, 24 (3), 324–338.
- **and Anja Sautmann**, “Adaptive Experiments for Policy Choice,” 2019.
- Kitagawa, Toru and Aleksey Tetenov**, “Equality-minded treatment choice,” 2017.
- **and —**, “Who should be treated? Empirical welfare maximization methods for treatment choice,” *Econometrica*, 2018, 86 (2), 591–616.
- Koedel, Cory, Kata Mihaly, and Jonah E. Rockoff**, “Value-Added Modeling: A Review,” *Economics of Education Review*, 2015, 47, 180–195.
- Korda, Nathaniel, Emilie Kaufmann, and Remi Munos**, “Thompson Sampling for 1-Dimensional Exponential Family Bandits,” in “Advances in Neural Information Processing Systems 26 (NIPS)” 2013, pp. 1448–1456.
- Lai, Tze Leung and Herbert Robbins**, “Asymptotically Efficient Adaptive Allocation Rules,” *Advances in Applied Mathematics*, 1985, 6 (1), 4–22.
- Lattimore, Tor and Csaba Szepesvári**, *Bandit Algorithms*, Cambridge University Press, 2019.
- Lucas, Adrienne M. and Isaac M. Mbiti**, “Effects of School Quality on Student Achievement: Discontinuity Evidence from Kenya,” *American Economic Journal: Applied Economics*, 2014, 6 (3), 234–63.
- Manski, Charles F.**, “Statistical Treatment Rules for Heterogeneous Populations,” *Econometrica*, 2004, 72 (4), 1221–1246.
- **and John V. Pepper**, “Monotone Instrumental Variables, with an Application to the Returns to Schooling,” *Econometrica*, July 2000, 68 (4), 997–1012.

- Mazzonna, Fabrizio and Franco Peracchi**, “Ageing, cognitive abilities and retirement,” *European Economic Review*, 2012, 56 (4), 691–710.
- Nie, Xinkun, Xiaoying Tian, Jonathan Taylor, and James Zou**, “Why Adaptively Collected Data Have Negative Bias and How to Correct for It,” in “*Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS)*” 2018.
- OECD**, “*Knowledge and Skills for Life*,” Technical Report 2001.
- Oosterbeek, Hessel, Nienke Ruis, and Inge Wolf**, “Using Admission Lotteries to Estimate the Heterogeneous Effects of Elite Schools,” 2020. Tinbergen Institute Discussion Paper, 2020-018/V.
- Perchet, Vianney, Philippe Rigollet, Sylvain Chassang, and Erik Snowberg**, “Batched bandit problems,” *Annals of Statistics*, 2016, 44 (2), 660–681.
- Pop-Eleches, Cristian and Miguel Urquiola**, “Going to a Better School: Effects and Behavioral Responses,” *American Economic Review*, 2013, 103 (4), 1289–1324.
- Richey, Jeremiah**, “An Odd Couple: Monotone Instrumental Variables and Binary Treatments,” *Econometric Reviews*, June 2016, 35 (6), 1099–1110.
- Rohwedder, Susann and Robert J Willis**, “Mental Retirement.,” *The Journal of Economic Perspectives*, 2010, 24 (1), 119–138.
- Rothstein, Jesse**, “Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement,” *The Quarterly Journal of Economics*, 2010, 125 (1), 175–214.
- , “Measuring the Impacts of Teachers: Comment,” *American Economic Review*, 2017, 107 (6), 1656–84.
- Rowling, Joanne K**, *Harry Potter and the Chamber of Secrets*, New York: Scholastic, Inc., 1999.
- Rubin, Donald B.**, “Matching to Remove Bias in Observational Studies,” *Biometrics*, 1973, 29 (1), 159–183.
- Russo, Daniel, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen**, “A Tutorial on Thompson Sampling,” *Foundations and Trends® in Machine Learning*, 2017, 11 (11), 1–96.
- Schaie, K Warner**, “The hazards of cognitive aging.,” *The Gerontologist*, 1989, 29 (4), 484–93.
- Schiltz, Fritz, Deni Mazrekaj, Daniel Horn, and Kristof De Witte**, “Does It Matter When your Smartest Peers Leave your Class? Evidence from Hungary,” *Labour Economics*,

2019, 59, 79–91. *Special Issue on European Association of Labour Economists, 30th annual conference, Lyon, France, 13-15 September 2018.*

Scott, Steven L., “A modern Bayesian look at the multi-armed bandit,” *Applied Stochastic Models in Business and Industry*, 2010, 26, 639–658.

Slivkins, Aleksandrs, *Introduction to Multi-Armed Bandits* 2019.

Stine-Morrow, Elizabeth A L, “The Dumbledore Hypothesis of Cognitive Aging,” *Current Directions in Psychological Science*, dec 2007, 16 (6), 295–299.

Stock, James H, Jonathan H Wright, and Motohiro Yogo, “A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments,” *Journal of Business & Economic Statistics*, 2002, 20, 518–529.

Thompson, William R., “On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples,” *Biometrika*, 1933, 25 (3-4), 285–294.

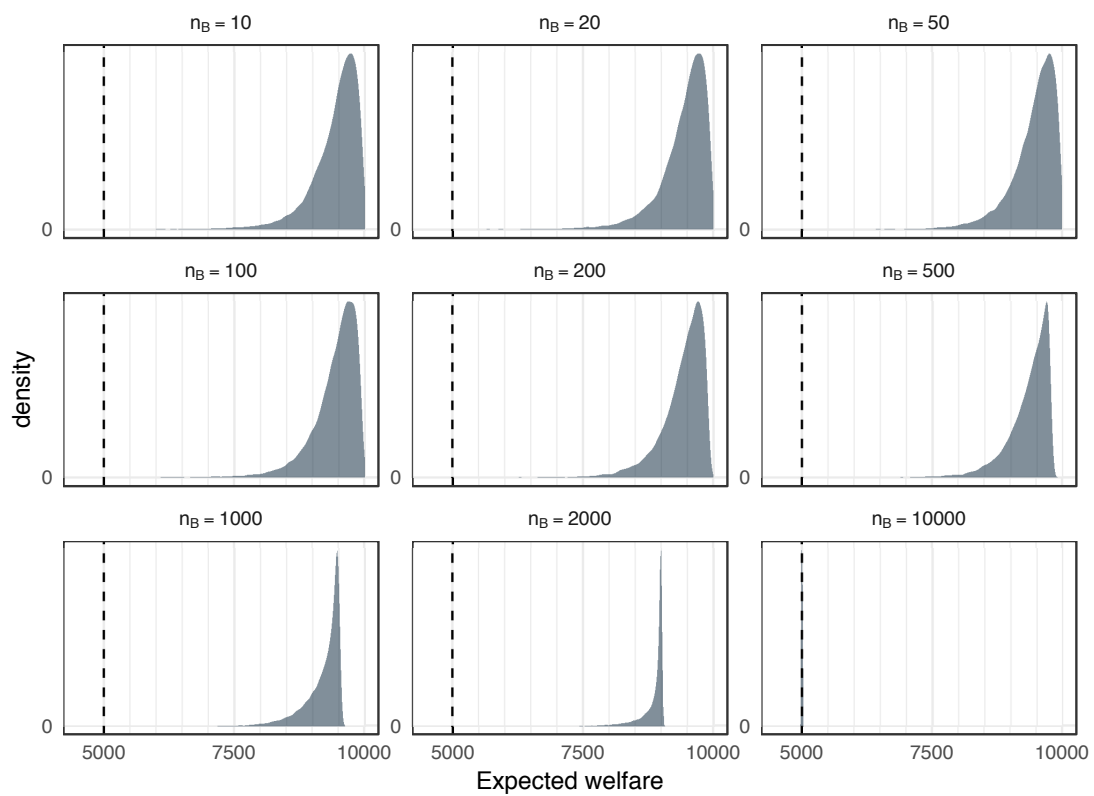
Villar, Sofía S., Jack Bowden, and James Wason, “Multi-armed Bandit Models for the Optimal Design of Clinical Trials: Benefits and Challenges,” *Statistical Science*, 2015, 30 (2), 199–215.

Appendix A

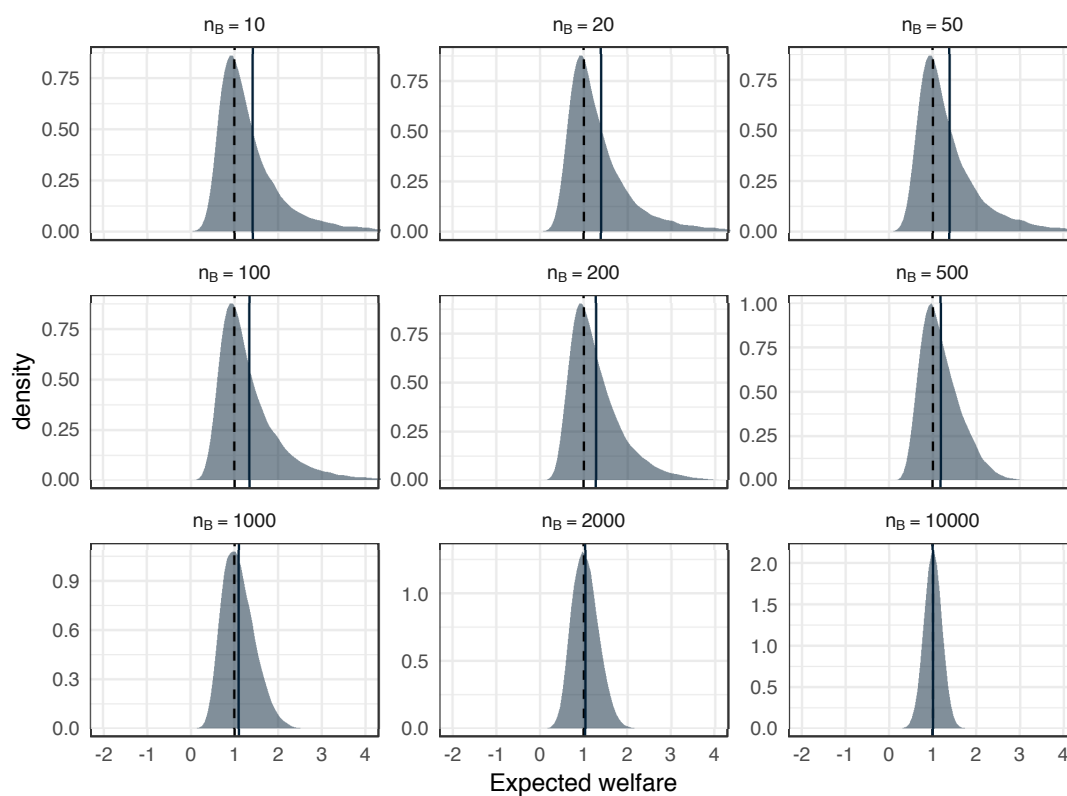
Appendix for Chapter 1

A.1 Simulation distributions

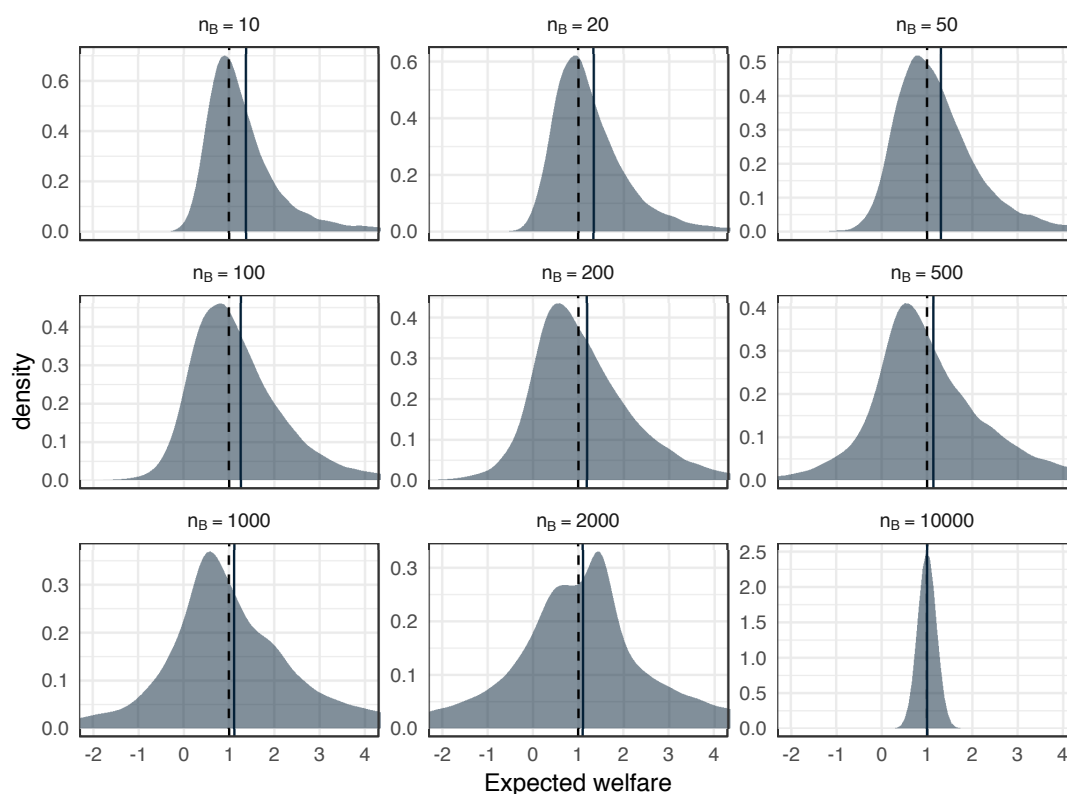
This section presents the whole simulation distributions for expected welfare and various estimators to complement the summary numbers in the main text.

Figure A.1: Distribution of welfare by batch size

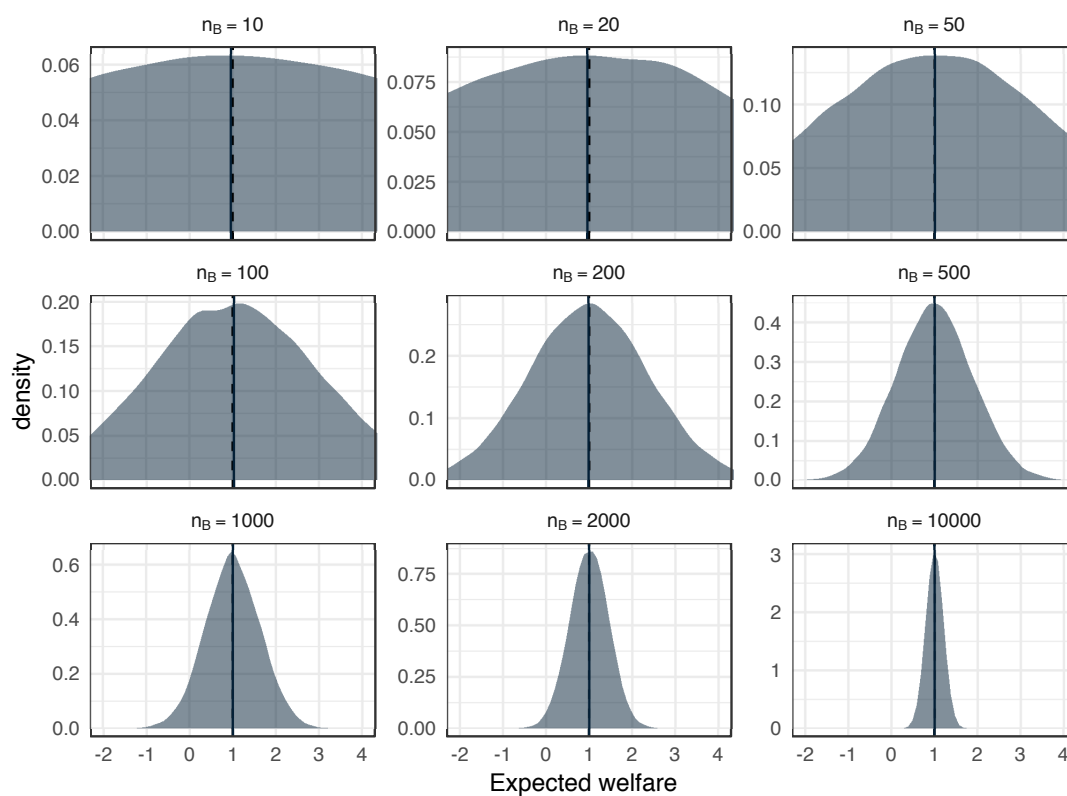
Notes: Each panel shows the distribution of the achieved welfare by the bandit algorithm with the corresponding batch size ($\sigma = 10$). Quicker adaptivity (smaller batch size) leads to higher achievable welfare but also higher variance.

Figure A.2: Distribution of $\hat{\tau}_0$ by batch size

Notes: Each panel shows the distribution of the standard treatment effect estimate for the bandit algorithm with the corresponding batch size ($\sigma = 10$). The dashed line shows the true treatment effect, while the solid line corresponds to the expected value of the estimates. Quicker adaptivity (smaller batch size) leads to a more volatile estimate with larger bias.

Figure A.3: Distribution of $\hat{\tau}_{IPW}$ by batch size

Notes: Each panel shows the distribution of the inverse-propensity-weighted treatment effect estimate for the bandit algorithm with the corresponding batch size ($\sigma = 10$). The dashed line shows the true treatment effect, while the solid line corresponds to the expected value of the estimates. Quicker adaptivity (smaller batch size) leads to larger bias. The variance is larger compared to $\hat{\tau}_0$, especially for larger batch sizes.

Figure A.4: Distribution of $\hat{\tau}_{FB}$ by batch size

Notes: Each panel shows the distribution of the treatment effect estimate calculated on the first batch of the bandit algorithm with the corresponding batch size ($\sigma = 10$). The dashed line shows the true treatment effect, while the solid line corresponds to the expected value of the estimates. The estimator is unbiased but really volatile, especially for smaller batch sizes.

A.2 Detailed simulation results

This section presents all the simulation results of expected welfare, expected bias and expected welfare for the various strategies in different setups. The welfare-estimation plots in the main text are based on these numbers.

Table A.1: Expected welfare for different strategies ($n = 10,000$)

allocation	Batch size									
	10	20	50	100	200	500	1000	2000	5000	10,000
$\sigma = 1$										
TS	9987	9985	9974	9950	9900	9750	9500	9000	7500	5000
ETC	9465	9847	9973	9950	9900	9750	9500	9000	7500	5000
LTS-0%					9851	9702	9455	8960	7475	5000
LTS-1%				9851	9802	9655	9410	8920	7450	5000
LTS-2%			9776	9752	9704	9560	9320	8840	7400	5000
LTS-5%		9490	9477	9455	9410	9275	9050	8600	7250	5000
LTS-10%	8995	8991	8980	8960	8920	8800	8600	8200	7000	5000
LTS-15%	8495	8493	8483	8465	8430	8325	8150	7800	6750	5000
LTS-20%	7996	7994	7985	7970	7940	7850	7700	7400	6500	5000
$\sigma = 2$										
TS	9957	9957	9953	9940	9898	9750	9500	9000	7500	5000
ETC	7858	8663	9580	9892	9899	9750	9500	9000	7500	5000
LTS-0%					9850	9702	9455	8960	7475	5000
LTS-1%				9846	9801	9655	9410	8920	7450	5000
LTS-2%			9767	9748	9703	9560	9320	8840	7400	5000
LTS-5%		9479	9471	9452	9410	9275	9050	8600	7250	5000
LTS-10%	8986	8984	8976	8958	8920	8800	8600	8200	7000	5000
LTS-15%	8489	8487	8479	8464	8430	8325	8150	7800	6750	5000
LTS-20%	7991	7990	7983	7969	7940	7850	7700	7400	6500	5000
$\sigma = 5$										
TS	9797	9800	9801	9798	9778	9691	9482	8998	7500	5000
ETC	6186	6710	7713	8412	9098	9615	9494	9000	7500	5000
LTS-0%					9750	9656	9442	8959	7475	5000
LTS-1%				9736	9714	9615	9399	8919	7450	5000
LTS-2%			9660	9656	9630	9527	9311	8839	7400	5000
LTS-5%		9395	9394	9386	9357	9252	9044	8600	7250	5000
LTS-10%	8922	8925	8921	8912	8883	8785	8597	8200	7000	5000
LTS-15%	8439	8442	8438	8429	8403	8315	8148	7800	6750	5000
LTS-20%	7954	7955	7951	7943	7920	7843	7699	7400	6500	5000
$\sigma = 10$										
TS	9372	9382	9389	9392	9383	9321	9178	8820	7469	5000
ETC	5626	5840	6375	6946	7533	8496	9004	8896	7498	5000
LTS-0%					9371	9308	9163	8800	7451	5000
LTS-1%				9358	9351	9288	9140	8774	7430	5000
LTS-2%			9302	9309	9301	9234	9082	8713	7384	5000

Table A.1: Expected welfare for different strategies ($n = 10,000$) (*continued*)

allocation	Batch size									
	10	20	50	100	200	500	1000	2000	5000	10,000
LTS-5%		9096	9106	9111	9097	9026	8871	8508	7241	5000
LTS-10%	8694	8707	8713	8711	8695	8623	8476	8139	6995	5000
LTS-15%	8260	8278	8280	8276	8258	8193	8060	7758	6747	5000
LTS-20%	7819	7829	7829	7826	7809	7750	7634	7370	6498	5000
$\sigma = 15$										
TS	8878	8873	8881	8896	8891	8828	8709	8400	7258	5000
ETC	5441	5615	5950	6285	6781	7594	8175	8429	7453	5000
LTS-0%					8884	8825	8703	8392	7250	5000
LTS-1%				8875	8871	8812	8691	8379	7238	5000
LTS-2%			8830	8849	8837	8779	8656	8343	7209	5000
LTS-5%		8678	8693	8708	8696	8638	8508	8199	7104	5000
LTS-10%	8356	8374	8386	8392	8377	8318	8196	7907	6900	5000
LTS-15%	7989	8014	8025	8030	8012	7955	7842	7579	6679	5000
LTS-20%	7609	7622	7629	7632	7616	7563	7464	7233	6450	5000
$\sigma = 20$										
TS	8359	8353	8369	8386	8380	8331	8226	7957	6968	5000
ETC	5312	5445	5696	5983	6328	6969	7577	7936	7300	5000
LTS-0%					8375	8327	8223	7954	6965	5000
LTS-1%				8373	8368	8321	8216	7946	6958	5000
LTS-2%			8340	8351	8345	8299	8194	7924	6941	5000
LTS-5%		8222	8243	8253	8244	8193	8090	7824	6869	5000
LTS-10%	7964	7983	8005	8014	7997	7946	7847	7598	6714	5000
LTS-15%	7660	7699	7712	7713	7700	7649	7556	7328	6532	5000
LTS-20%	7340	7364	7377	7375	7364	7319	7235	7032	6334	5000
$\sigma = 25$										
TS	7923	7936	7942	7945	7939	7901	7804	7578	6704	5000
ETC	5248	5373	5591	5804	6088	6632	7120	7531	7104	5000
LTS-0%					7935	7901	7802	7575	6702	5000
LTS-1%				7938	7930	7893	7797	7571	6699	5000
LTS-2%			7924	7922	7913	7878	7782	7556	6688	5000
LTS-5%		7829	7847	7856	7835	7797	7703	7483	6637	5000
LTS-10%	7603	7637	7653	7661	7646	7602	7507	7304	6516	5000
LTS-15%	7349	7401	7412	7413	7398	7355	7268	7080	6368	5000
LTS-20%	7084	7116	7130	7130	7115	7075	6994	6828	6200	5000
$\sigma = 30$										
TS	7566	7551	7585	7585	7590	7548	7473	7245	6485	5000
ETC	5260	5302	5466	5664	5899	6337	6827	7159	6896	5000
LTS-0%					7577	7548	7471	7244	6484	5000
LTS-1%				7566	7574	7543	7468	7240	6481	5000
LTS-2%			7567	7554	7562	7530	7456	7230	6473	5000
LTS-5%		7520	7506	7497	7501	7464	7394	7172	6434	5000
LTS-10%	7316	7347	7357	7338	7345	7301	7231	7025	6337	5000
LTS-15%	7091	7140	7141	7130	7134	7096	7026	6839	6212	5000

Table A.1: Expected welfare for different strategies ($n = 10,000$) (*continued*)

allocation	Batch size									
	10	20	50	100	200	500	1000	2000	5000	10,000
LTS-20%	6871	6895	6901	6889	6894	6855	6790	6623	6067	5000

Notes: TS: Thompson sampling, ETC: Explore-then-commit, LTS-X%: Limited Thompson sampling with X% limitation. Expected welfare is calculated as the average of the sum of outcomes ($\sum_{i=1}^n Y$) across the simulation runs. Number of simulations = 10,000 for $\sigma < 10$, 20,000 for $10 \leq \sigma < 20$ and 50,000 for $\sigma \geq 20$.

Table A.2: Bias for different strategies ($n = 10,000$)

strategy	Batch size									
	10	20	50	100	200	500	1000	2000	5000	10000
$\sigma = 1$										
TS	0.127	0.072	0.009	0.001	0.000	0.000	0.000	0.000	0.000	0.000
TS-IPW	0.177	0.169	0.047	0.001	0.000	0.000	0.000	0.000	0.000	0.000
TS-FB	0.018	-0.003	0.003	0.001	0.002	0.000	-0.001	0.000	0.000	0.000
ETC	0.008	0.000	0.002	0.001	0.000	0.000	0.000	0.000	0.000	0.000
LTS-0.5%					0.002	0.000	0.001	0.000	-0.001	0.000
LTS-1%				0.001	0.000	0.000	-0.001	-0.001	0.001	0.000
LTS-2%			0.001	-0.001	0.000	0.000	-0.001	-0.001	0.000	0.000
LTS-5%		0.000	0.000	0.000	0.000	-0.001	-0.001	0.000	0.000	0.000
LTS-10%	0.000	0.000	0.000	0.000	0.000	-0.001	0.000	0.000	0.000	0.000
LTS-15%	0.000	0.000	0.000	-0.001	0.000	0.000	0.000	0.000	0.000	0.000
LTS-20%	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$\sigma = 2$										
TS	0.196	0.174	0.109	0.043	0.003	-0.002	-0.001	0.000	0.000	0.000
TS-IPW	0.219	0.231	0.244	0.195	0.055	-0.002	-0.001	0.000	0.000	0.000
TS-FB	-0.002	-0.005	0.001	0.006	0.000	-0.002	-0.002	-0.001	0.000	0.000
ETC	0.000	-0.003	0.003	0.001	-0.002	-0.002	-0.001	0.000	0.000	0.000
LTS-0.5%					0.003	0.003	0.002	-0.001	0.001	0.000
LTS-1%				0.000	0.000	0.002	-0.003	-0.001	0.002	0.000
LTS-2%			0.001	0.001	-0.001	0.000	-0.002	-0.001	0.000	0.000
LTS-5%		-0.001	0.001	0.001	0.000	-0.001	-0.001	-0.001	0.000	0.000
LTS-10%	-0.001	0.000	0.001	0.000	0.000	-0.001	-0.001	-0.001	0.000	0.000
LTS-15%	-0.001	0.000	0.000	0.000	0.000	-0.001	-0.001	0.000	0.000	0.000
LTS-20%	0.000	0.000	0.001	0.000	0.000	-0.001	0.000	-0.001	0.000	0.000
$\sigma = 5$										
TS	0.313	0.302	0.260	0.213	0.148	0.052	0.013	0.003	0.001	0.001
TS-IPW	0.305	0.302	0.287	0.284	0.255	0.198	0.103	0.030	0.000	0.001
TS-FB	-0.012	-0.010	0.022	-0.007	-0.003	0.004	0.002	0.003	0.000	0.001
ETC	0.009	0.009	0.007	0.000	0.004	0.006	0.004	0.003	0.001	0.001
LTS-0.5%					0.004	0.007	0.005	-0.005	-0.006	0.001
LTS-1%				0.003	0.007	0.000	0.000	-0.005	-0.007	0.001

Table A.2: Bias for different strategies ($n = 10,000$) (*continued*)

strategy	Batch size									
	10	20	50	100	200	500	1000	2000	5000	10000
LTS-2%			0.002	0.005	0.004	-0.005	-0.004	-0.002	-0.004	0.001
LTS-5%		0.000	0.004	0.001	0.001	-0.005	-0.002	0.001	-0.001	0.001
LTS-10%	-0.001	0.003	0.001	0.001	0.000	-0.004	0.000	0.000	-0.001	0.001
LTS-15%	0.000	0.002	0.000	0.001	-0.001	-0.002	0.001	0.001	-0.001	0.001
LTS-20%	0.001	0.002	0.001	0.000	-0.001	-0.001	0.000	0.002	-0.001	0.001
$\sigma = 10$										
TS	0.419	0.394	0.383	0.342	0.279	0.182	0.099	0.035	0.003	0.002
TS-IPW	0.375	0.337	0.306	0.261	0.190	0.140	0.115	0.100	-0.007	0.002
TS-FB	-0.041	-0.043	0.008	0.030	-0.015	0.003	0.004	0.001	-0.001	0.002
ETC	-0.066	-0.023	0.002	0.007	-0.004	0.003	0.003	0.001	0.000	0.002
LTS-0.5%					-0.006	-0.001	-0.004	-0.010	0.001	0.002
LTS-1%				-0.010	-0.003	-0.001	-0.001	-0.005	-0.003	0.002
LTS-2%			-0.010	-0.002	-0.008	0.001	-0.005	-0.006	-0.007	0.002
LTS-5%		-0.004	-0.006	-0.003	0.003	-0.001	-0.005	0.003	-0.005	0.002
LTS-10%	-0.005	-0.004	-0.003	0.002	-0.001	-0.004	-0.001	0.002	-0.004	0.002
LTS-15%	-0.004	-0.005	0.000	0.001	0.001	-0.002	0.000	0.001	-0.003	0.002
LTS-20%	-0.002	-0.004	0.000	0.001	0.001	-0.001	-0.001	0.001	-0.002	0.002
$\sigma = 15$										
TS	0.516	0.509	0.462	0.424	0.376	0.267	0.184	0.092	0.017	0.002
TS-IPW	0.443	0.404	0.315	0.246	0.182	0.096	0.060	0.035	0.046	0.002
TS-FB	0.033	0.085	0.009	-0.019	0.029	-0.008	0.003	-0.002	0.000	0.002
ETC	0.045	0.067	0.015	-0.019	0.018	0.001	-0.001	0.001	0.002	0.002
LTS-0.5%					0.002	-0.002	0.005	0.008	0.007	0.002
LTS-1%				0.011	0.002	0.002	0.008	0.010	0.004	0.002
LTS-2%			-0.001	0.007	0.005	0.004	0.005	0.008	0.003	0.002
LTS-5%		0.003	-0.003	0.005	0.001	0.010	0.005	0.004	-0.001	0.002
LTS-10%	0.003	-0.001	-0.001	0.000	-0.001	0.008	0.001	0.002	-0.001	0.002
LTS-15%	0.001	-0.001	-0.002	-0.001	-0.002	0.005	0.002	0.003	0.001	0.002
LTS-20%	-0.001	0.002	-0.002	0.003	-0.002	0.003	0.001	0.000	0.000	0.002
$\sigma = 20$										
TS	0.541	0.545	0.507	0.478	0.421	0.322	0.226	0.131	0.025	-0.003
TS-IPW	0.437	0.401	0.305	0.220	0.156	0.082	0.039	0.029	0.023	-0.003
TS-FB	-0.040	0.009	-0.015	0.011	-0.020	-0.013	0.005	0.003	-0.002	-0.003
ETC	-0.027	-0.008	-0.012	0.003	-0.016	-0.010	0.000	-0.002	-0.004	-0.003
LTS-0.5%					0.002	-0.002	-0.003	0.005	0.000	-0.003
LTS-1%				-0.002	0.003	-0.001	-0.003	0.005	0.000	-0.003
LTS-2%			0.001	-0.002	0.001	0.005	-0.002	0.006	0.003	-0.003
LTS-5%		0.004	0.003	0.000	0.002	0.002	0.001	0.003	0.002	-0.003
LTS-10%	-0.005	0.000	0.002	0.002	-0.001	0.003	0.000	0.001	0.002	-0.003
LTS-15%	-0.005	0.001	0.002	0.000	0.000	0.001	0.001	0.001	0.002	-0.003
LTS-20%	-0.004	-0.001	0.000	-0.001	-0.001	-0.001	0.000	0.001	0.002	-0.003
$\sigma = 25$										

Table A.2: Bias for different strategies ($n = 10,000$) (*continued*)

strategy	Batch size									
	10	20	50	100	200	500	1000	2000	5000	10000
TS	0.603	0.588	0.558	0.511	0.463	0.358	0.260	0.159	0.036	0.006
TS-IPW	0.484	0.409	0.314	0.217	0.131	0.061	0.027	0.018	-0.006	0.006
TS-FB	0.002	0.025	0.032	0.019	0.007	0.005	-0.003	0.007	-0.004	0.006
ETC	-0.048	0.049	-0.001	0.000	-0.003	0.001	0.006	0.007	0.000	0.006
LTS-0.5%					0.003	0.003	-0.006	-0.009	-0.004	0.006
LTS-1%				0.003	-0.001	-0.001	-0.004	-0.009	-0.007	0.006
LTS-2%			0.010	0.003	-0.001	0.001	-0.003	-0.008	-0.002	0.006
LTS-5%		0.001	0.007	0.003	0.002	0.001	-0.004	-0.003	-0.004	0.006
LTS-10%	-0.002	0.003	0.000	0.004	0.002	0.000	-0.002	-0.001	-0.003	0.006
LTS-15%	0.000	0.003	0.005	0.004	0.001	0.002	-0.002	0.001	-0.002	0.006
LTS-20%	0.002	0.000	0.005	0.004	0.002	0.002	-0.002	0.003	-0.003	0.006
$\sigma = 30$										
TS	0.628	0.591	0.583	0.527	0.477	0.386	0.287	0.176	0.039	-0.002
TS-IPW	0.498	0.401	0.312	0.197	0.117	0.061	0.039	0.021	-0.012	-0.002
TS-FB	-0.069	-0.018	0.016	0.030	0.000	-0.014	0.013	-0.008	-0.001	-0.002
ETC	0.109	0.011	-0.032	-0.002	0.020	0.001	0.010	-0.002	-0.002	-0.002
LTS-0.5%					0.000	0.004	0.002	0.001	-0.011	-0.002
LTS-1%				0.003	-0.003	0.007	0.002	0.000	-0.010	-0.002
LTS-2%			0.002	-0.007	-0.005	0.002	-0.002	-0.003	-0.008	-0.002
LTS-5%		0.014	0.000	-0.004	-0.005	0.002	-0.005	0.000	-0.005	-0.002
LTS-10%	0.010	0.005	0.002	-0.007	0.001	-0.003	-0.001	0.001	-0.003	-0.002
LTS-15%	0.008	0.007	0.000	-0.006	-0.003	-0.002	-0.001	0.001	-0.002	-0.002
LTS-20%	0.006	0.007	0.000	-0.006	0.000	-0.001	-0.003	0.002	-0.002	-0.002

Notes: TS: Thompson sampling with $\hat{\tau}_0$, TS-IPW: Thompson sampling with $\hat{\tau}_{IPW}$, TS-FB: Thompson sampling with $\hat{\tau}_{FB}$, ETC: Explore-then-commit with $\hat{\tau}_0$, LTS-X%: Limited Thompson sampling with X% limitation and $\hat{\tau}_{IPW}$. Number of simulations = 10,000 for $\sigma < 10$, 20,000 for $10 \leq \sigma < 20$ and 50,000 for $\sigma \geq 20$.

Table A.3: MSE for different strategies ($n = 10,000$)

strategy	Batch size									
	10	20	50	100	200	500	1000	2000	5000	10000
$\sigma = 1$										
TS	0.118	0.068	0.036	0.020	0.010	0.004	0.002	0.001	0.001	0.000
TS-IPW	0.185	0.154	0.062	0.022	0.010	0.004	0.002	0.001	0.001	0.000
TS-FB	0.400	0.198	0.081	0.040	0.021	0.008	0.004	0.002	0.001	0.000
ETC	0.201	0.099	0.040	0.020	0.010	0.004	0.002	0.001	0.001	0.000
LTS-0.5%					0.019	0.020	0.018	0.016	0.010	0.000
LTS-1%				0.010	0.010	0.010	0.009	0.008	0.005	0.000
LTS-2%			0.005	0.005	0.005	0.005	0.005	0.004	0.003	0.000
LTS-5%		0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.001	0.000

Table A.3: MSE for different strategies ($n = 10,000$) (*continued*)

strategy	Batch size									
	10	20	50	100	200	500	1000	2000	5000	10000
LTS-10%	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.000
LTS-15%	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.000
LTS-20%	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.000
$\sigma = 2$										
TS	0.248	0.192	0.104	0.059	0.037	0.017	0.009	0.004	0.002	0.002
TS-IPW	0.319	0.346	0.384	0.379	0.201	0.027	0.009	0.004	0.002	0.002
TS-FB	1.576	0.807	0.323	0.160	0.081	0.032	0.016	0.008	0.003	0.002
ETC	0.793	0.411	0.162	0.082	0.040	0.017	0.009	0.004	0.002	0.002
LTS-0.5%					0.079	0.081	0.073	0.065	0.040	0.002
LTS-1%				0.039	0.040	0.039	0.036	0.032	0.021	0.002
LTS-2%			0.020	0.020	0.020	0.020	0.018	0.016	0.011	0.002
LTS-5%		0.008	0.009	0.008	0.009	0.008	0.008	0.007	0.005	0.002
LTS-10%	0.004	0.005	0.005	0.005	0.004	0.004	0.004	0.004	0.003	0.002
LTS-15%	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.002	0.002
LTS-20%	0.003	0.003	0.003	0.003	0.002	0.002	0.002	0.002	0.002	0.002
$\sigma = 5$										
TS	0.637	0.548	0.388	0.259	0.159	0.076	0.049	0.028	0.013	0.010
TS-IPW	0.701	0.731	0.762	0.902	1.208	1.821	1.697	0.546	0.016	0.010
TS-FB	9.965	4.994	1.994	0.989	0.515	0.197	0.101	0.051	0.020	0.010
ETC	5.109	2.567	1.019	0.507	0.261	0.104	0.055	0.028	0.013	0.010
LTS-0.5%					0.407	0.454	0.427	0.400	0.254	0.010
LTS-1%				0.224	0.221	0.227	0.223	0.204	0.130	0.010
LTS-2%			0.115	0.119	0.119	0.118	0.112	0.106	0.068	0.010
LTS-5%		0.050	0.052	0.051	0.051	0.050	0.048	0.045	0.031	0.010
LTS-10%	0.027	0.028	0.027	0.027	0.027	0.027	0.026	0.024	0.019	0.010
LTS-15%	0.021	0.020	0.020	0.019	0.019	0.019	0.019	0.018	0.015	0.010
LTS-20%	0.015	0.016	0.015	0.015	0.016	0.015	0.015	0.015	0.013	0.010
$\sigma = 10$										
TS	1.355	1.072	0.887	0.656	0.440	0.238	0.141	0.088	0.052	0.040
TS-IPW	1.461	1.294	1.384	1.515	1.722	2.508	3.616	4.936	3.608	0.040
TS-FB	39.354	19.621	8.090	4.034	1.988	0.802	0.401	0.202	0.080	0.040
ETC	19.620	9.940	4.059	2.010	1.001	0.407	0.210	0.111	0.053	0.040
LTS-0.5%					0.989	1.047	1.129	1.199	0.929	0.040
LTS-1%				0.628	0.624	0.646	0.661	0.673	0.492	0.040
LTS-2%			0.380	0.380	0.373	0.382	0.384	0.364	0.268	0.040
LTS-5%		0.186	0.191	0.183	0.182	0.183	0.175	0.166	0.123	0.040
LTS-10%	0.105	0.106	0.105	0.105	0.103	0.100	0.099	0.095	0.075	0.040
LTS-15%	0.083	0.075	0.076	0.077	0.075	0.073	0.073	0.070	0.059	0.040
LTS-20%	0.062	0.061	0.061	0.062	0.060	0.059	0.060	0.058	0.051	0.040
$\sigma = 15$										
TS	2.571	2.109	1.566	1.173	0.842	0.477	0.292	0.178	0.109	0.091
TS-IPW	2.666	2.407	2.188	2.220	2.376	2.788	3.776	5.158	5.206	0.091

Table A.3: MSE for different strategies ($n = 10,000$) (*continued*)

strategy	Batch size									
	10	20	50	100	200	500	1000	2000	5000	10000
TS-FB	90.993	44.865	17.786	8.885	4.576	1.777	0.900	0.452	0.181	0.091
ETC	45.523	22.524	8.949	4.504	2.271	0.916	0.480	0.253	0.121	0.091
LTS-0.5%					1.484	1.486	1.520	1.625	1.294	0.091
LTS-1%				1.053	1.028	1.011	1.021	1.027	0.796	0.091
LTS-2%			0.684	0.676	0.674	0.661	0.656	0.645	0.483	0.091
LTS-5%		0.373	0.373	0.366	0.366	0.359	0.345	0.334	0.253	0.091
LTS-10%	0.226	0.222	0.224	0.221	0.221	0.219	0.209	0.203	0.163	0.091
LTS-15%	0.180	0.165	0.166	0.164	0.163	0.164	0.158	0.154	0.131	0.091
LTS-20%	0.137	0.136	0.137	0.135	0.134	0.136	0.130	0.129	0.115	0.091
$\sigma = 20$										
TS	4.045	3.241	2.359	1.846	1.301	0.791	0.490	0.312	0.191	0.160
TS-IPW	4.121	3.481	3.039	2.952	2.943	3.234	3.647	4.817	4.828	0.160
TS-FB	160.604	79.991	31.699	15.892	7.963	3.204	1.585	0.812	0.321	0.160
ETC	79.498	39.556	15.826	7.942	4.007	1.641	0.835	0.446	0.215	0.160
LTS-0.5%					1.943	1.844	1.821	1.873	1.480	0.160
LTS-1%				1.463	1.384	1.337	1.317	1.310	0.991	0.160
LTS-2%			1.016	0.992	0.972	0.946	0.912	0.885	0.666	0.160
LTS-5%		0.596	0.587	0.573	0.573	0.556	0.539	0.513	0.394	0.160
LTS-10%	0.375	0.373	0.367	0.370	0.371	0.358	0.351	0.333	0.275	0.160
LTS-15%	0.307	0.284	0.286	0.282	0.284	0.277	0.271	0.262	0.227	0.160
LTS-20%	0.238	0.237	0.236	0.238	0.236	0.230	0.228	0.221	0.201	0.160
$\sigma = 25$										
TS	5.833	4.810	3.600	2.714	1.956	1.181	0.757	0.486	0.299	0.249
TS-IPW	5.831	5.002	4.337	4.033	3.712	3.741	4.020	4.813	4.535	0.249
TS-FB	251.082	124.151	49.360	24.801	12.551	5.030	2.480	1.239	0.497	0.249
ETC	124.807	62.439	24.887	12.560	6.382	2.570	1.310	0.689	0.335	0.249
LTS-0.5%					2.490	2.297	2.154	2.144	1.604	0.249
LTS-1%				1.927	1.804	1.713	1.636	1.552	1.172	0.249
LTS-2%			1.397	1.340	1.300	1.251	1.201	1.109	0.843	0.249
LTS-5%		0.840	0.824	0.806	0.811	0.785	0.754	0.701	0.543	0.249
LTS-10%	0.563	0.548	0.544	0.543	0.544	0.527	0.512	0.484	0.401	0.249
LTS-15%	0.466	0.425	0.429	0.425	0.429	0.420	0.409	0.393	0.341	0.249
LTS-20%	0.362	0.359	0.361	0.360	0.361	0.358	0.351	0.339	0.305	0.249
$\sigma = 30$										
TS	9.055	7.077	5.094	3.642	2.670	1.693	1.085	0.701	0.432	0.363
TS-IPW	9.056	7.285	6.168	4.952	4.528	4.539	4.597	5.099	4.519	0.363
TS-FB	364.392	178.634	72.500	35.187	18.079	7.213	3.605	1.810	0.723	0.363
ETC	180.242	89.158	35.970	18.149	9.123	3.720	1.891	1.001	0.480	0.363
LTS-0.5%					3.065	2.808	2.590	2.392	1.862	0.363
LTS-1%				2.493	2.289	2.151	2.018	1.845	1.395	0.363
LTS-2%			1.835	1.748	1.693	1.622	1.528	1.393	1.051	0.363
LTS-5%		1.145	1.114	1.081	1.059	1.060	1.007	0.920	0.727	0.363

Table A.3: MSE for different strategies ($n = 10,000$) (*continued*)

strategy	Batch size									
	10	20	50	100	200	500	1000	2000	5000	10000
LTS-10%	0.772	0.765	0.755	0.743	0.751	0.731	0.705	0.660	0.561	0.363
LTS-15%	0.646	0.604	0.604	0.602	0.606	0.587	0.573	0.548	0.487	0.363
LTS-20%	0.516	0.515	0.513	0.513	0.518	0.504	0.494	0.479	0.441	0.363

Notes: TS: Thompson sampling with $\hat{\tau}_0$, TS-IPW: Thompson sampling with $\hat{\tau}_{IPW}$, TS-FB: Thompson sampling with $\hat{\tau}_{FB}$, ETC: Explore-then-commit with $\hat{\tau}_0$, LTS-X%: Limited Thompson sampling with X% limitation and $\hat{\tau}_{IPW}$. Number of simulations = 10,000 for $\sigma < 10$, 20,000 for $10 \leq \sigma < 20$ and 50,000 for $\sigma \geq 20$.

Table A.4: Expected welfare for different strategies ($\sigma = 10$)

allocation	Batch size									
	10	20	50	100	200	500	1000	2000	5000	10,000
$n = 2000$										
TS	1631	1640	1638	1627	1605	1526	1367	1000		
ETC	1120	1196	1268	1363	1469	1550	1442	1000		
LTS-0.5%					1604	1525	1367	1000		
LTS-1%				1625	1603	1524	1366	1000		
LTS-2%			1630	1621	1599	1521	1363	1000		
LTS-5%		1612	1609	1602	1581	1504	1351	1000		
LTS-10%	1558	1566	1565	1557	1536	1466	1323	1000		
LTS-15%	1502	1510	1510	1503	1483	1419	1290	1000		
LTS-20%	1444	1449	1447	1442	1424	1367	1253	1000		
$n = 10000$										
TS	9372	9382	9389	9392	9383	9321	9178	8820	7469	5000
ETC	5626	5840	6375	6946	7533	8496	9004	8896	7498	5000
LTS-0.5%					9371	9308	9163	8800	7451	5000
LTS-1%				9358	9351	9288	9140	8774	7430	5000
LTS-2%			9302	9309	9301	9234	9082	8713	7384	5000
LTS-5%		9096	9106	9111	9097	9026	8871	8508	7241	5000
LTS-10%	8694	8707	8713	8711	8695	8623	8476	8139	6995	5000
LTS-15%	8260	8278	8280	8276	8258	8193	8060	7758	6747	5000
LTS-20%	7819	7829	7829	7826	7809	7750	7634	7370	6498	5000
$n = 20000$										
TS	19,268	19,272	19,288	19,298	19,292	19,243	19,119	18,791	17,465	14,998
ETC	11,181	11,714	12,871	13,743	15,346	17,252	18,468	18,757	17,496	15,000
LTS-0.5%					19,258	19,210	19,079	18,737	17,399	14,949
LTS-1%				19,202	19,204	19,153	19,017	18,666	17,329	14,899
LTS-2%			19,063	19,075	19,072	19,014	18,869	18,510	17,183	14,800
LTS-5%		18,583	18,597	18,598	18,588	18,520	18,368	18,009	16,740	14,500
LTS-10%	17,692	17,702	17,712	17,706	17,692	17,622	17,477	17,141	15,995	14,000
LTS-15%	16,765	16,779	16,781	16,773	16,758	16,692	16,561	16,259	15,247	13,500

Table A.4: Expected welfare for different strategies ($\sigma = 10$) (*continued*)

allocation	Batch size									
	10	20	50	100	200	500	1000	2000	5000	10,000
LTS-20%	15,819	15,827	15,832	15,822	15,810	15,750	15,635	15,371	14,498	13,000
$n = 40000$										
TS	39,152	39,167	39,182	39,210	39,197	39,166	39,058	38,757	37,460	34,998
ETC	22,203	23,374	25,845	28,008	30,444	34,583	37,421	38,571	37,486	35,000
LTS-0.5%					39,104	39,072	38,954	38,625	37,297	34,849
LTS-1%				38,977	38,972	38,932	38,803	38,460	37,127	34,700
LTS-2%			38,655	38,672	38,660	38,607	38,464	38,108	36,782	34,400
LTS-5%		37,582	37,598	37,608	37,586	37,518	37,366	37,008	35,739	33,500
LTS-10%	35,690	35,704	35,715	35,712	35,691	35,619	35,474	35,141	33,994	32,000
LTS-15%	33,763	33,779	33,782	33,776	33,757	33,690	33,559	33,260	32,246	30,500
LTS-20%	31,821	31,829	31,832	31,825	31,807	31,747	31,633	31,372	30,498	29,000

Notes: TS: Thompson sampling, ETC: Explore-then-commit, LTS-X%: Limited Thompson sampling with X% limitation. Expected welfare is calculated as the average of the sum of outcomes ($\sum_{i=1}^n Y$) across the simulation runs. Number of simulations = 20,000 for $n = 10,000$, and 10,000 otherwise.

Table A.5: Bias for different strategies ($\sigma = 10$)

strategy	Batch size									
	10	20	50	100	200	500	1000	2000	5000	10000
$n = 2000$										
TS	0.527	0.514	0.427	0.347	0.253	0.111	0.033	0.002		
TS-IPW	0.429	0.370	0.242	0.164	0.092	0.025	0.019	0.002		
TS-FB	-0.001	0.076	0.001	0.022	0.002	0.001	0.002	0.002		
ETC	0.010	0.065	0.001	0.025	0.013	0.003	0.003	0.002		
LTS-0.5%					-0.003	-0.008	0.002	0.002		
LTS-1%				-0.008	-0.005	-0.006	-0.002	0.002		
LTS-2%			-0.001	-0.003	0.006	-0.008	0.007	0.002		
LTS-5%		0.008	-0.003	-0.001	0.000	-0.011	-0.001	0.002		
LTS-10%	-0.001	0.007	-0.002	-0.004	0.000	0.000	-0.002	0.002		
LTS-15%	0.000	0.007	-0.005	-0.002	0.001	0.002	-0.003	0.002		
LTS-20%	0.005	0.007	-0.006	0.002	-0.001	0.003	-0.001	0.002		
$n = 10000$										
TS	0.419	0.394	0.383	0.342	0.279	0.182	0.099	0.035	0.003	0.002
TS-IPW	0.375	0.337	0.306	0.261	0.190	0.140	0.115	0.100	-0.007	0.002
TS-FB	-0.041	-0.043	0.008	0.030	-0.015	0.003	0.004	0.001	-0.001	0.002
ETC	-0.066	-0.023	0.002	0.007	-0.004	0.003	0.003	0.001	0.000	0.002
LTS-0.5%					-0.006	-0.001	-0.004	-0.010	0.001	0.002
LTS-1%				-0.010	-0.003	-0.001	-0.001	-0.005	-0.003	0.002
LTS-2%			-0.010	-0.002	-0.008	0.001	-0.005	-0.006	-0.007	0.002
LTS-5%		-0.004	-0.006	-0.003	0.003	-0.001	-0.005	0.003	-0.005	0.002

Table A.5: Bias for different strategies ($\sigma = 10$) (*continued*)

strategy	Batch size									
	10	20	50	100	200	500	1000	2000	5000	10000
LTS-10%	-0.005	-0.004	-0.003	0.002	-0.001	-0.004	-0.001	0.002	-0.004	0.002
LTS-15%	-0.004	-0.005	0.000	0.001	0.001	-0.002	0.000	0.001	-0.003	0.002
LTS-20%	-0.002	-0.004	0.000	0.001	0.001	-0.001	-0.001	0.001	-0.002	0.002
$n = 20000$										
TS	0.378	0.368	0.343	0.312	0.267	0.184	0.108	0.041	0.005	0.001
TS-IPW	0.352	0.341	0.310	0.271	0.207	0.192	0.163	0.135	0.078	0.009
TS-FB	-0.047	0.027	0.015	-0.032	0.022	-0.006	0.001	0.003	0.002	0.001
ETC	0.000	0.023	-0.004	-0.008	0.014	-0.004	0.000	0.001	0.002	0.001
LTS-0.5%					-0.010	0.000	-0.010	-0.009	0.009	0.009
LTS-1%				-0.003	-0.003	-0.002	-0.004	-0.005	0.012	0.007
LTS-2%			-0.005	-0.003	0.001	0.005	0.000	-0.001	0.010	0.001
LTS-5%		0.002	-0.003	-0.002	0.000	0.005	0.000	0.000	0.000	0.000
LTS-10%	0.001	0.002	-0.003	-0.001	0.002	0.001	0.001	-0.001	-0.001	0.001
LTS-15%	0.001	0.001	-0.004	-0.002	0.001	0.001	-0.001	0.000	-0.002	0.001
LTS-20%	0.000	-0.001	-0.001	0.001	0.001	0.002	-0.002	-0.002	0.000	0.000
$n = 40000$										
TS	0.305	0.306	0.300	0.289	0.251	0.189	0.117	0.047	0.002	0.000
TS-IPW	0.291	0.294	0.281	0.285	0.265	0.228	0.205	0.201	0.069	0.011
TS-FB	-0.078	-0.027	0.047	0.036	0.005	0.000	0.003	0.003	-0.005	0.001
ETC	-0.001	-0.005	0.061	0.025	0.010	0.001	-0.001	0.000	-0.002	0.000
LTS-0.5%					0.003	0.003	0.006	0.000	0.002	-0.001
LTS-1%				0.000	0.002	0.001	0.007	-0.003	0.005	0.002
LTS-2%			0.000	0.001	0.001	0.002	0.001	-0.001	-0.002	-0.001
LTS-5%		0.001	0.000	0.004	0.003	0.000	0.001	0.000	-0.003	-0.001
LTS-10%	-0.001	0.000	0.003	0.002	0.003	0.001	0.001	-0.002	-0.002	0.000
LTS-15%	-0.001	0.000	0.002	0.001	0.001	0.001	0.000	-0.002	-0.002	0.001
LTS-20%	-0.001	0.000	0.003	0.002	0.001	0.000	-0.001	-0.002	-0.002	0.001

Notes: TS: Thompson sampling with $\hat{\tau}_0$, TS-IPW: Thompson sampling with $\hat{\tau}_{IPW}$, TS-FB: Thompson sampling with $\hat{\tau}_{FB}$, ETC: Explore-then-commit with $\hat{\tau}_0$, LTS-X%: Limited Thompson sampling with X% limitation and $\hat{\tau}_{IPW}$. Number of simulations = 20,000 for $n = 10,000$, and 10,000 otherwise.

Table A.6: MSE for different strategies ($\sigma = 10$)

strategy	Batch size									
	10	20	50	100	200	500	1000	2000	5000	10000
$n = 2000$										
TS	3.045	2.203	1.399	0.992	0.616	0.343	0.243	0.200		
TS-IPW	3.210	2.629	2.303	2.401	2.464	2.604	2.227	0.200		
TS-FB	40.238	19.600	7.975	4.059	1.995	0.789	0.409	0.200		
ETC	20.402	10.048	4.107	2.064	1.047	0.462	0.271	0.200		

Table A.6: MSE for different strategies ($\sigma = 10$) (*continued*)

strategy	Batch size									
	10	20	50	100	200	500	1000	2000	5000	10000
LTS-0.5%					2.147	1.994	1.560	0.200		
LTS-1%				1.620	1.509	1.375	1.087	0.200		
LTS-2%			1.176	1.120	1.089	0.943	0.769	0.200		
LTS-5%		0.703	0.688	0.681	0.651	0.566	0.471	0.200		
LTS-10%	0.440	0.451	0.443	0.444	0.432	0.386	0.335	0.200		
LTS-15%	0.365	0.348	0.345	0.352	0.338	0.308	0.279	0.200		
LTS-20%	0.289	0.297	0.290	0.298	0.284	0.267	0.253	0.200		
$n = 10000$										
TS	1.355	1.072	0.887	0.656	0.440	0.238	0.141	0.088	0.052	0.040
TS-IPW	1.461	1.294	1.384	1.515	1.722	2.508	3.616	4.936	3.608	0.040
TS-FB	39.354	19.621	8.090	4.034	1.988	0.802	0.401	0.202	0.080	0.040
ETC	19.620	9.940	4.059	2.010	1.001	0.407	0.210	0.111	0.053	0.040
LTS-0.5%					0.989	1.047	1.129	1.199	0.929	0.040
LTS-1%				0.628	0.624	0.646	0.661	0.673	0.492	0.040
LTS-2%			0.380	0.380	0.373	0.382	0.384	0.364	0.268	0.040
LTS-5%		0.186	0.191	0.183	0.182	0.183	0.175	0.166	0.123	0.040
LTS-10%	0.105	0.106	0.105	0.105	0.103	0.100	0.099	0.095	0.075	0.040
LTS-15%	0.083	0.075	0.076	0.077	0.075	0.073	0.073	0.070	0.059	0.040
LTS-20%	0.062	0.061	0.061	0.062	0.060	0.059	0.060	0.058	0.051	0.040
$n = 20000$										
TS	1.102	1.007	0.721	0.565	0.409	0.230	0.132	0.079	0.043	0.027
TS-IPW	1.187	1.178	1.080	1.181	1.374	2.142	3.360	5.423	4.824	1.119
TS-FB	40.125	20.155	7.858	3.982	1.955	0.795	0.394	0.200	0.078	0.039
ETC	20.110	9.920	3.951	2.013	0.986	0.407	0.205	0.104	0.045	0.027
LTS-0.5%					0.630	0.668	0.685	0.759	0.735	0.526
LTS-1%				0.383	0.368	0.388	0.394	0.425	0.378	0.266
LTS-2%			0.216	0.217	0.216	0.216	0.222	0.220	0.194	0.140
LTS-5%		0.098	0.101	0.100	0.098	0.098	0.095	0.095	0.085	0.065
LTS-10%	0.053	0.053	0.054	0.055	0.054	0.053	0.052	0.052	0.047	0.038
LTS-15%	0.042	0.038	0.039	0.038	0.037	0.037	0.037	0.037	0.035	0.030
LTS-20%	0.031	0.030	0.031	0.031	0.030	0.030	0.030	0.030	0.029	0.026
$n = 40000$										
TS	0.709	0.704	0.570	0.496	0.372	0.222	0.129	0.076	0.040	0.023
TS-IPW	0.762	0.821	0.806	0.996	1.164	1.754	2.879	4.926	5.361	1.328
TS-FB	40.465	19.838	7.935	3.991	2.001	0.806	0.397	0.202	0.080	0.040
ETC	19.978	10.142	4.042	2.023	0.983	0.405	0.208	0.104	0.043	0.023
LTS-0.5%					0.376	0.402	0.400	0.413	0.432	0.383
LTS-1%				0.214	0.217	0.219	0.215	0.218	0.221	0.189
LTS-2%			0.117	0.120	0.118	0.115	0.118	0.115	0.112	0.099
LTS-5%		0.050	0.053	0.051	0.051	0.051	0.050	0.049	0.047	0.042
LTS-10%	0.027	0.027	0.028	0.028	0.028	0.028	0.027	0.027	0.025	0.023
LTS-15%	0.022	0.019	0.020	0.020	0.020	0.020	0.020	0.019	0.018	0.017

Table A.6: MSE for different strategies ($\sigma = 10$) (*continued*)

strategy	Batch size									
	10	20	50	100	200	500	1000	2000	5000	10000
LTS-20%	0.016	0.015	0.016	0.016	0.016	0.016	0.016	0.015	0.015	0.014

Notes: TS: Thompson sampling with $\hat{\tau}_0$, TS-IPW: Thompson sampling with $\hat{\tau}_{IPW}$, TS-FB: Thompson sampling with $\hat{\tau}_{FB}$, ETC: Explore-then-commit with $\hat{\tau}_0$, LTS-X%: Limited Thompson sampling with X% limitation and $\hat{\tau}_{IPW}$. Number of simulations = 20,000 for $n = 10,000$, and 10,000 otherwise.

Appendix B

Appendix for Chapter 2

B.1 Comparison of methodologies in the literature

Table B.1: Comparing the methodologies of the literature

	Rohwedder and Willis (2010)	Mazzonna and Peracchi (2012)	Bonsang et al. (2012)
$f(R_i; \beta)$	$\tilde{\beta} \mathbf{1}(R_i > 0)$	βR_i	$\tilde{\beta} \mathbf{1}(R_i \geq 1)$
X_i^*	-	age, gender, country dummies	age, age squared, individual fixed effects
Z_i	normal & early eligibility dummies (no variation within country-gender cells)	normal & early eligibility dummies (some variation within country-gender cells)	social security eligibility & normal retirement age
used data	2004 waves of SHARE & HRS & ELSA	2004 wave of SHARE	6 waves (1998-2008) of HRS

Notes: Using a simple retirement dummy instead of years in retirement is equivalent to estimating the average effect conditional on the average time spent in retirement in the sample (or the average time excluding fresh retirees as in the case of Bonsang et al. (2012)), i.e. $\tilde{\beta} = \beta \bar{R}_i$.

B.2 Additional tables for the replication exercises

This appendix contains summary tables of the first stage regressions for the replications exercises. It also contains a robustness check summary, replicating the method of Bonsang et al. (2012) on various subsamples.

Table B.2: Comparing the methodology of Rohwedder and Willis (2010) by two versions of the instrumental variable: first stage

	(1) Rohwedder and Willis (2010)	(2) Mazzonna and Peracchi (2012)
Eligible for early benefits	0.323*** (0.028)	0.246*** (0.027)
Eligible for full benefits	0.165*** (0.014)	0.185*** (0.014)
Constant	0.375*** (0.027)	0.439*** (0.025)
Observations	4,464	4,464
Adjusted R^2	0.0643	0.065

Notes: Both results are from the first stage estimation of $S_i = \alpha + \beta \mathbf{1}(R_i > 0) + u_i$ where the retirement dummy is instrumented by early and normal eligibility dummies. The corresponding coefficients for early and full benefits in Rohwedder and Willis (2010) are 0.19*** and 0.16**, respectively, with the adjusted R^2 being 0.059 on a sample of 8,828 observations.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table B.3: Moving from the strategy of Rohwedder and Willis (2010) to that of Mazzonna and Peracchi (2012): first stage

	(1) aged 60-64	(2) aged 50-70	(3) + worked at 50	(4) + age	(5) + country
Years after early eligibility	0.303*** (0.071)	0.016 (0.033)	0.183*** (0.013)	0.129*** (0.013)	0.033 (0.025)
Years after normal eligibility	0.580*** (0.079)	0.581*** (0.033)	0.266*** (0.013)	0.144*** (0.015)	0.166*** (0.018)
Age				0.200*** (0.013)	0.274*** (0.024)
Constant	6.437*** (0.38)	8.310*** (0.18)	3.696*** (0.071)	-8.659*** (0.83)	-11.793*** (1.42)
Country dummies	No	No	No	No	Yes
Observations	4,052	17,448	14,052	14,052	14,052
Adjusted R^2	0.0546	0.1561	0.4513	0.4599	0.4784

Notes: All results are from the first stage estimation of $S_i = \alpha + \beta R_i + \mathbf{X}_i^{*'} \gamma^* + u_i$ where the retirement dummy is instrumented by early and normal eligibility dummies.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table B.4: First stages, FE-IV estimation mimicing [Bonsang et al. \(2012\)](#)

	(1) Retired	(2) Retired for at least 1 year
Eligible for normal retirement	0.104*** (0.0072)	0.110*** (0.0076)
Eligible for early retirement	0.059*** (0.007)	0.059*** (0.007)
Age	0.0005 (0.006)	0.005 (0.006)
Age (sq.)	0.00007 (0.00005)	0.00004 (0.00005)
Observations	41,476	37,374
Within- R^2	0.0575	0.0689

Notes: The results are from the first stage estimation of $S_i = \alpha + \beta g(R_i) + u_i$ where $g(R_i) = \mathbf{1}(YR_i > 0)$ or $\mathbf{1}(YR_i \geq 1)$ and these retirement dummies are instrumented by early and normal eligibility dummies. The corresponding coefficients for the eligibility dummies are 0.11*** and 0.07*** in [Bonsang et al. \(2012\)](#) on a sample of 54,377 observations with a within- R^2 of 0.242.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table B.5: Replication of [Bonsang et al. \(2012\)](#) on various subsamples

	(1) wave 1-2	(2) wave 2-4	(3) wave 1-2-4
Retired for at least 1 year	0.394 (0.65)	0.248 (0.19)	0.127 (0.15)
Age	0.105*** (0.038)	0.201*** (0.026)	0.202*** (0.024)
Age (sq.)	-0.001** (0.0003)	-0.002*** (0.0002)	-0.001*** (0.0002)
Observations	24,470	19,362	19,746
Weak IV F statistic	8.28	78.28	106.80

Notes: The results are from the second stage estimation of $S_i = \alpha + a_i + \beta g(R_i) + u_i$ where $g(R_i) = \mathbf{1}(R_i \geq 1)$ and this dummy is instrumented by early and normal eligibility dummies. The corresponding first stage estimates are summarized in [Table B.6](#). Weak IV F statistic is calculated according to [Angrist and Pischke \(2008\)](#). [Stock et al. \(2002\)](#) suggest that an F below 10 should make us worry about the potential bias in the IV estimation.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table B.6: First stages, replication of [Bonsang et al. \(2012\)](#) on various subsamples

	(1) wave 1-2	(2) wave 2-4	(3) wave 1-2-4
Eligible for normal retirement	0.104*** (0.0072)	0.110*** (0.0076)	0.133*** (0.0097)
Eligible for early retirement	0.059*** (0.007)	0.059*** (0.007)	0.058*** (0.009)
Age	0.001 (0.006)	0.005 (0.006)	0.027*** (0.009)
Age (sq.)	0.0001 (0.0000)	0.0000 (0.0000)	-0.0001 (0.0001)
Observations	41,476	37,374	19,746

Notes: The results are from the first stage estimation of $S_i = \alpha + \beta g(R_i) + u_i$ where $g(R_i) = \mathbf{1}(YR_i \geq 1)$ and this dummy is instrumented by early and normal eligibility dummies.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

B.3 Detailed estimation tables for my strategy

This section contains the estimation results of my strategy for each cognitive score, for each period, along with the corresponding first stages. The main results of these tables are summarized in Figure 2.2 in Section 2.5.

Table B.7: Panel estimation: change in total word recall score between wave 1 and 4:
first stage

	(1) Years in retirement	(2) Years in retirement	(3) Years in retirement
Years after early eligibility	0.270*** (0.015)	0.249*** (0.015)	0.215*** (0.019)
Years after normal eligibility	0.071*** (0.015)	0.094*** (0.015)	0.132*** (0.019)
Years elapsed	0.270*** (0.067)	0.255*** (0.066)	0.377*** (0.096)
Female		-0.408*** (0.052)	-0.366*** (0.051)
Constant	-0.594 (0.47)	-0.218 (0.47)	-0.041 (0.67)
Country dummies	No	No	Yes
Observations	6,394	6,394	6,394

Notes: The results are from the first stage estimation of $\Delta S_i = \alpha^* + \beta \Delta R_i + \mathbf{W}_i^{*'} \boldsymbol{\nu} + \Delta \tilde{u}_i$ where $\Delta M_i = M_{i,w=4} - M_{i,w=1}$, S_i = Total word recall_{*i*} and ΔR_i is instrumented by distance from early and normal retirement age.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table B.8: Panel estimation: change in total word recall score between wave 1 and 2

	(1) 2SLS	(2) 2SLS	(3) 2SLS	(4) OLS	(5) OLS
Years in retirement	-0.019 (0.013)	-0.018 (0.013)	-0.017 (0.013)	-0.017* (0.0090)	-0.015 (0.013)
Years elapsed	0.001 (0.025)	0.002 (0.025)	0.085** (0.040)	0.084** (0.040)	0.084** (0.040)
Female		0.016 (0.020)	0.010 (0.020)	0.010 (0.020)	0.010 (0.020)
Age at first wave					0.000 (0.0024)
Constant	0.016 (0.059)	0.008 (0.060)	-0.094 (0.11)	-0.094 (0.11)	-0.067 (0.18)
Country dummies	No	No	Yes	Yes	Yes
Observations	9,256	9,256	9,256	9,256	9,256
Weak IV F statistic	4272.83	4353.52	4332.50		

Notes: All results are from the estimation of $\Delta S_i = \alpha^* + \beta \Delta R_i + W_i^{*'} \nu + \Delta \tilde{u}_i$ where $\Delta M_i = M_{i,w=2} - M_{i,w=1}$ and $S_i = \text{Total word recall}_i$. For 2SLS, the second stage regressions are reported where ΔR_i is instrumented by the distance from early and normal retirement age, and W include years elapsed, female dummy and country dummies. For OLS, W additionally include age at first wave. The corresponding first stage estimates are summarized in Table B.9.

Weak IV F statistic is calculated according to Angrist and Pischke (2008). Stock et al. (2002) suggest that an F below 10 should make us worry about the potential bias in the IV estimation.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table B.9: Panel estimation: change in total word recall score between wave 1 and 2: first stage

	(1) Years in retirement	(2) Years in retirement	(3) Years in retirement
Years after early eligibility	0.063*** (0.0041)	0.052*** (0.0042)	0.073*** (0.0060)
Years after normal eligibility	0.061*** (0.0041)	0.073*** (0.0043)	0.054*** (0.0059)
Years elapsed	0.388*** (0.021)	0.380*** (0.021)	0.374*** (0.033)
Female		-0.173*** (0.018)	-0.146*** (0.018)
Constant	0.070 (0.056)	0.226*** (0.058)	0.518*** (0.096)
Country dummies	No	No	Yes
Observations	9,256	9,256	9,256

Notes: The results are from the first stage estimation of $\Delta S_i = \alpha^* + \beta \Delta R_i + W_i^{*'} \nu + \Delta \tilde{u}_i$ where $\Delta M_i = M_{i,w=4} - M_{i,w=1}$, $S_i = \text{Total word recall}_i$ and ΔR_i is instrumented by distance from early and normal retirement age.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table B.10: Panel estimation: change in total word recall score between wave 2 and 4

	(1) 2SLS	(2) 2SLS	(3) 2SLS	(4) OLS	(5) OLS
Years in retirement	-0.027*** (0.0079)	-0.028*** (0.0079)	-0.045*** (0.0081)	-0.028*** (0.0057)	-0.012 (0.0082)
Years elapsed	0.040 (0.042)	0.041 (0.042)	-0.029 (0.053)	-0.043 (0.052)	-0.037 (0.052)
Female		-0.003 (0.023)	0.008 (0.023)	0.010 (0.023)	0.003 (0.023)
Age at second wave					-0.008*** (0.0028)
Constant	-0.119 (0.18)	-0.119 (0.18)	0.011 (0.23)	0.020 (0.23)	0.463* (0.28)
Country dummies	No	No	Yes	Yes	Yes
Observations	8,095	8,095	8,095	8,095	8,095
Weak IV <i>F</i> statistic	4185.47	4205.43	4081.89		

Notes: All results are from the estimation of $\Delta S_i = \alpha^* + \beta \Delta R_i + \mathbf{W}_i^{*'} \boldsymbol{\nu} + \Delta \tilde{u}_i$ where $\Delta M_i = M_{i,w=4} - M_{i,w=2}$ and $S_i = \text{Total word recall}_i$. For 2SLS, the second stage regressions are reported where ΔR_i is instrumented by the distance from early and normal retirement age, and \mathbf{W} include years elapsed, female dummy and country dummies. For OLS, \mathbf{W} additionally include age at first wave. The corresponding first stage estimates are summarized in Table B.11.

Weak IV *F* statistic is calculated according to Angrist and Pischke (2008). Stock et al. (2002) suggest that an *F* below 10 should make us worry about the potential bias in the IV estimation.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table B.11: Panel estimation: change in total word recall score between wave 2 and 4: first stage

	(1) Years in retirement	(2) Years in retirement	(3) Years in retirement
Years after early eligibility	0.157*** (0.0086)	0.147*** (0.0087)	0.125*** (0.012)
Years after normal eligibility	0.072*** (0.0088)	0.083*** (0.0090)	0.107*** (0.012)
Years elapsed	-0.055 (0.058)	-0.048 (0.058)	0.234*** (0.072)
Female		-0.189*** (0.032)	-0.168*** (0.032)
Constant	1.310*** (0.25)	1.420*** (0.25)	0.863*** (0.32)
Country dummies	No	No	Yes
Observations	8,095	8,095	8,095

Notes: The results are from the first stage estimation of $\Delta S_i = \alpha^* + \beta \Delta R_i + \mathbf{W}_i^{*'} \boldsymbol{\nu} + \Delta \tilde{u}_i$ where $\Delta M_i = M_{i,w=4} - M_{i,w=2}$, $S_i = \text{Total word recall}_i$ and ΔR_i is instrumented by distance from early and normal retirement age.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table B.12: Panel estimation: change in numeracy score between wave 1 and 4

	(1) 2SLS	(2) 2SLS	(3) 2SLS	(4) OLS	(5) OLS
Years in retirement	-0.008 (0.0053)	-0.008 (0.0052)	-0.008 (0.0055)	-0.009** (0.0040)	-0.010 (0.0059)
Years elapsed	-0.001 (0.029)	-0.001 (0.029)	0.048 (0.045)	0.049 (0.045)	0.049 (0.045)
Female		-0.001 (0.023)	-0.002 (0.023)	-0.002 (0.023)	-0.002 (0.024)
Age at first wave					0.000 (0.0029)
Constant	0.035 (0.19)	0.035 (0.19)	-0.252 (0.31)	-0.252 (0.31)	-0.260 (0.36)
Country dummies	No	No	Yes	Yes	Yes
Observations	6,420	6,420	6,420	6,420	6,420
Weak IV <i>F</i> statistic	3728.74	3787.09	3668.76		

Notes: All results are from the estimation of $\Delta S_i = \alpha^* + \beta \Delta R_i + W_i^{*'} \nu + \Delta \tilde{u}_i$ where $\Delta M_i = M_{i,w=4} - M_{i,w=1}$ and $S_i = \text{Numeracy}_i$. For 2SLS, the second stage regressions are reported where ΔR_i is instrumented by the distance from early and normal retirement age, and W include years elapsed, female dummy and country dummies. For OLS, W additionally include age at first wave. The corresponding first stage estimates are summarized in Table B.13.

Weak IV *F* statistic is calculated according to Angrist and Pischke (2008). Stock et al. (2002) suggest that an *F* below 10 should make us worry about the potential bias in the IV estimation.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table B.13: Panel estimation: change in numeracy score between wave 1 and 4: first stage

	(1) Years in retirement	(2) Years in retirement	(3) Years in retirement
Years after early eligibility	0.270*** (0.015)	0.250*** (0.015)	0.215*** (0.019)
Years after normal eligibility	0.070*** (0.015)	0.093*** (0.015)	0.132*** (0.019)
Years elapsed	0.275*** (0.067)	0.260*** (0.066)	0.382*** (0.096)
Female		-0.409*** (0.052)	-0.366*** (0.051)
Constant	-0.627 (0.46)	-0.250 (0.46)	-0.074 (0.67)
Country dummies	No	No	Yes
Observations	6,420	6,420	6,420

Notes: The results are from the first stage estimation of $\Delta S_i = \alpha^* + \beta \Delta R_i + W_i^{*'} \nu + \Delta \tilde{u}_i$ where $\Delta M_i = M_{i,w=4} - M_{i,w=1}$, $S_i = \text{Numeracy}_i$ and ΔR_i is instrumented by distance from early and normal retirement age.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table B.14: Panel estimation: change in numeracy score between wave 1 and 2

	(1) 2SLS	(2) 2SLS	(3) 2SLS	(4) OLS	(5) OLS
Years in retirement	-0.026** (0.013)	-0.028** (0.013)	-0.028** (0.013)	-0.020** (0.0090)	-0.012 (0.013)
Years elapsed	0.027 (0.025)	0.029 (0.025)	0.018 (0.040)	0.015 (0.040)	0.013 (0.040)
Female		0.029 (0.020)	0.031 (0.020)	0.032 (0.020)	0.030 (0.020)
Age at first wave					-0.002 (0.0024)
Constant	-0.039 (0.059)	-0.053 (0.060)	0.023 (0.11)	0.019 (0.11)	0.154 (0.18)
Country dummies	No	No	Yes	Yes	Yes
Observations	9,295	9,295	9,295	9,295	9,295
Weak IV F statistic	4282.96	4362.29	4341.05		

Notes: All results are from the estimation of $\Delta S_i = \alpha^* + \beta \Delta R_i + W_i^{*'} \nu + \Delta \tilde{u}_i$ where $\Delta M_i = M_{i,w=2} - M_{i,w=1}$ and $S_i = \text{Numeracy}_i$. For 2SLS, the second stage regressions are reported where ΔR_i is instrumented by the distance from early and normal retirement age, and W include years elapsed, female dummy and country dummies. For OLS, W additionally include age at first wave. The corresponding first stage estimates are summarized in Table B.15.

Weak IV F statistic is calculated according to Angrist and Pischke (2008). Stock et al. (2002) suggest that an F below 10 should make us worry about the potential bias in the IV estimation.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table B.15: Panel estimation: change in numeracy score between wave 1 and 2: first stage

	(1) Years in retirement	(2) Years in retirement	(3) Years in retirement
Years after early eligibility	0.063*** (0.0041)	0.052*** (0.0042)	0.073*** (0.0060)
Years after normal eligibility	0.061*** (0.0041)	0.073*** (0.0043)	0.053*** (0.0059)
Years elapsed	0.390*** (0.021)	0.382*** (0.021)	0.375*** (0.033)
Female		-0.172*** (0.018)	-0.144*** (0.018)
Constant	0.064 (0.056)	0.218*** (0.058)	0.515*** (0.096)
Country dummies	No	No	Yes
Observations	9,295	9,295	9,295

Notes: The results are from the first stage estimation of $\Delta S_i = \alpha^* + \beta \Delta R_i + W_i^{*'} \nu + \Delta \tilde{u}_i$ where $\Delta M_i = M_{i,w=2} - M_{i,w=1}$, $S_i = \text{Numeracy}_i$ and ΔR_i is instrumented by distance from early and normal retirement age.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table B.16: Panel estimation: change in numeracy score between wave 2 and 4

	(1) 2SLS	(2) 2SLS	(3) 2SLS	(4) OLS	(5) OLS
Years in retirement	-0.003 (0.0020)	-0.003 (0.0020)	-0.005** (0.0020)	-0.005*** (0.0014)	-0.004* (0.0021)
Years elapsed	0.030*** (0.010)	0.031*** (0.010)	0.018 (0.013)	0.018 (0.013)	0.018 (0.013)
Female		-0.005 (0.0056)	-0.006 (0.0056)	-0.006 (0.0056)	-0.006 (0.0057)
Age at second wave					0.000 (0.00069)
Constant	-0.123*** (0.043)	-0.122*** (0.043)	-0.031 (0.057)	-0.031 (0.057)	-0.011 (0.069)
Country dummies	No	No	Yes	Yes	Yes
Observations	8,110	8,110	8,110	8,110	8,110
Weak IV <i>F</i> statistic	4172.20	4190.63	4072.09		

Notes: All results are from the estimation of $\Delta S_i = \alpha^* + \beta \Delta R_i + W_i^{*'} \nu + \Delta \tilde{u}_i$ where $\Delta M_i = M_{i,w=4} - M_{i,w=2}$ and $S_i = \text{Numeracy}_i$. For 2SLS, the second stage regressions are reported where ΔR_i is instrumented by the distance from early and normal retirement age, and W include years elapsed, female dummy and country dummies. For OLS, W additionally include age at first wave. The corresponding first stage estimates are summarized in Table B.17.

Weak IV *F* statistic is calculated according to Angrist and Pischke (2008). Stock et al. (2002) suggest that an *F* below 10 should make us worry about the potential bias in the IV estimation.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table B.17: Panel estimation: change in numeracy score between wave 2 and 4: first stage

	(1) Years in retirement	(2) Years in retirement	(3) Years in retirement
Years after early eligibility	0.157*** (0.0086)	0.147*** (0.0087)	0.125*** (0.012)
Years after normal eligibility	0.072*** (0.0088)	0.083*** (0.0090)	0.107*** (0.012)
Years elapsed	-0.062 (0.058)	-0.055 (0.058)	0.231*** (0.072)
Female		-0.184*** (0.032)	-0.164*** (0.032)
Constant	1.338*** (0.25)	1.446*** (0.25)	0.872*** (0.32)
Country dummies	No	No	Yes
Observations	8,110	8,110	8,110

Notes: The results are from the first stage estimation of $\Delta S_i = \alpha^* + \beta \Delta R_i + W_i^{*'} \nu + \Delta \tilde{u}_i$ where $\Delta M_i = M_{i,w=4} - M_{i,w=2}$, $S_i = \text{Numeracy}_i$ and ΔR_i is instrumented by distance from early and normal retirement age.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table B.18: Panel estimation: change in fluency score between wave 1 and 4

	(1) 2SLS	(2) 2SLS	(3) 2SLS	(4) OLS	(5) OLS
Years in retirement	-0.038*** (0.0053)	-0.039*** (0.0053)	-0.041*** (0.0055)	-0.024*** (0.0040)	-0.003 (0.0060)
Years elapsed	0.008 (0.030)	0.008 (0.030)	0.058 (0.046)	0.046 (0.046)	0.046 (0.046)
Female		0.018 (0.024)	0.042* (0.024)	0.045* (0.024)	0.036 (0.024)
Age at first wave					-0.014*** (0.0030)
Constant	0.076 (0.20)	0.071 (0.20)	-0.313 (0.31)	-0.311 (0.31)	0.540 (0.36)
Country dummies	No	No	Yes	Yes	Yes
Observations	6,368	6,368	6,368	6,368	6,368
Weak IV <i>F</i> statistic	3733.22	3793.56	3670.54		

Notes: All results are from the estimation of $\Delta S_i = \alpha^* + \beta \Delta R_i + W_i^{*'} \nu + \Delta \tilde{u}_i$ where $\Delta M_i = M_{i,w=4} - M_{i,w=1}$ and $S_i = \text{Fluency}_i$. For 2SLS, the second stage regressions are reported where ΔR_i is instrumented by the distance from early and normal retirement age, and W include years elapsed, female dummy and country dummies. For OLS, W additionally include age at first wave. The corresponding first stage estimates are summarized in Table B.19.

Weak IV *F* statistic is calculated according to Angrist and Pischke (2008). Stock et al. (2002) suggest that an *F* below 10 should make us worry about the potential bias in the IV estimation.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table B.19: Panel estimation: change in fluency score between wave 1 and 4: first stage

	(1) Years in retirement	(2) Years in retirement	(3) Years in retirement
Years after early eligibility	0.270*** (0.015)	0.248*** (0.015)	0.213*** (0.019)
Years after normal eligibility	0.072*** (0.015)	0.095*** (0.015)	0.135*** (0.019)
Years elapsed	0.263*** (0.067)	0.247*** (0.067)	0.371*** (0.097)
Female		-0.416*** (0.052)	-0.373*** (0.051)
Constant	-0.541 (0.47)	-0.152 (0.47)	0.008 (0.67)
Country dummies	No	No	Yes
Observations	6,368	6,368	6,368

Notes: The results are from the first stage estimation of $\Delta S_i = \alpha^* + \beta \Delta R_i + W_i^{*'} \nu + \Delta \tilde{u}_i$ where $\Delta M_i = M_{i,w=4} - M_{i,w=1}$, $S_i = \text{Fluency}_i$ and ΔR_i is instrumented by distance from early and normal retirement age.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table B.20: Panel estimation: change in fluency score between wave 1 and 2

	(1) 2SLS	(2) 2SLS	(3) 2SLS	(4) OLS	(5) OLS
Years in retirement	-0.021* (0.011)	-0.021* (0.011)	-0.026** (0.012)	-0.015* (0.0081)	-0.003 (0.011)
Years elapsed	0.044* (0.023)	0.044* (0.023)	0.155*** (0.036)	0.150*** (0.036)	0.147*** (0.036)
Female		0.005 (0.018)	0.011 (0.018)	0.011 (0.018)	0.009 (0.019)
Age at first wave					-0.003 (0.0022)
Constant	-0.084 (0.054)	-0.086 (0.055)	-0.587*** (0.099)	-0.594*** (0.099)	-0.404*** (0.16)
Country dummies	No	No	Yes	Yes	Yes
Observations	9,200	9,200	9,200	9,200	9,200
Weak IV F statistic	4250.45	4330.57	4310.36		

Notes: All results are from the estimation of $\Delta S_i = \alpha^* + \beta \Delta R_i + W_i^{*'} \nu + \Delta \tilde{u}_i$ where $\Delta M_i = M_{i,w=2} - M_{i,w=1}$ and $S_i = \text{Fluency}_i$. For 2SLS, the second stage regressions are reported where ΔR_i is instrumented by the distance from early and normal retirement age, and W include years elapsed, female dummy and country dummies. For OLS, W additionally include age at first wave. The corresponding first stage estimates are summarized in Table B.21.

Weak IV F statistic is calculated according to Angrist and Pischke (2008). Stock et al. (2002) suggest that an F below 10 should make us worry about the potential bias in the IV estimation.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table B.21: Panel estimation: change in fluency score between wave 1 and 2: first stage

	(1) Years in retirement	(2) Years in retirement	(3) Years in retirement
Years after early eligibility	0.062*** (0.0041)	0.051*** (0.0042)	0.072*** (0.0060)
Years after normal eligibility	0.062*** (0.0042)	0.074*** (0.0043)	0.055*** (0.0060)
Years elapsed	0.388*** (0.021)	0.379*** (0.021)	0.375*** (0.033)
Female		-0.173*** (0.018)	-0.145*** (0.018)
Constant	0.074 (0.056)	0.230*** (0.058)	0.519*** (0.096)
Country dummies	No	No	Yes
Observations	9,200	9,200	9,200

Notes: The results are from the first stage estimation of $\Delta S_i = \alpha^* + \beta \Delta R_i + W_i^{*'} \nu + \Delta \tilde{u}_i$ where $\Delta M_i = M_{i,w=2} - M_{i,w=1}$, $S_i = \text{Fluency}_i$ and ΔR_i is instrumented by distance from early and normal retirement age.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table B.22: Panel estimation: change in fluency score between wave 2 and 4

	(1) 2SLS	(2) 2SLS	(3) 2SLS	(4) OLS	(5) OLS
Years in retirement	-0.036*** (0.0072)	-0.036*** (0.0072)	-0.031*** (0.0073)	-0.016*** (0.0052)	0.002 (0.0075)
Years elapsed	-0.014 (0.038)	-0.015 (0.038)	0.192*** (0.048)	0.179*** (0.048)	0.185*** (0.048)
Female		0.012 (0.021)	0.021 (0.021)	0.023 (0.021)	0.016 (0.021)
Age at second wave					-0.008*** (0.0025)
Constant	0.132 (0.16)	0.129 (0.16)	-0.693*** (0.21)	-0.686*** (0.21)	-0.222 (0.25)
Country dummies	No	No	Yes	Yes	Yes
Observations	8,070	8,070	8,070	8,070	8,070
Weak IV <i>F</i> statistic	4173.04	4193.86	4070.78		

Notes: All results are from the estimation of $\Delta S_i = \alpha^* + \beta \Delta R_i + W_i^{*'} \nu + \Delta \tilde{u}_i$ where $\Delta M_i = M_{i,w=4} - M_{i,w=2}$ and $S_i = \text{Fluency}_i$. For 2SLS, the second stage regressions are reported where ΔR_i is instrumented by the distance from early and normal retirement age, and W include years elapsed, female dummy and country dummies. For OLS, W additionally include age at first wave. The corresponding first stage estimates are summarized in Table B.23.

Weak IV *F* statistic is calculated according to Angrist and Pischke (2008). Stock et al. (2002) suggest that an *F* below 10 should make us worry about the potential bias in the IV estimation.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table B.23: Panel estimation: change in fluency score between wave 2 and 4: first stage

	(1) Years in retirement	(2) Years in retirement	(3) Years in retirement
Years after early eligibility	0.157*** (0.0086)	0.146*** (0.0087)	0.124*** (0.012)
Years after normal eligibility	0.072*** (0.0088)	0.084*** (0.0090)	0.108*** (0.012)
Years elapsed	-0.049 (0.058)	-0.042 (0.058)	0.240*** (0.072)
Female		-0.192*** (0.032)	-0.172*** (0.032)
Constant	1.289*** (0.25)	1.401*** (0.25)	0.848*** (0.32)
Country dummies	No	No	Yes
Observations	8,070	8,070	8,070

Notes: The results are from the first stage estimation of $\Delta S_i = \alpha^* + \beta \Delta R_i + W_i^{*'} \nu + \Delta \tilde{u}_i$ where $\Delta M_i = M_{i,w=4} - M_{i,w=2}$, $S_i = \text{Fluency}_i$ and ΔR_i is instrumented by distance from early and normal retirement age.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Appendix C

Appendix for Chapter 3

C.1 Additional Tables

Section C.1.1 presents additional summary statistics on the variables we use in our school value-added models. Section C.1.2 presents validity checks for the 10th-grade sample. Section C.1.3 presents additional results and compares the non-parametric bounds to the school value-added estimates.

C.1.1 Data

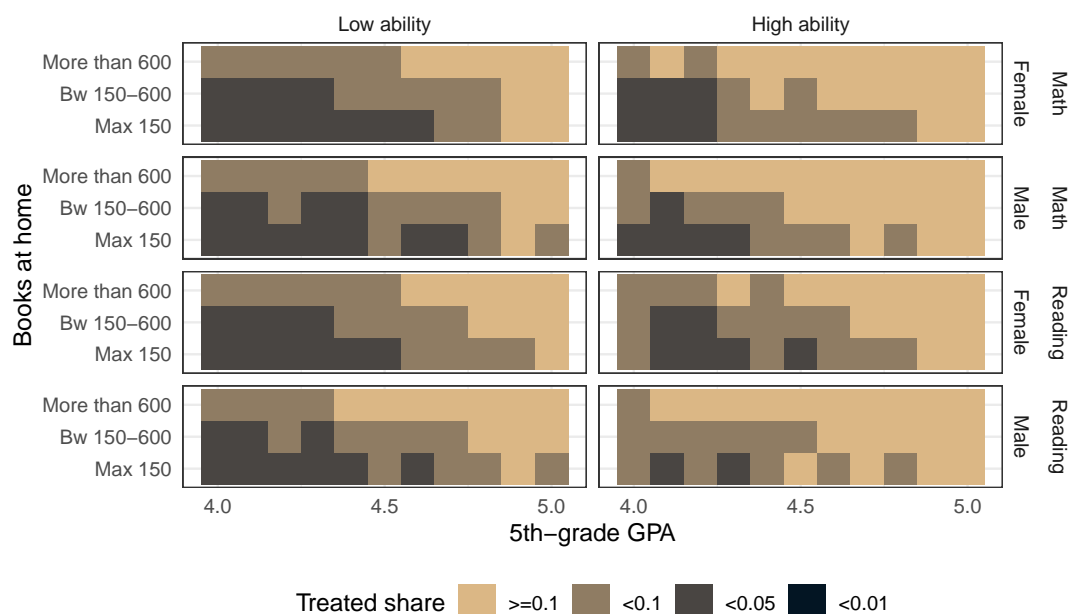
Table C.1: Additional summary statistics

	8th-grade sample			10th-grade sample		
	Elite-school students	Non-elite-school students	Total	Elite-school students	Non-elite-school students	Total
	(1)	(2)	(3)	(4)	(5)	(6)
<i>A. Pre-treatment student characteristics</i>						
Primary education (father)	0.02 (0.12)	0.06 (0.23)	0.05 (0.22)	0.01 (0.11)	0.05 (0.21)	0.04 (0.20)
Secondary education (father)	0.48 (0.50)	0.69 (0.46)	0.67 (0.47)	0.48 (0.50)	0.69 (0.46)	0.66 (0.47)
Tertiary education (father)	0.50 (0.50)	0.25 (0.43)	0.28 (0.45)	0.51 (0.50)	0.26 (0.44)	0.29 (0.46)
Primary education (mother)	0.02 (0.12)	0.07 (0.25)	0.06 (0.23)	0.02 (0.13)	0.06 (0.23)	0.05 (0.22)
Secondary education (mother)	0.40 (0.49)	0.62 (0.49)	0.59 (0.49)	0.39 (0.49)	0.61 (0.49)	0.58 (0.49)
Tertiary education (mother)	0.58 (0.49)	0.32 (0.47)	0.35 (0.48)	0.59 (0.49)	0.34 (0.47)	0.37 (0.48)
Disadvantaged	0.00 (0.07)	0.03 (0.17)	0.03 (0.16)	0.00 (0.06)	0.01 (0.12)	0.01 (0.12)
<i>B. School location</i>						
Capital or county capital	0.59 (0.49)	0.41 (0.49)	0.44 (0.50)	0.59 (0.49)	0.64 (0.48)	0.63 (0.48)
Town	0.41 (0.49)	0.40 (0.49)	0.40 (0.49)	0.41 (0.49)	0.36 (0.48)	0.36 (0.48)
Village	0.00 (0.00)	0.19 (0.39)	0.16 (0.37)	0.00 (0.00)	0.00 (0.07)	0.00 (0.06)
Number of students	16,702	109,494	126,196	8,850	63,112	71,962

Notes: The table presents the means and standard deviations of student characteristics for each sample. Columns (1) and (4) focus on students who did not enroll in an elite school, columns (2) and (5) focus on students who enrolled in an elite school, and columns (3) and (6) focus on the entire sample.

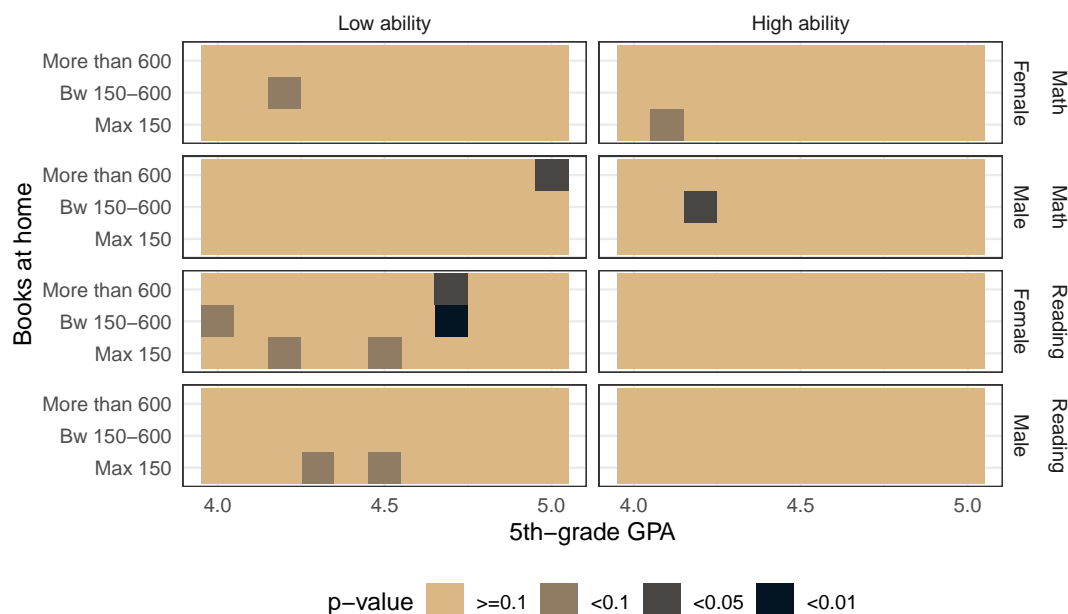
C.1.2 Validity check

Figure C.1: Validity check: Elite-school enrollment and student characteristics – 10th-grade sample



Notes: The figure presents the share of students who enrolled in an elite school by student characteristics. Each cell shows the share of elite-school students for a combination of 5th-grade GPA and the number of books at home. In the top panels (bottom) high-ability/low-ability is defined as having 6th-grade mathematics (reading) test score above/below the median. Sample: 10th-grade sample, N = 71,962.

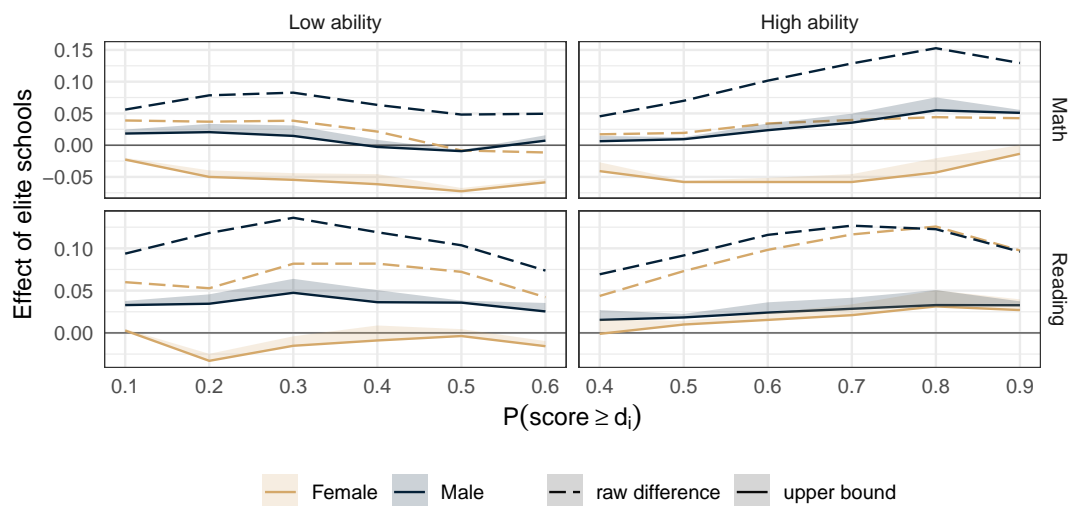
Figure C.2: Validity check: The p-values of the Kolmogorov-Smirnov test – 10th-grade sample



Notes: The figure displays the p-values of the one-sided Kolmogorov-Smirnov test. The Kolmogorov-Smirnov test tests the equality of the distributions of elite-school and non-elite-school students' 6th-grade standardized test scores. Each cell shows the p-value for a combination of 5th-grade GPA and the number of books at home. In the top panels (bottom) high-ability/low-ability is defined as having 6th-grade mathematics (reading) test score above/below the median. Sample: 10th-grade sample, N = 71,962.

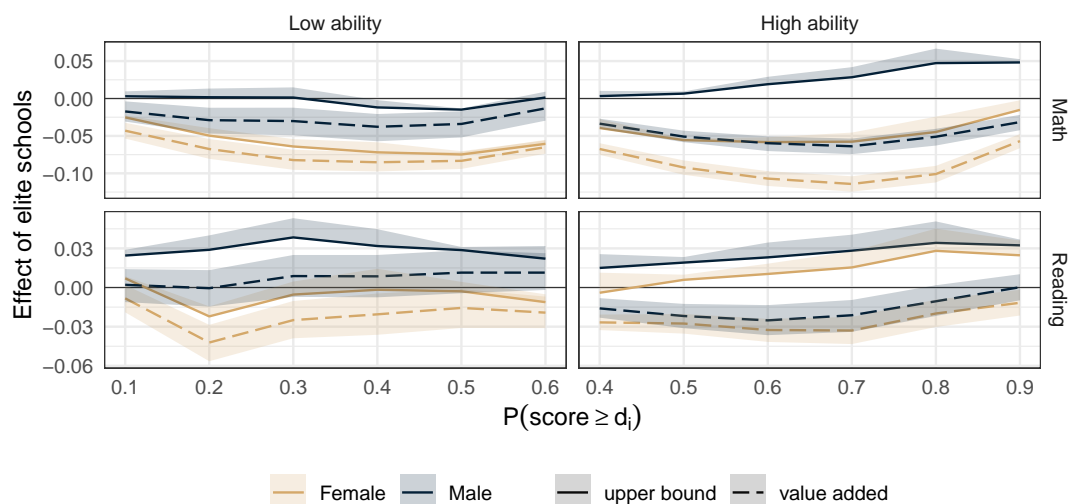
C.1.3 Results

Figure C.3: The effect of elite-school enrollment on the distribution of elite-school students' 8th-grade standardized test scores: Non-comprehensive schools



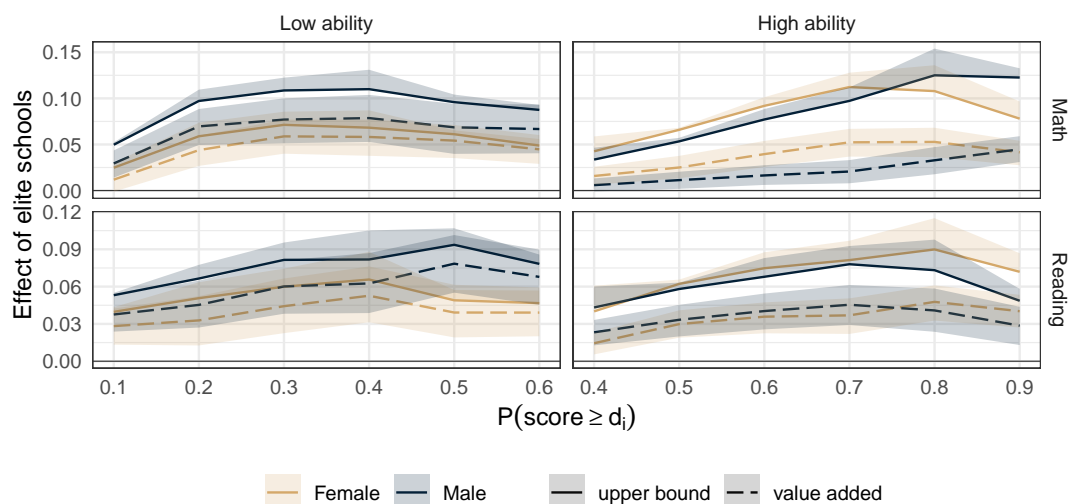
Notes: The figure presents our upper-bound estimates of the effect of elite-school enrollment on the distribution of elite-school students' 8th-grade standardized test scores (solid lines). The dashed line denotes the raw difference between the outcomes of elite-school and non-elite-school students. We report the estimates separately by gender and low- and high-ability students. Low- and high-ability students are defined by whether the students' 6th-grade standardized test score is below or above the median. The shaded area represents the area between the upper confidence band (95%) and the upper bound estimate itself. The 95% confidence intervals are based on 1,000 bootstrap draws. Sample: 8th-grade sample – non-comprehensive schools, N = 124,189.

Figure C.4: The effect of elite-school enrollment on the distribution of elite-school students' 8th-grade standardized test scores: Bounds and school VA



Notes: The figure presents our upper-bound estimates of the effect of elite-school enrollment on the distribution of elite-school students' 8th-grade standardized test scores (solid lines) along with the school value-added estimates of full model (dashed lines). The figure presents the estimates for the deciles of the outcome distribution. We report the estimates separately by gender and ability. Low- and high-ability students are defined by whether the students' 6th-grade standardized test score is below or above the median. The shaded area represents the 95% confidence intervals (for the bound estimate, only the upper confidence bound is plotted) based on 1,000 bootstrap draws. Sample: 8th-grade sample, $N = 126,196$.

Figure C.5: The effect of elite-school enrollment on the distribution of elite-school students' 10th-grade standardized test scores: Bounds and school VA



Notes: The figure presents our upper-bound estimates of the effect of elite-school enrollment on the distribution of elite-school students' 10th-grade standardized test scores (solid lines) along with the school value-added estimates of full model (dashed lines). The figure presents the estimates for the deciles of the outcome distribution. We report the estimates separately by gender and ability. Low- and high-ability students are defined by whether the students' 6th-grade standardized test score is below or above the median. The shaded area represents the 95% confidence intervals (for the bound estimate, only the upper confidence bound is plotted) based on 1,000 bootstrap draws. Sample: 10th-grade sample, $N = 71,962$.

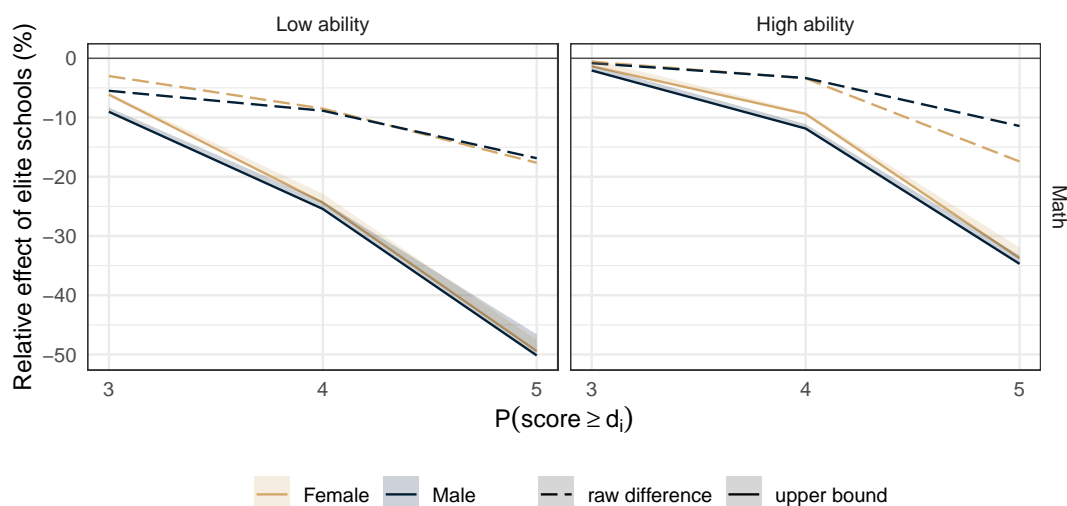
C.2 The relative effects of elite-school enrollment

This Appendix presents non-parametric bounds on the relative ATET, i.e., the relative effect of enrollment in an elite school for elite-school students. We define the relative ATET as follows:

$$\text{Relative ATET} = \frac{\mathbb{P}[Y(1) > \gamma | D = 1]}{\mathbb{P}[Y(0) > \gamma | D = 1]} - 1 = \frac{\tau(\gamma)}{\mathbb{P}[Y(1) > \gamma | D = 1] - \tau(\gamma)}, \quad \forall \gamma.$$

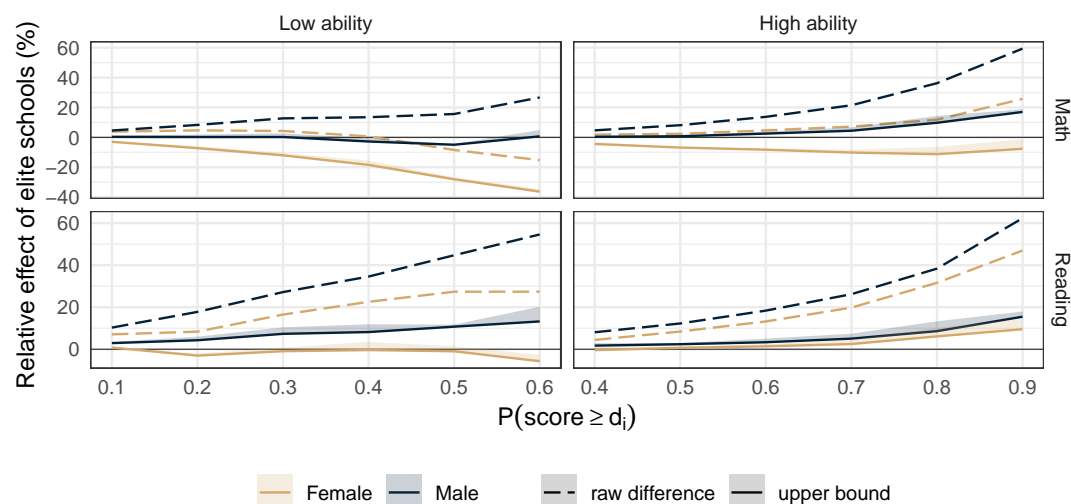
In this Appendix, we present the relative ATET estimates for each figure presented in the main text.

Figure B1: The relative effect of elite-school enrollment on the distribution of elite-school students' 8th-grade mathematics grade



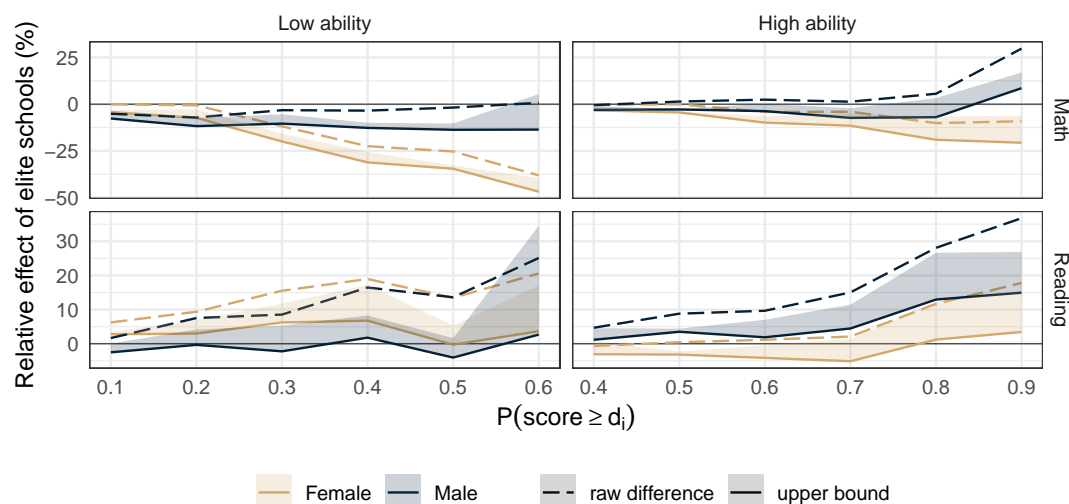
Notes: The figure presents our upper-bound estimates of the **relative** effect of elite-school enrollment on the distribution of elite-school students' 8th-grade mathematics grade (solid lines). The dashed line denotes the raw difference between the outcomes of elite-school and non-elite-school students. We report the estimates separately by gender and low- and high-ability students. Low- and high-ability students are defined by whether the students' 6th-grade mathematics test score is below or above the median. Students' 6th-grade mathematics grade is measured on the scale of 1–5. The shaded area represents the area between the upper confidence band (95%) and the upper bound estimate itself. The 95% confidence intervals are based on 1,000 bootstrap draws. Sample: 8th-grade sample, $N = 126,196$.

Figure B2: The relative effect of elite-school enrollment on the distribution of elite-school students' 8th-grade standardized test scores



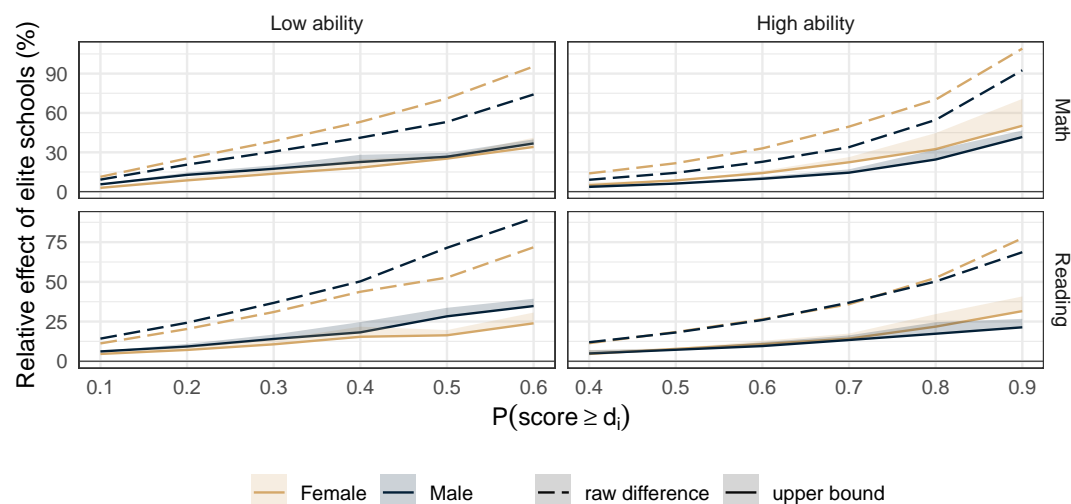
Notes: The figure presents our upper-bound estimates of the **relative** effect of elite-school enrollment on the distribution of elite-school students' 8th-grade mathematics grades (solid lines). The dashed line denotes the raw difference between the outcomes of elite-school and non-elite-school students. We report the estimates separately by gender and low- and high-ability students. Low- and high-ability students are defined by whether the students' 6th-grade standardized test score is below or above the median. The shaded area represents the area between the upper confidence band (95%) and the upper bound estimate itself. The 95% confidence intervals are based on 1,000 bootstrap draws. Sample: 8th-grade sample, $N = 126,196$.

Figure B3: The relative effect of elite-school enrollment on the distribution of elite-school students' 8th-grade standardized test scores: Comprehensive schools



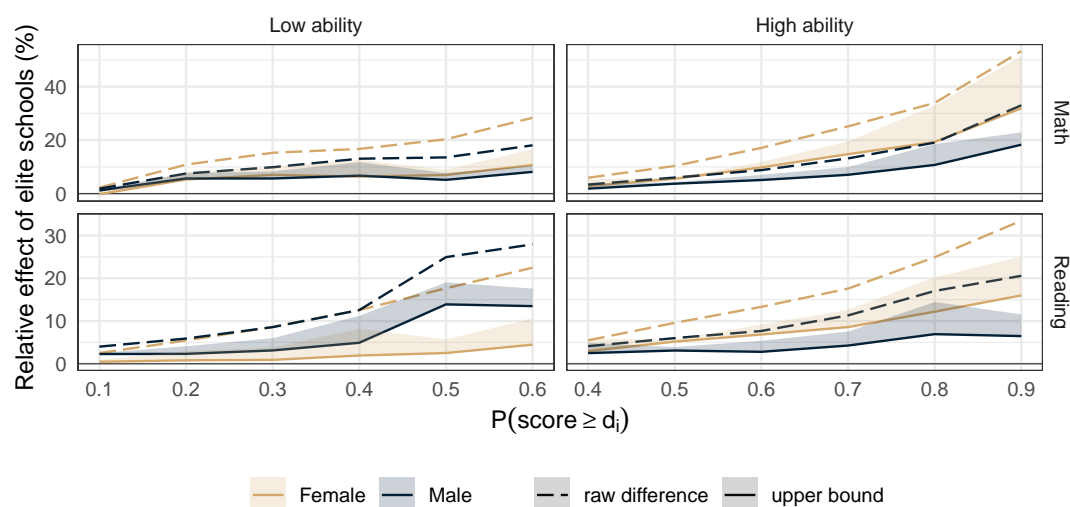
Notes: The figure presents our upper-bound estimates of the **relative** effect of elite-school enrollment on the distribution of elite-school students' 8th-grade mathematics grades (solid lines). The dashed line denotes the raw difference between the outcomes of elite-school and non-elite-school students. We report the estimates separately by gender and low- and high-ability students. Low- and high-ability students are defined by whether the students' 6th-grade standardized test score is below or above the median. The shaded area represents the area between the upper confidence band (95%) and the upper bound estimate itself. The 95% confidence intervals are based on 1,000 bootstrap draws. Sample: 8th-grade sample – comprehensive schools, $N = 111,501$.

Figure B4: The relative effect of elite-school enrollment on the distribution of elite-school students' 10th-grade standardized test scores



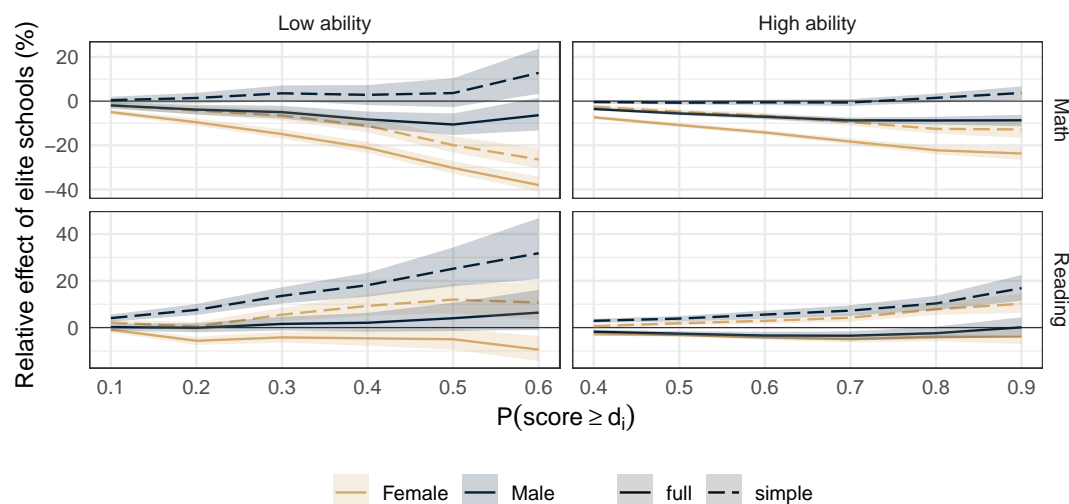
Notes: The figure presents our upper-bound estimates of the **relative** effect of elite-school enrollment on the distribution of elite-school students' 10th-grade standardized test scores (solid lines). The dashed line denotes the raw difference between the outcomes of elite-school and non-elite-school students. We report the estimates separately by gender and low- and high-ability students. Low- and high-ability students are defined by whether the students' 6th-grade standardized test score is below or above the median. The shaded area represents the area between the upper confidence band (95%) and the upper bound estimate itself. The 95% confidence intervals are based on 1,000 bootstrap draws. Sample: 10th-grade sample, N = 71,962.

Figure B5: The relative effect of elite-school enrollment on the distribution of elite-school students' 10th-grade standardized test scores: Elite secondary grammar schools



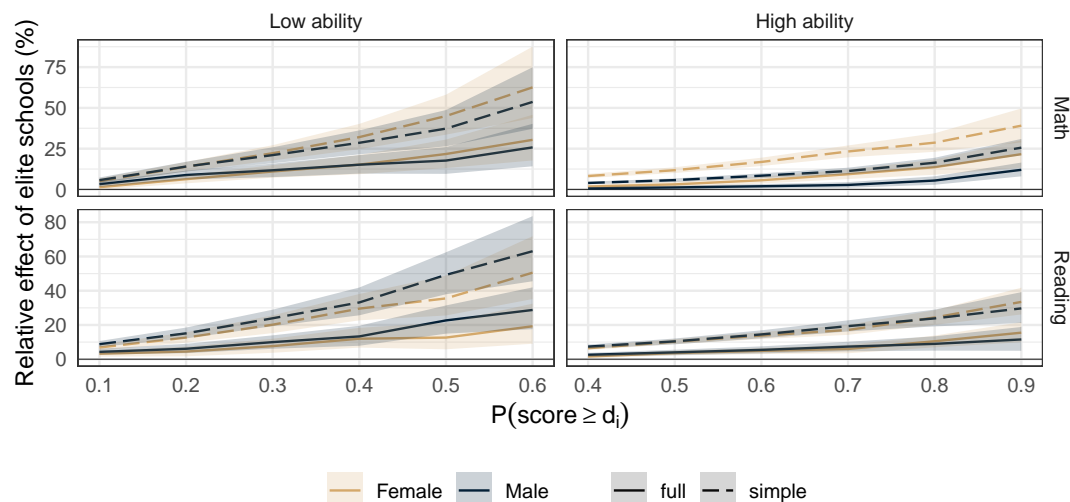
Notes: The figure presents our upper-bound estimates of the effect of elite-school enrollment on the distribution of elite-school students' 10th-grade standardized test scores (solid lines). The dashed line denotes the raw difference between the outcomes of elite-school and non-elite-school students. We report the estimates separately by gender and low- and high-ability students. Low- and high-ability students are defined by whether the students' 6th-grade standardized test score is below or above the median. The shaded area represents the area between the upper confidence band (95%) and the upper bound estimate itself. The 95% confidence intervals are based on 1,000 bootstrap draws. Sample: 10th-grade sample – elite secondary grammar schools, N = 21,384.

Figure B6: The relative effect of elite-school enrollment on the distribution of elite-school students' 8th-grade standardized test scores: School VA



Notes: The figure presents the school value-added estimates of the **relative** effect of elite-school enrollment on the distribution of elite-school students' 8th-grade standardized test scores. The figure presents school VA estimates for the deciles of the outcome distribution. The top (bottom) panels focus on mathematics (reading). The left (right) panels focus on students whose 6th-grade test score is above (below) the median. The dashed lines refer to the estimates of the simple school VA model (6th-grade standardized test score, cohort fixed effects) and the solid lines refer to the full school VA model (6th-grade standardized test score, cohort fixed effects, 5th-grade GPA, number of books at home, parental education, disadvantaged status, county of the school, type of the settlement where the school is located). The shaded areas represent the 95% confidence intervals around the school VA estimates. The confidence intervals are based on 1,000 bootstrap draws. Sample: 8th-grade sample, N = 126,196.

Figure B7: The relative effect of elite-school enrollment on the distribution of elite-school students' 10th-grade standardized test scores: School VA



Notes: The figure presents the school value-added estimates of the **relative** effect of elite-school enrollment on the distribution of elite-school students' 8th-grade standardized test scores. The figure presents school VA estimates for the deciles of the outcome distribution. The top (bottom) panels focus on mathematics (reading). The left (right) panels focus on students whose 6th-grade test score is above (below) the median. The dashed lines refer to the estimates of the simple school VA model (6th-grade standardized test score, cohort fixed effects) and the solid lines refer to the full school VA model (6th-grade standardized test score, cohort fixed effects, 5th-grade GPA, number of books at home, parental education, disadvantaged status, county of the school, type of the settlement where the school is located). The shaded areas represent the 95% confidence intervals around the school VA estimates. The confidence intervals are based on 1,000 bootstrap draws. Sample: 10th-grade sample, N = 71,962.

C.3 Data (for online publication)

This Appendix describes our data. We begin, in Section C.3.1, by describing our sample restrictions. In Section ??, we explain how we construct our variables.

C.3.1 Sample restrictions

Table C1 presents the evolution of the sample size as a result of our sample restrictions.

Table C1: Evolution of the sample size

	2010	2011	2012	2013	2014	Total
A. 8th-grade sample						
raw	104,266	96,843	92,966	89,913	87,542	471,530
w/o missing test score	96,212	89,005	85,245	81,919	80,065	432,446
w/o missing variables	76,875	70,777	68,278	66,636	65,651	348,217
w/o missing after imputation	91,372	84,562	81,441	75,971	74,637	407,983
w/o missing history	62,637	57,651	56,943	53,660	53,703	284,594
final sample	27,328	25,550	25,326	24,275	23,717	126,196
B. 10th-grade sample						
raw			102,037	95,649	90,188	287,874
w/o missing test score			90,315	83,554	78,727	252,596
w/o missing variables			72,697	68,041	64,963	205,701
w/o missing after imputation			84,911	78,561	74,433	237,905
w/o missing history			54,062	48,945	47,365	150,372
final sample			25,371	23,519	23,072	71,962

Notes: The first row shows the number of students in our raw data, the National Assessment of Basic Competencies (NABC), in each of the relevant years and grades. We document how much of them we lose due to missing test scores and missing background variables. We win back a part of this loss by imputing background variables (see Appendix C.3.3 for more detail), resulting in a sample of more than 80% of the whole cohorts. Unfortunately, we can only link 60-70% of these students to their 6th-grade results. Restricting the sample to those for whom elite school seems to be a relevant option (having good grades and being in a school in 6th grade from which at least one student goes into an elite school in our sample period) further decreases the size: we end up with about 25% of the cohorts. Our samples are highly selective, but that makes them more relevant for our question.

C.3.2 Variable description

This Appendix describes the construction of variables in Tables 3.2 and C.1.

- **Number of books at home** is categorical variable with three values: max. 150 books, between 150 and 600 books, and more than 600 books;

- **Mother's education** is a categorical variable with three values: primary, secondary, and tertiary education;
- **Father's education** is a categorical variable with three values: primary, secondary, and tertiary education;
- **Disadvantaged status** is a binary variable, which takes a value of one when a student has a disadvantaged status (i.e., comes from a low-income family);
- **The type of the settlement** where the school is located is a categorical variable with four values: village, town, county capital, capital;
- **County** of the school is a categorical variable with 20 values;

C.3.3 Imputation

Student's characteristics (number of books at home, mother's education, father's education, disadvantaged status) are gained from an extensive background survey that complements the NABC and which the students fill out together with their parents on a voluntary basis (the average completing rate is around 75 percent, see Table C1). As these characteristics should be mainly constant over time, we could exploit the longitudinal aspect of our data to fill out missing values in one year from the corresponding questionnaire of another year. The imputation is done by following a before-after approach: in the 8th-grade sample, we first look for a value in the 6th-grade questionnaire, or if that is still missing, we rely on the 10th-grade questionnaire; in the 10th-grade sample, we impute from grade 8, or if that is still missing, we go on to grade 6.