universität wien

DISSERTATION / DOCTORAL THESIS

Titel der Dissertation /Title of the Doctoral Thesis

„Essays in Applied Microeconomics“

verfasst von / submitted by

Nóra Kungl

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Doctor of Philosophy (PhD)

Wien, 2024 / Vienna 2024

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 794 370 140
Dissertationsgebiet lt. Studienblatt / field of study as it appears on the student record sheet:	Volkswirtschaftslehre
Betreut von / Supervisors:	Univ-Prof. Philipp Schmidt-Dengler, PhD Univ.-Prof. Dr. Wieland Müller

Acknowledgements

First, I would like to thank my supervisors, Philipp Schmidt-Dengler and Wieland Müller for letting me pursue my interests and supporting me along the way.

I am thankful to the VGSE faculty, and faculty members of the Department of Economics at the University of Vienna for all their helpful comments and guidance throughout my PhD studies. Discussions with Omar Bamieh, Simon Hess, Lennart Ziegler, Stephanos Vlachos, Dijana Zejcirovic, Karl Schlag, Daniel Garcia, and Christian Koch were particularly helpful. I extend my thanks to Ana Begona Ania Martinez for her kind support and guidance, both in teaching and in completing my dissertation.

I thank Anita Győrfi, Péter Elek and Dániel Prinz for the joint work on the last chapter of this thesis. Thanks to Anita for the countless motivating discussions, and thanks to Péter for guiding my first steps in the field of health economics.

I also thank Hubert János Kiss and the Eltecon faculty for introducing me to the world of economics with devotion and kindness. Their teaching and mentoring played a significant role in shaping my thinking and attitude towards research.

Moreover, I am thankful to the LMEC community at the University of Bologna which also greatly contributed to my journey towards earning a PhD.

Special thanks go to Le ragazze, Ana, Farah, and Negar, for sharing in both the struggles and the joys, not only during our Master's program but also throughout our PhD years. I am grateful for our friendship and the adventures ahead.

The list would not be complete without thanking my VGSE colleagues and friends. Our discussions not only enriched the content of this thesis, but helped whenever challenges emerged. Their company, the joint lunch and coffee breaks made the PhD experience less lonely. I also share great memories with the PhD group at CEU, and I thank them for these.

Last but not least, I thank my family for their continuous support and guidance throughout my studies, and more importantly, in life. Anya, Apa, Ádám, köszönöm! Finally, I thank Niccolò for being the biggest support at the finish line.

Contents

List of Figures

List of Tables

Introduction

This thesis consists of three chapters applying applied microeconomics tools in the fields of education and health economics, utilizing unique administrative datasets. All chapters shed light on how individuals respond to incentives in different institutional environments, uncovering unintended consequences, and unravelling disparities. Rooted in the context of Hungary, a country characterized by pronounced educational and healthcare inequalities, these investigations not only contribute to academic discourse but also bear significant policy implications.

The first two chapters investigate fraudulent behaviour in the Hungarian standardized student assessments (NABC, National Assessment of Basic Competencies). While such testing is introduced to provide valuable feedback on the education system's state, higher stakes can potentially distort quality signals and overall data quality.

In the first chapter, I study the prevalence of test score manipulation and find no evidence of systematic manipulation. I argue that this occurs because of the unique testing environment where the low-stake testing is paired with strict quality assurance. This means that compared to previously studied tests there is not only less incentive to cheat, but it is also more costly. I employ three methods: the algorithm by Jacob and Levitt (2003), a clustering technique and a simpler screening based on summary statistics. By revealing the limitations of these methods through simulation exercises, this chapter underscores the complexity of uncovering fraudulent practices.

In the second chapter, I provide suggestive evidence that a less costly form of manipulation, namely test pool manipulation, is present in the testing. The manipulation of the test pool happens through teachers encouraging non-participation of low-performing students. I exploit a policy change which introduced higher stakes for schools and employ a difference-in-differences estimation strategy. First, in line with the incentives of the new policy, I find that post-policy absence rates increased particularly in schools at risk of not meeting the minimum requirement. Finally, using multiple imputation I aim to quantify schools' gains from these absences. I find that although the variation in absences is large, schools do not benefit substantially from it.

The third chapter, joint work with Péter Elek, Anita Győrfi and Dániel Prinz, focuses on geographic and socioeconomic variations in healthcare use. Exploiting migration across

regions in Hungary we show that place-specific factors account for 66% and 31% of the variation in outpatient and drug spending, respectively, but play no role in inpatient care use. Notably, place effects explain 80% of outpatient spending variation for non-employed working-age individuals and those below the first quartile of the wage distribution, but less than 40% for individuals with above-median wage incomes. We also find a positive association between place effects and outpatient capacity, especially for low-income individuals. These results suggest that even in a system with universal coverage, access to healthcare can significantly vary, particularly for vulnerable groups.

While seemingly disparate, these chapters converge thematically in their exploration of how individuals and institutions respond to incentives and policies. The thesis highlights how administrative data can be harnessed to inform policymaking, with a particular emphasis on designing systems resilient to manipulation and mitigating existing inequalities.

In his 2000 Nobel lecture, James Heckman linked the birth of microeconometrics to the post-World War II production of firm- and individual-level micro data and how those revealed previously unknown dimensions behind the aggregate data (Heckman, 2001). Since then, the past two decades have witnessed a rise in the usage of large-scale administrative datasets. Administrative data allowed researchers to ask questions that were not explored before, shaping the direction of economic research with a shift to more policy-relevant work (Card et al., 2010; Currie et al., 2020; Einav and Levin, 2014; Nagaraj and Tranchero, 2023).[1]

This thesis utilizes two administrative datasets: (1) a three-way panel dataset of all students subject to standardized testing, including item-level responses, standardised results, and demographics, complemented with a family survey, and (2) an individual-level panel dataset that covers monthly healthcare, labor market and demographic information for the years 2009–2017 on a random 50% sample of the 2003 population of Hungary. While the third chapter provides an example of how large-scale, linked admin data can be utilized to identify causal effects, the first two chapters raise caution about the use of such data. While admin data suffers less from attrition and missing values than e.g., survey data, data quality might still be a concern. When there are incentives to misreport, or if the institutional environment prompts strategic responses from the economic agents, researchers must be aware and take this into account.

[1]For a historical overview on the rise of empirical work and the changing interplay of theoretical and applied economics, coupled with an increasing interest in policy-relevant research, see Backhouse and Cherrier (2017).

References

Backhouse, R. E., and Cherrier, B. (2017). The Age of the Applied Economist: The Transformation of Economics Since the 1970s. *History of Political Economy*, *49*, 1–33.

Card, D., Chetty, R., Feldstein, M. S., and Saez, E. (2010). Expanding Access to Administrative Data for Research in the United States. *American Economic Association, Ten Years and Beyond: Economists Answer NSF's Call for Long-Term Research Agendas*.

Currie, J., Kleven, H., and Zwiers, E. (2020). Technology and Big Data Are Changing Economics: Mining Text to Track Methods. *110*, 42–48.

Einav, L., and Levin, J. (2014). Economics in the Age of Big Data. *Science*, *346*(6210), 1243089.

Heckman, J. J. (2001). Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture. *Journal of Political Economy*, *109*(4), 673–748.

Nagaraj, A., and Tranchero, M. (2023). How Does Data Access Shape Science? Evidence from the Impact of US Census's Research Data Centers on Economics Research. *NBER Working Paper*, *No. 31372*.

Chapter 1

Detecting Test Score Manipulation in Standardized Testing*

1.1 Introduction

Accountability systems aim to assess school and teaching quality, mostly based on large-scale testing of students. With the increased attention on the importance and effectiveness of educational investment, standardized tests became more and more prevalent in education systems around the world (Figlio and Loeb, 2011). This has resulted in large and high-quality data sets, both at international and national levels. Testing and schooling data have been used to show that education is an important determinant of economic growth (Hanushek and Kimko, 2000), and on a more micro level, test scores can predict labour market outcomes such as employment and wages (Meghir and Palme, 2005; Murnane et al., 1995). Thus, it is particularly important to use reliable and uncorrupted data when studying such economic questions and drawing policy conclusions.

However, it has also been shown that fraudulent behaviour is widespread in education. Evidence points towards the presence of manipulation across countries, levels of the education system, and the parties involved (Ajzenman, 2021; Battistin and Neri, 2023; Borcan et al., 2017; Cagala et al., 2021; Jacob and Levitt, 2003a; Lin and Levitt, 2020). Students might cheat to get a better grade, teachers might help students to appear as better teachers, and school principals consent to such behaviour to secure more funding or achieve a higher ranking.

In this paper, I use data from the Hungarian standardized testing (National Assessment of Basic Competencies, NABC) which experienced only a few cheating scandals since its start in 2001. Despite a few cases coming to light, the question remains whether schools in Hungary systematically engage in such behaviour. Based on anecdotal evidence, I focus on cheating happening at the class level with the help of the teacher, and

*This chapter builds on the master thesis "Who Looks Suspicious? An Exploration of Cheating Behaviour in the National Assessment of Basic Competencies Tests in Hungary" by Kungl (2019).

I refer to it as class-level or teacher cheating in the following.

The paper consists of two parts. While I investigate whether there is any cheating in Hungary, in the second part I also shed light on the potential limitations of detection methods. I draw my conclusions with these limitations in mind.

In the first part (Section 1.5) I employ the algorithm developed in the seminal paper of Jacob and Levitt (2003a). In contrast to previously published studies, I find no evidence of test score manipulation. I argue this is because of differences in the structure and organization of the tests, providing both fewer opportunities and fewer incentives for cheating than the previously studied testing systems in the US and Italy. In fact, contrary to most standardized tests in the US, the Hungarian NABC is not a high-stakes test.[1] Yet, it devotes more attention to quality assurance to reduce the possibility of cheating. The use of different test books and completely centralised corrections make teacher- or school-level cheating more challenging. This is also the most apparent difference compared to the low-stake testing system in Italy which is characterized by a high occurrence of cheating practices.

To assess the robustness of the findings I apply two other detection methods: (1) the simple summary statistics-based screening conducted by the Hungarian Educational Authority each year, and (2) a clustering technique developed by Quintano et al. (2009) for the Italian Testing Authority (INVALSI). Both methods confirm that substantial cheating in the data is unlikely.

In the second part (Section 1.6) I compare the three methods and discuss their limitations. To shed light on their most crucial assumptions and the consequent limitations, I simulate different cheating scenarios and assess the performance of the Jacob-Levitt algorithm. I show that the algorithm performs better (finds more of the cheater classes) in cases when manipulation happens at the top of the distribution, i.e., when the cheater classes cheat to an extent that could place them in the top 25% of classes. However, I also show that the performance declines with the level of overall manipulation in the data. This is because the algorithm relies on estimating a multinomial logit model on the observed (and potentially corrupted) sample to predict the probabilities of correct answers and assess the likelihood of the observed patterns. High levels of manipulation in the data would corrupt these estimations. These results suggest that the algorithm is most suitable for finding extreme cases of cheating at the top of the distribution when the overall cheating occurrence is relatively low.

The results align with theoretical work on unethical behaviour, its sources and potential preventive measures. The idea of why agents would engage in fraudulent behaviour, or commit a crime was formalised by Becker (1968), and can be applied also in this setting. Cheating will happen if the expected benefit of the crime is higher than its expected cost,

[1] Generally, testing is considered high-stakes if schools' funding, teachers' salaries, or students' grades depend on the test results. None of this is the case in Hungary.

taking into account both the probability of getting caught and the size of the punishment. While it is not straightforward to quantify what schools can gain by such manipulation, understanding the incentive structure and enforcement system can help in understanding the motivations behind cheating.

Test score manipulation is often seen as an unintended consequence of high-stakes accountability systems that introduce direct gains in the form of funding, teacher bonuses, or student grade promotion. Teachers' strategic behaviour is taken as an example by Holmstrom and Milgrom (1991) in a multitask principal-agent problem. When teachers have to choose between teaching tasks and their salary is based on the standardized test results, they will "teach to the test", i.e., focus on those skills that are tested. The same reasoning applies to more outright cheating violations. However, despite having a low-stakes testing system, previous empirical studies found proof of severe manipulations in Italy as well (Angrist et al., 2017; Bertoni et al., 2013; Quintano et al., 2009).

Recently, more attention is devoted to the role of quality assurance measures which can make fraudulent behaviour more costly. A number of studies explore how randomly assigned monitoring is associated with a decrease in performance, and argue it is because of the reduced prevalence of manipulation (Bertoni et al., 2021, 2013; Borcan et al., 2017; Lin and Levitt, 2020; Lucifora and Tonello, 2020). Theoretical work also shows that deterrence is crucial, and screening processes can act as a deterrence device (Block et al., 1981; Dionne et al., 2009; Lazear, 2006). Block et al. (1981) shows how the optimal price of a cartel depends on the effort level of antitrust enforcement and the size of the penalties. Lazear (2006) argues that the optimal rules for high-stakes testing depend on the cost of learning and monitoring, and incentives have to be designed to motivate those whose costs are high. In the context of insurance frauds Dionne et al. (2009) provides theory and application that the optimal auditing strategy is to first screen and then investigate the suspicious cases.

This paper contributes to the literature on test score manipulation by teachers in two ways (a detailed overview of the literature can be found in Section 1.2). First, it studies cheating in a low-stakes setting and shows that manipulation is less likely if incentives are low and the cost of certain manipulation is high. Second, I use several detection methods on the same data and provide a systematic overview of the advantages and disadvantages of these methods. In addition, I complement this with two simulation exercises to demonstrate how the performance of the Jacob-Levitt algorithm is affected by the potential limitations. Specifically, I look at how it performs when cheating does not happen at the top, and how its performance depends on the overall level of manipulation in the data.

The paper also relates to a broader literature on studying fraudulent behaviour as a strategic response to incentives in various settings. Observational studies are methodologically relevant since they present detection methods or quasi-experimental settings to

uncover frauds (see e.g., Bø et al. (2001) and Bíró et al. (2022) on tax evasion, Dellavigna and Ferrara (2010) on illegal arms trade, Dahl et al. (2023) on strategic voting, or the extensive literature on cartel detection). Experimental studies tend to focus on the possible mechanisms and incentives in a cleaner setting, where cheating is relatively easier to identify (for an overview and meta-study see Abeler et al. (2019) and Jacobsen et al. (2018), while related examples are discussed in Section 1.6.3).

The remainder of the paper is organized as follows. First, Section 1.2 reviews the literature on detecting cheating, distinguishing between the different types of detection methods. Second, in Section 1.3 I introduce the Hungarian school system and the National Assessment of Basic Competencies test, while Section 1.4 describes the available dataset and provides summary statistics on the data. The main part of the chapter, Section 1.5 is divided into subsections presenting all three detection methods and the corresponding results. Section 1.6 compares these methods, and evaluates the Jacob-Levitt method with two simulation exercises. Finally, Section 1.7 concludes and discusses drawbacks of the study and directions for further research.

1.2 Related Literature - Detection Methods

With the increasing availability of data and the development of statistical methods, the literature on cheating in tests developed substantially in the past decades. Detection algorithms have been developed to explore both student-level[2] and class-level fraudulence - this study focuses on the latter. Based on the applied approaches this literature can be divided into three main groups: (1) studies aiming to uncover suspicious patterns in item responses, (2) studies using a fuzzy clustering approach, and (3) studies using discontinuity analysis. In contrast to the first two, discontinuity analysis has lower data requirements (instead of item-level data only student-level test scores are necessary), but can be used only in a system with a well-defined performance threshold.

A large body of the related literature is based on the seminal paper of Jacob and Levitt (2003a) which developed the very first method aiming to assess cheating based on suspicious patterns in responses within a class. They found that unexpected test score fluctuations *(Indicator 1)* and suspicious answer sequences *(Indicator 2)* together imply a (minimum) 4-5 percent occurrence of cheating in elementary school classes of the Chicago school district annually. The unexpected test score gains indicator has a high value if ranking gains in a given year are relatively higher than ranking gains in the following

[2]Student-level cheating includes cooperation of students during test taking (without the consent of a teacher), copying from their neighbours, or any other individually executed misbehaviour. For completeness, see the literature on students' cheating (Arnold, 2016; Borisova and Peresetsky, 2016; Lin and Levitt, 2020). More recent studies focus on the prevalence of student cheating in online tests and related preventive measures as online testing became more common because of COVID-19 (Bilen and Matros, 2021; Humbert et al., 2022; Janke et al., 2021).

period. The indicator of suspicious answer strings is based on four different measures of cheating: (1) the most unusual block of answers, (2) within-classroom correlation in student answers, (3) variance in the within-classroom correlations across questions and (4) comparison of answers of students with the exact same scores. Classes with high values of both indicators are identified as potential cheaters.[3]

Their method has been slightly modified by several researchers to fit the contexts of different tests. Ferrer-Esteban (2013) used the same algorithm to identify cheating classes in the Italian standardized tests (INVALSI, National Institute for the Evaluation of the Education System) and through a logistic regression found that score manipulation is more likely to happen in socially homogeneous classrooms where teachers identify themselves better with the students. Gustafsson and Deliwe (2017) analysing the Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ) and Trends in International Mathematics and Science Study (TIMSS) tests and having data only on whether responses of students were correct or not, used only Jacob and Levitt's fourth measure, and modified it so that it adjusts for class size and the number of test items. Focusing on regional and across-country differences in the 15 participating countries, they found that cheating is prevalent, but their proposed method for score adjustments does not affect substantially the ranking of countries.

Using Hungarian data, Horn (2012) estimated cheating prevalence with a variation of the Jacob-Levitt algorithm, however, several limitations arose. First, because of limited data availability, he only analysed the year 2008, and Indicator 1 could not be calculated (no past and future scores available). Instead, he proposed a measure of unusually small standard errors, based on Item Response Theory (Rasch, 1980). Second, Indicator 2 is modified such that not only multiple-choice items are considered, but this way the ability to distinguish between different wrong answers was lost. Finally, the paper only presents the uncorrected share of suspicious classes, i.e., not taking into account that some schools would be above the cutoff of both indicators even in the absence of cheating. This way he identifies 1,5-1,8% of classes as potential cheaters. As a contribution, the more recent dataset allows me to implement Jacob and Levitt's method without modifications.

In Italy large disparities were observed between the results of the more strictly monitored international test TIMSS and the results of their national testing (INVALSI). As a result, not only cheating detection, but adjustment practices and other possible interventions received a particular attention by the educational authority. The detection method officially used by the INVALSI employs a clustering technique and was developed by Quintano et al. (2009). Their procedure is comprised of detecting classes with high average scores and low within-class variability through factorial analysis, and then using a fuzzy clustering algorithm to calculate the probability of belonging to a set of outliers. This way a correction factor is obtained for each class and the ranking of schools is made

[3]For details, see Section 5.

based on these adjusted scores. The same approach is used both by Angrist et al. (2017) and Battistin et al. (2017), but their final indicator is a dummy variable of cheating instead of a continuous probability of cheating. Their findings confirm regional differences in cheating behaviour in Italy, while Angrist et al. (2017) also finds that small classes are more likely to cheat. Most recently, Longobardi et al. (2018) proposed a combination of the fuzzy clustering technique and a statistical model-based approach.

A more recent line of research aims to detect cheating using discontinuity analysis. Dee et al. (2019) and Dee et al. (2011) find that in the New York Regents Examination there are significant discontinuities in the score distribution, i.e. manipulation is most frequent just below the performance threshold, and it benefits low-achieving students. Diamond and Persson (2016) show similar test score manipulation just above discrete grade cutoffs in a Swedish nationwide mathematics test. Battistin and Neri (2023) measures score inflation by the excess mass of students at cutoffs (bunching) in order to provide evidence of an increase in housing prices and changes in area demographics in areas with more grade inflation. Compared to the above methods, a discontinuity analysis can be conducted only in settings where there is a predefined achievement cutoff for students which is rarely the case for standardised tests.

In addition to the above methods, natural experiments of external monitoring can also help in uncovering cheating cases. Several studies exploit the particular feature of the Italian testing that, since 2012, incorporates a natural experiment by randomly assigning external examiners to classes (Angrist et al., 2017; Battistin et al., 2017; Pereda-Fernández, 2019). These studies assume cheating does not happen in the monitored classes and thus, they can be used as a control group of the unmonitored classes. Bertoni et al. (2013) focusing on the effect of external examiners found that their presence, through reducing cheating, is associated with a 5.5-8.5% decrease in the ratio of correct answers. In a follow-up paper, Bertoni et al. (2021) found that external examiners still affect the manipulation probability one year later, but the effect disappears after two years. Borcan et al. (2017) has similar findings using Romanian data.

I contribute to this literature in two ways. I use several detection methods on the same data, and based on the comparison of the results I provide a systematic overview of the advantages and disadvantages of these methods. In addition, I complement this with two simulation exercises to demonstrate how the performance of the Jacob-Levitt algorithm is affected by the potential limitations. Specifically, I look at how it performs when cheating does not happen at the top, and how its performance depends on the overall level of manipulation in the data.

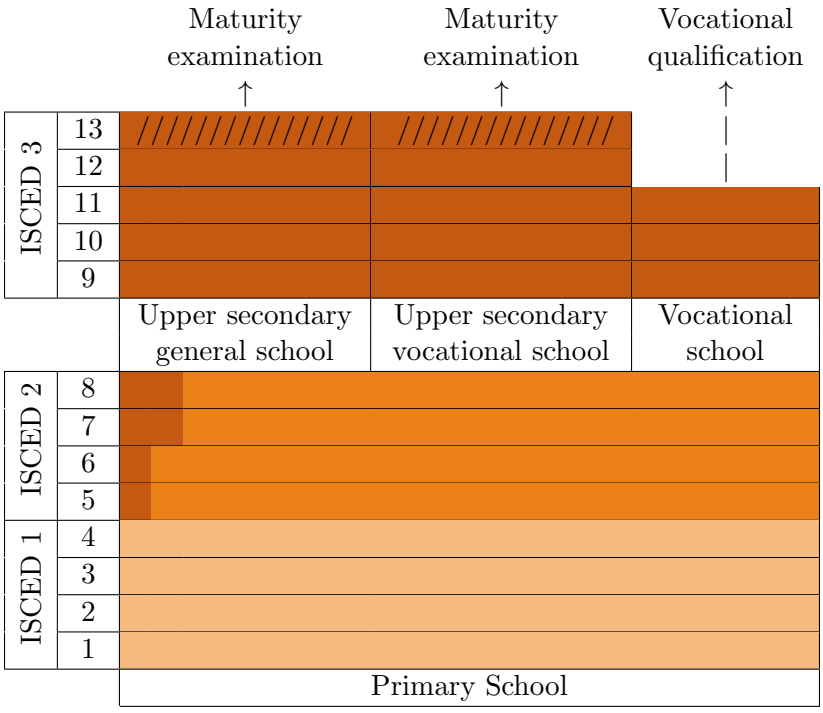
1.3 Institutional Background

To understand the incentives and factors that can influence the occurrence of cheating in the Hungarian testing system, in this section I introduce the most important institutional features of the education system. First, I outline the structure of the school system, and then describe the role that the National Assessment of Basic Competencies (NABC) test plays in it.

The Hungarian School System

School attendance in Hungary is compulsory between the ages of 6 and 16, but students typically attend school for 12 years, until the age of 18. The school leaving age was 18 for a long time, but it has been decreased to 16 in 2012 as a controversial policy change generating debates among teachers and policy experts. Throughout the 10-12 years of studying students attend schools at three different levels: lower and higher primary education and secondary education which is summarized in Table 1.1.

Table 1.1: The Structure of the Hungarian School System



Note: The table presents the structure of the Hungarian school system. Grades can be seen in column (2), with the corresponding ISCED (International Standard Classification of Education) categorization in column (1). The colouring within rows does not represent the share of students enrolled in the given type of school.

Primary education lasts for eight years and is divided into two parts, lower primary education (grades 1st-4th), and higher primary education (grades 5th-8th). In the first

four years classes have the same head teacher who teaches all subjects for the students. Schools in smaller villages might only teach until 4th grade, and then students have to continue their studies in another school (usually in the closest town where such a school is available). However, most classes remain together also for the next four years, getting a new head teacher and different subjects being taught by different teachers.

Secondary education covers three types of schools -secondary general schools, secondary vocational schools and vocational schools - which can accept students based on a general admission test written by all 8th graders on the same day in January. Upper secondary general schools (also called gymnasiums) last for four years and teach general subjects, aiming to prepare students for the maturity examination which is the entry requirement for higher education. There are specialized classes which teach a given subject (e.g. mathematics, IT, humanities or science) in an increased number of hours. A particular type is the language-specialized classes which are often 5-year long, starting with a (0th) training year only focusing on the language. Moreover, there are some 8-year and 6-year gymnasiums as well which select the best students even earlier than 8th grade (students can enrol based on an admission test taken in 5th grade and 7th grade, respectively). Upper secondary vocational school prepares students also for the maturity examination, but it is less common for these students to apply to universities. In vocational schools students learn a profession, and get a vocational qualification upon finishing.

At all levels of the education system, school choice is free. Parents can freely choose any primary school for their children, up to the availability of places. While students must be accepted to schools in the same district as their residence, parents can also choose schools further from their home, subject to capacity. In practice, even primary schools exercise selection in informal ways (Kertesi and Kézdi, 2005a). At the upper secondary level, high schools have the right to use ranking and selection procedures during admissions (deferred acceptance mechanism). In the 8th grade, every student participates in a nationwide entrance examination, and schools decide how they weigh this result alongside oral exams and interviews. The experience shows that free school choice contributed to the high level of segregation in Hungarian schools (Kertesi and Kézdi, 2005b, 2011; Kisfalusi et al., 2021).

Evaluations of the system are partly based on national and international tests which aim to give a clear picture of the abilities of students and the quality of teaching in each school, and the country as a whole. Hungarian schools participate in the PISA, TIMSS, PIRLS international tests[4], but for the evaluation of the whole education system, NABC gives the most detailed picture since it is a nationwide test involving all schools (instead of a representative sample of them).[5]

[4]PISA: Programme for International Student Assessment, TIMSS: Trends in International Mathematics and Science Study, PIRLS: Progress in International Reading Literacy Study

[5]PISA test participants are selected in a two-stage sampling procedure. First, a representative sample of at least 150 schools is selected in each country, and then on average 42 students are selected within

The National Assessment of Basic Competencies (NABC)

The NABC is a test aiming to assess reading and numeracy skills of students, focusing solely on basic competencies and how students can apply them in real-life settings, and does not aim to test the material of the curriculum. Accordingly, students do not get a grade for their performance, but they can check their results online once the national, regional and school-level reports have been published (usually at the beginning of the following calendar year, i.e. eight months after the testing took place).

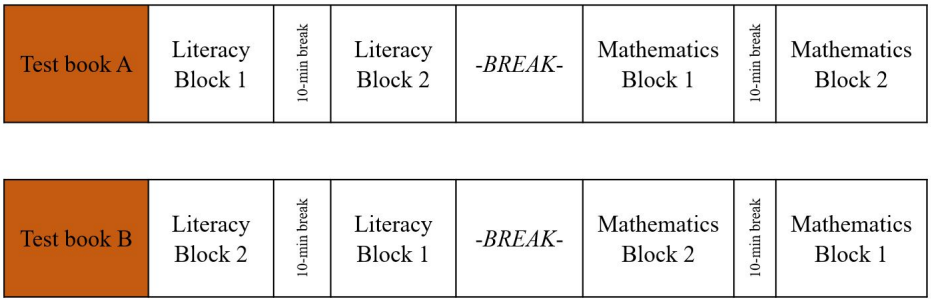
The test was first conducted in 2001 with the participation of 20 students in 5th and 9th grade in every school. Since 2008, all the students have to complete the test at the end of the 6th, 8th and 10th grades in mathematics and reading (and since 2015 in a foreign language as well). 2008 was also the first year when school-level results have been made publicly available, and unique student identifiers have been introduced which remain the same throughout the years and enable tracking student performance.

The introduction of the unique identifiers made it necessary to bring all scores on a common ability scale which enables direct comparisons among all years and all grades. The way of analysis was renewed in 2010, but the ability scores obtained with this new method are available from 2008. The basis of the standardization became the literacy and mathematics results of the 6th-grade students in the year 2008, with the average set at 1500 and the standard deviation at 200. This means that ability scores from the same test (i.e. literacy or mathematics) can be compared among any two years and any two grades and allows for analysis of individual development, and the trends of development between grades. In addition to the ability scores, ability levels are also calculated based on the difficulty levels of the test items: the ability level of a student is the difficulty level at which they are expected to solve at least half of the exercises.

These ability levels serve as the basis for defining the minimum requirement for the schools to which the only direct consequence of the testing is attached. The *minimum level* is defined by law (Ministry of Human Capacities Decree 20/2012 (VIII.31)). According to this law, an action plan has to be prepared if at least half of the students did not reach the 2nd ability level in 6th grade and the 3rd ability level in 8th and 10th grade. It can be seen how low this requirement is if we compare it to the definition of the *base level* which is considered to be the necessary level to be able to gain further knowledge in the next grades and which is defined as the 3rd ability level in 6th grade and the 4th ability level for 8th graders and 10th graders. On average, around 8-10% of the schools are below the minimum requirement and have to prepare an action plan. This means that only a small fraction of schools are affected by direct incentives (since the preparation of an action plan is costly, and more importantly, it is a stigma for the schools), and thus, the NABC

each school. The selected students are assigned sampling weights in order to represent all PISA-eligible 15-year-olds.

Figure 1.1: The Structure of the NABC Test Books



Note: The figure shows the structure of the two NABC test books solved by students seated next to each other. The test comprises four 45-minute sections. The test books contain the same exercises, but the order of the test blocks is varied.

cannot be considered as a high-stake test.[6]

Besides not being a high-stakes test, there are other differences in the structure of the tests and the organization of the testing (compared to the Iowa Test of Basic Competencies, analysed by Jacob and Levitt, 2003). The test is conducted on a given day at the end of the school year (in May), there is no possibility of retaking it or taking it earlier than the given day. As Figure 1.1 shows, the test is divided into four parts, each lasting for 45 minutes. There are A and B test books so that neighbouring students are not solving the same exercises at the same time and the possibility of cooperation between students is reduced. The mathematics test consists of around 70 items, while the literacy test has somewhat more, around 100 items. These items can be either multiple choice questions (true-false and ordering exercises included) or open-ended questions, on average in a 7:3 ratio.

Special attention is devoted to quality assurance both during the test-taking and during the assessment of the test. A test conductor – a teacher of the school - is present in each classroom. They all work following uniform, centrally written guidelines, and they are responsible for the seating, handing out the exercise sheets, and monitoring students during test taking. In addition, during the first years of the testing (years 2008-2010) external monitors were sent to each school to monitor whether schools followed the rules when conducting the tests. In the following two years (2011 and 2012) external monitors were present in about 10% of the schools, while the practice was discontinued in 2013. Finally, contrary to many national student assessments, the correction of test books, recording and cleaning of the data is entirely done by external staff of the Ministry of

[6]For more on the Hungarian accountability system see Balázsi and Ostorics (2020). Note that some argue that free school choice creates an environment resembling hard accountability systems (Tóth, 2015). And while the importance of reputation creates indirect incentives to perform well on the test, it is still a low-stakes test compared to high-stakes tests where direct financial incentives are involved, or where students' grade promotion or graduation depends on it. *When talking about hard and soft accountability systems, other factors beyond testing are considered as well, i.e. they are broader concepts related to accountability than high-stake and low-stake testing.

Education which makes it less likely that teachers can alter the answers of students after the test.

These institutional differences are summarized in Table 1.2 and guide us in identifying the most likely form of cheating. First, since students have no direct interest in a good performance, I do not expect individual cheating to be prevalent. Schools and teachers have more incentives to achieve high scores since it can improve their reputation and attract more students. Second, since teachers have access to the test books only for a short period of time after the test has been conducted, it is not likely that they could alter the answers given by the students (as it is the case in other countries' national tests). If they would like to make sure students perform better than they actually could (i.e. they cheat), they have to do it during the test either by not monitoring the students very strictly and letting them cooperate, or by helping them individually, or collectively. This latter form seems to be the most efficient way if the aim is to reach a better overall performance in the class, and anecdotes are also supporting this hypothesis.

While there is only anecdotal evidence of cheating practices, with very few severe cases documented in the press[7][8], some studies assess teachers' views on the NABC testing (Tóth, 2011, 2015; Tóth and Csapó, 2022).

In their surveys of a representative sample of teachers in 2010, Tóth (2011) and Tóth and Csapó (2022) ask about the importance and utilization of the NABC results, the associated pressure, and preparation for the testing (focusing on ISCED levels 1-2, and 2-3, respectively). Both studies find that teachers rather do not think that the tests are a good way to objectively measure school performance, however, they rather agree that it incentivizes schools to exert more effort. Notably, teachers feel more pressure to do well on the NABC test relative to international tests (e.g. PISA). When asked about the pressure coming from the interested parties (government, school principals, parents, students), they rank school principals and themselves the highest, and students the last. There is evidence of teaching to the test since more than 90% of teachers claim that they prepare for the tests, and many use similar exercises as the NABC. However, they reject the statement that lower-performing students would not participate in the test (89% rather disagree or disagree).

[7] "World-beating mathematicians in Diosjeno" reports about a case when a school in a village with 2,800 inhabitants in one of the poorest regions of Hungary performed better on the 6th-grade testing than elite schools in Budapest. It also reports that most of those students continued their studies in vocational schools. Source: https://atlatszo.hu/kozugy/2017/10/04/vilagvero-matekosok-diosjenon-tobben-ketelkednek-a-kompetencia-felmeres-tisztasagaban/

[8] Teacher cheating is generally understudied in other fields (e.g., psychology, pedagogy) as well, since most studies focus on student cheating and related perceptions. There is only one survey, conducted by the (since dissolved) Association of Teachers and Lecturers in the United Kingdom which asked teachers about their own cheating. 35% of the participating 512 teachers stated they would be willing to cheat because of the increasing pressure to improve student performance, while some of them admitted teacher cheating is already present in their school. Source: https://www.theguardian.com/education/2012/apr/02/teachers-under-pressure-to-cheat.

Table 1.2: Comparison of the NABC, ITBS and INVALSI Tests

	NABC, Hungary	**ITBS, Chicago, US**	**INVALSI, Italy**
Target group of the test	6th, 8th and 10th grade	3-8th grade	2nd, 5th, 8th, 10th, 13th grade
Frequency	Each year (but students observed every 2nd year)	Each year	Each year (but students observed every 2nd-3rd year)
Fields	Reading, mathematics, (foreign language)	Reading, mathematics	Reading, mathematics, (English)
Test items	Multiple choice questions, open-ended questions	Multiple choice questions	Multiple choice questions
	A and B test books, with different orders of exercises		
Correction	Centrally	Centrally, but teachers check answer sheets	By teachers of the same school
External inspection	2008-2010: each school 2011-2012: 10% of schools 2013-: none	-	Since 2010: 7% of schools
Background information	From Student surveys and School surveys	Only gender, race and free lunch eligibility; and school-level characteristics	Student survey
School system	Change of school after 8th grade	No change during these years	Change of school at least once
Stake	Low-stake; 2013: min. requirement sets stakes for low-performing schools	High-stake; 1996: grade promotion of students, and probation for low-performing schools	Low-stake

Note: NABC: National Assessment of Basic Competencies, ITBS: Iowa Tests of Basic Skills, INVALSI: Italian National Evaluation Institute of the Ministry of Education

Tóth (2015) conducts semi-structured interviews with 70 teachers in 2012, focusing only on the sources of pressure. Teachers think that testing results affect the school choice of parents - interestingly, these beliefs were more present among primary school teachers. Teachers also expressed concerns regarding the publicity of results - arguing value-added should be published because raw results cover crucial differences in the student body of schools.[9] These studies indicate that although teachers have concerns about the effective use of the test results, they feel pressured to perform well and take the assessment seriously.

1.4 Dataset

I combine two student-level datasets from the Ministry of Education. The student-level researcher database contains results and background information on each individual student. Data on socioeconomic background is obtained from the student questionnaires that have to be filled together with the parents at home. Although this questionnaire is not compulsory, the participation rate is high and stable (around 80% in all grades). The student-level report files are administered during the correction process (coding of test books) and contain item-level data, i.e., each student's answers to every single test item.

The data is available for the years 2008-2017 for 6th, 8th and 10th grade.[10] However, since the algorithm developed by Jacob and Levitt (2003) requires both previous and future scores, the analysis is limited to 8th graders in the years 2010 to 2015 (see Figure A1.2 in the Appendix). Throughout these six years, 560 497 students were enrolled in the 8th grade, and they are all included in the initial raw datasets. However, not all of them participated in the test, either because they were exempted or absent for other unknown reasons. Students can be exempted if they are disabled, special needs students, temporarily injured or have language difficulties. Figure 1.2 shows the typical structure of the dataset (a representative class) for a given year. As it can be seen, around 3% of the students are exempted, and most of them are not taking the test.[11] More important is the relatively high share of students who are not exempted, but do not participate in the test (9.13%) which can be because of an illness, or just the student deciding to skip the day. Moreover, it is a reportedly frequent form of score manipulation to ask low-achieving students to stay home on the day of the test.[12]

Because of the features of the algorithm, not only absent students cannot be included in the analysis, but I cannot use the data of those students who missed the test in 6th

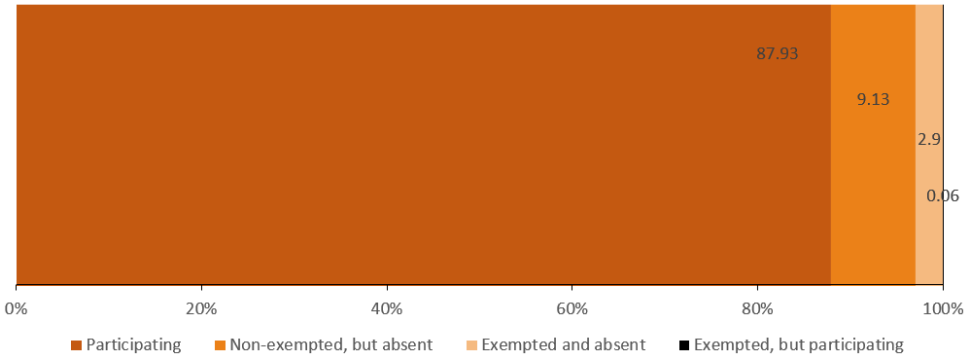
[9]Note, that since 2015 the Educational Authority also publishes the "list of highest value-added" and "list of highest improvements relative to previous year".

[10]For a yearly breakdown of the number of participating institutions and students in each grade, see Table A1.1 in the Appendix.

[11]Some of the exempted students still participate in the test, but it is a negligible ratio (0.04% of the whole population). They are going to receive their individual results, but will not be included in the overall reports of the school. For reasons of exemptions see Appendix Section A1.1.

[12]For more on test pool manipulation in the Hungarian testing, see Chapter 2 of this thesis.

Figure 1.2: A Representative Class of the Sample



Note: Exempted students include students with Special Education Needs, language difficulties, or temporarily injured. Percentages are based on all 8th graders in 2012.

and/or 10th grade. These include not only students who skip the test day because of illness or other reasons, but it is safe to assume that most of them were not even enrolled in the given grade. This can be because of grade repeating, or in grade 10 because they left the school system completely when turned 16 (school leaving age). Finally, since the algorithm has no power if there are not enough observations in a class, I drop all the classes which did not have at least 10 participating students after these restrictions. This results in losing around 25% of the classes and around 33% of the participating students.[13]

Table 1.3 contains summary statistics for the analysed period (2010-2015), for the whole population and the resulting subsample after dropping the observations which cannot be included in the analysis. It can be seen that dropping absent students and small classes naturally results in higher average class size and lower absence and exemption rates. The ratio of students who are too old for the given grade, or those who have repeated a grade before is lower for similar reasons. For example, if they repeated a grade before, they might participated in the test more than 2 years earlier, and thus, their 6th-grade result is not available. There is also a higher chance that they will repeat a class again, or leave the school system when turning 16 years old, which means their 10th-grade result can be missing. However, differences are less extreme in the two groups' other characteristics, such as the average performance, participation rate in the student questionnaire, or family background. This somewhat mitigates the problem of losing a lot of observations since it implies that we might lose cheating and non-cheating classes proportionally.

[13]I address the question in the simulation exercise in Section 1.6.2, while in Appendix Section A1.2 I discuss other potential concerns regarding the large share of data which cannot be included in the analysis.

Table 1.3: Summary Statistics, All Years Included

Classroom characteristics	Whole sample	Analysis sample
Class size, including absent students as well	20.21	22.57
% absent students in the class (on the test day)	9.05%	6.74%
Average number of participating students	-	16.70
% exempted students in the class of the student	2.82%	0.70%
% Special Need (SNI) students in the class of the student	2.74%	0.64%
% Male students	51.48%	50.55%
Average mathematics score	1592.30	1622.17
Average literacy score	1542.97	1576.42
Average of mathematics grade in previous school year	3.19	3.32
% students being too old for 8th grade	13.27%	8.43%
% students who answered at least 5 questions of the student questionnaire	79.59%	82.65%
* % students who have ever repeated a grade before	7.94%	4.91%
* Average family situation (1-5 scale)	2.95	3.00
* % students who get reduced-price or free lunch	38.30%	33.26%
* Median education of mothers in the class	14.37	14.67
* Median education of fathers in the class	14.26	14.48
* Mode of education of mothers in the class	14.33	14.66
* Mode of education of fathers in the class	14.22	14.44
% classes belonging to low-achieving schools (i.e. action plan needed)	9.55%	1.66%
Number of students	560 497	343 408
Number of classes	27 738	20 566

Note: The table presents summary statistics of class characteristics. Column (2) presents means for all classes in 8th grade throughout the years 2010-2015, while means in column (3) are computed only on the sample of classes included in the Jacob-Levitt analysis. The number of students includes absent students for the whole sample, but only participating students for the analysis sample.

*Questionnaire data: Only those students can be included who answered the related question in the student questionnaire.

1.5 Analysis: Test Score Manipulation in the Hungarian Testing

1.5.1 The Algorithm Developed by Jacob and Levitt (2003)

To detect potential cheating behaviour in the available data, as a first step I use the method developed by Jacob and Levitt (2003). Their algorithm relies on two indicators, the first one capturing unexpected test score fluctuations, while the second the suspicious answer strings in a class. According to their identification strategy, it is likely that some kind of cheating occurred in a class if both indicators are estimated to be high, i.e. if students had an outstanding performance compared to their previous and future performances and at the same time, high correlation in student answers is paired with high variance in correlations across questions.

Figure 1.3: A Potential Cheater Class

Multiple choice items													
64	65	70	71	72	73	74	75	76	78	79	80	81	82
1	2	3	1	6	4	2	5	2	1	1	3	1	1
4	2	3	1	6	4	2	5	2	1	1	3	1	1
1	1	3	1	6	4	2	5	2	1	1	1	1	4
3	2	3	1	6	4	2	5	2	1	1	1	1	4
1	4	3	1	6	4	2	5	2	1	1	1	1	4
1	2	3	1	6	4	2	5	2	1	1	1	1	4
1	2	3	1	6	4	2	5	2	1	1	1	1	4
3	2	3	1	6	4	2	5	2	1	1	1	1	4
1	2	3	1	6	4	2	5	3	1	1	1	1	4
1	2	3	1	6	4	2	5	2	4	1	1	1	4
2	2	3	1	6	4	2	5	2	1	1	2	1	4
2	2	3	1	6	4	2	5	2	4	1	2	1	4
1	2	3	1	6	4	2	5	2	1	1	3	1	4
1	2	3	1	6	4	2	5	2	1	1	3	1	4
1	1	3	1	6	4	2	5	2	1	1	3	1	4
1	1	3	1	6	4	2	5	2	1	1	3	1	4
1	2	3	1	6	4	2	5	2	1	1	3	1	4
3	2	3	1	6	4	2	5	2	1	1	3	1	4
1	1	3	1	6	4	2	5	2	1	1	3	1	4
1	1	3	1	6	4	2	5	2	1	1	3	1	4
1	2	3	1	6	4	2	5	2	1	1	3	1	4
1	2	3	1	6	4	2	5	2	1	1	3	1	4
1	2	3	1	6	4	2	5	2	1	1	3	1	4
1	1	3	1	6	4	2	5	2	1	1	3	1	4
1	2	3	1	6	4	2	5	3	1	1	3	1	4
1	2	3	1	6	4	2	5	3	1	1	3	1	4
1	3	3	1	6	4	2	5	3	3	1	3	1	4
1	3	3	1	6	4	2	5	3	3	1	3	1	4
3	2	3	1	6	4	2	5	2	1	3	3	1	4
1	1	3	1	6	4	2	5	2	1	1	1	3	4
1	1	3	1	6	4	2	5	2	1	1	1	4	4
1	1	3	1	6	4	2	5	2	1	1	3	4	4

Note: The figure shows a potential cheater class, highlighting the most unusual block of answers within the class. The example comes from the 2012 literacy test, the class is identified as suspicious using 95th and 90th percentile cutoffs for Indicator 1 and 2, respectively.

Figure 1.3 shows an example for such a suspicious pattern of answers. These are the

last exercises of the literacy test (64-82nd items) in 2012, so they are supposed to be the most difficult. It can be seen that there was an ordering exercise (from 1 to 6) which was solved by the whole class correctly.

In the following paragraphs I discuss each of the calculated indicators and measures in detail.[14]

Indicator 1 - Unexpected Test Score Fluctuations

The idea behind Indicator 1 is that cheating results in large test score gains (compared to the results in 6th grade) which cannot be sustained in 10th grade, i.e. performance either declines from 8th to 10th grade, or at least it does not improve as much as the national average.

First, I calculate the test score gains for each student from 6th to 8th grade, and from 8th to 10th grade, using the standardised ability scores calculated by the Ministry of Education. Then using these, I calculate average test score gains in 8th and 10th grade for each class. Having the average gains for each class, classes can be ranked and a variable containing their percentile rank can be created – for both 8th and 10th grade (pctrank_gain8, pctrank_gain10). Finally, indicator 1 is calculated using the following formula:

$$IND1_{class} = (pctrank\_gain8_{class})^2 + (1 - pctrank\_gain10_{class})^2 \tag{1.1}$$

which is high if a class had a relatively large average gain from 6th grade to 8th grade, but had a relatively small gain from the 8th to 10th grade (see signs in the formula). Taking the squared terms gives relatively more weight to the largest gains in 8th grade and the smallest gains (i.e. greatest declines) in 10th grade. Note that gains are not necessarily positive and students can change classes between the years, but that does not affect any of the calculations, the composition of the classes in 8th grade is used throughout the whole analysis.

Indicator 2 - Suspicious Answer Strings

Assuming teachers are helping the students by sharing answers with the whole class, or giving hints to the whole class, Indicator 2 aims to find patterns in the data which can be a sign of such cheating. It is constructed with a combination of four different measures.

Measure 1 - the most unlikely block of identical answers

For the 1st measure we need to calculate the likelihood of the observed data, asking the question what is the likelihood that all these students in the same class gave the same answer on all these consecutive questions. Probabilities are needed because checking only for answering patterns could lead us to too strong conclusions. For example, it should not be suspicious if high-ability students of a class answer all questions correctly, even though it produces a large block of identical answers.

For this reason we estimate a multinomial logit model for every test item, in which the

[14]For a detailed description of the approach see Jacob and Levitt (2002).

dependent variable is the students' answer to the given item. As explanatory variables, I included previous year's maths grade, gender, school size, previous and future NABC scores and a dummy indicating whether the student was exempted or not. These are available for all students in the subsample created for the analysis, thus, I do not lose more observations.

With the estimation of the model we can obtain the probability of the student choosing the actually chosen answer which will then be used to calculate the probability of the actual answer strings, and block of answers. This is done by simply taking the product of the chosen answers in the given string for each student, and then taking the product of these probabilities across students who were classmates and gave the same answers in the string. As Jacob and Levitt (2003), I also calculated the probabilities for 3-7 long answer strings[15]. Finally, M1class is the lowest predicted probability among all strings and blocks in the class.

Measure 2 – within-class correlation

Measure 2 is based on the same multinomial logit model as Measure 1, but it uses its residuals to capture how unexpected the student's actual response was. The residuals are first summed across students in the same class (for each possible outcome of each item), then these sums are combined the following way:

$$v_{item,class} = \frac{\sum_{outcomes} e^2_{outc,item,class}}{\text{number of students in the class}} \tag{1.2}$$

where e is the sum of students' residuals within a class.

The final measure, $M2_{\text{class}}$ is the classroom average of the above variance ($v_{i,c}$) across all items. This measure is thus adjusted not only for class size but item number as well. It is high if many responses are the same within the class, and even higher if these responses are unlikely according to the estimated multinomial logit model.

Measure 3 - variance of within-class correlation

Measure 3 is the cross-question variance of the within-class correlation and it is calculated as:

$$M3_{\text{class}} = \frac{\sum_{item}(v_{\text{item,class}} - \bar{v}_{\text{class}})^2}{\text{number of items}} \tag{1.3}$$

where $\bar{v}_{\text{class}}$ is actually $M2_{\text{class}}$.

The reason for introducing this measure is that a high M2 (i.e. high within-class correlation) can be a result not only of cheating, but e.g. teaching to the test during the school year, and practising certain types of exercises which make students in the same classroom particularly good at these. M3 is thus high if within-class correlation (M2) is

[15]Since the NABC also contains open-ended questions which cannot be analysed by the Jacob-Levitt algorithm, the consecutive items in the data might not be consecutive items in the test (there could be open-ended questions in between the multiple-choice items). But since the majority of the items are multiple choice questions, it can be assumed that they are 'close enough' to each other.

higher on a particular set of questions than on others.

Measure 4 - comparing response pattern to students with the same overall performance

Measure 4 compares the answers of students to all other students with the exact same final scores. Although it is not specified in the original paper what performance level has been used, in order to have enough observations at each score level, I am simply using the raw test scores (the ability scores have too high variation).

First, we need to calculate the mean responses $\bar{q}_i^A$ at each aggregate score level (for each item)[16], and then $Z_{\text{student}} = \sum_{item}(q_{\text{item,stud}} - \bar{q}_i^A)^2$ shows how much a student's response pattern differs from other students' who have the exact same final score. This measure is high if for example the student answers correctly to items which are relatively difficult for the students who are at the same aggregate score level as they are, or if they do not answer correctly to a relatively easy question. The final measure is obtained by the following formula:

$$M4_{\text{class}} = \sum_{student} (Z_{\text{student}} - \bar{Z}^A) \tag{1.4}$$

i.e. subtracting the mean (score-level) deviation from the student's own deviation, and summing this within the classes.

To obtain Indicator 2 - similarly to the first indicator - all classes are ranked based on the above 4 measures, and their percentile ranks are used:

$$IND2_{\text{class}} = (pctrank\_M1_{\text{class}})^2 + (pctrank\_M2_{\text{cl}})^2 + (pctrank\_M3_{\text{cl}})^2 + (pctrank\_M4_{\text{cl}})^2 \tag{1.5}$$

As previously mentioned, Jacob and Levitt (2003) argue that the above two indicators should be correlated if there was cheating in a classroom. Thus, a class is identified as a potential cheater if both indicators are high enough. Since it is hard to tell what is the threshold above which these should be considered high, they use three possible cutoffs for both indicators, at the 80th, 90th and 95th percentiles, and I follow their approach also in this.

I conduct the analysis for 8th graders in years 2010-2015. Since the number of test items is not the same in each test, the raw indicators and measures cannot be compared across years and subjects, and thus, I analyse each year and subject test separately. In the followings, I present the analysis for the year 2014, but Appendix Table A1.5 and Appendix Figure A1.6 show that the results are similar across years.

[16]Mean response is the mean of 'correctness' of answers, i.e. also can be seen as the fraction of students who answered the item correctly at the given score-level.

Estimated Prevalence of Cheating

Table 1.4 shows the average percentages of suspicious classes (i.e. classes scoring high on both indicators), depending on the chosen cutoffs.

Table 1.4: Percentage of Classes Scoring High on Both Indicators, 2014

Cutoff for Indicator 2 (Suspicious Answer Strings)	Cutoff for Indicator 1 (Large Test Score Fluctuation)		
	80th pctile	*90th pctile*	*95th pctile*
Mathematics			
80th pctile	4.03	1.85	0.80
90th pctile	2.04	0.83	0.46
95th pctile	1.02	0.40	0.18
Reading			
80th pctile	3.11	1.51	0.83
90th pctile	1.54	0.77	0.40
95th pctile	0.71	0.28	0.18

Note: The table shows the percentage of classes that score high on both indicators, using 80th, 90th and 95th percentiles as cutoffs for 'high' values. Sample size is $N_{class} = 3,379$.

However, it is important to account for the fact that classes could end up having high values on both indicators even by chance, or in the absence of cheating. Thus, the estimated prevalence of cheating is obtained by the following formula:

$$\hat{n}_{\text{cheat}} = n_{\text{hh}} - S_{\text{nc}} * A_{\text{nc}} \tag{1.6}$$

where n_{hh} is the ratio of classes having high values on both indicators, S_{nc} is the probability that a non-cheating class has a high value of Indicator 1, and A_{nc} is the probability that a non-cheating class has a high value of Indicator 2.

The probabilities that a non-cheating class would have a high value on Indicator 1 or Indicator 2 are estimated by the observed relationship between the two indicators on the - supposedly - cheating-free part of the data. Assuming that cheaters are in the last quarter of the distributions, and in the absence of cheating we would observe the same patterns in the 4th quarter as in the cheating-free 3rd quarter. We estimate S_{nc} by the average fraction of high test score gains (Indicator 1) among classes with typical answer strings (Indicator 2), and A_{nc} by the average fraction of unusual answer strings among classes with low test score fluctuation.[17]

The averages of the overall cheating frequencies estimated this way are summarized in Table 1.5. The obtained cheating ratios are negative, meaning there are even fewer classes scoring high on both indicators than one would expect in the absence of cheating.

[17]To illustrate this, Table A1.3 and Table A1.4 of the Appendix present the relationship between the

Table 1.5: Overall Prevalence of Cheating, 2014

Cutoff for Indicator 2 (Suspicious Answer Strings)	Cutoff for Indicator 1 (Large Test Score Fluctuation)		
	80th pctile	*90th pctile*	*95th pctile*
Mathematics			
80th pctile	−0.94	−0.67	−0.46
90th pctile	−0.65	−0.52	−0.22
95th pctile	−0.28	−0.25	−0.14
Reading			
80th pctile	−1.62	−0.93	−0.39
90th pctile	−0.86	−0.47	−0.22
95th pctile	−0.63	−0.41	−0.16

Note: The table shows the overall prevalence of cheating after correcting for the number of classes that would be above the cutoffs even in the absence of manipulation. Sample size is $N_{class} = 3,379$.

Thus, I find no evidence of test score manipulation using the method of Jacob and Levitt (2003a). There are two possible explanations for the obtained negative rates: (i) there is no cheating in the Hungarian NABC, or (ii) the algorithm developed by Jacob and Levitt (2003a) is not suitable to detect cheating in the Hungarian context. To get a better understanding of these possibilities, I apply other detection methods and run simulations.

1.5.2 Screening of the Hungarian Educational Authority

Although their findings are not published and it does not affect the evaluation of schools, the Hungarian Ministry of Education also conducts a screening to detect potential cheating schools. First, when encoding the students' answers, the encoders have to administer every case where they found at least three open-ended questions with the same word-by-word answer from at least five students within the same class. Subsequently, a class-level statistical analysis is performed to detect suspicious answer strings. Schools are notified by the Ministry if any of their classes were found to be suspicious, either during the encoding process or based on the algorithm.

The statistical analysis is based on five different indices. The first index ranks classes based on the number of items that were solved by at least 90% of the students in the class, and considers suspicious the top 3% of classes. The second index follows exactly the same steps as the first one but takes into account only the multiple-choice tests. The third index checks the 15 most difficult test items, and if at least one of them was solved by 90% of the students in the class, the class is considered suspicious. The fourth index considers classes suspicious if there are 2-4 multiple-choice items that were solved by less

two indicators in 2014.

than 10% of the students in the class (usually one such item appears in the class, while if there are at least 5 such items, it rather signals an overall bad performance than cheating). Since a truly high-achieving class could easily satisfy two of the above conditions to be considered suspicious, there is a fifth, so-called control index. This fifth condition is fulfilled if there is at least one open-ended item for which at least 90% of the class gave the same 'typical wrong answer'[18]. Schools are notified based on the statistical screening if a class is suspicious according to at least two indices of the first four and according to the control index, or if there are 10 items with a 'typical wrong answer' rate higher than 90%.

Table 1.6: Results of Conducting the Statistical Analysis of the Hungarian Ministry of Education

		%Suspicious classes	
	Based on	**Maths**	**Reading**
Index 1	items which were solved by at least 90% of the class	3%*	3%*
Index 2	multiple choice items which were solved by at least 90% of the class	3%*	3%*
Index 3	15 most difficult test items, at least 1 solved by 90% of the class	8.32%	2.50%
Index 4	2-4 multiple choice items which were solved by less than 10% of the class	21.12%	7.37%
Index 5 (control)	open-ended item, at least 90% of the class giving the same 'typical wrong answer'	0.2%	1.04%

Note: The table shows the definition of each index, and the corresponding shares of flagged classes by subject in year 2014. *by definition

Based on the above description, I have performed the same statistical analysis. Table 1.6 summarizes not only the definition of the indices, but the corresponding share of suspicious classes as well. Since the control index is always required for a class to be flagged as suspicious overall, the 0.2% and 1.04% represent an upper bound. In mathematics 4.8% of classes are flagged as suspicious based on at least two of the first four indices. When combined with the control index, this results in 0.11% of classes being notified due to suspected cheating (5 classes overall). For the reading test, 3.19% of classes are flagged based on the first four indices, while the overall occurrence is 0.22% (10 classes).[19]There are no classes which are found suspicious because they had at least 10 open-ended test items where more than 90% of students gave the 'typical wrong answer' (the maximum number of such test items in one class was 3, both in the mathematics and reading tests).

[18]The correction key contains the typical wrong answers as well, which are encoded differently than the unusual wrong answers. They are identified during the pre-testing period.

[19]Corresponding shares are higher in terms of schools.

1.5.3 Clustering Method by Quintano et al. (2009)

Quintano et al. (2009) developed a detection method for the Italian testing authority (Italian National Evaluation Institute of the Ministry of Education, INVALSI) which uses a clustering algorithm to classify classes in groups. It is based on four summary statistics: average score, standard deviation of scores, average rate of missing answers, and an answer homogeneity index. First, principal component analysis is used, and then a fuzzy clustering algorithm is applied to identify a suspicious group of classes and correct for the bias in the data.

As the previous approaches, this method was also developed to detect class-level cheating. The four measures are all calculated within classes: within-class average score, within-class standard deviation of scores, within-class average percent missing, and within-class index of answer homogeneity. Naturally, we expect a cheating class would be characterized by a high average score paired with a low standard deviation of scores. Assuming that, with the help of teachers, students are able to answer questions they would otherwise not even attempt, missing answers should be less common relative to non-cheater classes. Finally, answers in a cheating class are expected to be more homogeneous as well. The final measure uses item-level data to compute an index of answer homogeneity the following way. First, a Gini measure of homogeneity, E_{qc}, is computed for every test item (question) q:

$$E_{qc} = 1 - \sum_{s=1}^{S} \left(\frac{n_s}{N_c} \right)^2 \tag{1.7}$$

where $\frac{n_s}{N_c}$ is the ratio of students in class c who gave answer s to question q.
From this, the average Gini corresponds to the within-class index of answer homogeneity:

$$\bar{E_c} = \frac{\sum_{q=1}^{Q} E_{qc}}{Q} \tag{1.8}$$

To perform the clustering, first, a principal component analysis (PCA) has to be conducted in order to reduce the dimensionality of the data. Then, the k-means clustering algorithm can be applied for the classification of classes. The dimensionality reduction will enable us to present the results in a more comprehensive way and facilitate interpretation (see later Figure 1.4).

Before presenting the results of these two steps, I give a short overview of clustering techniques. The aim of k-means clustering is to partition the data into k clusters, such that the centroids best represent the points in the cluster. This representation is measured by the squared Euclidean distances between the centroid and each point in the given cluster. The criteria is to minimize the so called within-cluster sum of squares (WCSS)

$$\sum_{i=1}^{k} \sum_{x \in S_i} |x_i - \mu_i|^2$$

where μ_i is the cluster centroid of cluster S_i. The outcome of a standard (hard) k-means

clustering is the assignment of each data point to a cluster. Quintano et al. (2009) however uses fuzzy k-means clustering, which assigns a membership probability to each cluster for every data point.[20] The purpose of this is to use the assignment probabilities in a correction procedure and recover the undistorted test scores.

Table 1.7 and 1.8 present the results of the principal component analysis. It can be seen that the first two components explain 78% of the overall variance.[21] Table 1.8 shows the correlations between these two components and the four summary measures. We can see that Component 1 negatively correlates with the class average and positively with the Gini-measure, while Component 2 is strongly correlated with the standard deviation of the class average. Thus, a low score on Component 1 implies high class average and high answer homogeneity (low Gini), and a low score on Component 2 implies low within-class standard deviation. The suspicious cluster will be identified by applying this interpretation, which already suggests that it will be characterised by negative scores on both factors.

Table 1.7: Principal Component Analysis (PCA), Eigenvalues

Component	**Eigenvalue**		
	Total	**Proportion**	**Cumulative**
1	2.065	0.516	0.516
2	1.061	0.265	0.781
3	0.703	0.176	0.957
4	0.171	0.043	1.00

Note: The table presents the results of the principal component analysis applied to the 4 summary statistics of class characteristics. Column (1) shows the Eigenvalues of the correlation matrix, Column (2) the share of variance explained by each component, and Column (3) the cumulative share of explained variance. The PCA is performed on the 2014 mathematics data, excluding classes with $N_{partic} < 10$, and thus, the sample size is $N_{class} = 4,071$.

Table 1.8: Principal Component Analysis (PCA), Correlations

Summary measures	**Component**	
	1	**2**
Class average	-0.610	0.127
Standard deviation of average	0.156	0.909
Class non-response rate	0.431	-0.370
Index of answer homogeneity	0.646	0.146

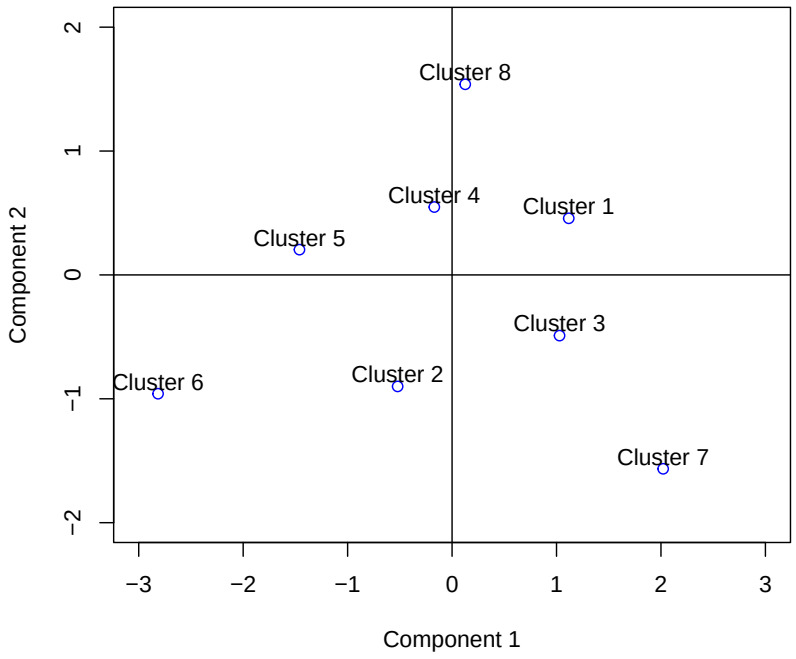
Note: The table presents correlations between the first two components and the summary measures. The PCA is performed on the 2014 mathematics data, excluding classes with $N_{partic} < 10$, and thus, the sample size is $N_{class} = 4,071$.

[20]Note that Angrist et al. (2017) (also in the Italian setting) applies hard clustering to assign each class to exactly one cluster, instead of assigning membership probabilities to all clusters.

[21]Note that in Quintano et al. (2009) the first two components explained more than 90% of the variance.

I follow Quintano et al. (2009) in applying the fuzzy k-means clustering such that it partitions the data into 8 clusters. Figure 1.4 shows the cluster centroids projected on the factorial plane, i.e. with Component 1 on the x-axis and Component 2 on the y-axis. Following the above interpretation, Cluster 6 stands out as having an outlier profile. This cluster consists of 290 classes which corresponds to 7.13% of the analysis sample and 6.4% of the total sample of classes. Classes in this cluster are characterized by high class average, low within-class standard deviation and high answer homogeneity, which potentially raise concerns about test score manipulation. Note however that truly outstanding classes with very high-performing students can easily exhibit the same characteristics since solving the test almost perfectly necessarily decreases standard deviation and increases answer homogeneity.[22]

Figure 1.4: Cluster Centroids



Note: The figure shows the cluster centroids obtained by the fuzzy k-means clustering algorithm with 8 clusters $s = 8$ and standard value of fuzzy partitioning degree $r = 2$. Cluster sizes are as follows: 577, 539, 605, 753, 502, 290, 301, 504 for Clusters 1-8, respectively. Data: 2014 mathematics test, excluding classes with $N_{partic} < 10$, with final sample size $N_{class} = 4,071$

While this method identifies approximately 7% of classes as suspicious, it is also important to note some of the potential limitations. This approach does not allow for empty clusters, i.e. no cheaters, since k-means clustering techniques do not perform well in cases when cluster sizes are different. In addition, there are potential parameters of the clustering process that

[22]Wrong answers can vary more, while there is only one correct answer. For a possible correction for false positives see Longobardi et al. (2018), and for further discussion see the next subsections and Section 1.7.

could be optimized, such as the number of clusters. While it is possible to assign classes to clusters, the average membership degrees in each cluster are between 0.54 and 0.58. In the following, I compare the results obtained by the different methods and discuss the most important limitations.

1.5.4 Summary and Discussion

In the above sections, I investigate test score manipulation in the Hungarian standardized testing applying three different detection methods, and find no evidence of systematic cheating. The algorithm of Jacob and Levitt (2003a) flags approximately 4% of classes as suspicious. However, this corresponds to fewer classes scoring high on both indicators than we would expect based on other parts of the distributions. Note that even without the benchmark, 4% of classes being above the 80th percentile cutoffs corresponds exactly to what would be expected by chance. The simple method of the Hungarian Ministry of Education flags 0.1% and 0.32% of classes. Finally, the clustering method identifies 7% of classes as suspicious, however, the suspicious cluster is the smallest. The rare overlap between the flagged classes (see Section 1.6.1) suggests that there are no extreme outliers that could be confidently identified.

The outcome of the detection methods is supported by the normal distribution of raw scores as well.[23] Overall, descriptives, results of the detection methods and the incentive structure all point in the direction that we should not be concerned about substantial cheating.

I argue that the absence of cheating at the top, or the lower prevalence of cheating compared to other testing systems, can be explained by differences in the structure and organization of the testing. While the Hungarian NABC is a low-stakes test for students and entails only reputational concerns for schools and teachers, it is also characterized by strict quality assurance (external monitoring, A and B test books, new test items each year, testing on a single day, and centralized correction) which makes it costly to cheat.

The results provide suggestive evidence that the presence of cheating can be limited by thoughtfully designing the incentive structure. When the incentives to manipulate test scores are low, and the associated costs are high, cheating is less likely to happen. This is in line with the literature investigating the impact of varied incentives on cheating behaviour. Several studies focus on the effect of direct monetary incentives on misreporting across various settings (Balafoutas et al., 2020; Martinelli et al., 2018). More closely related are studies in the education economics literature which evaluate or exploit policy changes aimed at reducing cheating. Dee et al. (2019) find that the introduction of centralized correction of tests resulted in reduced cheating, while others assess how external monitoring can mitigate manipulation concerns (Bertoni et al., 2013; Borcan et al., 2017; Lucifora and Tonello, 2020).

However, it is important to note that incentive schemes successfully reducing test score manipulation might give rise to other strategic behaviours. In Chapter 2 of this dissertation, I

[23]See Appendix Figure A1.7. In contrast, studies investigating cheating in the Italian testing found the smoking gun already when looking at the distribution of test scores. "Looking only at the second year of primary classes, the considerable presence of outlier classes has produced an unimodal distribution where the mode is equal to the top score" (Quintano et al., 2009, p. 156), meaning many students had perfectly solved answer sheets.

provide evidence on test pool manipulation in Hungary, i.e., students selectively participating in the test. This has been observed in other testings as well, in the form of grade retentions, exemptions, or longer disciplinary suspensions (Figlio, 2006; Figlio and Getzler, 2002; Jacob, 2005). The above-mentioned study by Lucifora and Tonello (2020) not only shows that external monitoring is an effective way to deter cheating, but it might prompt higher absence rates. Additionally, they find that sanctions do not reduce cheating, but limit absences, and in the presence of reputational concerns both monitoring and sanctioning have a larger effect.

Moreover, the methods have two important limitations which cannot be overlooked in the Hungarian context.[24] First, if cheating happened in all years since the testing was introduced, and some students are more likely to be in cheater schools throughout all their studies, the Jacob-Levitt method cannot find them. The concern can be somewhat mitigated by the fact that I focus on 8th graders and students typically change schools between 8th and 10th grade. Second, all detection methods assume that cheating happens at the top. However, in the absence of high stakes, but with reputational concerns, schools might prioritize avoiding the bottom of rankings rather than aiming for the top. This would lead to cheating in the bottom and middle of the distribution rather than at the top. The question arises whether we should only care about cheating at the top. While this might be of lesser importance from a ranking perspective, it becomes important as underperforming schools face sanctions. Methodologically, another question is whether the presence of cheating at the bottom could affect detection at the top (e.g. by corrupting the benchmark).

In the following section, I provide an overview of the methods and explore some of the suggested limitations using simulations.

1.6 Evaluation of Detection Methods

1.6.1 Comparison of the Three Methods

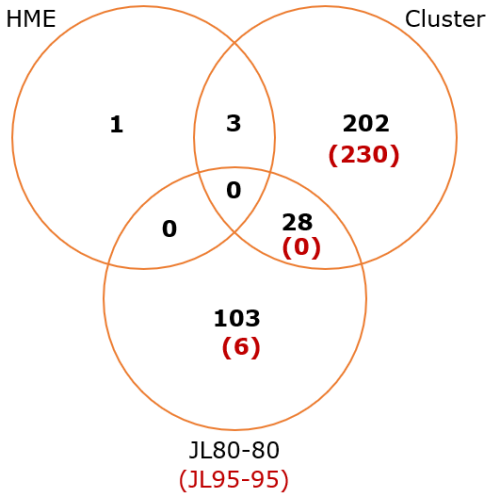
As we have seen above, none of the applied detection methods suggest the presence of systematic test score manipulation in the Hungarian standardized testing. However, it is still worth comparing the results, and how the group of classes identified as suspicious overlap in the three cases. When considering all the measures used as part of the detection methods (for an overview see Appendix Table A1.7), except for correlations between direct performance measures (e.g. number of items solved by at least 90% of the class, class average), no strong correlations are found.

One would think that although the number of suspicious classes is low, those that are flagged have to be extreme cases found by all methods. When comparing the set of classes identified as suspicious with the different methods, there is no evidence of that. Figure 1.5 shows the overlap between the classes flagged as suspicious by the three methods in the 2014 mathematics test. It can be seen that HME found 4 cheater classes, and while 3 out of the 4 are also in the suspicious

[24] But note that Jacob and Levitt (2003a) also use a setup where sanctioning of the worst-performing schools was introduced.

cluster, none of them were above the 80th percentile cutoffs in the Jacob-Levitt case. There are also very few classes that are found both by the clustering and the Jacob-Levitt method. This might happen because in contrast to the other two methods, Jacob and Levitt (2003a) can take into account student ability which enables a differentiation between the truly good classes and the cheaters.

Figure 1.5: Comparison of Suspicious Cases



Note: The figure shows the overlap in the identified suspicious classes using the three different methods (HME: Hungarian Ministry of Education, Cluster: Quintano et al. (2009), JL: Jacob and Levitt (2003a)). Numbers in the Venn diagram represent the number of classes. In case of the Jacob-Levitt method, black numbers correspond to the usage of 80th percentile cutoffs, while red numbers in parentheses correspond to the usage of 95th percentile cutoffs.

Given these findings, we need to acknowledge the drawbacks of these detection methods. There is no one-size-fits-all approach, and the assumptions made by each method may differ based on the expected level and form of cheating. Additionally, the comparison between these methods is not straightforward due to variations in data requirements, sample sizes, and the possibility of no cheating.

Table 1.9 summarizes the most important advantages and drawbacks of each method. The expected characteristics of suspicious classes are similar. All methods are looking for cheating at the top of the distribution, assuming cheater classes reach a high average score with low within-classroom variation in the answers.

While Jacob and Levitt (2003a) relies on two indices to decrease the possibility of false positives, the method still has its limitations. Using past and future test scores enables accounting for student ability, but as a result, we are unable to detect cases when manipulation happened in all years. It also has higher data requirements which makes its application more restricted.

Finally, it is important to note how all these methods rely on assumptions about the expected cheating occurrence in the data. The chosen cutoffs or the chosen number of clusters determine

the upper bound on the level of cheating that can be identified. This requires an in-depth understanding of the institutional context. Similarly, it is also assumed that cheating happens only at the top. In the next section, I aim to shed light on the potential consequences of such assumptions by comparing the performance of the Jacob-Levitt algorithm in five scenarios in which cheating happens at different parts of the distribution.

Table 1.9: Summary of Detection Methods

	HME	**Clustering**	**Jacob-Levitt**
Characteristics of suspicious classes	high test score	high test score	high test score fluctuation
	same wrong answer	answer homogeneity	answer homogeneity
		few missing answers	unlikely answers
Pros/Cons	minimal data requirement	minimal data requirement	high data requirement (past and future scores)
	prone to false positives	prone to false positives	accounts for student ability
		no empty cluster	cannot find those who always cheat
Assumptions	cutoffs, number of items	number of clusters	cutoffs, cheating in Q4, benchmark is Q3

Note: The table summarizes the main characteristics and assumptions of the different detection methods applied in this paper. HME: screening of the Hungarian Ministry of Education, Clustering: method developed by Quintano et al. (2009), and the algorithm by Jacob and Levitt (2003a).

1.6.2 Simulations: Jacob-Levitt Method

To assess how the algorithm performs in different scenarios, I run simulations. I add artificial cheating to the data by changing answers to the correct solutions in a subset of schools, on a subset of answers. Then, I run the Jacob-Levitt algorithm on these manipulated data sets. To measure the algorithm's performance, I use the share of correctly identified classes, the share of correctly identified cheaters and non-cheaters separately, and the share of false positives and false negatives. I consider cheaters those classes that are above the 80th percentile cutoff according to both the 'large test score fluctuation' and the 'suspicious answer string' indicators.[25]

I consider five scenarios which are summarized in Table 1.10. I divide the classes into four equal-sized groups (quarters) to consider the worst performers (bottom 25%), the mid-performers

[25]Note that in the full analysis, this step is followed by a correction (according to the benchmark). Since the correction only applies to the total number of cheater classes, cheater classes cannot be identified. All classes above the cutoffs are cheaters with a probability of $\frac{1}{N}$, where N is the number of classes above the cutoff on both indicators.

(middle 50%) and the highest performers (top 25%). Scenario 1 is the 'textbook example of cheating' that was mostly assumed by the development of different detection algorithms as well: when those at the bottom and middle of the performance distribution cheat such that they get on the top of the distribution. Considering different scenarios is important for two reasons. First, as mentioned above, schools might not aim to be at the top, but to avoid being at the bottom. Second, while all detection methods focus on detecting cheating at the top, the experimental literature shows both theoretically and empirically that individuals do not tend to cheat to the maximum possible extent (Abeler et al., 2019; Gneezy et al., 2018).

Table 1.10: Simulation Scenarios

(1)	(2)	(3)	(4)	(5)	(6)
Scenario	**Overall occurrence**	**Which quarter**	**Points needed**	**# items changed**	**Effective cheating**
1. Textbook cheating	15% (20% of each 3 Qs)	$1 \rightarrow 4$	14	26	14.02 (.11)
		$2 \rightarrow 4$	8	18	7.67 (.05)
		$3 \rightarrow 4$	3	8	2.97 (.05)
2. Quartile jump	15% (20% of each 3 Qs)	$1 \rightarrow 2$	4	7	3.84 (.05)
		$2 \rightarrow 3$	3	7	3.17 (.04)
		$3 \rightarrow 4$	3	8	2.97 (.05)
3. Bottom-to-Mid	15% (60% of the 1st Q)	$1 \rightarrow 2$	4	7	3.84 (.02)
4. Bottom-to-Top	15% (60% of the 1st Q)	$1 \rightarrow 4$	14	26	14.02 (.04)
5. Mid-to-Top	15% (30% of each 2 Qs)	$2 \rightarrow 4$	8	18	7.67 (.04)
		$3 \rightarrow 4$	3	8	3.24 (.03)

Note: The table shows the five different simulation scenarios. Overall occurrence shows the share of classes where manipulation was added, and is the same in all scenarios. Classes are ranked according to their average mathematics score, and grouped into four equal-sized groups. Column (3) shows in which quarter(s) manipulation is added, and all else equal, in which quarter the affected classes would end up with the added manipulation. Column (4) shows the points needed to make the jump between the given quarters, while Column (5) shows how many test items should be manipulated to gain the necessary points (since some answers were correct even without the manipulation). Column (6) shows the average points gained (and standard deviations) for each manipulated quarter, i.e. checks whether the necessary gain was achieved.

I make three important assumptions when simulating the cheating scenarios. First, the procedure assumes that teachers and school principals are aware of their students' abilities and how their school performance compares to other schools (in which quarter they are). This is a reasonable assumption since 8th-grade students' performance was once assessed 2 years earlier in 6th grade, and since the student body of schools does not vary a lot from year to year, i.e. it is possible to predict how the current 8th graders will perform based on how the school's 8th graders did last year. Most importantly, I assume that cheating is 'naive' in the sense that it does not anticipate others to cheat. This means that anticipating correctly the level of their own and other schools' performance, cheating teachers choose the number of exercises to manipulate while taking all else constant. This is important because introducing dynamic incentives could

result in a trickle-down effect where everyone has to cheat to the maximum extent and reach the perfect score. Finally, I assume that the necessary level of manipulation is calculated as the difference between the minimum score in the 'destination quarter' and the mean of the 'origin quarter'. Appendix Table A1.8 shows summary statistics of raw scores by quarters. As an example, imagine Scenario 1, and a class in the 3rd quarter aiming to get in the top quarter. The average score in the 3rd quarter is 34, while the minimum score in the 4th quarter is 37. Thus, as an expectation, they will aim to gain 3 points with the manipulation.[26]

The steps of the simulation are as follows:

1. Random selection of classes where manipulation will be added to the data (depending on the scenario, see Table 1.10).
2. Adding manipulation, i.e. changing the students' answers to correct answers in the selected classes. The number of test items changed depends on the Scenario and in which quarter the school is (how much cheating is needed). The test items changed are selected randomly in each class, and following the logic of Jacob and Levitt, it is always consecutive test items that are selected within a class.
3. Dropping too small classes, according to the criteria of Jacob and Levitt (2003), i.e. keeping classes with at least 10 participating students whose past and future scores are also available.
4. Running the Jacob-Levitt algorithm.

For each Scenario, this is repeated 100 times (100 rounds).

Considering the whole set of classes is important, because it is a major limitation of many detection methods that they do not have power in small classes, and thus, small classes can never be subject to the screening or detection process. It can be argued that cheating in small classes does less harm than in small classes, however, it still should not be ignored. This way, my simulation also sheds light on how the exclusion of small – and potentially cheater – classes matters.

For the simulations I use the data on the 2010 mathematics test, assuming it is relatively cheating-free data. This is not only supported by the above analysis, but even within the observation period, it is the year when external monitors were still present in all schools. In the literature it is generally argued that external monitoring reduces cheating (Bertoni et al., 2013; Longobardi et al., 2018; Pereda-Fernández, 2019).[27]

[26] In practice I had to select more test items for manipulation than the points needed based on these expectations. This is parallel with cheating in reality because teachers also have to provide the correct answer for more questions since some students would be able to answer correctly even without any help. To continue the above example, to gain 3 points in the 3rd quarter, 8 test items have to be manipulated. Table A1.9 shows the average scores obtained on multiple-choice questions by quarter. The mean divided by the total number of multiple-choice questions (39) corresponds to the probability of randomly selecting a correct test item. Thus, in the 3rd quarter, the number of items to be changed is obtained by $\frac{3}{1-24.82/39} = 8$.

[27] Using a simulated dataset would have challenges in this specific context. Since the detection method requires item-level data, individual responses would need to be generated as well. If this is performed by a multinomial logit model, the detection process (also relying on multinomial logits) could lead to a tautology.

The level of cheating has two dimensions: (1) how many classes are cheating (extensive margin), and (2) how much manipulation is added by the cheater classes, i.e., how many points they gain (intensive margin). I refer to the first as overall occurrence, following the terminology used in Section 1.5.1. This can be seen in Column (2) of Table 1.10. The second involves how many students get help on how many test items. For low-ability classes, there is more room for cheating than for high-ability classes. If teachers give the correct answers to a certain number of questions, the effective average gain will be less than the number of manipulated answers since some students know the answer even without the teacher's help. Column (6) in Table 1.10 presents the level of effective cheating, i.e., the average points gained, which corresponds to the second dimension of the level of cheating. In the following simulations, first, I keep the level of overall occurrence constant (at 15%) with varying scenarios. Second, I vary the overall occurrence while keeping the 'Textbook' scenario fixed.

Overall Occurrence Fixed, Varying Scenarios

For comparability, I keep the level of overall manipulation constant across scenarios by always adding manipulation to 15% of classes.

Figures 1.6–1.8 present the averages of 100 simulation rounds for each Scenario.[28] Figure 1.6 shows the share of classes that were correctly identified, separately for cheaters and non-cheaters, while Figure 1.7 shows the share of correctly identified cheater classes both relative to the whole sample and the analysis sample. It can be seen that the algorithm does not perform well overall. It finds 7.6% of cheater classes within the analysis sample when cheating happens at the bottom, and to an extent that it reaches the top. This is the largest jump to be made, as shown in Table 1.10 the average gain has to be 14 points which can be reached by changing 26 test items.[29]. However, it can also be seen that the performance is much worse if considering the overall sample. This is because small classes are overrepresented in the bottom quarter, and thus, they are more likely not to be subject to the screening at all. As expected, it performs the worst when cheating happens at the bottom, without getting to the top (Scenarios 2 and 3). In these cases, only 4-5% of cheaters are found in the analysis sample, and around 3% in the whole sample. This provides evidence that the algorithm performs better at looking for cheating at the top.

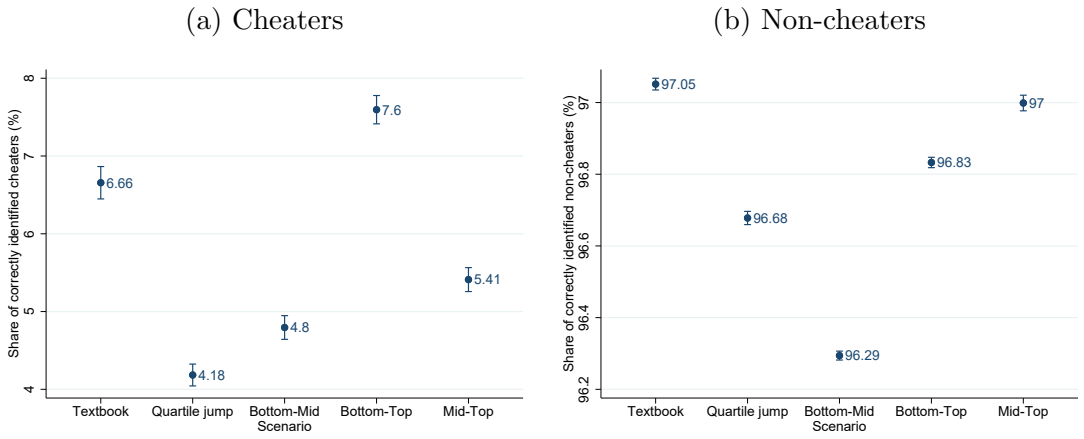
Next, I focus on another aspect of the algorithm's performance, the number and share of false positives. We tend to focus on the false negatives because we do not want the cheaters to get away with their cheating easily. However, false cheating accusations can be equally harmful. This was seen in Italy when the first version of Quintano et al. (2009)'s method was applied to correct the school rankings, while it was very prone to false positives.[30] Using algorithms prone to false positives undermines any incentives to improve.

[28]The corresponding results are also summarized in Appendix Tables A1.10 and A1.11.

[29]In the bottom quarter students answer correctly on average 46% of the questions. Thus, to gain 14 points on average, 26 test items have to be manipulated.

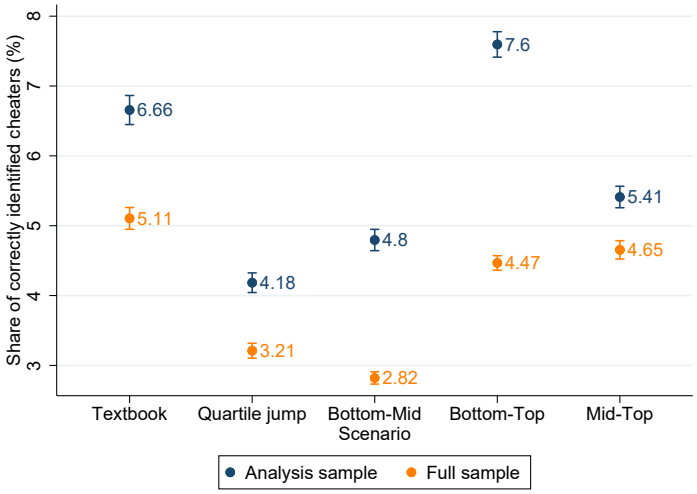
[30]See Section 1.6.3 for discussion and an example.

Figure 1.6: Share of Correctly Identified Classes in the Analysis Sample



Note: The figures show for each scenario the average share of correctly identified classes in 100 simulation rounds, with 95% confidence intervals. Panel (a) presents the share of correctly identified cheater classes, while panel (b) presents the share of correctly identified non-cheater classes. Share is considered relative to the number of cheater classes and non-cheater classes in the analysis sample, respectively. The analysis sample consists of classes with at least 10 students who participated in the test in all grades (6, 8, 10). Overall number of classes in the analysis sample is the same across scenarios and rounds: $N_{class} = 3,920$. The number of cheater and non-cheater classes depends on how many small classes were selected as cheaters. The average number of cheater classes in the analysis sample is between 432 and 630. The corresponding results are also summarized in Appendix Table A1.10.

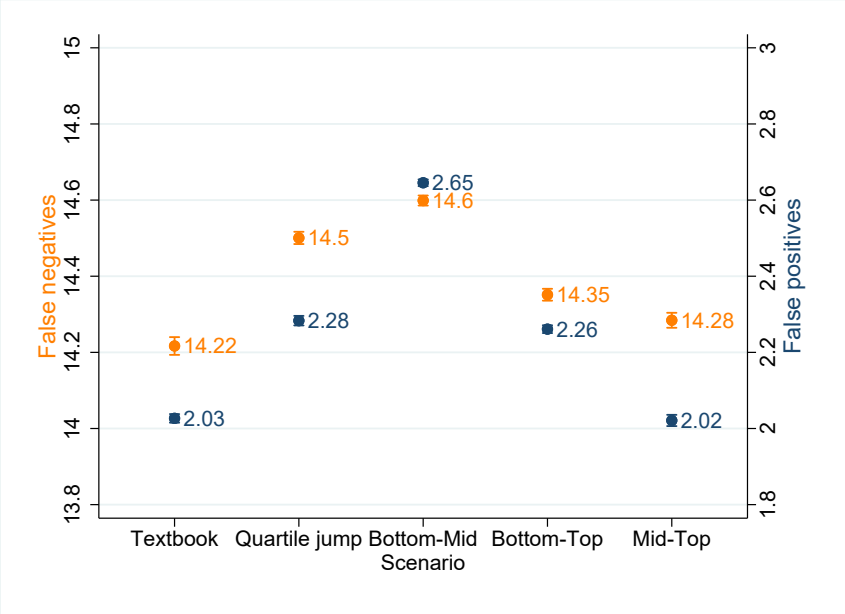
Figure 1.7: Share of Correctly Identified Cheater Classes in the Full Sample



Note: The figure shows the average share of correctly identified cheater classes in 100 simulation rounds, with 95% confidence intervals. It compares the shares considering only the analysis sample (blue), and considering the full sample (orange). The corresponding results are also summarized in Table A1.10.

Figure 1.8 presents the average shares of false positives and false negatives in the full sample of classes. As in the last columns of Table A1.10, small classes are considered to be classified as non-cheaters by the algorithm. The results do not show substantial differences across the scenarios. The share of correctly classified positives is below 1%.[31] The share of false negatives is around 2-3%, while the share of false positives is around 14%. [32]

Figure 1.8: Share of False Negatives and False Positives



Note: The figure shows average shares of false negatives and false positives from the 100 simulation rounds, for each cheating scenario. False negatives (orange): manipulated classes that are not classified as suspicious by the Jacob-Levitt algorithm. False positives (blue): non-manipulated classes that are classified as suspicious. Shares are computed relative the number of all classes (in the full sample) thus, the share of false positives, share of false negatives and share of correctly identified classes add up to 100. Simulation is performed on data from the 2010 mathematics test. Overall number of classes $N = 4,886$. The corresponding results are summarized in Table A1.11.

Overall, we do not see large differences according to any of the performance measures. The algorithm does not perform well in any of the scenarios (finds 21-37 of the approximately 730 cheater classes). Even when taking into account that classes at the bottom are more likely to be small, and thus, be dropped from the analysis sample, it barely finds 10% of the remaining cheaters. This might be the case because the level of added cheating is relatively large (15%). Jacob and Levitt (2003a) performs a simulation where manipulation is added to a single class each round, and the performance of the algorithm is measured by whether this one class was

[31] In case all cheaters would be found, this value would be 15%. Considering the classes dropped, the maximum value could be between 7% and 12% (when 46.5% and 82.2% of cheater classes are in the analysis sample).

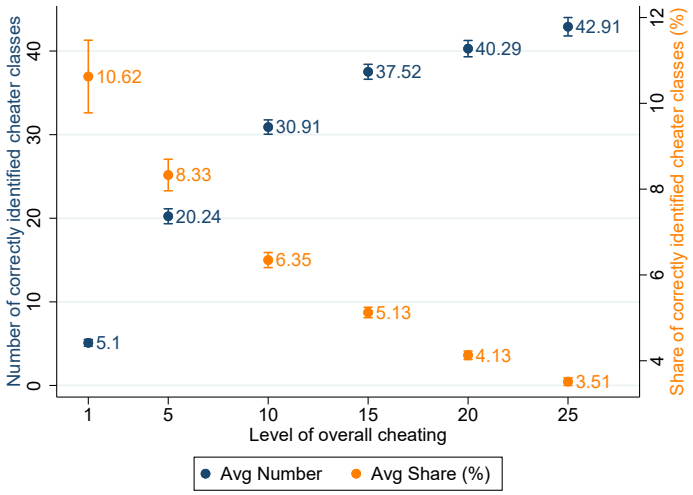
[32] Note that the simulation is not able to account for the correction described in subsection 5.1 which adjusts for the number of classes that would be above the cutoffs by both indicators even in the absence of cheating. When using the 80th percentile cutoffs, this is around 4% of the analysis sample (as Tables 1.4 and 1.5 showed), and the share of false negatives should be compared to this.

found or not. This, however, does not take into account the fact that the algorithm relies on estimating a multinomial logit model on the observed (and potentially corrupted) sample to predict the probabilities of correct answers. In case there is substantial cheating in the data, the corrupted test items are going to seem easier overall, and the predicted probability of correct answers increases, making those correct answers less suspicious.

Varying Overall Occurrence, Fixed Cheating Scenario ('Textbook Cheating')

To consider this, I run an additional simulation of Scenario 1 with different levels of overall occurrence (1, 5, 10 15, 20 and 25%). Figure 1.9 shows the number and share of correctly identified cheater classes, depending on the level of overall cheating. It can be seen that while the number of correctly identified cheaters is increasing with the level of overall cheating, the share of correctly identified cheaters is declining. This suggests that the higher levels of cheating in the data can substantially affect the assessment of answering probabilities which are a crucial part of the Jacob-Levitt method. Note that using the 80th percentile cutoffs, the highest possible level of cheating to be found is 20% (in case the top 20% of classes by the two indicators would coincide). This again points towards the importance of the initial assumptions about the occurrence of cheating.

Figure 1.9: Varying Levels of Cheating



Note: The figure shows the average number (blue) and share (orange) of correctly identified cheater classes with 95% confidence intervals, depending on the level of overall added manipulation. The overall level is defined as the share of manipulated classes relative to the total number of classes (e.g. if overall cheating is 5%, 244 classes are cheaters, since the overall sample of classes N=4,886). The share of correctly identified cheater classes (orange) equals the number of correctly identified classes divided by the number of cheater classes. Scenario 1 is implemented on the initial data from the 2010 mathematics test. For each level, 100 simulation rounds were conducted.

1.6.3 Discussion

Section 1.6 offers a glimpse into how the suggested limitations come into play. The simulations show that the performance of the Jacob-Levitt algorithm diminishes significantly as the level of overall manipulation increases, and it performs even worse when cheating schools do not end up at the top of the ranking. This suggests that the method is most suitable to find the few extreme cheaters when cheating is not widespread.

While I am focusing on the performance of the detection methods, exploring further questions within these simulations can help to understand how the limitations matter. First, one could look at which classes are consistently flagged as suspicious, and whether any unmanipulated classes are found suspicious in all scenarios. As a robustness check, these classes could be excluded from the analysis since they might be genuinely engaging in cheating. Second, investigating how the values of the cheating measures and indicators change, and how the cutoffs are affected by the added manipulation would shed light on the underlying mechanism.

Additional simulations can enable modifications and corrections of the algorithm. In order to uncover less extreme cheating (not at the top), the role of cutoffs and benchmarks should be studied. For example, how do the results change if we assume cheating is not at the top, and the cheating-free part of the distribution is not the 3rd quarter. Moreover, it is also worth investigating how the algorithm's performance is affected by absences, especially when the test pool is deliberately manipulated.[33] An additional measure could be introduced to find classes with large test score gains and low participation rates.

In a rare collaboration with the testing administration (Chicago Public Schools, CPS), Jacob and Levitt (2003a) also validated their results by a follow-up testing where suspected cheaters were unable to sustain their performance in a monitored environment. In another paper (Jacob and Levitt, 2003b) they provide more details on the follow-up audits and retesting, arguing that it also allows them to suggest potential improvements to their method (which included some arbitrary assumptions on functional forms and weighting). They suggest that, instead of combining the indicators, it might be sufficient to consider suspicious those classes that score high on any of the indicators. This would be similar to the screening method of the Hungarian Ministry of Education (see Section 1.5.2). They assess how effectively each measure predicts test score declines (in the retest), and find that M4 predicts them the best, M2 is the second best, while M3 has no significant relationship with the decline.[34] This implies Indicator 2 should be created by giving a larger weight to M4, and even consider eliminating M3. However, to my knowledge, none of these modifications were explored later on.

Another interesting avenue for future research would be to estimate the trade-off between simpler and more complex detection techniques. It should be assessed whether the use of more complex methods is justified by their better performance. If more complex methods are only able to find extreme outliers, while having a high data requirement, then it might be better to opt

[33] See Appendix Section A1.2 for a discussion on non-random absences, and Chapter 2 of the thesis on test pool manipulation.

[34] M1: most unusual block of answers, M2: Within-classroom correlation in answers, M3: cross-question variance in the within-classroom correlations, M4: comparison of students with the exact same final scores.

for the simpler statistics. This is especially true when such methods are used solely for screening and not directly for punishment. If screenings are followed by in-depth investigations, then the higher occurrence of false positives might not be a problem. The cost of screening approaches and follow-up investigations should also be assessed, and weighted against each other.

It is also important to note that any punishment based on statistical probabilities has to be carefully designed. For example, after applying the clustering method for the testing, the Italian testing authority published a corrected ranking of schools which resulted in a storm of indignation by those who were falsely accused of cheating.[35] If the method is prone to false positives, true high-performers will be punished. Moreover, if schools are aware that unexpected performances might be under suspicion, low-performers will have less incentive to improve. The Dutch childcare benefit scandal also serves as a cautionary example of how the blind reliance on algorithms might backfire (Peeters and Widlak, 2023). At the beginning of the 2010s, as a response to benefit frauds, the Dutch tax authority introduced anti-fraud measures which included the application of a self-learning algorithm to flag suspicious claims for childcare benefits. While the scandal erupted because discriminatory profiling was part of the algorithm, it is also important to notice the complete absence of checks and balances. This, again, underlines the importance of follow-up investigations and other preventive measures.

Finally, the poor performance of the algorithm and the long list of limitations naturally raise the question of why detection methods should still be employed. Screening and detection processes can act as a deterrence device by influencing the (perceived) probability of detection. This aligns with findings in many other contexts. Dionne et al. (2009) show that, for insurance frauds, the optimal auditing strategy is to first screen insurance claims using specific indicators, and then refer the flagged suspicious cases to a special investigative unit. Block et al. (1981) show theoretically how the optimal price of a cartel depends on the effort level of antitrust enforcement and the size of the penalties, while Laine et al. (2020) provide experimental evidence on the role of varying detection risk and penalty size. Kleven et al. (2011) study tax evasion under different audit regimes in a large-scale field experiment in Denmark, and find that self-reported income increases under audit threats, and it is positively related to the probability of auditing. Cronert (2022) shows how even unenforced regulation can have an effect because organizational factors, norms and culture all matter. Theoretical and experimental work indicates that keeping enforcement efforts low or revealing enforcement choices might be optimal for authorities, depending on the sophistication and preferences of offenders (Buechel et al., 2020; Calford and DeAngelo, 2023; Lazear, 2006). These findings suggest that even in the absence of punishment – i.e., only screening and notifying suspicious schools – there can be sufficient deterrence power.

[35] Source: "Invalsi test suspects fraud for the most brilliant institutes, the ministry refuses" (Test Invalsi, sospetto brogli per gli istituti più brillanti Il ministero rifiuta i compiti) https://www.ilgiorno.it/milano/cronaca/2013/10/06/961194-invalsi-brogli.shtml

1.7 Conclusion

Standardized testing data provides important information to researchers, policy makers, teachers and parents about the state of the education system, and thus, it is crucial to ensure the reliability of such data. This paper investigates potential cheating behaviour in a low-stake testing environment in Hungary, and provides an overview of different detection methods and their limitations.

To detect manipulations in the Hungarian standardized testing (NABC) data, as a first step I employ the method developed by Jacob and Levitt (2003a). Their algorithm relies on two indicators, the first one capturing unexpected test score fluctuations, while the second suspicious answering patterns in a class. According to their identification strategy, it is likely that some kind of cheating occurred in a class if both indicators are estimated to be high, i.e. if students had an outstanding performance compared to their previous and future performances, and at the same time, high correlation in student answers is paired with high variance in correlations across questions.

Using this established method, no test score manipulation is found in the Hungarian testing. To assess the robustness of this result, I apply two additional detection methods: the screening of the Hungarian Educational Authority, and a clustering method proposed by Quintano et al. (2009). Based on the combined results, I conclude that systematic cheating, particularly to an extreme extent that would put cheater classes at the top of the ranking, is unlikely.

I argue this is because the low-stake nature of the Hungarian testing is paired with strict quality assurance which makes it costly to cheat. While previous studies analysed tests where correction was administered within the schools, in Hungary, tests are centrally corrected, making it less likely that teachers can alter the answers of students after completion. The usage of new test items each year, and A and B test books reduces the possibility of coordination during test taking. A more direct comparison with other countries is however not possible because of the different tests.

The results provide suggestive evidence that the presence of cheating can be limited by thoughtfully designing the incentive structure. When the incentives to manipulate test scores are low, and the associated costs are high, cheating is less likely to happen. As a policy implication, the findings underscore the significance of two key factors in ensuring the reliability of testing data: avoiding high stakes and reducing the possibility of manipulation (e.g. by centralised correction or varied test items).

However, it is important to note that all the applied methods have limitations. The second part of the paper reviews the advantages and disadvantages of the screening methods and, with a simulation exercise, sheds light on some of the possible drawbacks. I show that the Jacob-Levitt algorithm performs better in cases when manipulation happens at the top of the distribution. However, its effectiveness diminishes significantly as the level of overall manipulation in the data increases. This is because the algorithm relies on estimating the likelihood of answers based on the same observed data that is subject to these manipulations. By this, it overlooks the possibility that substantial cheating could bias these estimates.

In summary, the research findings indicate that there is no one-size-fits-all approach for detecting cheating in educational settings. The choice of method depends on the specific institutional setting and the expected level and form of cheating. Moreover, the methods themselves strongly depend on the assumptions made on the expected cheating behaviour, including assumptions on the form of manipulation (how do they manipulate) and on the extent of manipulation (how many schools cheat and to what extent). Therefore, further validation and benchmarking are necessary to enhance the effectiveness of these detection methods.

One possible avenue for validation and benchmarking is to draw upon the cartel detection literature. This field has developed techniques for identifying collusive behaviour and conducted systematic simulation studies (Huber and Imhof, 2019; Imhof et al., 2018), which could provide insights and methodologies that can be applied to the detection of cheating in educational settings. Additionally, external monitoring, as proposed by Longobardi et al. (2018) or Pereda-Fernández (2019), can be used to establish benchmark distributions for cheating detection.[36] Alternatively, randomized retesting, as suggested by Jacob and Levitt (2003a), can serve as a valuable approach to validate the effectiveness of detection methods.

Exploring validation techniques, or conducting simulation studies to systematically investigate the potential drawbacks this paper only shed light on, are potential avenues for further research within the literature on detection methods. These could enhance our understanding of cheating behaviours and improve detection techniques.

Despite their limitations, detection methods remain valuable tools in the hands of policy makers and testing authorities. Even in the absence of monetary punishment or direct enforcement, these methods can serve a crucial role in maintaining integrity in educational settings. First, they serve as deterrence devices, increasing the perceived probability of getting caught and deterring potential cheating. Second, they also function as screening mechanisms, helping authorities in allocating their resources effectively. They may offer a cost-effective alternative to extensive external monitoring, e.g., by allowing targeted monitoring only in the most suspicious schools. Future research should assess the costs and benefits of both screening methods and external monitoring. Exploring their trade-off is essential for understanding how these preventive measures can complement and reinforce each other, thereby ensuring the integrity of educational assessments.

[36]For more on the natural experiment of external monitoring incorporated in the Italian testing, see Bertoni et al. (2013) and Lucifora and Tonello (2020). In the Hungarian context, monitoring practices were in place until 2012. For more, see Appendix section A1.3.

References

Abeler, J., Nosenzo, D., and Raymond, C. (2019). Preferences for Truth-Telling. *Econometrica*, *87*(4), 1115–1153.

Ajzenman, N. (2021). The Power of Example: Corruption Spurs Corruption. *American Economic Journal: Applied Economics*, *13*(2), 230–257.

Angrist, B. J. D., Battistin, E., and Vuri, D. (2017). In a Small Moment: Class Size and Moral Hazard in the Italian Mezzogiorno. *American Economic Journal: Applied Economics*, *9*(4), 216–249.

Arnold, I. J. M. (2016). Cheating at Online Formative Tests: Does it Pay Off? *Internet and Higher Education*, *29*, 98–106.

Balafoutas, L., Czermak, S., Eulerich, M., and Fornwagner, H. (2020). Incentives for Dishonesty: An Experimental Study with Internal Auditors. *Economic Inquiry*, *58*(2), 764–779.

Balázsi, I., and Ostorics, L. (2020). The Hungarian Educational Assessment System. In Harju-Luukkainen, H., McElvany, N., and Stang, J. (Eds.) *Monitoring Student Achievement in the 21st Century: European Policy Perspectives and Assessment Strategies* (pp. 157–169). Springer.

Battistin, E., De Nadai, M., and Vuri, D. (2017). Counting Rotten Apples: Student Achievement and Score Manipulation in Italian Elementary Schools. *Journal of Econometrics*, *200*(2), 344–362.

Battistin, E., and Neri, L. (2023). School Performance, Score Inflation and Neighborhood Development. *Journal of Labor Economics*, *41*(3).

Becker, G. S. (1968). Crime and Punishment: An Economic Approach. *Journal of Political Economy*, *76*(2), 169–217.

Bertoni, M., Brunello, G., Benedetto, M. A. D., and Paola, M. D. (2021). Does Monitoring Deter Future Cheating? The Case of External Examiners in Italian Schools. *Economics Letters*, *201*, 109742.

Bertoni, M., Brunello, G., and Rocco, L. (2013). When the Cat is Near, the Mice Won't Play: The Effect of External Examiners in Italian Schools. *Journal of Public Economics*, *104*, 65–77.

Bilen, E., and Matros, A. (2021). Online Cheating Amid COVID-19. *Journal of Economic Behavior and Organization*, *182*, 196–211.

Bíró, A., Prinz, D., and Sándor, L. (2022). The Minimum Wage, Informal Pay, and Tax Enforcement. *Journal of Public Economics*, *215*, 104728.

Block, M. K., Nold, F. C., and Sidak, J. G. (1981). The Deterrent Effect of Antitrust Enforcement. *Journal of Political Economy*, *89*(3), 429–445.

Bø, E. E., Slemrod, J., and Thoresen, T. O. (2001). Taxes on the Internet: Deterrence Effects of Public Disclosure. *American Economic Journal: Economic Policy*, *7*(1), 36–62.

Borcan, O., Lindahl, M., and Mitrut, A. (2017). Fighting Corruption in Education: What Works and Who Benefits? *American Economic Journal: Economic Policy*, *9*(1), 180–209.

Borisova, E., and Peresetsky, A. (2016). Do Secrets Come Out? Statistical Evaluation of Student Cheating. *Applied Econometrics*, 119–130.

Buechel, B., Feess, E., and Muehlheusser, G. (2020). Optimal Law Enforcement with Sophisticated and Naïve Offenders. *Journal of Economic Behavior and Organization*, *177*, 836–857.

Cagala, T., Glogowsky, U., and Rincke, J. (2021). Detecting and Preventing Cheating in Exams: Evidence From a Field Experiment. *Journal of Human Resources*, 0620–10947R1.

Calford, E. M., and DeAngelo, G. (2023). Ambiguity and Enforcement. *Experimental Economics*, *26*(2), 304–338.

Cronert, A. (2022). When the Paper Tiger Bites: Evidence of Compliance With Unenforced Regulation Among Employers in Sweden. *Regulation and Governance*, *16*(4), 1141–1159.

Dahl, G. B., Engelberg, J., Lu, R., and Mullins, W. (2023). Cross-State Strategic Voting. *NBER Working Paper*, *No. 30972*.

Dee, T. S., Dobbie, W., Jacob, B. A., and Rockoff, J. (2019). The Causes and Consequences of Test Score Manipulation: Evidence from the New York Regents Examinations. *American Economic Journal: Applied Economics*, *11*(3), 382–423.

Dee, T. S., Jacob, B. A., Rockoff, J. E., and McCrary, J. (2011). Rules and Discretion in the Evaluation of Students and Schools: The Case of the New York Regents Examinations. *SSRN Electronic Journal*.

Dellavigna, S., and Ferrara, E. L. (2010). Detecting Illegal Arms Trade. *American Economic Journal: Economic Policy*, *2*(4), 26–57.

Diamond, R., and Persson, P. (2016). The Long-term Consequences of Teacher Discretion in Grading of High-stakes Tests. *NBER Working Paper*, *No. 22207*.

Dionne, G., Giuliano, F., and Picard, P. (2009). Optimal Auditing with Scoring: Theory and Application to Insurance Fraud. *Management Science*, *55*(1), 58–70.

Ferrer-Esteban, G. (2013). Rationale and Incentives for Cheating in the Standardised Tests of the Italian Assessment System. *Fondazione Giovanni Agnelli Working Paper*, *N. 50*.

Figlio, D., and Loeb, S. (2011). School Accountability. In Hanushek, E. A., Machin, S., and Woessmann, L. (Eds.) *Handbook of the Economics of Education* (Vol. 3, pp. 383–421). Elsevier.

Figlio, D. N. (2006). Testing, Crime and Punishment. *Journal of Public Economics*, *90*(4), 837–851.

Figlio, D. N., and Getzler, L. S. (2002). Accountabilty, Ability and Disability: Gaming the System. *NBER Working Paper*, *No. 9307*.

Gneezy, U., Kajackaite, A., and Sobel, J. (2018). Lying Aversion and the Size of the Lie. *American Economic Review*, *108*(2), 419–453.

Gustafsson, M., and Deliwe, C. N. (2017). Rotten Apples or Just Apples and Pears? Understanding Patterns Consistent With Cheating in International Test Data. *Stellenbosch Economic Working Papers*, *No. 17/2017*.

Hanushek, E. A., and Kimko, D. D. (2000). Schooling, Labor-Force Quality, and the Growth of Nations. *The American Economic Review*, *90*(5), 1184–1208.

Holmstrom, B., and Milgrom, P. (1991). Multitask Principal–Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. *The Journal of Law, Economics, and Organization*, *7*, 24–52.

Horn, D. (2012). Catching Cheaters in Hungary: Estimating the Ratio of Suspicious Classes on the National Assessment of Basic Competencies Tests. *Research Centre for Economic and Regional Studies of Hungarian Academy of Sciences and Eötvös Lóránd University*.

Huber, M., and Imhof, D. (2019). Machine Learning With Screens for Detecting Bid-Rigging Cartels. *International Journal of Industrial Organization*, *65*, 277–301.

Humbert, M., Lambin, X., and Villard, E. (2022). The Role of Prior Warnings When Cheating is Easy and Punishment Is Credible. *Information Economics and Policy*, *58*, 100959.

Imhof, D., Karagök, Y., and Rutz, S. (2018). Screening for Bid Rigging - Does It Work? *Journal of Competition Law and Economics*, *14*(2), 235–261.

Jacob, B. A. (2005). Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools. *Journal of Public Economics*, *89*, 761–795.

Jacob, B. A., and Levitt, S. D. (2002). Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating. *NBER Working Paper*, *No. 9413*.

Jacob, B. A., and Levitt, S. D. (2003a). Rotten Apples: an Investigation of the Prevalence and Predictors of Teacher Cheating. *The Quarterly Journal of Economics*, 843–877.

Jacob, B. A., and Levitt, S. D. (2003b). Catching Cheating Teachers: The Results of an Unusual Experiment in Implementing Theory. *Brookings-Wharton Papers on Urban Affairs*, *2003*(1), 185–220.

Jacobsen, C., Fosgaard, T. R., and Pascual-Ezama, D. (2018). Why Do We Lie? a Practical Guide To the Dishonesty Literature. *Journal of Economic Surveys*, *32*(2), 357–387.

Janke, S., Rudert, S. C., Änne Petersen, Fritz, T. M., and Daumiller, M. (2021). Cheating in the Wake of COVID-19: How Dangerous Is Ad-Hoc Online Testing for Academic Integrity? *Computers and Education Open*, *2*, 100055.

Kertesi, G., and Kézdi, G. (2005a). Általános iskolai szegregáció, I. rész. Okok és következmények [Segregation in the Primary-School System, I. Causes and Consequences]. *Közgazdasági Szemle (Economic Review-monthly of the Hungarian Academy of Sciences)*, *52*(4), 317–355.

Kertesi, G., and Kézdi, G. (2005b). Általános iskolai szegregáció, II. rész. Az általános iskolai szegregálódás folyamata Magyarországon és az iskolai teljesítménykülönbségek [Primary-School Segregation II. The Process of Primary-School Segregation in Hungary and Performance Differences Between Schools]. *Közgazdasági Szemle (Economic Review-monthly of the Hungarian Academy of Sciences)*, *52*(5), 462–479.

Kertesi, G., and Kézdi, G. (2011). The Roma/Non-Roma Test Score Gap in Hungary. *American Economic Review*, *101*(3), 519–525.

Kisfalusi, D., Janky, B., and Takács, K. (2021). Grading in Hungarian Primary Schools: Mechanisms of Ethnic Discrimination Against Roma Students. *European Sociological Review*, *37*(6), 899–917.

Kleven, H. J., Knudsen, M. B., Kreiner, C. T., Pedersen, S., and Saez, E. (2011). Unwilling or Unable to Cheat? Evidence from a Tax Audit Experiment in Denmark. *Econometrica*, *79*(3), 651–692.

Laine, T., Silander, T., and Sakamoto, K. (2020). What Distinguishes People Who Turn Into Tax Evaders When Properly Incentivized From Those Who Don't? An Experimental Study Using Hypothetical Scenarios. *Journal of Behavioral and Experimental Economics*, *85*.

Lazear, E. P. (2006). Speeding, Terrorism, and Teaching to the Test. *The Quarterly Journal of Economics*, *121*(3), 1029–1061.

Lin, M.-J., and Levitt, S. D. (2020). Catching Cheating Students. *Economica*, *87*(348), 885–900.

Longobardi, S., Falzetti, P., and Pagliuca, M. M. (2018). Quis Custiodet Ipsos Custodes? How to Detect and Correct Teacher Cheating in Italian Student Data. *Statistical Methods and Applications*, *27*(3), 515–543.

Lucifora, C., and Tonello, M. (2015). Cheating and Social Interactions. Evidence From a Randomized Experiment in a National Evaluation Program. *Journal of Economic Behavior and Organization*, *115*, 45–66.

Lucifora, C., and Tonello, M. (2020). Monitoring and Sanctioning Cheating at School: What Works? Evidence from a National Evaluation Program. *Journal of Human Capital*, *14*, 584–616.

Martinelli, C., Parker, S. W., Pérez-Gea, A. C., and Rodrigo, R. (2018). Cheating and Incentives: Learning from a Policy Experiment. *American Economic Journal: Economic Policy*, *10*(1), 298–325.

Meghir, C., and Palme, M. (2005). Educational Reform, Ability, and Family Background. *The American Economic Review*, *95*(1), 414–424.

Murnane, R. J., Willett, J. B., and Levy, F. (1995). The Growing Importance of Cognitive Skills in Wage Determination. *The Review of Economics and Statistics*, *77*(2), 251–266.

Peeters, R., and Widlak, A. C. (2023). Administrative Exclusion in the Infrastructure-Level Bureaucracy: The Case of the Dutch Daycare Benefit Scandal. *Public Administration Review*, *83*(4), 863–877.

Pereda-Fernández, S. (2019). Teachers and Cheaters: Just an Anagram? *Journal of Human Capital*, *13*(4), 635–669.

Quintano, C, Castellano, R, and Longobardi, S. (2009). A Fuzzy Clustering Approach to Improve the Accuracy of Italian Student Data. An Experimental Procedure to Correct the Impact of Outliers on Assessment Test Scores. *Statistica e Applicazioni*, *7 (2)*, 149–171.

Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests.* University of Chicago Press.

Tóth, E. (2011). Pedagógusok nézetei a tanulói teljesítmény-mérésekről [Educators' Views on Student Assessments]. *Magyar Pedagógia*, *111*(3), 225–249.

Tóth, E. (2015). Az országos kompetenciamérés hatása a tanítási munkára pedagógusinterjúk alapján [The Impact of the National Assessment of Basic Competencies on Teaching Practices, Based on Teacher Interviews]. *Magyar Pedagógia*, *115*(2), 115–138.

Tóth, E., and Csapó, B. (2022). Teachers' Beliefs About Assessment and Accountability. *Educational Assessment, Evaluation and Accountability*, *34*(4), 459–481.

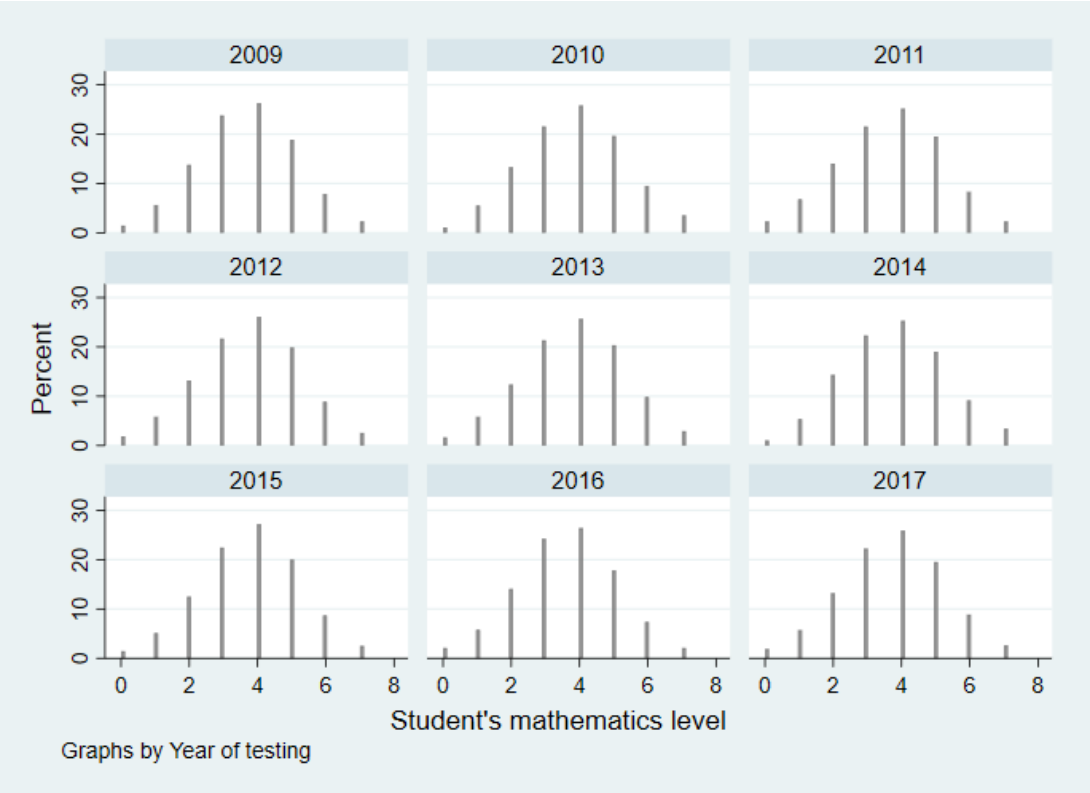
Appendix

A1.1 Institutional Background II

Ability Scores and Ability Levels

The introduction of the unique identifiers made it necessary to bring all scores on a common ability scale which enables direct comparisons among all years and all grades. The basis of the standardization became the literacy and mathematics results of 6th-grade students in the year 2008, with the average set at 1500 and the standard deviation at 200. This means that ability scores from the same test (i.e. literacy or mathematics) can be compared among any two years and any two grades, and allows for analysis of individual development, and the trends of development between grades.

Figure A1.1: Distribution of Mathematics Ability Levels



Exemptions

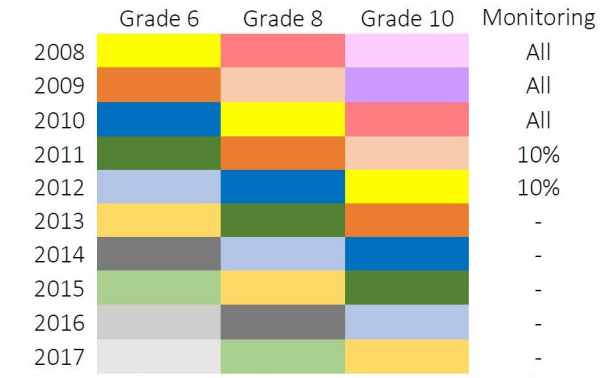
Schools have to indicate in their November data provision (online) which students will be exempted and why, but this can be modified until the testing day. School coordinators have to indicate reasons for absences on the attendance sheet following the below coding.

Code 1	Students with a certificate (doctor's) about their disability (physical, sensory, mental disorders)
Code 2	Temporary injury (e.g. broken arm)
Code 3	Language difficulties (studies in Hungarian for less than a year)

There are other students whose scores are not taken into account when school averages are calculated, but who cannot be exempted from the test:

b, t, m (i, l, c)	Students with integration, learning (e.g. dyslexia) or conduct disorder Included in school averages as well
P	Other developmental disorders, Not included in school averages

Figure A1.2: The Structure of the Data



Note: The figure presents the structure of the data, highlighting the cohorts for which both past and future scores are available. The same colours indicate the same cohorts of students.

Table A1.1: The Hungarian School System in Numbers

	Institutions	**School sites**	**Classes**	**Students**
6th grade				
2008	2 351	2 909	5 051	107 654
2009	2 255	2 849	4 763	100 620
2010	2 195	2 797	4 628	96 898
2011	2 180	2 779	4 535	94 047
2012	2 126	2 714	4 453	92 082
2013	2 128	2 702	4 445	93 907
2014	2 150	2 684	4 438	92 544
2015	2 199	2 688	4 463	91 956
2016	2 187	2 669	4 418	90 834
2017	2 216	2 671	4 428	91 599
8th grade				
2008	2 444	2 966	5 202	108 194
2009	2 347	2 925	5 006	104 230
2010	2 285	2 880	4 981	104 266
2011	2 266	2 844	4 726	96 843
2012	2 222	2 794	4 587	92 966
2013	2 216	2 776	4 484	89 913
2014	2 232	2 766	4 461	87 542
2015	2 279	2 769	4 500	88 967
2016	2 278	2 756	4 495	88 382
2017	2 297	2 751	4 485	87 990
10th grade				
2008	1 035	1 166	4 081	112 409
2009	995	1 158	3 961	108 960
2010	978	1 153	3 874	107 274
2011	972	1 148	3 792	102 705
2012	957	1 147	3 767	102 037
2013	955	1 138	3 662	95 649
2014	963	1 120	3 648	90 188
2015	954	1 114	3 793	85 683
2016	712	1 123	3 789	85 061
2017	696	1 115	3 680	84 957

Note: The table presents the number of institutions, schools, classes and students enrolled in each year and grade in the Hungarian school system, for the years 2008-2017.

A1.2 Missing Not at Random

Observing high rates of non-random absences in the NABC data raises the question of how this can affect the detection process. Since most detection methods have to rely on somewhat arbitrary cutoffs when determining the "too suspicious" levels, ranking of the classes matters. If missing is not at random, this can result in a high rate of false negatives or false positives. If cheaters are more likely to be missing or dropped from the dataset, a higher number of non-cheaters will likely be incorrectly identified as cheaters. On the other hand, if non-cheaters are more likely to be missing, this can result in a higher number of false negatives.

Table A1.2 presents the ratio of dropped observations because of missing values, both in my analysis and in Jacob and Levitt (2003)'s. It can be clearly seen that a large amount of data is lost because of missing future scores. The reason for this is that 'future score' in the context of the NABC only means 10th-grade scores, while the school leaving age is 16 years and students can potentially leave the education system before the 10th grade. It is also quite intuitive that during the two years between the observations more changes can happen than during the one year in case of the ITBS (e.g. more students can be lost because of leaving the Hungarian school system, or just having to repeat a grade).

Table A1.2: Share of Dropped Observations

	ITBS, Chicago US	**NABC, Hungary**
Missing students – in the 'baseline' year	13%	8.53%
Missing students – only in the past or in the future	12%	23%
Small classes (with less than 10 'valid students' after the above exclusions)	3% of students	8% of students

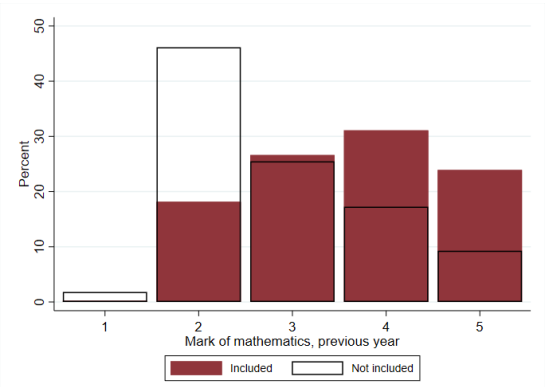
Note: The table presents the ratio of dropped observations because of missing values in the current analysis using NABC data, and in the ITBS data used by Jacob and Levitt (2003).

Although I have shown in Section 1.4 that class-level averages and characteristics in the analysed subsample are similar to the characteristics observed in the whole population, here I also look at differences between students. Since low-performing students are expected to be more likely to finish studying at the age of 16, I look at the differences in school performance. Figure A1.3 presents the distribution of mathematics grades prior to the testing year. Panel (a) shows that students included in the analysis are more on the right tail of the distribution, and a particularly high ratio (around 46%) of dropped students has just reached the passing grade. Panel (b) focuses on students who disappear from the system (at least on the test day) between 8th and 10th grade, and shows similar patterns as the previous figure.

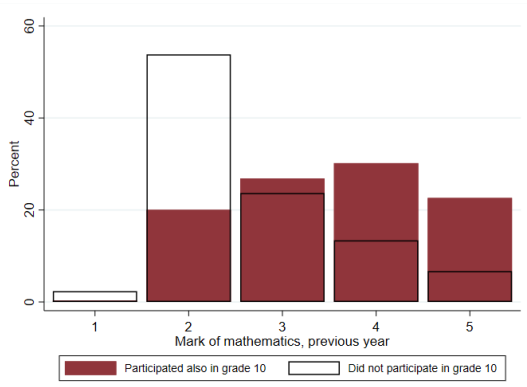
Thus, the data suggest some evidence that missing might not be at random. It follows that it would be important to further investigate how missing data can affect the validity of the Jacob-Levitt algorithm.

Figure A1.3: Distribution of Previous Year's Mathematics Marks

(a) Included vs Dropped



(b) Participation in 10th Grade



Note: The figures present the distribution of previous year's mathematics grade by participation. Figure (a) presents the distribution of mathematics grades for the students included in the analysis, and the students who were dropped because of missing data. Figure (b) presents the distribution of mathematics grades, focusing on a subset of the dropped observations, students who did not participate in the testing in 10th grade. The blue columns show the score distribution of students who participated both in 8th and 10th grade, while the transparent columns belong to students who participated in 8th grade only, and not in the 10th. Both histograms are based on 8th graders in 2014. *In Hungary a 5-point grading scale is used, 1 being the worst, 5 the best grade.

A1.3 External Monitoring

Besides using approaches which proved to be successful in detecting cheating in other settings, it can be useful to look at the raw data focusing on policy changes which can affect the prevalence of cheating. Such a policy is the practice of sending external monitors to schools which has been shown to be associated with a decrease in test scores in Italy (Bertoni et al., 2021, 2013; Lucifora and Tonello, 2015, 2020).

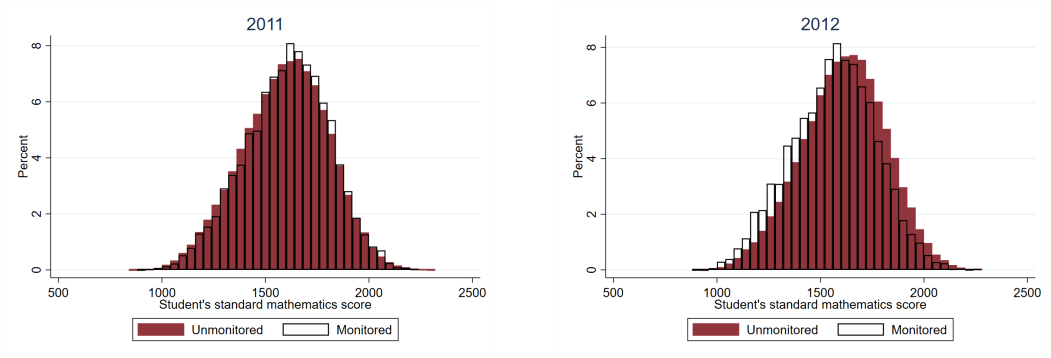
The Hungarian Educational Authority had a similar practice until 2012. More precisely, there was an external monitor in each school until 2010, while approximately 10% of the schools were monitored in 2011 and 2012. This means that we can observe groups of schools which became unmonitored at different points in time. In the following, I try to exploit this heterogeneity in the data.

The next figures compare the score distribution of monitored and unmonitored classes in the two years when both groups are observed. If external monitoring is reducing cheating, then one would expect unmonitored classes to have higher scores than monitored ones who could not cheat. In 2011 we observe the opposite, monitored classes were doing better, while in 2012 - in line with our expectations - unmonitored classes indeed performed better than monitored ones (Kolmogorov-Smirnov tests suggest imbalances as well).

However, it is important to note that even though according to the Educational Authority monitors are sent randomly to schools[37], there seems to be a difference between the two years

[37]In 2011 the regional directorates of the Ministry of Education, in 2012 county-level government offices

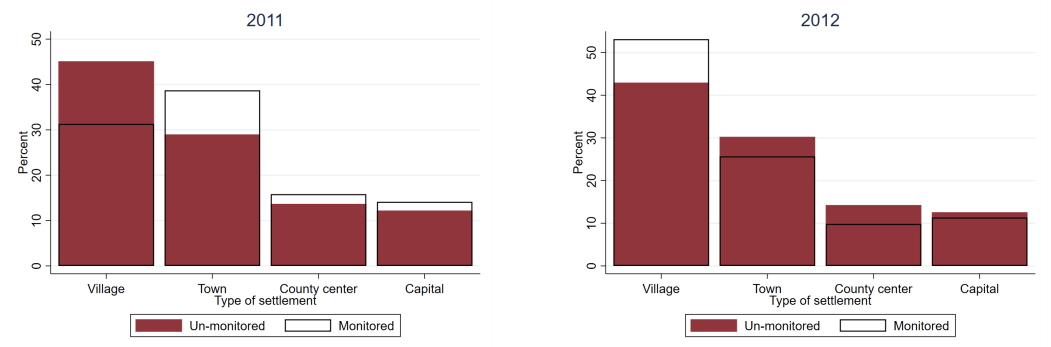
Figure A1.4: Score Distributions of Monitored and Unmonitored Classes



Note: The figures show the distribution of the average scores on the mathematics test for monitored and unmonitored classes in 2011 and 2012.

in terms of the allocation of monitors. Figure A1.5 suggests that in 2011 schools in cities were more likely to be monitored than village schools, and for 2012 this turned around. Thus, it will be crucial for any analyses to deal with this randomization problem.

Figure A1.5: Type of Settlement of Monitored and Unmonitored Classes



Note: The figures show the distribution of settlement types for the monitored and unmonitored classes separately.

determined which schools will be monitored, but they had no guideline from the Ministry of Education regarding the selection process.

A1.4 Additional Results, Jacob-Levitt Algorithm

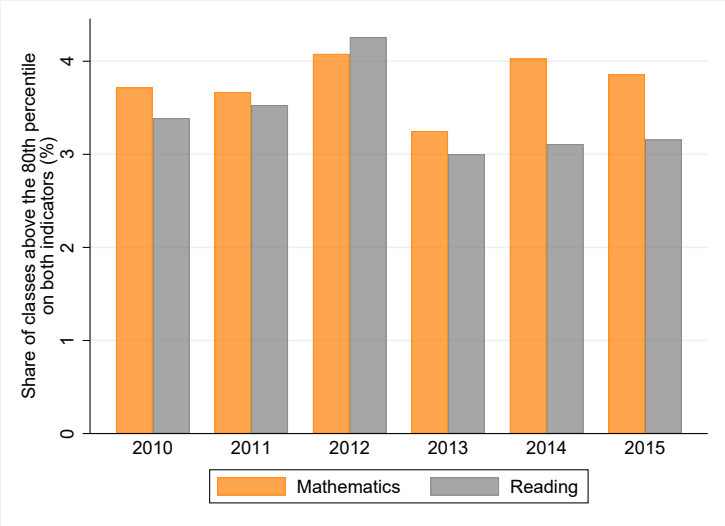
Table A1.3: The Relationship Between the Two Indicators, 2014 (I)

Mathematics		IND2 falls within the range		
		0-25th pctile	25-50th pctile	50-75th pctile
Percent of	80th pctile	19.70	17.88	21.95
classes with	90th pctile	11.33	8.75	11.10
IND1 above:	95th pctile	6.65	3.82	5.55
Reading		IND2 falls within the range		
		0-25th pctile	25-50th pctile	50-75th pctile
Percent of	80th pctile	20.81	22.19	20.10
classes with	90th pctile	9.98	11.71	10.36
IND1 above:	95th pctile	5.05	5.80	5.18

Table A1.4: The Relationship Between the Two Indicators, 2014 (II)

Mathematics		IND1 falls within the range		
		0-25th pctile	25-50th pctile	50-75th pctile
Percent of	80th pctile	18.60	19.00	22.69
classes with	90th pctile	8.62	9.37	12.21
IND2 above:	95th pctile	4.19	4.81	5.92
Reading		IND1 falls within the range		
		0-25th pctile	25-50th pctile	50-75th pctile
Percent of	80th pctile	16.75	23.67	23.55
classes with	90th pctile	8.13	11.71	11.96
IND2 above:	95th pctile	3.69	5.80	6.66

Figure A1.6: Yearly Share of Suspicious Classes, Jacob-Levitt Algorithm



Note: The figure shows the share of classes above the 80th percentile cutoff on both indicators by year and test type.

Table A1.5: Percentage of Classes Scoring High on Both Indicators and Corrected Prevalence, 2010-2015

2010	**Classes above cutoffs (%)**			**Overall (corrected) prevalence**		
Cutoff for	Cutoff for Indicator 1			Cutoff for Indicator 1		
Indicator 2	*80th pctile*	*90th pctile*	*95th pctile*	*80th pctile*	*90th pctile*	*95th pctile*
Mathematics						
80th pctile	3.72	1.86	0.69	-0.09	0.21	0.07
90th pctile	1.91	0.97	0.48	0.00	0.14	0.17
95th pctile	0.87	0.51	0.28	-0.04	0.12	0.13
Reading						
80th pctile	3.39	1.33	0.64	-1.00	-1.00	-0.35
90th pctile	1.68	0.74	0.36	-0.72	-0.53	-0.18
95th pctile	0.71	0.28	0.18	-0.58	-0.41	-0.112
2011	**Classes above cutoffs (%)**			**Overall (corrected) prevalence**		
Cutoff for	Cutoff for Indicator 1			Cutoff for Indicator 1		
Indicator 2	*80th pctile*	*90th pctile*	*95th pctile*	*80th pctile*	*90th pctile*	*95th pctile*
Mathematics						
80th pctile	3.67	1.48	0.45	0.38	0.04	-0.15
90th pctile	1.85	0.64	0.20	0.21	-0.08	-0.10
95th pctile	0.84	0.25	0.06	0.07	-0.09	-0.08
Reading						
80th pctile	3.53	1.62	0.73	-0.62	-0.60	-0.29
90th pctile	1.65	0.78	0.28	-0.59	-0.42	-0.27
95th pctile	0.73	0.28	0.14	-0.35	-0.30	-0.12

2012	**Classes above cutoffs (%)**			**Overall (corrected) prevalence**		
Cutoff for	Cutoff for Indicator 1			Cutoff for Indicator 1		
Indicator 2	*80th pctile*	*90th pctile*	*95th pctile*	*80th pctile*	*90th pctile*	*95th pctile*
Mathematics						
80th pctile	4.08	1.57	0.74	0.52	0.03	0.01
90th pctile	2.10	0.83	0.44	0.30	0.05	0.08
95th pctile	1.10	0.38	0.24	0.19	-0.01	0.05
Reading						
80th pctile	4.26	1.98	0.77	0.55	0.25	-0.01
90th pctile	2.13	0.86	0.27	0.16	-0.06	-0.15
95th pctile	1.15	0.44	0.18	0.16	-0.02	-0.03

2013	**Classes above cutoffs (%)**			**Overall (corrected) prevalence**		
Cutoff for	Cutoff for Indicator 1			Cutoff for Indicator 1		
Indicator 2	*80th pctile*	*90th pctile*	*95th pctile*	*80th pctile*	*90th pctile*	*95th pctile*
Mathematics						
80th pctile	3.25	1.41	0.74	-0.47	-0.20	0.13
90th pctile	1.50	0.49	0.25	-0.27	-0.28	-0.04
95th pctile	0.40	0.12	0.06	-0.50	-0.27	-0.08
Reading						
80th pctile	3.00	1.16	0.43	-1.98	-1.24	-0.76
90th pctile	1.26	0.37	0.09	-1.09	-0.77	-0.47
95th pctile	0.46	0.12	0.00	-0.73	-0.45	-0.28

2014	**Classes above cutoffs (%)**			**Overall (corrected) prevalence**		
Cutoff for	Cutoff for Indicator 1			Cutoff for Indicator 1		
Indicator 2	*80th pctile*	*90th pctile*	*95th pctile*	*80th pctile*	*90th pctile*	*95th pctile*
Mathematics						
80th pctile	4.03	1.85	0.80	−0.94	−0.67	−0.46
90th pctile	2.04	0.83	0.46	−0.65	−0.52	−0.22
95th pctile	1.02	0.40	0.18	−0.28	−0.25	−0.14
Reading						
80th pctile	3.11	1.51	0.83	−1.62	−0.93	−0.39
90th pctile	1.54	0.77	0.40	−0.86	−0.47	−0.22
95th pctile	0.71	0.28	0.18	−0.63	−0.41	−0.16

2015	**Classes above cutoffs (%)**			**Overall (corrected) prevalence**		
Cutoff for	Cutoff for Indicator 1			Cutoff for Indicator 1		
Indicator 2	*80th pctile*	*90th pctile*	*95th pctile*	*80th pctile*	*90th pctile*	*95th pctile*
Mathematics						
80th pctile	3.86	1.82	0.82	-0.32	-0.07	-0.15
90th pctile	1.82	0.97	0.33	-0.29	0.02	-0.16
95th pctile	0.70	0.36	0.15	-0.45	-0.16	-0.12
Reading						
80th pctile	3.16	1.58	0.82	-0.86	-0.50	-0.23
90th pctile	1.34	0.58	0.30	-0.72	-0.49	-0.24
95th pctile	0.49	0.21	0.12	-0.51	-0.30	-0.14

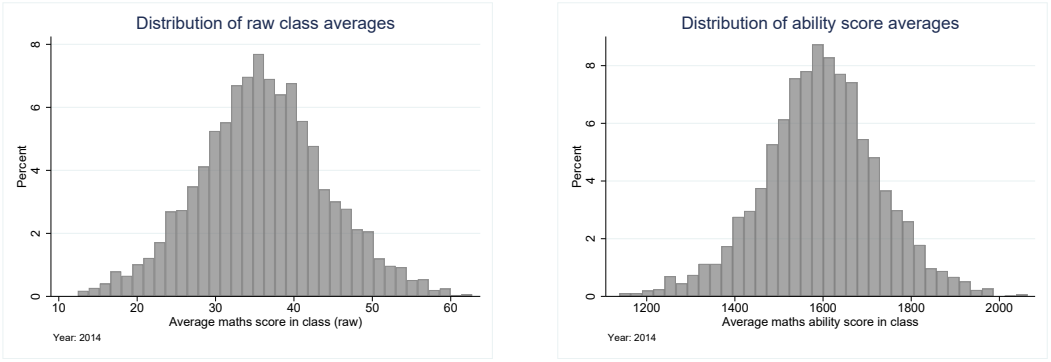
Note: The table shows the percentage of classes that score high on both indicators (left panel) and the overall –corrected– prevalences (right panel), using 80th, 90th and 95th percentiles as cutoffs for 'high' values, separately for mathematics and reading tests in all years.

Table A1.6: Robustness: Jacob-Levitt Results, Using Ability Scores

2014	**Classes above cutoffs (%)**			**Overall (corrected) prevalence**		
Cutoff for	Cutoff for Indicator 1			Cutoff for Indicator 1		
Indicator 2	*80th pctile*	*90th pctile*	*95th pctile*	*80th pctile*	*90th pctile*	*95th pctile*
Mathematics						
80th pctile	4.16	2.25	1.02	-1.09	-0.12	-0.06
90th pctile	2.25	0.99	0.52	-0.54	-0.27	-0.05
95th pctile	1.14	0.52	0.28	-0.26	-0.10	-0.01
Reading						
80th pctile	3.73	1.73	0.68	-0.13	-0.40	-0.41
90th pctile	1.88	0.92	0.40	-0.08	-0.16	-0.15
95th pctile	1.02	0.34	0.09	0.02	-0.21	-0.192

Note: The table shows the percentage of classes that score high on both indicators, using 80th, 90th and 95th percentiles as cutoffs for 'high' values, separately for mathematics and reading tests in all years. Relative to the baseline specification, here Indicator 1 is calculated using ability scores (taking into account the difficulty of the test items), instead of the raw scores.

Figure A1.7: Score Distributions



Note: The figures show the raw score distribution and the ability score distribution of 8th graders in 2014.

Table A1.7: Measures Used by the Detection Methods

Hungarian Ministry of Education	
Index 1	Number of items, solved by min.90% of the class
Index 2	Number of multiple choice items, solved by min. 90% of the class
Index 3	90% of the class solved (min.) one of the 15 most difficult test items
Index 4	2-4 multiple choice items were solved by less than 10% of the class
Index 5 (control)	Same 'typical wrong answer' for 90% of the class (open-ended)
Clustering	
1	Average score (class)
2	Within-class standard deviation of scores
3	Average rate of missing answers
4	Answer homogeneity (average Gini)
Jacob-Levitt	
(Indicator 1) Large test score fluct.	Average score gains across years
(Indicator 2) Suspicious answer strings	Most unusual block of answers
	Within-classroom correlation in answers
	Cross-question variance in the within-classroom correlations
	Comparison of students with the exact same final scores

Note: The table gives an overview of all measures used by the three detection methods applied in the paper.

A1.5 Simulation Appendix

Table A1.8: Summary Statistics on Raw Scores, by Quarters

Quarter	Observations	Min	Mean	(sd)	Max
1	1,224	7.20	**22.74**	(3.69)	27.10
2	1,221	27.11	**29.45**	(1.28)	31.69
3	1,220	31.7	**34.06**	(1.47)	36.85
4	1,221	36.85	**41.74**	(4.21)	59.63

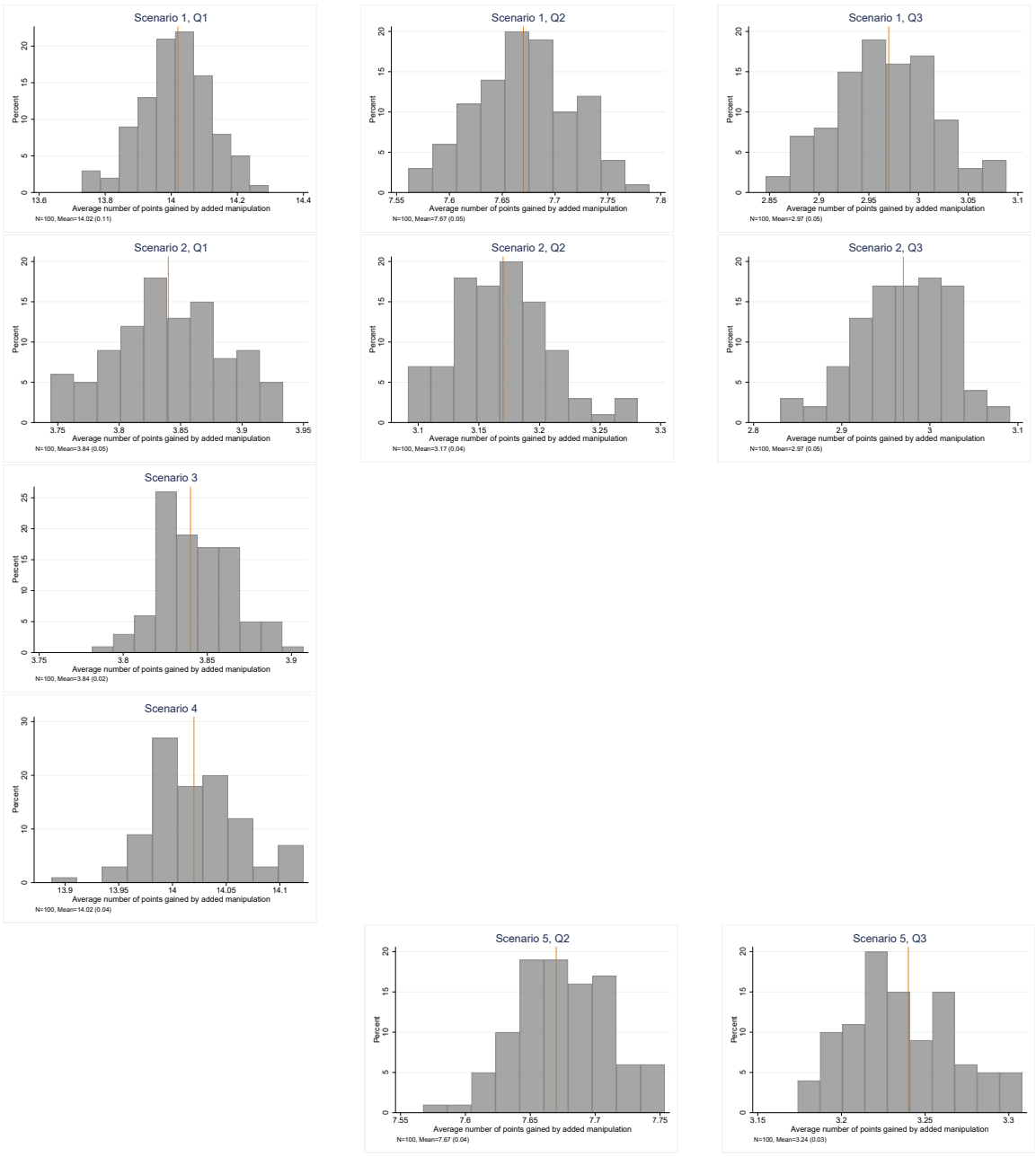
Note: The table shows summary statistics on raw scores by quarters in 2010. All classes included, $N_{class} = 4,886$.

Table A1.9: Average Score on Multiple Choice Questions, by Quarters

Quarter	Observations	Mean (sd)
1	1,224	18.11 (2.43)
2	1,221	22.08 (1.11)
3	1,220	24. 82 (1.42)
4	1,221	29.13 (2.71)

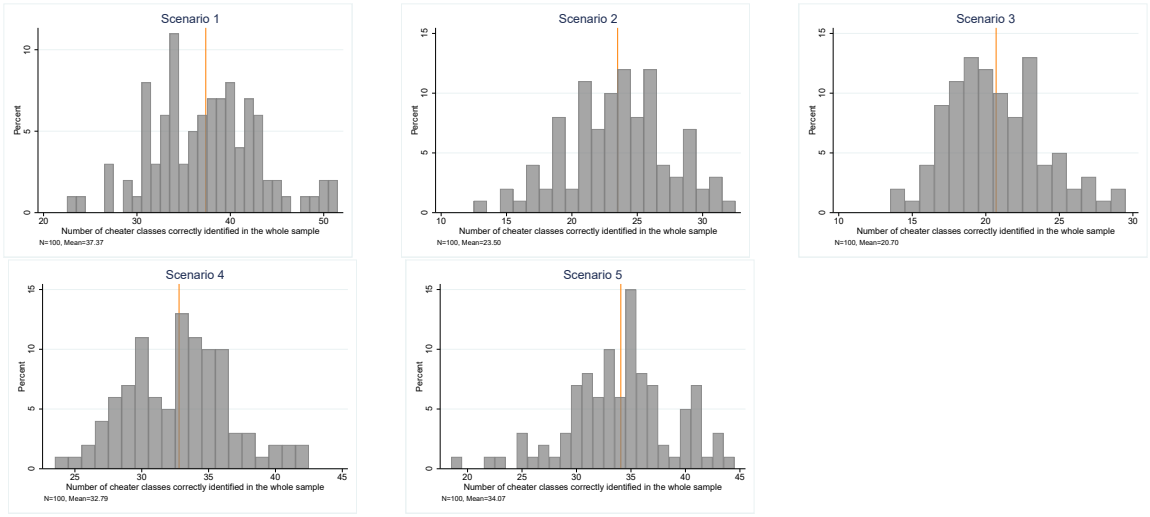
Note: The table shows summary statistics on multiple-choice scores, by quarters in the year 2010. This corresponds to the probability of randomly selecting a correct test item in the given quarter. All classes included, $N_{class} = 4,886$.

Figure A1.8: Distribution of 'Effective Cheating', by Scenario and Quarter



Note: The figures show the distribution of 'effective cheating', i.e. the average number of points gained through the artificial cheating, by scenario and quarter. Quarters are defined according to the classes' average mathematics score in 2014.

Figure A1.9: Distribution of the Number of Artificial Cheaters Found in the Simulation Rounds, by Scenario



Note: The figures show the distribution of the number of (artificial) cheaters found in the 100 simulation rounds for each scenario separately. The overall level of cheating added to the sample is 15%, i.e. 15% of classes cheat to some extent (see Table 1.10 for more details). The number of artificial cheater classes ranges between 675 and 684, depending on the scenario.

Table A1.10: Simulation Results

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Whole Sample	Analysis Sample	%	Correctly identified, AS	%	Correctly identified, WS	%
1. Textbook cheating							
Cheater	732	562	76.75	37	6.65	37	5.11
Non-cheater	4154	3358	80.84	3259	97.05	4055	97.62
Total	4886	3920	80.23	3297	84.1	4092	83.76
2. Quartile jump							
Cheater	732	562	76.75	24	4.18	24	3.21
Non-cheater	4154	3358	80.84	3247	96.68	4042	97.31
Total	4886	3920	80.23	3270	83.42	4066	83.22
Scenario 3							
Cheater	734	432	58.85	21	4.79	21	2.82
Non-cheater	4152	3488	84.01	3359	96.29	4023	96.89
Total	4886	3920	80.23	3379	86.21	4043	82.76
Scenario 4							
Cheater	734	432	58.85	33	7.59	33	4.47
Non-cheater	4152	3488	84.01	3378	96.83	4042	97.34
Total	4886	3920	80.23	3410	87	4074	83.39
Scenario 5							
Cheater	732	630	86.03	34	5.41	34	4.65
Non-cheater	4154	3290	79.21	3191	97	4055	97.62
Total	4886	3920	80.23	3226	82.28	4089	83.69

Note: The table presents averages of the 100 simulation rounds, for each cheating scenario. S1: 'Textbook cheating', manipulation happens in the bottom three quarters such that they reach the level of the top quarter. S2: 'Quartile jump', the bottom three quarters cheat such that they reach the next quarter. S3: 'Bottom-to-Mid', manipulation in the bottom quarter such that they reach the second quarter. S4: 'Bottom-to-Top', manipulation in the bottom quarter such that they reach the top quarter. S5: 'Mid-to-Top', manipulation in the second and third quarter such that they reach the top quarter. Column (1) presents the number of classes in the whole sample. Column (2) presents the number of classes in the analysis sample (i.e. after dropping small classes, where $N_{partic} < 10$), while Column (3) shows their share relative to the total number of classes in the respective category. Columns (4) and (5) show the number and share of correctly identified classes in the analysis sample, by categories. Columns (6) and (7) show the number and share of correctly identified classes in the whole sample by categories, considering small classes are never classified as cheaters.

Table A1.11: False Positives

	(1) # classes	(2) % classes, WS
Scenario 1		
Correctly classified positives	37	.76
Correctly classified negatives	4055	82.99
False positives	695	14.22
False negatives	99	2.03
Scenario 2		
Correctly classified positives	24	.48
Correctly classified negatives	4042	82.74
False positives	709	14.5
False negatives	112	2.28
Scenario 3		
Correctly classified positives	21	.42
Correctly classified negatives	4023	82.33
False positives	713	14.6
False negatives	129	2.65
Scenario 4		
Correctly classified positives	33	.67
Correctly classified negatives	4042	82.72
False positives	701	14.35
False negatives	110	2.26
Scenario 5		
Correctly classified positives	34	.7
Correctly classified negatives	4055	83
False positives	698	14.28
False negatives	99	2.02

Note: The table presents averages of the 100 simulation rounds, for each cheating scenario. Column (1) shows the number of correctly identified classes, and the number of false positives and false negatives. Column (2) shows the corresponding shares. Numbers and shares are computed on the whole sample, i.e. taking into account small classes that are never classified as cheaters because they are not part of the analysis sample. Simulation is performed on data from the 2012 mathematics test. Overall number of classes $N = 4,521$.

Chapter 2

Cheating on Standardized Tests: Test Pool Manipulation

2.1 Introduction

In Chapter 1, I study the prevalence of test score manipulation in the Hungarian standardized student assessment in the period 2010-2015. Applying the detection algorithm developed in Jacob and Levitt (2003)[1] I find no evidence of score manipulation. I argue this is because of the unique testing environment where low-stakes testing is paired with strict quality assurance. This means that compared to previously studied tests (in the US and Italy) there is not only less incentive to manipulate the answers, but it is also more costly.

However, the question emerges, whether schools engage in another type of manipulation which is less costly, or less risky. As a starting point, there is considerable anecdotal evidence in Hungary that instead of manipulating the test scores, schools manipulate the test pool by sending low-performing students home on testing day.

The literature documents test pool manipulation through exemptions, additional grade repetitions, or testing day absences, leading to the underrepresentation of low-performing students in the testing Jacob (2005), Lucifora and Tonello (2020), and Machin and Sandi (2020). It is relevant in an international context as well. In the past decades, every PISA result publication was followed by cheating accusations because of alleged selective testing. In 2003 the UK was excluded from the analysis because of low participation rates, while in 2015 they were again the country with the highest national exclusion rate. China was accused of cherry-picking its most developed and educated provinces every year since they joined the testing. Most recently, in 2018 controversies emerged in Sweden because it was revealed that many schools excluded immigrant children from the testing, even though

[1]The detection algorithm of Jacob and Levitt (2003) relies on two indicators capturing unexpected test score fluctuations and suspicious answer strings within classrooms.

the test is compulsory for all those who have learnt Swedish for more than a year.[2]

This chapter aims to assess the prevalence of test pool manipulation, and its potential consequences in a low-stake testing environment, using data from the Hungarian standardized testing. I propose a difference-in-differences design, exploiting changes in the incentives for manipulation which affect schools only. In 2013, the – previously non-binding – minimum requirement of the testing was redefined. Schools became obligated to prepare an action plan if more than 50% of their students fall below a specified ability level.

First, in line with the incentives of the new policy, I find that post-policy absence rates increased particularly in schools at risk of not fulfilling the minimum requirement. Second, using multiple imputation I quantify schools' gains from these manipulations and absences. I find that although the variation in absences is large, schools do not benefit substantially from it. In the Appendix I provide an overview of other potential factors that could affect absences besides manipulation.

Since Holmstrom and Milgrom (1991) and Baker et al. (1994), it is well-established in the economics literature that high-powered incentives can distort the effort exerted and result in unintended consequences. One of the most well-known examples described in these theoretical works is high-stakes testing in education, more specifically, teacher responses to financial incentives. The empirical literature also widely documents the presence of teachers' strategic reactions and manipulation in high-stake accountability systems Borcan et al. (2017), Diamond and Persson (2016), Jacob (2005), and Jacob and Levitt (2003).

It has also been shown that low-stakes testing (i.e. in the absence of financial incentives) may be as effective as high-stakes testing in improving outcomes if it creates sufficient reputational concern for schools Cilliers et al. (2021) and Figlio and Loeb (2011). However, if there is an incentive to improve, it might lead to the same distortions documented in high-stakes testing. Still, except for the Italian low-stake setting Bertoni et al. (2013), Longobardi et al. (2018), and Lucifora and Tonello (2020), there is little evidence of this. In this paper, I show in a traditionally low-stake setting, how changes in incentives and perceived stakes (perceived pressure) matter, even in the absence of monetary incentives or direct punishment.

Besides data accuracy problems, the presence of cheating in schools raises further questions in the long run. If parents' school choice or the policy makers' funding decision depends on the corrupted information, it can lead to inefficiencies and poor policy conclu-

[2]Sources: Schneider, M. (2019, December 10), *The Strange Case of 'China' and Its Top PISA Rankings — How Cherry-Picking Regions to Take Part Skews Its High Scores.* The 74. https://www.the74million.org/article/schneider-the-strange-case-of-china-and-its-top-pisa-rankings-how-cherry-picking-regions-to-take-part-skews-its-high-scores/ and Radio Sweden (2021, April 29). *Too many students excluded from taking PISA test, report finds.* https://sverigesradio.se/artikel/too-many-students-excluded-from-taking-pisa-test-report-finds

sions. Looking at long-run effects, Dee et al. (2019) find heterogeneous effects of teacher cheating on the cheated students' educational outcome: while their probability of high school graduation increased, they were less likely to take advanced coursework. Battistin and Neri (2023) show how score inflation leads to residential sorting, and consequently, affects housing prices and local economic activities in the long run. Other literature also studies the link of academic cheating to workplace dishonesty and corruption, showing they reinforce each other Ajzenman (2021), Borcan et al. (2017), and Orosz et al. (2018).

This chapter contributes to two strands of the literature. It adds to the literature on strategic pooling by quantifying the distortions caused. Moreover, it contributes to the mostly experimental literature on the effects of non-monetary incentives on cheating and shows how schools strategically react to policy changes even in a low-stake testing environment.

Studies on strategic pooling tend to exploit policy changes in a similar fashion, but document various ways of manipulating the test pool. Jacob (2005) shows that introducing high-stakes testing results in systematic grade retention. Cullen and Reback (2006), Figlio and Getzler (2002), and Jacob (2005) all find that in order to exclude low-performing students from the testing pool, schools tend to classify them as Special Needs students who are exempted from the testing. Figlio (2006) and Machin and Sandi (2020) document that schools even use suspensions on disciplinary grounds to avoid the participation of low-performing students. Finally, although they do not consider the possibility of strategic pooling, Cuesta et al. (2020) also provide evidence that low-performing students are underrepresented on test days. This is the only related study which aims to quantify the biases caused.

While there is experimental evidence that low-stake or non-monetary incentives matter in cheating decisions, it is also important to bring closer the experimental results to the observational data and real-world settings. Field experiments mostly focus on individual-level cheating, and how it is affected by financial incentives or monitoring Azar and Applebaum (2020), Cagala et al. (2021), and Martinelli et al. (2018). Studies of the low-stake assessment system in Italy are the closest to my setting Bertoni et al. (2013), Longobardi et al. (2018), and Lucifora and Tonello (2020), however, it is usually argued that manipulation is present because of shirking of teachers. This chapter is among the few studies to show that reputational concerns are sufficient to distort incentives in standardized testing.

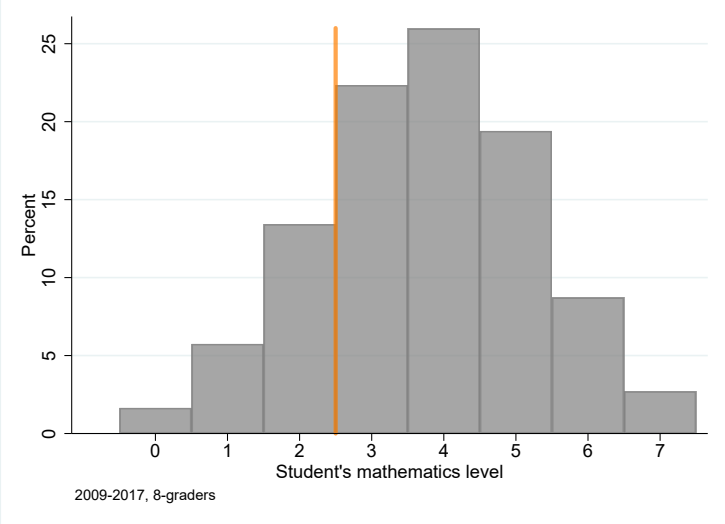
The rest of the chapter is organized as follows. In Section 2.2 I describe the institutional background and the data. Section 2.3 presents the empirical strategy to detect test pool manipulation and summarizes the results. In Section 2.4 I quantify the potential biases caused by absences. Finally, Section 2.5 concludes and discusses directions for further research.

2.2 Institutional Background and Data

The National Assessment of Basic Competencies (NABC) test is a PISA-like test administered in Hungary to evaluate school quality and student performance. For more details on the Hungarian education system and the NABC, see Section 1.3 in Chapter 1.

In this chapter I focus on the sole accountability requirement of the test which is a nationally defined minimum requirement. Under this requirement, if at least half of the students do not reach a certain ability level[3], the school must prepare an action plan. The minimum requirement was introduced in 2008, however, it was not adjusted when a new point system (with more ability levels) was introduced in 2010. Consequently, between 2010 and 2012, the minimum requirement was not binding. The law was modified in 2012 September[4], and was in effect from the 2013 testing. It states that an action plan has to be prepared by the school if at least half of the students fail to reach the 2nd ability level in 6th grade and the 3rd ability level in 8th and 10th grade. This requirement is remarkably low compared to the definition of the base level, which is considered necessary for acquiring further knowledge in subsequent grades. The base level is defined as the 3rd ability level in 6th grade and the 4th ability level in 8th and 10th grades. Figure 2.1 shows the ability distribution of 8th graders, denoting the required minimum level.

Figure 2.1: Distribution of Ability Levels Among 8th Graders



Note: The figure shows the distribution of ability levels among 8th graders in the time period 2009-2017. The orange line indicates the minimum level which has to be achieved by at least 50% of 8th graders within school since 2013. $N = 772,783$

[3]A student's ability level corresponds to the difficulty level at which the student is expected to solve at least half of the exercises correctly. Item difficulty is calculated based on Item Response Theory (Rasch, 1980).

[4]Ministry of Human Capacities Decree 20/2012 (VIII.31) on the Educational Institutions Operation and the Use of Names of the Public Education Institutions. (In Hungarian: 20/2012 (VIII.31) EMMI rendelet a nevelési-oktatási intézmények működéséről és a köznevelési intézmények névhasználatáról)

Each year around 8-10% of the schools are below the minimum requirement and have to prepare the action plan. If the minimum level is not reached in three consecutive years, a new action plan will be prepared and implemented by the educational provider. Preparing an action plan is costly, but most importantly it is considered to be a stigma for the school.

Another important change in 2013 was the centralisation of school governance. This included transferring the responsibility of governing and funding decisions from local governments to a central organization (Klebelsberg Institution Governance Centre, KLIK), but went hand-in-hand with an increase in private and church-owned schools. Appendix Figure A2.6 shows the change in ownership structure through time.

The centralisation happened unexpectedly and created a general uncertainty among school principals, teachers and parents. The government takeover of schools and the surrounding uncertainty potentially made the adjustment of the minimum requirement more salient. Schools afraid of retaliation, further school closures or school mergers by KLIK might have reacted by paying more attention to the testing results - which could be the base of comparisons across schools.[5]

In the current analysis, I focus on 8th graders for two reasons. First, because both past and future test scores are available for them. Second, the difference between school and student incentives is the most striking in this grade. Most students switch schools between the 8th and 9th grades, thus, they have no interest in the school's future ranking. Moreover, since the results are published only the following year, they cannot be held accountable for their performance.

I use the researcher database of the Hungarian Ministry of Education which contains individual and school-level test scores, and answers from institutional, school site and student surveys. First, the school-level analysis relies on the balanced sample of 2,453 schools. This is obtained by linking schools based not only on their school identifiers (which might change) but also based on their addresses.[6] In the second part of the paper, imputations are performed at the student level, utilizing a 3-wave panel dataset of 700,000 students between 2008 and 2017.

Overall absence rate is usually around 9%[7], and absences are indeed considered to be a problem by the Ministry of Education. They correct for absences by a mean imputation at the class level, i.e. assuming the absent students would have performed as their participating classmates on average. However, this means schools still have incentives to send home those who would perform below the class average.

[5]Hermann and Semjén (2021) study the effect of the centralisation on inequalities, and find evidence of equalization of resources (per-student school expenditures in municipalities), but no effect on student outcomes.

[6]For more on the improved linking of schools, see Appendix Section A2.2.

[7]It varies a lot across grades, e.g. it is around 8% for 8th graders, while in some years it is even above 12% for 10th graders.

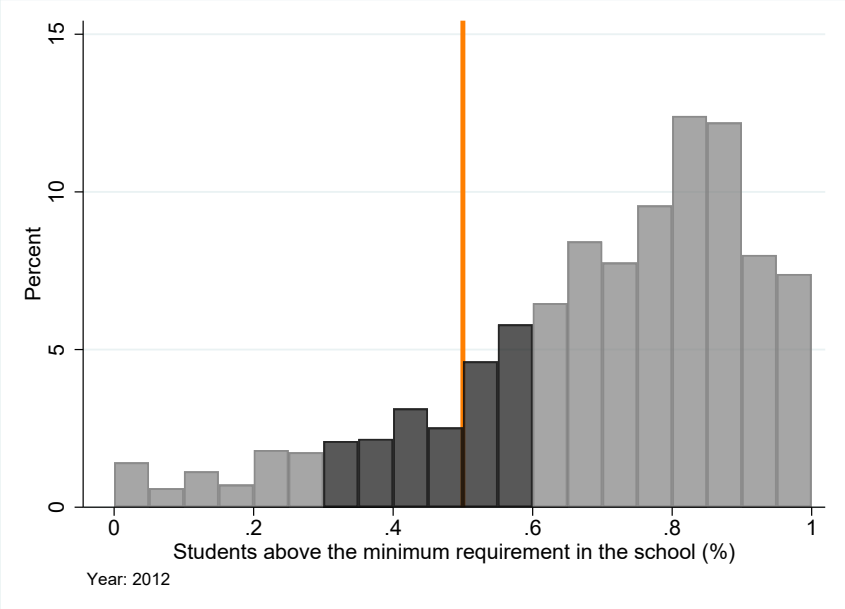
2.3 Evidence of Strategic Pooling

2.3.1 Empirical Strategy

To identify strategic pooling I exploit a policy change which introduced higher stakes in the testing, and affected schools only (i.e. did not change the incentives for students). I use difference-in-differences estimation where treatment intensity, i.e. the "bite" of the policy differs across the treated schools.[8]

As described above, the NABC testing redefined its minimum requirement in 2013. This requires schools to have more than 50% of their students above the given ability levels. Figure 2.1 shows the ability distribution of 8th-grade students, while Figure 2.2 shows the distribution of schools' "achievement rate" (the percentage of students who reached the required ability level). The orange line in Figure 2.2 indicates the 50% minimum requirement. Thus, schools that fulfilled the requirement are located on the right of the orange line, while those to the left are required to prepare an action plan.

Figure 2.2: Distribution of "Achievement Rate"

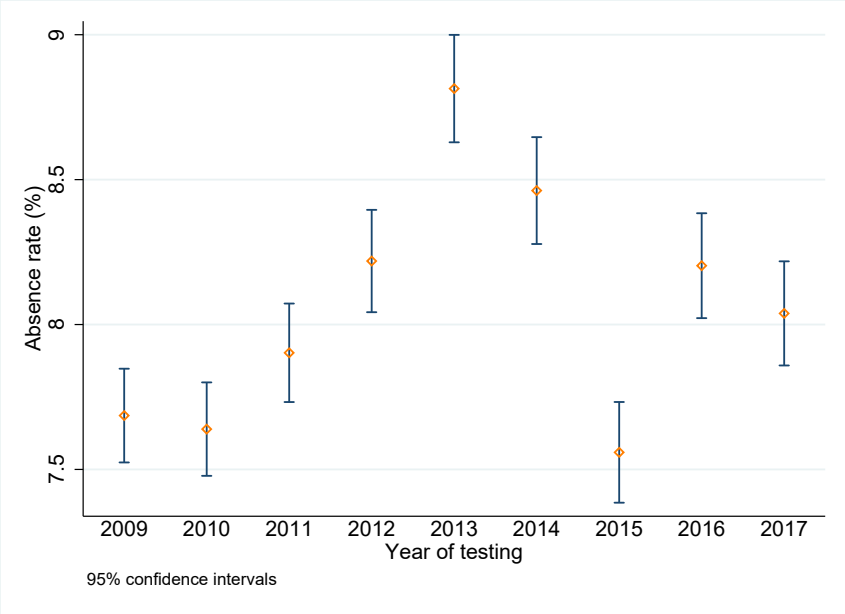


Note: The figure shows the distribution of achievement rate of schools, i.e. the share of students that (would have) reached the minimum level in the school in 2012. The orange line indicates the 50% minimum requirement introduced in 2013. Schools to the left would not have fulfilled the requirement, while those to the right would have fulfilled it. The darker grey shade shows the approx. 20% of schools that are around the minimum requirement. $N_{schools} = 2,812$.

[8]Other studies with similar identifications in other fields include Card and Krueger (1994), Clarke (2017), and Duflo (2001)).

First, descriptive statistics help in identifying patterns of interest. As it can be seen in Figure 2.3, there seems to be an increase in 2013, but it is not persistently high in the following years. Second, Table 2.1 shows that absent and participating students are different along many dimensions: e.g. 8.65% of participating students is a grade repeater, in contrast to 26% of absent students, while the average mathematics mark in 7th grade was 3.38 and 2.82, respectively.[9]

Figure 2.3: Absence Rate of 8th Graders on the Testing Day



Note: The figure shows the overall absence rate of 8th graders on the testing day between 2009 and 2017.

But it is important to note that these differences, in themselves, are not proof of manipulation. Low-ability students or low-SES students are more likely to be absent both on normal school days and on test days, without the teachers asking them to do so. The question is whether we can disentangle the role of individual characteristics (students staying home by themselves), and school-level test pool manipulation. The policy change described above can be exploited since it only affects the incentives for schools, but not for the students. If strategic pooling is not happening, we should not observe any change in the absence rates after the introduction of the minimum requirement.

The following equation can be estimated at the school level:

$$Absrate_{st} = \alpha_0 + \alpha_1 X_{st} + \beta Post_t + t + \phi + \epsilon_{st} \tag{2.1}$$

where $Post_t$ is a dummy for post-intervention periods, X_{st} is a vector of school characteristics, t is time trend, and ϕ is a set of fixed effects (e.g. region, school type).

[9]Simple t-tests are all significantly different at the 1% level, but within-cluster correlation (classes, schools) should be taken into account.

Table 2.1: Summary Statistics

mean *(sd)*	Absent N=6,100	Participant N=80,110	All N=86,210
Grade repeater (%)	25.972	8.645	9.87
	(43.852)	(28.102)	(29.826)
Prev. maths grade	2.823	3.377	3.339
	(1.085)	(1.092)	(1.101)
SNE (%)	8.311	4.182	4.474
	(27.608)	(20.017)	(20.673)
ILC (%)	11.836	6.915	7.264
	(32.306)	(25.372)	(25.954)
Disadvantaged (%)	19.18	9.154	9.863
	(39.375)	(28.837)	(29.817)
Maths score (6th grade)	1433.874	1498.222	1494.415
	(200.677)	(188.565)	(189.91)
Maths score (10th grade)	1602.254	1663.659	1660.855
	(228.035)	(206.89)	(208.296)

Note: The table presents summary statistics separately for the group of absent and participating students, and the whole sample of eligible students in 8th grade in 2014. Exempted students are excluded. *Mathematics grade: 1 (worst) to 5 (best). *SNE: Special Needs Education, *ILC: integration, learning or conduct disorder.

Since we can expect low-performing schools to be more affected by the introduction of the minimum requirement, the following equation allows us to capture differences depending on the incentives to manipulate the results:

$$Absrate_{st} = \alpha_0 + \alpha_1 X_{st} + \beta_1 Post_t + \beta_2 R_s + \beta_3 (R_s * Post_t) + t + \phi + \epsilon_{st} \tag{2.2}$$

where R_s is a dummy for those schools which have higher incentives to manipulate the results, e.g. because they are more at risk of not reaching the minimum requirement.

Unfortunately there is no spatial or time variation which could be used for a clear difference-in-differences design, in principle every school was affected by the policy. The question is how R_s could be defined to capture the "bite" of the policy (treatment intensity), and get around this problem. First, one can expect low-achieving schools below the threshold - which are directly affected by the policy - to have higher incentives to cheat.[10] If schools react with test pool manipulation to the new policy, we expect β to be positive in specification 2.1, and since schools at higher risk of failing are more likely to be affected by the policy, β_3 is expected to be positive in specification 2.2.

In the baseline specification I define control ($R_s = 0$) and treatment ($R_s = 1$) groups

[10] Alternatively, we can argue that only schools around the threshold are most affected, and the "bite" of the policy depends on the distance from the threshold. For example, if all students are above the minimum level, the threshold is irrelevant, while if all students are below, sending them home does not help.

based on being below or above the threshold in the pre-policy period. By this, I assume that in the short run schools had strategic pooling as their only "quick fix" to improve their results. This is a reasonable assumption for two reasons. First, they did not have time from the announcement of the law (2012 autumn) until the first test (May 2013) to change their student body or to improve their students' performance substantially. Second, as seen in previous results as well, test score manipulation is more costly than strategic pooling.

In further specifications the treatment group is defined by on which part of the "achievement" distribution the schools were before the policy was introduced, i.e. how far from the threshold they were. I either use the pre-policy year (2012)[11], or a longer pre-policy period.

I also estimate the following event study specification:

$$Absrate_{st} = \alpha_0 + \alpha_1 X_{st} + \delta_1 year_t + \delta_2 R_s + \delta_t (R_s * year_t) + \phi + \epsilon_{st} \tag{2.3}$$

2.3.2 Results

Figure 2.4 shows absence rates separately for schools below and above the threshold in 2012. In the baseline specification I take these schools as treated and non-treated, respectively. This treatment group consists of schools that would have been directly affected by the policy in case it had been implemented a year earlier. It can be seen that absences peaked in 2013 in these schools. Note that in order to avoid endogeneity, the year 2012 is omitted from the following estimations (since treatment and control groups are defined according to this year).

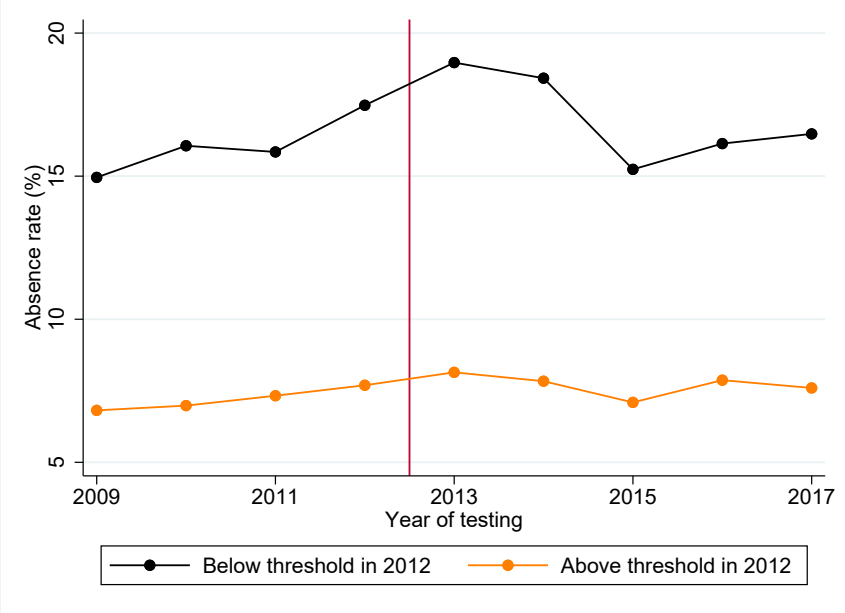
In line with the graphical illustrations, Column 1 of Table 2.2 shows that absence rates increased on average by 0.6 percentage points after the policy was introduced (estimation of Equation 1). This corresponds to a 7.7% increase, given the overall average of 7.82% absence rate. Column 2 shows the estimation results of Equation (2.2) with the baseline specification: the increase in post-policy absence rates in schools which would have been below the threshold in 2012 is higher than in the schools above (+1.56 percentage point).[12] Figure 2.5 illustrates the event study results, estimating Equation (2.3). It shows that the coefficient estimates are the highest in the first two years following the policy change.

Although one would expect that those just above the threshold in 2012, and those just below have a similarly high perceived risk of not fulfilling the requirement, the result in Column 3 shows that this is not the case. Defining the treatment group R_s as the schools with an achievement rate between 30% and 60% (illustrated by the dark shaded part in

[11] In this case, observations from the year 2012 are dropped from the estimation to avoid endogeneity.

[12] The results presented here are estimated on the balanced sample of schools. For details on constructing the balanced sample see Section A2.2 in the Appendix. For robustness checks with different samples, see Table A2.7 in the Appendix.

Figure 2.4: Absence Rates in the Treated and Non-treated Group



Note: The figure shows absence rates in the treated and non-treated groups separately. Treated: schools that would have been below the threshold in the pre-policy year 2012. Number of schools above the threshold in 2012: 2,618. Number of schools below the threshold in 2012: 194. Number of observations throughout the nine years: 24,290.

Figure 2.2), I find no significant difference between those around the threshold and those further from the threshold. Columns 4 and 5 define the treatment group as those schools that were at least once below the requirement in the pre-policy period, but not always. The idea behind this is again that the requirement might not be binding not only for the best schools but also for the worst ones. The worst schools should not be expected to bother with manipulation because they have a student body where e.g. all students would fail the test anyway. The estimated effect size is weakly significant and positive (0.80 percentage points).

Note that the estimated effects correspond to a lower bound on the prevalence of test pool manipulation. Since anecdotal evidence suggests the practice was present in earlier years as well (and even a minimum requirement was in place between 2008 and 2010), the treatment effects do not measure the actual level of test pool manipulation, but the reaction of schools to the changes in incentives.

The results hold when R_s is a continuous measure of the distance from the threshold (following Card and Krueger (1994), this measures the proportional increase in the achievement rate required to meet the introduced minimum requirement[13]). The results

[13]The GAP-method of Card and Krueger (1994) estimates Equation (2.2) where the GAP is defined as $R_s = \frac{min - x_{pre}}{x_{pre}}$ if $x_{pre} < min$ and 0 otherwise, i.e., it measures the proportional increase in x (achievement rate) required to meet the introduced minimum requirement. Estimated on a sample of 2,080 schools, the effect size is weakly significant 0.318** (0.147).

Table 2.2: Main Results: Treatment Effects on Absence Rates

Outcome	Absence rate				
	(1)	(2)	(3)	(4)	(5)
Treatment group	Overall	Below threshold in 2012	Around threshold in 2012	Sometimes below, pre-policy	Sometimes below, pre-policy
1.Treat#1.post2012		1.564***	0.209	0.804*	0.378
		(0.551)	(0.263)	(0.441)	(0.409)
1.post2012	0.614***				
	(0.0903)				
2012 included	✓	-	-	-	✓
Balanced sample		✓	✓	✓	✓
Observations	25,395	19,618	19,618	19,618	22,071
Number of schools	3,244	2,453	2,453	2,453	2,453
R-squared	0.234	0.232	0.231	0.231	0.232

Note: The table presents the estimated effects of the minimum requirement on the post-policy absence rates. Column (1) shows estimation results of Equation (2.1). Columns (2)-(5) present results of estimating Equation (2.2) with different treatment groups. Time-variant school-level controls: % Exemption, % Grade repeater, number of 8th graders in school. School and year fixed effects are included. Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

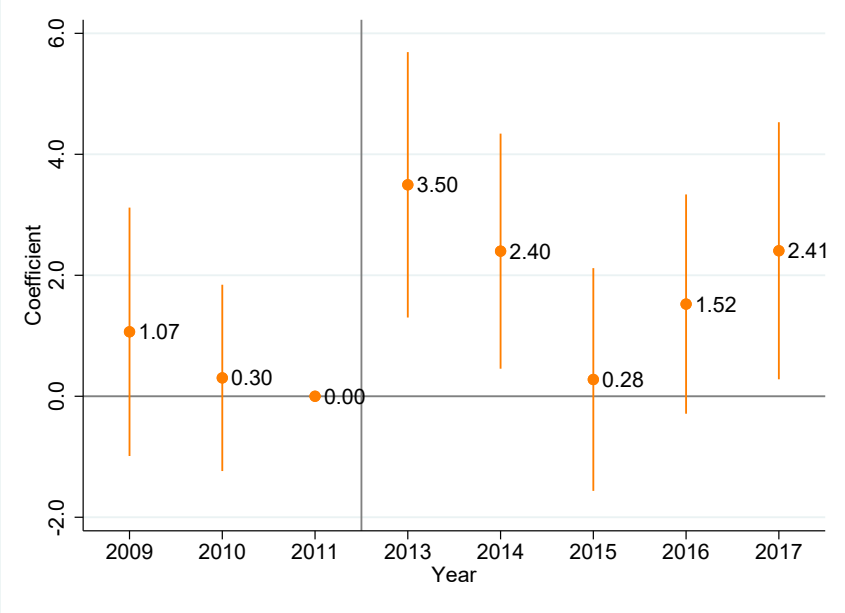
are driven by the schools at the bottom of the distribution. This can be seen in Figure 2.6 as well, where the estimated post-policy increases in absence rates are shown by deciles (according to achievement rate). Schools that were in the treatment group in the baseline specification (below the threshold in 2012, shown in Figure 2.4) are in the bottom two deciles.

Since in the year 2015 there is a drop in absence rates in all groups, and this might be because of administrative changes, Appendix Table A2.8 shows results when the years from 2015 onwards are excluded from the post-policy period. As expected, the significantly positive effect sizes become larger in this case.

Other robustness checks can be found in the Appendix. Tables A2.6 shows how the inclusion of fixed effects and the clustering of standard errors matter, while Table A2.7 shows that the results are robust to using different sample restrictions. Exploring the role of school size, Appendix Section A2.3 presents the results when reestimating Equations 2.1 and 2.2 on the sample of small and large schools separately, and using the number of students as outcome variables (see Tables A2.10–A2.12). Table A2.2 presents the estimated treatment effects on exemption rates.

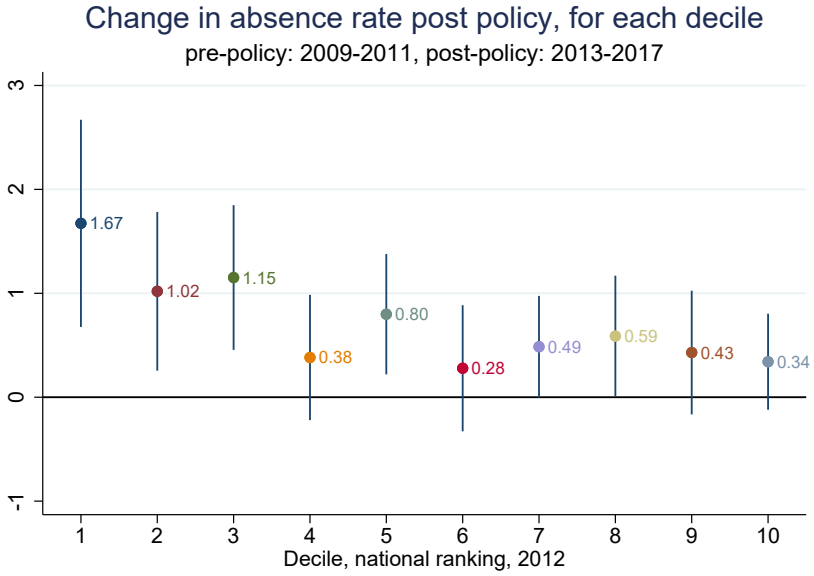
As alternative explanations, I provide a brief summary on 14-year-olds' healthcare use in Section A2.3, and summary statistics on weather conditions around the testing day in Section A2.3. Moreover, in Section A2.3 I explore how competition and local rankings might influence the propensity to engage in test pool manipulation, but I find no convincing evidence of this hypothesis.

Figure 2.5: Event Study Plot, Baseline Specification



Note: The figure shows coefficient estimates of the event study in Equation (2.3), using the baseline specification where R_s, i.e. the treated group is defined as those schools that did not fulfil the minimum requirement in 2012.

Figure 2.6: Coefficient Estimates of Post-policy Period, by Deciles



Note: The figure shows the effect of the policy by deciles, comparing absences in pre-policy years 2009-2011 and post-policy years 2013-2017. Schools are assigned to deciles according to their national ranking on the mathematics test in 2012. The coefficient estimates are obtained from regressions estimating absence rates on a post-policy dummy, separately on subsamples of each decile. Controls include the number of students, the share of exempted and grade-repeater students in the school, and school fixed effects are included. The sample is the balanced sample ($N_{schools} = 2,453$), but since the likelihood of being in the balanced sample differs across deciles, there are 214 schools in the lowest decile, while 254 schools in the highest decile.

2.4 Quantifying the Distortions

After documenting the presence of any kind of manipulation, the critical question emerges: how much does it matter? What level of manipulation matters for policy makers aiming to evaluate school quality, and how could corrupted data influence the school choice decisions of parents? If everyone is cheating to the same extent, rankings are unaffected. Some schools might falsely meet the minimum requirement, in which case spillover effects depend on potential benefits from preparing an action plan. On the extreme, if all schools are cheating to the maximum extent, resulting in perfect scores[14], the testing becomes uninformative, constituting a waste of resources. However, if cheating is non-random, it introduces further biases in the rankings, leading to further inefficiencies.

Early studies on test pool manipulation examine correlations to identify schools where cheating is more likely Figlio (2006), Figlio and Getzler (2002), and Jacob (2005), but they investigate the consequences of the manipulation or whether it matters at all. More recently Cuesta et al. (2020) document that low-performing students are underrepresented on test days in Chile, and they estimate the costs caused by the distorted quality signal in a particularly high-stakes accountability system.

Schools in Hungary can gain from the absences in two ways: (i) by reaching the minimum threshold they avoid the stigma associated with it, and the cost of preparing an action plan, and (ii) by signalling higher quality they gain more reputation, and can attract better students and better teachers in the long-run. Since the threshold to be reached is difficult to calculate in advance (it is defined in ability levels and points, not in raw scores), we can expect that schools either 'overshoot' and do much better than the minimum level, or cannot reach the threshold even by manipulation. Moreover, it has to be taken into account that the Hungarian Ministry of Education does correct for absences, using a mean imputation at the class level. This is also difficult to assess by teachers in advance, and limits the achievable gains.

Following Cuesta et al. (2020) I use a multiple imputation method to quantify the distortions caused in the schools' performance, and I compare it to the mean-imputed correction of the Ministry of Education.

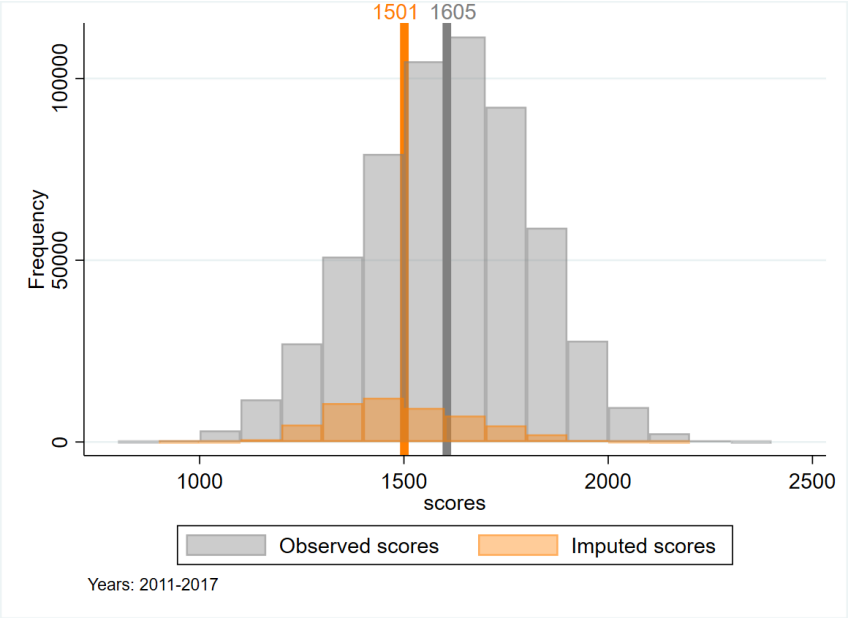
The mean imputation assumes that absent students would have performed as their participating classmates did on average. This means that class averages are unaffected, and (because of sorting within schools) school averages are slightly lower after the correction. The method of multiple imputation chained equations (MICE) allows for more flexibility by modelling each missing data conditional on the other variables in the data. It can be used in my setting because certain demographic variables (gender, age, previous maths grade in school, previous or future NABC participation and scores) are available

[14]For reference, see Quintano et al. (2009) for examples of skewed test score distributions in Italian primary schools

even for the absent students.

First, Figure 2.7 shows the distribution of observed and MICE-imputed test scores. It can be seen that (as expected) the average of the imputed scores is half a standard deviation lower than the observed average of the participating students.

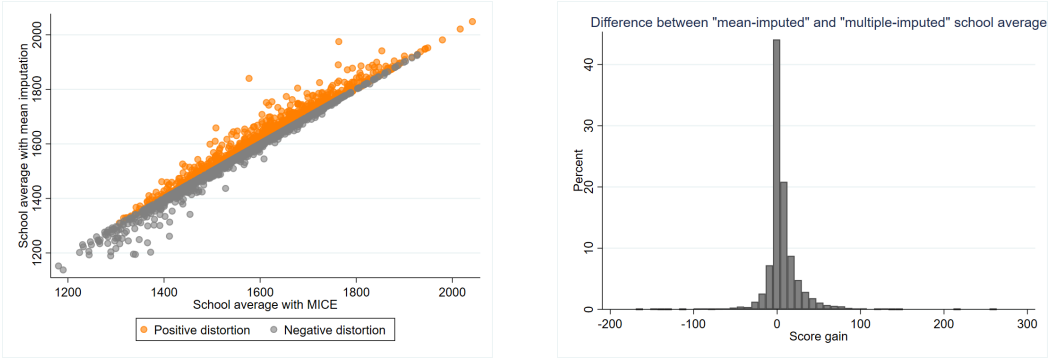
Figure 2.7: Observed and Imputed Test Scores



Note: The figure shows the distribution of observed and MICE-imputed test scores on a student level.

Second, I define a school's score gain as the difference between its MICE-imputed and mean-imputed test scores. Figure 2.8a shows the relationship between the two imputations, while Figure 2.8b shows the distribution of the obtained test score gains. It can be seen that most of the score gains are just slightly higher than 0, and they barely even reach the size of half a standard deviation (100). Note that although a MICE imputation is more precise than the mean imputation, it probably still overestimates the performance of absent students. The upper bound for the undistorted averages implies we have a lower bound for the score gains. According to back-of-the-envelope calculations reaching the threshold is rarely the result of absences (12-15 schools). Most schools do not reach the threshold despite high absence rates.

Figure 2.8: Imputation Results



(a) Relationship Between Imputed Scores

(b) Distribution of Score Gains

Note: The figures show the relationship between mean-imputed and MICE-imputed school averages. Panel (a) shows their relation on a scatter plot, indicating positive distortions (mean-imputed<MICE-imputed) by orange, and negative distortions (mean-imputed>MICE-imputed) by grey. Panel (b) shows the distribution of the distortions or score gains (the difference between the mean-imputed school average and the MICE-imputed school average).

2.5 Conclusion and Discussion

In this chapter, I provide evidence of test pool manipulation in a traditionally low-stakes testing environment, where free school choice creates reputational concerns for schools. To identify manipulation I exploit a policy change which introduced higher stakes in the testing through a minimum requirement for schools. I use difference-in-difference estimation where the "bite" of the otherwise universal policy differs across schools.

First, in line with the incentives of the new policy, I find that post-policy absence rates increased particularly in schools at risk of not fulfilling the minimum requirement. While absence rates increased on average by 0.6 percentage points after the policy was introduced, for the bottom 10% of schools the increase was 1.62 percentage points. Schools that would have been below the threshold in the pre-policy year (2012) increased their absence rate by 1.56 percentage points more than schools which would have been above the threshold. Note, however, that these estimates provide a lower bound on the extent of test pool manipulation, given anecdotal evidence on the pre-policy presence of the practice. In addition, I also discuss alternative explanations behind the increased absences (other policy changes, competition, health and weather conditions).

Second, I quantify schools' gains from test pool manipulation using multiple imputation. I find that although the variation in absences is large, schools do not benefit substantially from it. In particular, manipulations did not result in substantially more schools reaching the minimum threshold.

The results suggest that the net effect is small because schools with a large increase in their absence rate are likely to be the worst-performing schools which cannot reach

the threshold even when engaging in unethical activities. This might be because of the high level of segregation in the Hungarian school system Hermann and Kisfalusi (2023), Horn (2013), Horn et al. (2016), Lannert (2018), and Lőrincz and Antal-Fekete (2022). However, the manipulations could matter in cities with a more competitive schooling market, or in case school closure decisions would take into account NABC performance-[15] Further simulations are required to assess the role of manipulations depending on the structure of the local school market.

In Chapter 1, applying the detection algorithm developed in Jacob and Levitt (2003), I find no evidence of test score manipulation in the Hungarian NABC testing. This, together with the current results, points to the importance of costs and the potential benefits of manipulation. I argue that the presence of test pool manipulation and the absence of test score manipulation arises because of the unique testing environment where low-stake testing is paired with strict quality assurance, but no requirements on participation rate within schools. Compared to other tests studied, there is overall less incentive to cheat, but manipulation of test scores is more costly as well (see Chapter 1). Test pool manipulation is however less strictly monitored and punished.

To summarize, I conclude that standardized tests have to be carefully designed in order to ensure the reliability of their results. Avoiding high stakes and reducing the possibility of manipulation through various measures (e.g. centralised correction, using detection algorithms, or a minimum participation requirement) are both required to reach that goal.

For future research, the current analysis could be repeated using individual-level data instead of the school-level data, and predict the probability of being absent on the testing day for each student. In terms of assessing the extent of test pool manipulation more precisely, the effect of more recent (and stronger) policy changes could be estimated. Since 2018 not reaching the minimum requirement has become more costly for schools: in addition to preparing an action plan, they also have to send their teachers to training courses. At the same time, aggregate data show that –relative to earlier years– a larger fraction of students disappear from the testing between the 6th and 8th grade (see Figure A2.7). Student-level data would be necessary to assess whether these students dropped out from education, exemptions increased, or the differences are due to (manipulated) absences on the testing day.

Finally, it is also important to note that while the Educational Authority is aware of the possibility of such manipulations, the most recent changes in the testing protocol are expected to mitigate this problem. Since 2022, the NABC testing is computerized: students answer questions in a random order, using a computer in their school. Because of

[15]For a similar argument in the context of cross-state strategic voting in the US see Dahl et al. (2023), while in Appendix Section A2.3 I also discuss how within-school and across-school variation affect the potential gains and consequences of manipulation.

the limited availability of computers, the testing is scheduled throughout weeks, instead of a single testing day. Thus, it is not possible any more to make students miss the test just by letting them stay at home for one day, i.e., such manipulation becomes more costly. It will be, however, interesting to see whether the changes prompt new strategic responses by those schools and teachers that aim to be higher in the ranking.

References

Adamecz, A. (2023). Longer Schooling With Grade Retention: The Effects of Increasing the School Leaving Age on Dropping Out and Labour Market Success. *Economics of Education Review*, *97*, 102487.

Adamecz-Völgyi, A., Prinz, D., Szabó-Morvai, Á., and Vujić, S. (2021). The Labor Market and Fertility Impacts of Decreasing the Compulsory Schooling Age. *KRTK-KTI Working Papers*, (No. 40).

Ajzenman, N. (2021). The Power of Example: Corruption Spurs Corruption. *American Economic Journal: Applied Economics*, *13*(2), 230–257.

Azar, J., Marinescu, I., and Steinbaum, M. (2022). Labor Market Concentration. *Journal of Human Resources*, *57*, 167–199.

Azar, O. H., and Applebaum, M. (2020). Do Children Cheat to Be Honored? A Natural Experiment on Dishonesty in a Math Competition. *Journal of Economic Behavior and Organization*, *169*, 143–157.

Baker, G., Gibbons, R., and Murphy, K. J. (1994). Subjective Performance Measures in Optimal Incentive Contracts. *The Quarterly Journal of Economics*, *109*(4), 1125–1156.

Battistin, E., and Neri, L. (2023). School Performance, Score Inflation and Neighborhood Development. *Journal of Labor Economics*, *41*(3).

Bertoni, M., Brunello, G., and Rocco, L. (2013). When the Cat is Near, the Mice Won't Play: The Effect of External Examiners in Italian Schools. *Journal of Public Economics*, *104*, 65–77.

Borcan, O., Lindahl, M., and Mitrut, A. (2017). Fighting Corruption in Education: What Works and Who Benefits? *American Economic Journal: Economic Policy*, *9*(1), 180–209.

Cagala, T., Glogowsky, U., and Rincke, J. (2021). Detecting and Preventing Cheating in Exams: Evidence From a Field Experiment. *Journal of Human Resources*, 0620–10947R1.

Card, D., and Krueger, A. B. (1994). Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania. *The American Economic Review*, *84*(4), 772–793.

Cilliers, J., Mbiti, I. M., and Zeitlin, A. (2021). Can Public Rankings Improve School Performance? Evidence From a Nationwide Reform in Tanzania. *Journal of Human Resources*, *56*(3), 655–685.

Clarke, D. (2017). Estimating Difference-in-Differences in the Presence of Spillovers. *Munich Personal RePEc Archive*, *No. 81604*.

Cuesta, J. I., González, F., and Philippi, C. L. (2020). Distorted Quality Signals in School Markets. *Journal of Development Economics*, *147*, 102532.

Cullen, J. B., and Reback, R. (2006). Tinkering Toward Accolades: School Gaming Under a Performance Accountability System. *NBER Working Paper*, *No. 12286*.

Dahl, G. B., Engelberg, J., Lu, R., and Mullins, W. (2023). Cross-State Strategic Voting. *NBER Working Paper*, *No. 30972*.

Dee, T. S., Dobbie, W., Jacob, B. A., and Rockoff, J. (2019). The Causes and Consequences of Test Score Manipulation: Evidence from the New York Regents Examinations. *American Economic Journal: Applied Economics*, *11*(3), 382–423.

Depken, C. A. (1999). Free-Agency and the Competitiveness of Major League Baseball. *Review of Industrial Organization*, *14*(3), 205–217.

Diamond, R., and Persson, P. (2016). The Long-term Consequences of Teacher Discretion in Grading of High-stakes Tests. *NBER Working Paper*, *No. 22207*.

Duflo, E. (2001). Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment. *The American Economic Review*, *91*(4), 795–813.

Figlio, D., and Loeb, S. (2011). School Accountability. In Hanushek, E. A., Machin, S., and Woessmann, L. (Eds.) *Handbook of the Economics of Education* (Vol. 3, pp. 383–421). Elsevier.

Figlio, D. N. (2006). Testing, Crime and Punishment. *Journal of Public Economics*, *90*(4), 837–851.

Figlio, D. N., and Getzler, L. S. (2002). Accountabilty, Ability and Disability: Gaming the System. *NBER Working Paper*, *No. 9307*.

Hermann, Z. (2018). A családi pótlék iskolába járáshoz kötésének hatása az iskolába járásra és az iskolai teljesítményre. *Gyerekesély Műhelytanulmányok*, (No. 1).

Hermann, Z. (2019). The Impact of Decreasing Compulsory School-Leaving Age on Dropping Out of School. *Institute of Economics, Centre for Economic and Regional Studies*.

Hermann, Z., and Kisfalusi, D. (2023). School Segregation, Student Achievement, and Educational Attainment in Hungary. *International Journal of Comparative Sociology*.

Hermann, Z., and Semjén, A. (2021). The Effects of Centralisation of School Governance and Funding on Inequalities in Education Lessons From a Policy Reform in Hungary. *KRTK-KTI Working Papers*, (No. 38).

Holmstrom, B., and Milgrom, P. (1991). Multitask Principal–Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. *The Journal of Law, Economics, and Organization*, *7*, 24–52.

Horn, D. (2013). Diverging Performances: The Detrimental Effects of Early Educational Selection on Equality of Opportunity in Hungary. *Research in Social Stratification and Mobility*, *32*, 25–43.

Horn, D., Keller, T., and Robert, P. (2016). Early Tracking and Competition–a Recipe for Major Inequalities in Hungary. In Blossfeld, H.-P., Buchholz, S., Skopek, J., and Triventi, M. (Eds.) *Models of Secondary Education and Social Inequality: An International Comparison* (pp. 129–147).

Jacob, B. A. (2005). Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools. *Journal of Public Economics*, *89*, 761–795.

Jacob, B. A., and Levitt, S. D. (2003). Rotten Apples: an Investigation of the Prevalence and Predictors of Teacher Cheating. *The Quarterly Journal of Economics*, 843–877.

Köllő, J., and Sebők, A. (2023). The aftermaths of lowering the school leaving age–effects on roma youth. *KRTK-KTI Working Papers*, (No. 31).

Lannert, J. (2018). Nem gyermeknek való vidék: A magyar oktatás és a 21. századi kihívások. In Kolosi, T., and Tóth, I. G. (Eds.) *Társadalmi Riport 2018* (pp. 267–285). Tarki Tarsadalomkutatasi Intezet Zrt; TARKI.

Longobardi, S., Falzetti, P., and Pagliuca, M. M. (2018). Quis Custiodet Ipsos Custodes? How to Detect and Correct Teacher Cheating in Italian Student Data. *Statistical Methods and Applications*, *27*(3), 515–543.

Lucifora, C., and Tonello, M. (2020). Monitoring and Sanctioning Cheating at School: What Works? Evidence from a National Evaluation Program. *Journal of Human Capital*, *14*, 584–616.

Lőrincz, B., and Antal-Fekete, E. (2022). Oktatási egyenlőtlenségek, iskolai mobilitás és az oktatási rendszer átalakulása Magyarországon az 1980-as évektől napjainkig. In Kolosi, T., Szelényi, I., and Tóth, I. G. (Eds.) *Társadalmi Riport 2022* (pp. 207–223). TÁRKI.

Machin, S., and Sandi, M. (2020). Autonomous Schools and Strategic Pupil Exclusion. *The Economic Journal*, *130*, 125–159.

Martinelli, C., Parker, S. W., Pérez-Gea, A. C., and Rodrigo, R. (2018). Cheating and Incentives: Learning from a Policy Experiment. *American Economic Journal: Economic Policy*, *10*(1), 298–325.

Orosz, G., Tóth-Király, I., Bőthe, B., Paskuj, B., Berkics, M., Fülöp, M., and Roland-Lévy, C. (2018). Linking Cheating in School and Corruption. *European Review of Applied Psychology*, *68*(2), 89–97.

Quintano, C, Castellano, R, and Longobardi, S. (2009). A Fuzzy Clustering Approach to Improve the Accuracy of Italian Student Data. An Experimental Procedure to Correct the Impact of Outliers on Assessment Test Scores. *Statistica e Applicazioni*, *7 (2)*, 149–171.

Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests.* University of Chicago Press.

Rockerbie, D. W., and Easton, S. T. (2022). Race to the Podium: Separating and Conjoining the Car and Driver in F1 Racing. *Applied Economics*, *54* (54), 6272–6285.

Appendix

A2.1 Institutional Background

Free School Choice, Importance of NABC

Figure A2.1: Share of Students Enrolled in the Closest School to Their Home

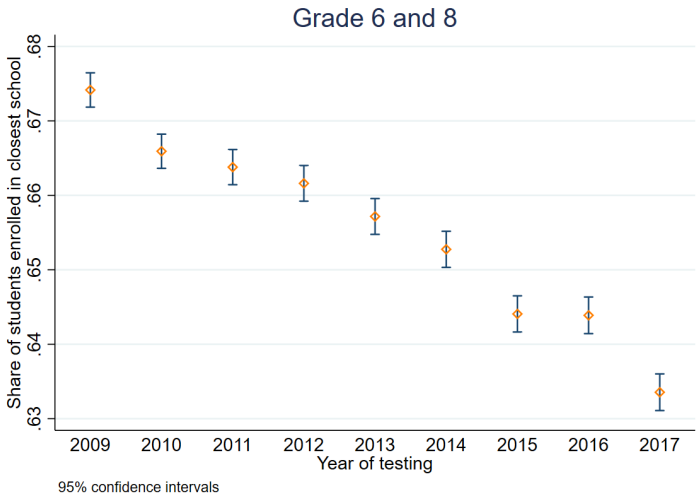


Table A2.1: Self-Declared Importance of NABC Results for Schools

	% Yes—Answer
Previous NABC results used	97.8
- to set pedagogical goals	89.7
- to describe school performance	82.5
- to evaluate students	24.6
- to evaluate teachers	31.7
Who were informed about the results	
- teachers	99.0
- provider	92.2
- students	90.3

Note: The table shows the self-declared importance of NABC results for schools. Data is from the institutional questionnaires filled out by school principals. Number of institutions 2008-2017: 22,819. Answer rate: 97.8%.

Exemptions

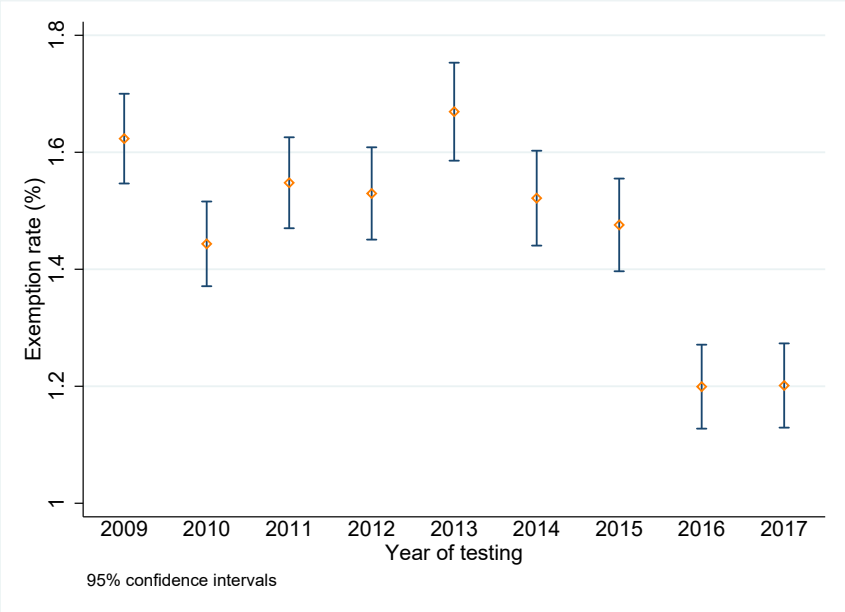
Schools have to indicate in their November data provision (online) which students will be exempted and why, but this can be modified until the testing day. School coordinators have to indicate reasons for absences on the attendance sheet following the below coding.

Code 1	Students with a certificate (doctor's) about their disability (physical, sensory, mental disorders)
Code 2	Temporary injury (e.g. broken arm)
Code 3	Language difficulties (studies in Hungarian for less than a year)

There are other students whose scores are not taken into account when school averages are calculated, but who cannot be exempted from the test:

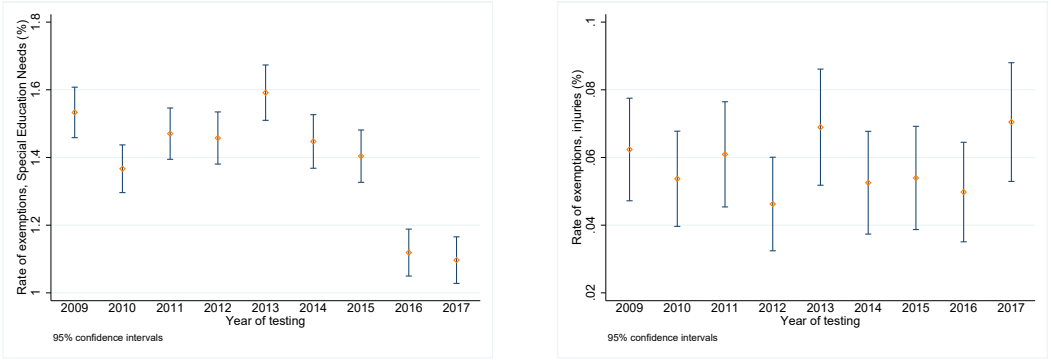
b, t, m (i, l, c)	Students with integration, learning (e.g. dyslexia) or conduct disorder Included in school averages as well
P	Other developmental disorders, Not included in school averages

Figure A2.2: Share of 8th Graders Exempted From the Testing



Note: The figure shows the share of exempted 8th graders throughout the years, with 95% confidence intervals.

Figure A2.3: Share of Exempted 8th Graders, by Reason of Exemptions



(a) Special Education Needs

(b) Temporary Injury

Note: The figure shows the share of exempted 8th graders throughout the years, with 95% confidence intervals.

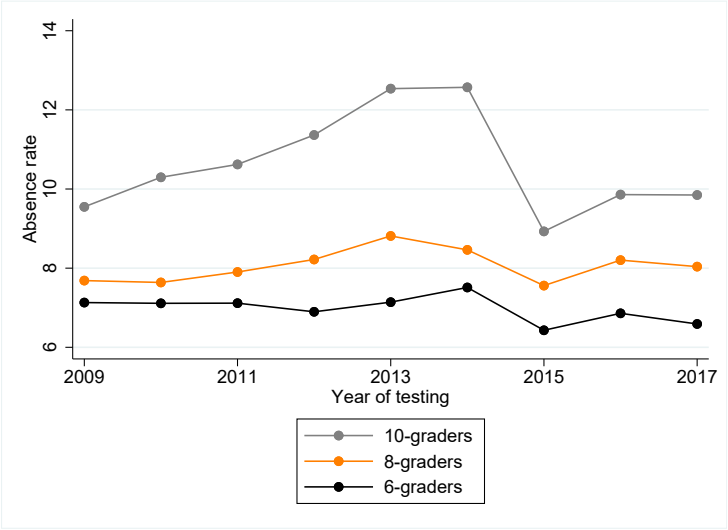
Table A2.2: Robustness: Treatment Effects on Exemption Rates

Outcome	Exemption rate				
	(1)	(2)	(3)	(4)	(5)
Treatment group	Overall	Below threshold in 2012	Around threshold in 2012	Sometimes below, pre-policy	Sometimes below, pre-policy
1.Treat#1.post2012		-0.749	0.244	0.145	0.182
		(0.500)	(0.193)	(0.392)	(0.352)
1.post2012	0.151**				
	(0.0625)				
2012 included	✓	-	-	-	✓
Balanced sample		✓	✓	✓	✓
Observations	25,395	19,618	19,618	19,618	22,071
Number of schools	3,244	2,453	2,453	2,453	2,453
R-squared	0.033	0.049	0.049	0.049	0.049

Note: The table presents the estimated effects of the minimum requirement on the post-policy exemption rates. Column (1) shows estimation results of Equation (2.1), exemption rate as the outcome variable. Columns (2)-(5) present results of estimating Equation (2.2) with different treatment groups. Time-variant school-level controls: % Grade repeater, number of 8th graders in school. School and year fixed effects are included. Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

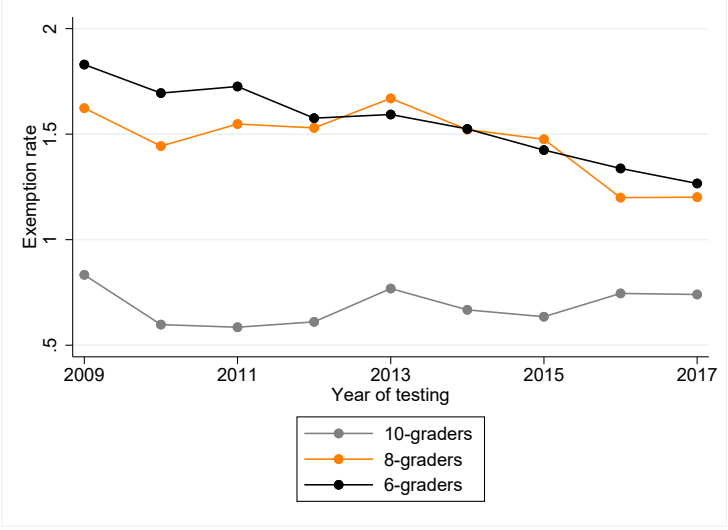
Other Tested Grades: 6th and 10th Grades

Figure A2.4: Share of Absent Students on the Testing Day, All Grades



Note: The figure shows the share of absent students on the testing day from 2009 to 2017.

Figure A2.5: Share of Students Exempted From the Testing, All Grades



Note: The figure shows the share of exempted students in each grade from 2009 to 2017.

Other Policy Changes in Hungary

During the studied time period, other policy changes also took place in Hungary that might affect absences: a change in the compulsory school leaving age, and child benefit became dependent on school attendance.

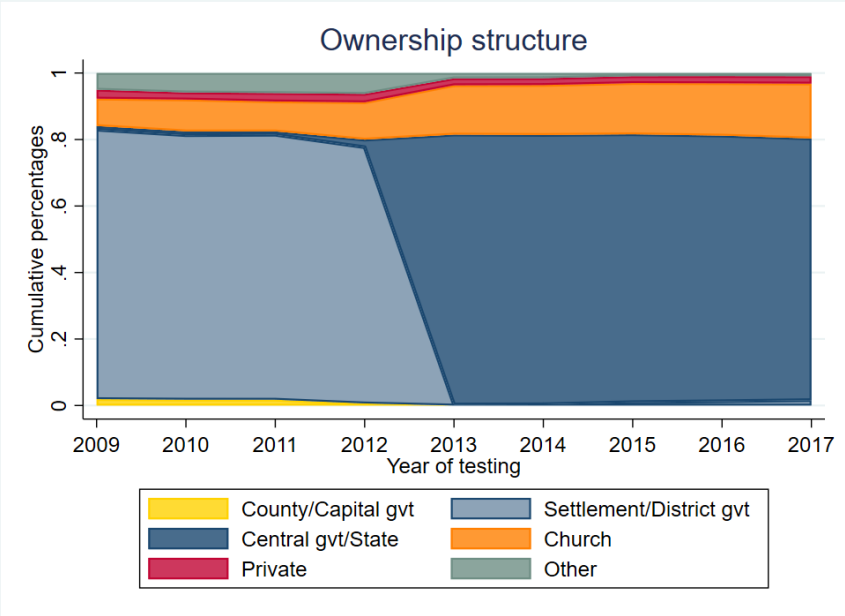
In 2012 compulsory school-leaving age was decreased from 18 to 16 years allowing earlier drop-outs. Adamecz-Völgyi et al. (2021) and Hermann (2019) show that the reform increased the likelihood of being neither in education nor in employment among 16-18-year-olds. Focusing on 17-year-old students, Köllő and Sebők (2023) show that school attendance decreased by more than 20 percentage points for Roma children and by 6 percentage points for non-Roma children. Studying the 1996 reform which increased the school leaving age from 16 to 18, Adamecz (2023) finds no effect on dropout probabilities and employment. Overall, these studies share the conclusion that compulsory schooling age legislation in itself is not successful in improving school completion outcomes.

This policy should mostly affect students after primary school (students in 8th grade are 13-14 years old), however, because of grade retention some 8th graders might be also affected. But note that students leaving the school system do not show up in the statistics as absent, as long as schools administer their dropouts. Since schools have to update their list of eligible students during the week before the test, only uncertain dropouts could affect absence rates.

Since 2010, child benefit suspension payments are tied to the school attendance of children. Initially, they implemented a policy where if a student accumulated more than 50 hours of unjustified absences, the child benefit would be suspended for 3 months. However, if there were no further unjustified absences afterwards, the benefits were paid out retrospectively. This policy was updated in 2012, stating that for absences exceeding 50 hours, the child benefit would be suspended indefinitely. Hermann (2018) studies the impact of the policy, focusing on short-term effects only (the period between 2010 and 2012) since later years would be affected by the school leaving age reform as well. He finds that the reform slightly decreased absences and grade retention, but did not affect student performance and dropouts. He also notes that students (and parents) easily gamed the system by getting doctor's notes even when they were not genuinely sick. In the context of my analysis, this policy should have the opposite effect, if any, and therefore, should not significantly influence the results.

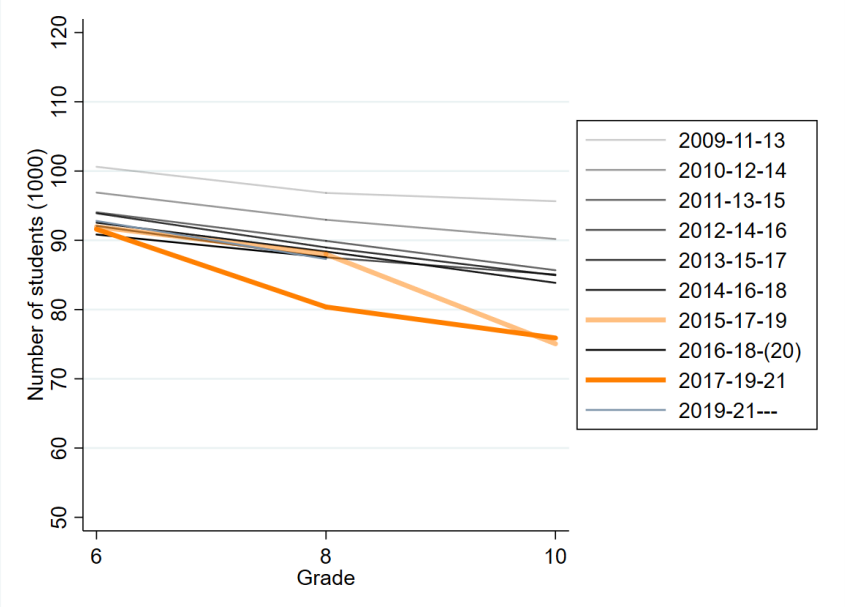
However, both for child benefit suspension and school leaving age, using daily absences would be a better benchmark.

Figure A2.6: Changes in Ownership Structure Through Time



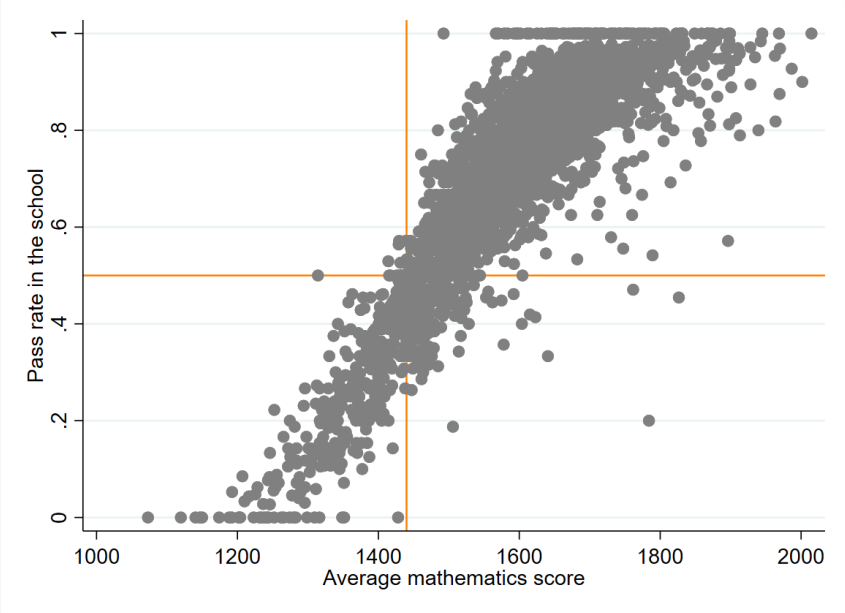
Note: The figure shows the changes in the ownership structure of schools through time, depicting owner types with different colours.

Figure A2.7: The Number of Students Eligible for Testing, by Cohorts



Note: The figure shows the number of students in each cohort tested in 6th, 8th and 10th grade, i.e., every second year. The label shows the years of testing for each cohort. Eligible students are those not exempted from the testing. The figure is recreated based on the idea of István Nahalka.

Figure A2.8: Two Dimensions of the Minimum Requirement



Note: The figure demonstrates how the minimum requirement can be captured by two dimensions. The x-axis shows the average mathematics score in the school, the vertical (orange) line representing the minimum level (1408 points). The y-axis shows the measure of the minimum requirement, i.e. the share of students who reached the minimum level in the school, the horizontal (orange) line representing the 50% cutoff. Note that schools in the bottom right area do not fulfil the minimum requirement, although their average test score is above the score corresponding to the minimum level.

A2.2 Panel Structure of the Data – the Balanced Sample

Schools are identified in the data with the combination of institution ID, site ID, and school type (primary school, 6-grade high school, or 8-grade high school). The institution ID can change when the provider of the institution changes, or when schools merge. Institutions might consist of several sites within a city, or even across villages. The site ID is assigned by the institutions each year, and thus, changes more often. The panel structure of the data can be improved by adding new IDs, based on the schools' addresses. Table A2.3 shows how the size of the balanced sample increases when linking schools based on their address as well, while Table A2.4 and Table A2.5 compare summary statistics for the balanced and the total sample.

Table A2.3: Size of the Balanced Sample

Identifier	N	Share
school IDs	1,939	47.04%
address	2,120	65.38%
combined ID	2,453	75.50%

Note: The table presents the size of the balanced sample (relative to the overall number of schools), depending on the way of identifying schools.

Table A2.4: Balance Table I

	Dropped	Balanced sample	Total
Average number of students	25.54	34.20	33.07
	(18.62)	(20.95)	(20.87)
Average number of participants	23.13	31.55	30.45
	(17.67)	(19.92)	(19.85)
Action plan	0.11	0.05	0.06
	(0.32)	(0.22)	(0.24)
% Absence	9.43	7.75	7.92
	(11.10)	(7.24)	(7.74)
% Exemption	2.04	1.26	1.33
	(7.77)	(3.56)	(4.19)
% Male	51.59	50.77	50.85
	(13.22)	(9.96)	(10.33)
Avg maths score	1575.00	1606.79	1603.58
	(127.87)	(115.47)	(117.17)
Avg ability level	3.49	3.72	3.70
	(0.93)	(0.84)	(0.85)
Observations	3,531	22,077	25,608

Note: The table presents summary statistics separately for schools in the balanced sample, and for schools excluded from the baseline specification.

Table A2.5: Balance Table II: Categorical Variables

	Categories	Dropped	Balanced sample	Total
Size	Less than 5 students	3.5	0.1	0.6
	Small school	51.1	29.3	32.1
	Medium	26.6	33.8	32.8
	Large	18.8	36.8	34.5
Type	Primary school	87.2	92.9	92.1
	8-year high school	3.5	3.2	3.2
	6-year high school	9.3	4.0	4.7
Locality	village	34.1	44.4	43.1
	city	34.9	29.6	30.3
	county center	17.9	13.5	14.1
	capital (Budapest)	13.2	12.5	12.6
Size of locality	Small village	11.2	8.9	9.2
	Middle-sized village	20.1	29.0	27.8
	Large village	2.8	6.5	6.1
	Small city	10.2	9.6	9.7
	Middle-sized city	20.5	14.5	15.2
	Large city	4.2	5.5	5.3
	County center	17.9	13.5	14.1
	Budapest	13.2	12.5	12.6
Region	Budapest	13.2	12.5	12.6
	Central Hungary	6.3	10.8	10.2
	Central Transdanubia	9.8	12.0	11.7
	Western Transdanubia	7.0	11.3	10.7
	Southern Transdanubia	14.7	10.1	10.7
	Northern Hungary	14.8	15.4	15.3
	Northern Great Plain	16.3	15.9	15.9
	Southern Great Plain	17.8	12.1	12.8
	Below threshold in 2012	11.7	6.2	6.7
	Around threshold in 2012	28.9	23.1	23.6
	Sometimes below	12.8	10.5	10.8
	Observations	3,531	22,077	25,608

Note: The table presents how schools are distributed across different types, separately for schools in the balanced sample, and for schools excluded from the baseline specification. All numbers are percentages (%). The last three rows show the shares of treated schools according to the different treatment definitions.

A2.3 Robustness

Alternative Specifications

Table A2.6: Robustness: Treatment Effects, Inclusion of Fixed Effects and Standard Errors

	(1)	(2)	(3)	(4)
Outcome	Absence rate			
(I) Below threshold	1.326**	1.410**	1.236***	1.263**
	(0.556)	(0.560)	(0.378)	(0.534)
(II) Around threshold	0.205	0.142	0.121	0.132
	(0.260)	(0.264)	(0.221)	(0.259)
(III) Sometimes below	1.017**	1.116**	0.842***	0.905**
	(0.444)	(0.448)	(0.303)	(0.435)
2012 included	-	-	-	-
Time-invariant controls		✓		
School FE			✓	✓
Standard error	robust	robust	conv	clustered
Observations	21,395	20,936	20,936	21,395
Number of schools			2,809	2,809

Notes: The table presents how the inclusion of school fixed effects and clustering of standard errors at the school level affects the estimated treatment effect ($\hat{\beta}_3$). Each row corresponds to a different treatment definition (presented in Columns of the baseline Table 2.2). Time-invariant school-level controls: training type, type of settlement, region. Time-variant school-level controls: % Exemption, % Grade repeater, number of 8th graders in school. Time-invariant controls: settlement type, school type, region. Year fixed effects are included in all specifications. Inclusion of school fixed effects and types of standard errors are indicated in the bottom rows. $N = 21,454, N_{school} = 3,061$. Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1.

Table A2.7: Robustness: Treatment Effects, Sample Restrictions

	(1)	(2)	(3)	(4)	(5)
Outcome	Absence rate				
(I) Below threshold	1.263**	1.564***	1.710***	1.704***	1.752***
	(0.534)	(0.551)	(0.574)	(0.576)	(0.576)
(II) Around threshold	0.132	0.209	0.152	0.211	0.168
	(0.259)	(0.263)	(0.272)	(0.262)	(0.272)
(III) Sometimes below	0.905**	0.804*	0.893*	0.908**	0.909**
	(0.435)	(0.441)	(0.460)	(0.461)	(0.459)
School FE	✓	✓	✓	✓	✓
Standard error	clustered	clustered	clustered	clustered	clustered
Balanced sample		✓	✓	✓	✓
Drop church_switch			✓	✓	✓
Drop small schools				✓	
Linear time trend					✓
Observations	21,395	19,618	18,354	18,327	18,354
Number of schools	2,809	2,453	2,295	2,294	2,295

Notes: The table presents robustness of the baseline results in Table 2.2. Now, instead of columns, rows correspond to the different treatment definitions. Time-variant school-level controls: % Exemption, % Grade repeater, number of 8th graders in school. School and year fixed effects are included. Cluster-robust standard errors in parentheses (clustered at the school level) *** p<0.01, ** p<0.05, * p<0.1

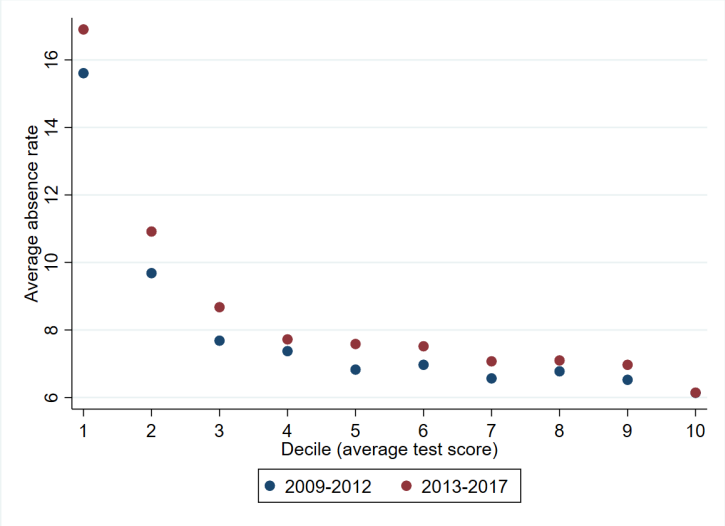
Column (1): estimated on the whole (unbalanced) sample. Column (2) corresponds to the main specification. Column (3) excludes schools that switched to church ownership around 2013. Column (4) additionally excludes the smallest schools, i.e., those with less than five students. Note that small schools are often not part of the balanced sample, and they often switched to church ownership. Column (5) adds a linear time trend.

Table A2.8: Robustness: Treatment Effects, Shorter Post-policy Period

Outcome	Absence rate				
Treatment group	(1) Overall	(2) Below threshold in 2012	(3) Around threshold in 2012	(4) Sometimes below, pre-policy	(5) Sometimes below, pre-policy
1.Treat#1.post2012		1.601***	0.00530	0.756	0.331
		(0.609)	(0.291)	(0.494)	(0.467)
1.post2012	0.533***				
	(0.101)				
2012 included	✓	-	-	-	✓
Balanced sample		✓	✓	✓	✓
Observations	19,824	14,712	14,712	14,712	17,165
Number of schools	3,204	2,453	2,453	2,453	2,453
R-squared	0.234	0.239	0.239	0.239	0.239

Note: The table presents the estimated effects of the minimum requirement on the post-policy absence rates, using a shorter post-policy period than in the baseline Table 2.2. The post-policy years are only 2013 and 2014. Column (1) shows estimation results of Equation (2.1). Columns (2)-(5) present results of estimating Equation (2.2) with different treatment groups. Time-variant school-level controls: % Exemption, % Grade repeater, number of 8th graders in school. School and year fixed effects are included. Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

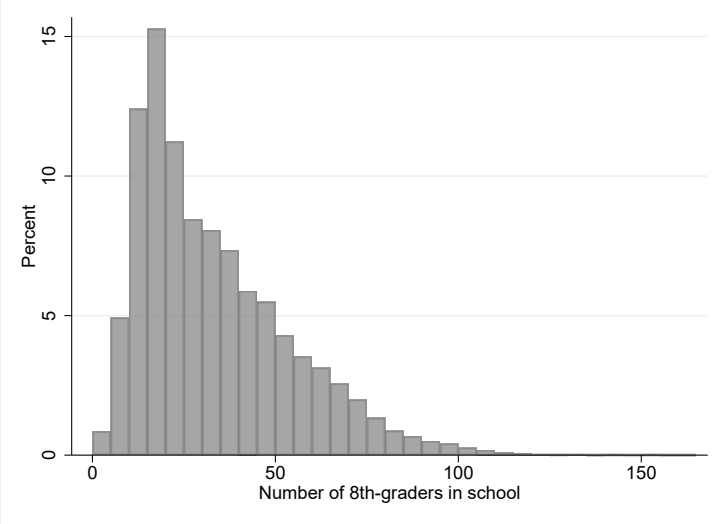
Figure A2.9: Absence Rate in Each Decile, Pre and Post Policy



Note: The figure shows the share of absent students in the pre-policy and post-policy periods, by decile. Pre-policy period: 2009-2012 (blue). Post-policy period: 2013-2017 (red).

Role of School Size

Figure A2.10: Distribution of School Size (Number of 8th Graders Enrolled in a School)



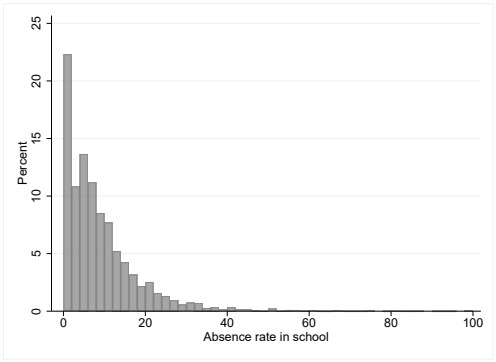
Note: The figure shows the distribution of school size, defined as the number of 8th graders enrolled in a school. Includes all schools in the period 2009-2017, $N = 25,559$.

Table A2.9: School Size: Number of 8th Graders Enrolled in a School, by Year

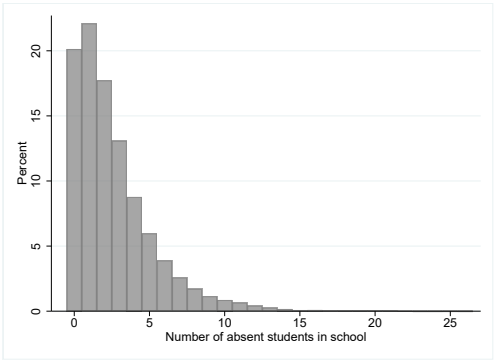
	(1) Mean	(2) Median	(3) SD	(4) Min	(5) Max	(6) N
2009	35.22	30	22.22	1	153	2,959
2010	35.82	30	22.65	1	149	2,911
2011	33.70	28	20.92	1	153	2,874
2012	32.85	28	20.36	1	134	2,830
2013	32.01	27	20.15	1	165	2,809
2014	31.28	26	20.01	1	134	2,799
2015	31.76	27	20.32	1	130	2,801
2016	31.70	27	20.29	1	147	2,788
2017	31.56	26	20.41	2	158	2,788
Total	32.91	28	20.91	1	165	25,559

Note: The table presents summary statistics on school size through time. School size is defined as the number of 8th graders enrolled in a school.

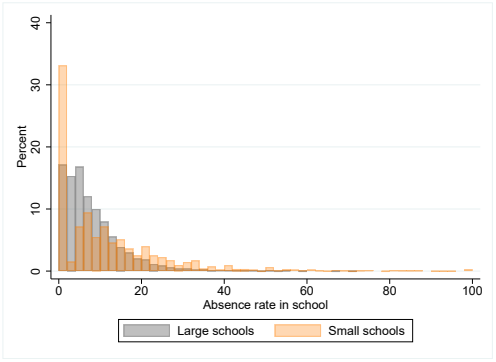
Figure A2.11: Distribution of Absences



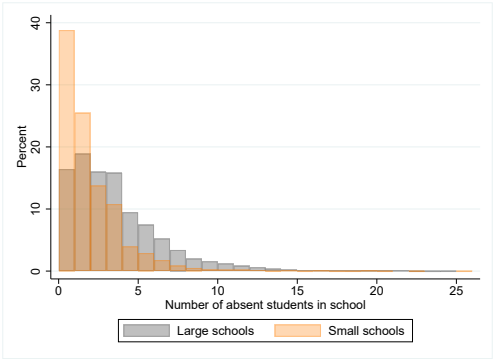
(a) Absence Rate



(b) Number of Absent Students



(c) Absence Rate, by Size



(d) Number of Absent Students, by Size

Note: The figure shows the distribution of absences across schools. The left panels show the distribution of absence rates, while the right panels show the distribution of the number of absent students. The bottom panels show the distribution separately for large (grey) and small (orange) schools. Small schools are defined (using the Educational Authority's categorization) as schools with less than 5 students or primary schools with 5-33 students, 8-year grammar schools with 5-83 students, 6-year grammar schools with 5-54 students. Primary schools are considered mid-size or large above 33 students enrolled, 8-year grammar schools above 83 students and 6-year grammar schools above 54 students. This definition takes into account all tested grades (6, 8, 10), and the type of schools.

To account for the differences in school size, in the baseline specifications I used absence rates as the outcome variable. Given that there are many small schools in Hungary (see distribution of school size in Figure A2.10, and the distribution of absences in Figure A2.11), however, raises another potential concern. One absent student corresponds to a higher absence rate in a smaller class or school. This raises the question of whether the observed effect sizes are driven by small schools, where the addition of just one absent student might result in a substantial change in absence rates.

Thus, I re-estimate the main specification presented in Table 2.2, using the number of absent students instead of absence rate as the outcome variable. Table A2.10 shows that the treatment effects in this case are negative and only weakly statistically significant or insignificant. Even though I control for school size (the number of students enrolled in

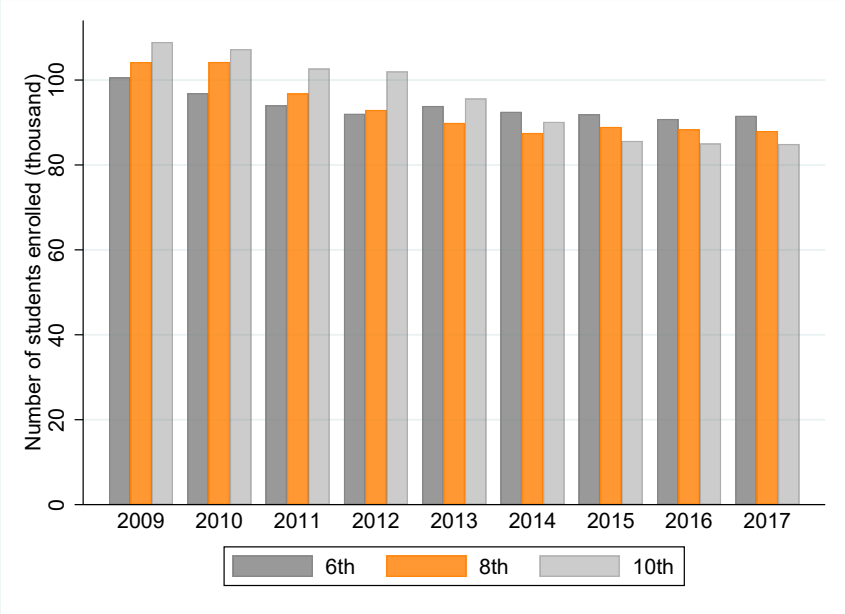
8th grade), the negative effects may still be attributed to the concurrent decline in the overall number of students (see Figures A2.12 and A2.13).

Table A2.10: Robustness: Number of Absent Students

Outcome	Number of absent students				
	(1)	(2)	(3)	(4)	(5)
Treatment group	Overall	Below threshold in 2012	Around threshold in 2012	Sometimes below, pre-policy	Sometimes below, pre-policy
1.Treat#1.post2012		-0.0376	-0.0882	-0.180*	-0.242**
		(0.148)	(0.0756)	(0.106)	(0.0997)
1.post2012	0.184***				
	(0.0266)				
2012 included	✓	-	-	-	✓
Balanced sample		✓	✓	✓	✓
Observations	25,395	19,618	19,618	19,618	22,071
Number of schools	3,244	2,453	2,453	2,453	2,453
R-squared	0.186	0.181	0.181	0.181	0.182

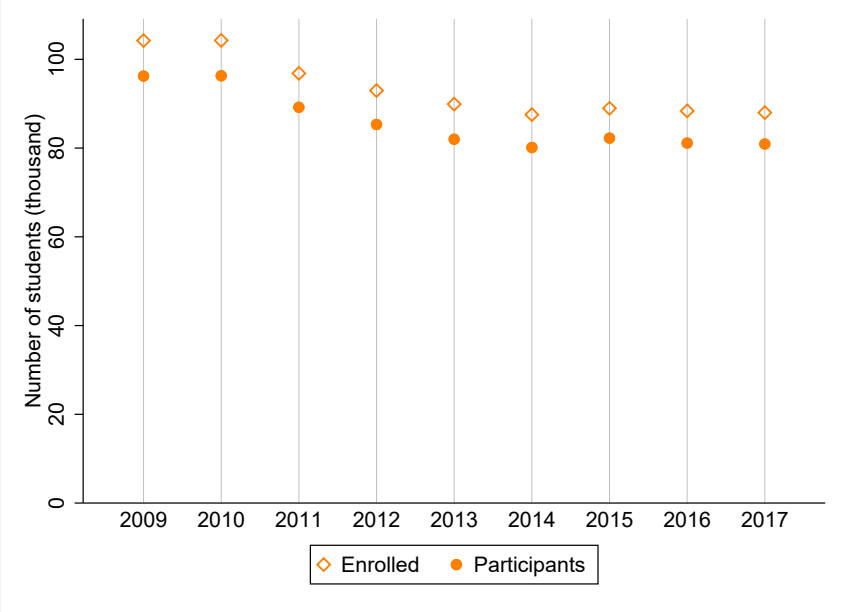
Note: The table presents the estimated effects of the minimum requirement on the post-policy absences. The outcome variable is the number of absent students instead of absence rates (as in Table 2.2). Column (1) shows estimation results of Equation (2.1). Columns (2)-(5) present results of estimating Equation (2.2) with different treatment groups, indicated in the top row. Time-variant school-level controls: % Exemption, % Grade repeater, number of 8th graders in school. School and year fixed effects are included. Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Figure A2.12: Total Number of Students Enrolled, by Grade and Year



Note: The figure shows the total number of students enrolled, by grade and year.

Figure A2.13: Total Number of Enrolled and Participating Students in 8th Grade, by Year



Note: The figure shows the number of enrolled and participating students in 8th grade, by year.

Next, to address potential differences in absence rates arising from varying denominators, I re-estimate the main specifications on the sample of small and large schools separately. Table A2.11 and Table A2.12 show that the treatment effects are larger (and significant) in case of large schools. This mitigates the concern that small schools' (mechanically) high absence rates drive the results.

Table A2.11: Robustness: Small vs Large schools I

Definition of Educational Authority

	Outcome	Absence rate				
		(1)	(2)	(3)	(4)	(5)
	Treatment group	Overall	Below threshold in 2012	Around threshold in 2012	Sometimes below, pre-policy	Sometimes below, pre-policy
Large	1.Treat#1.post2012		2.297***	0.809***	1.794***	1.393***
			(0.756)	(0.277)	(0.574)	(0.531)
	1.post2012	0.620***				
		(0.0866)				
	Observations	17,078	13,861	13,861	13,861	15,594
	Number of schools	2,071	1,733	1,733	1,733	1,733
	R-squared	0.171	0.171	0.170	0.171	0.175
Small	1.Treat#1.post2012		1.031	-0.699	0.151	-0.324
			(0.804)	(0.529)	(0.667)	(0.616)
	1.post2012	0.646***				
		(0.219)				
	Observations	8,317	5,757	5,757	5,757	6,477
	Number of schools	1,173	720	720	720	720
	R-squared	0.276	0.283	0.283	0.282	0.282
	2012 included	✓	-	-	-	✓
	Balanced sample		✓	✓	✓	✓

Note: The table presents the estimated effects of the minimum requirement on the post-policy absences, separately for small (top panel) and large (bottom panel) schools. Small schools are defined using the Educational Authority's categorization as schools with less than 5 students or primary schools with 5-33 students, 8-year grammar schools with 5-83 students, 6-year grammar schools with 5-54 students. Primary schools are considered mid-size or large above 33 students enrolled, 8-year grammar schools above 83 students and 6-year grammar schools above 54 students. This definition takes into account all tested grades (6, 8, 10), and the type of schools. Column (1) shows estimation results of Equation (2.1). Columns (2)-(5) present results of estimating Equation (2.2) with different treatment groups, indicated in the top row. Time-variant school-level controls: % Exemption, % Grade repeater, number of 8th graders in school. School and year fixed effects are included. Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Table A2.12: Robustness: Small vs Large Schools II

Below/Above Median School Size (Students Enrolled)

	Outcome	Absence rate				
		(1)	(2)	(3)	(4)	(5)
	Treatment group	Overall	Below threshold in 2012	Around threshold in 2012	Sometimes below, pre-policy	Sometimes below, pre-policy
Large	1.Treat#1.post2012		2.673**	0.971***	1.400	1.311
			(1.217)	(0.348)	(0.905)	(0.812)
	1.post2012	0.574***				
		(0.0945)				
	Observations	11,799	9,606	9,606	9,606	10,807
	Number of schools	1,440	1,201	1,201	1,201	1,201
	R-squared	0.141	0.142	0.141	0.141	0.142
Small	1.Treat#1.post2012		1.376**	-0.135	0.763	0.209
			(0.622)	(0.369)	(0.515)	(0.478)
	1.post2012	0.682***				
		(0.150)				
	Observations	13,596	10,012	10,012	10,012	11,264
	Number of schools	1,804	1,252	1,252	1,252	1,252
	R-squared	0.258	0.260	0.260	0.260	0.261
	2012 included	✓	-	-	-	✓
	Balanced sample		✓	✓	✓	✓

Note: The table presents the estimated effects of the minimum requirement on the post-policy absences, separately for small (top panel) and large (bottom panel) schools. Small and large schools are defined as schools below and above the median school size in the balanced sample, respectively (median schools size based on the number of students enrolled in 8th grade: 30). Compared to Table A2.11, this definition provides a more balanced partition in terms of sample size. Column (1) shows estimation results of Equation (2.1). Columns (2)-(5) present results of estimating Equation (2.2) with different treatment groups, indicated in the top row. Time-variant school-level controls: % Exemption, % Grade repeater, number of 8th graders in school. School and year fixed effects are included. Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Healthcare Use of Students

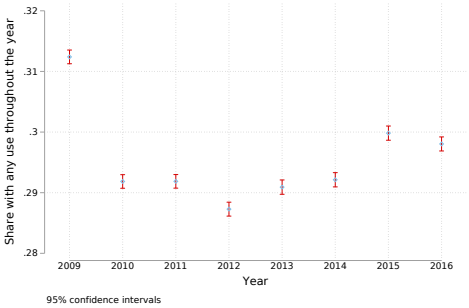
To provide an overview of the state of healthcare use by the tested students, I present aggregate statistics on 14-year-old children (although 14-year-olds might be in 7th grade and 9th grade as well, they are most likely to be in 8th grade). I use an individual-level administrative panel data set that covers monthly healthcare expenses for a random 50% sample of the Hungarian population (National Health Insurance Fund Administration). The data contains information on the number of GP, outpatient and inpatient visits, and corresponding monthly spending. Spending measures exhibit trends which suggest they depend on reimbursement schemes that can change throughout the years. Thus, I mostly focus on frequency measures. Since children's healthcare use is less common (compared to adults), I construct a dummy variable indicating whether any healthcare was used in a given month (based on GP, outpatient, and inpatient visits). Other dummies indicate whether a certain type of care was used (GP, outpatient care, inpatient care, prescribed pharmaceuticals). I also use the average number of GP visits (conditional on usage) and average outpatient spending (conditional on usage).

First, Figure A2.14 shows yearly healthcare use of 14-year-olds, and one can notice an increasing trend (starting in 2013) in the share of students using outpatient care (at least once in the calendar year). Similar patterns can be seen in Figure A2.15 which shows the share of students using healthcare in May (during the month when the NABC test is conducted). Note that while in 2013 there is a slight increase in the share of 14-year-olds who used inpatient care in May, the share is still below 1%. This is well below the share of absent students, while testing takes place on one day only, and these numbers represent inpatient care use in the whole month.

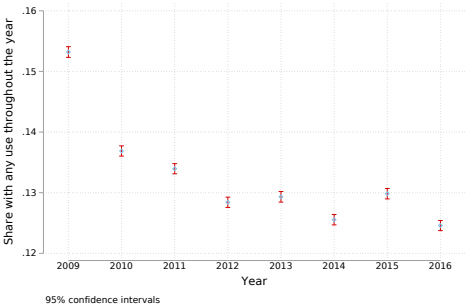
Second, Figure A2.16 shows healthcare use by cohorts (based on expected school-starting year). The cohort that is in 8th grade in 2013 started school in 2005. They do not seem to exhibit very different patterns than other cohorts: increase throughout the years, and a larger jump in outpatient care use around the 8th and 9th grade. However, because of the somewhat larger jump in outpatient visits in 2013 May, it might be worth further investigating why the cohort and/or the month would be special in terms of sicknesses, and whether these could be related to higher absence rates on the testing.

Overall, one can notice that there are yearly trends in spending, and there are age effects as well. From the age of 14, children tend to visit outpatient care more often (while there is no difference between 11, 12 and 13-year-olds, we see an increase from the age of 14).

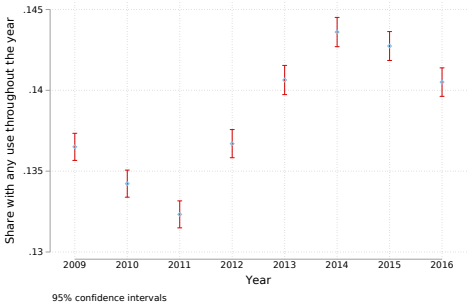
Figure A2.14: Yearly Healthcare Use of 14-Year-Olds



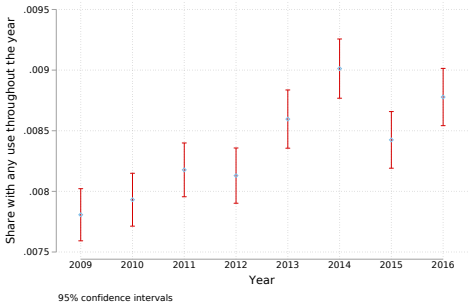
(a) GP Visit



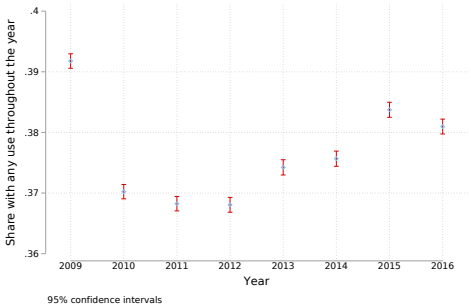
(b) Prescribed Pharmaceuticals



(c) Outpatient Care



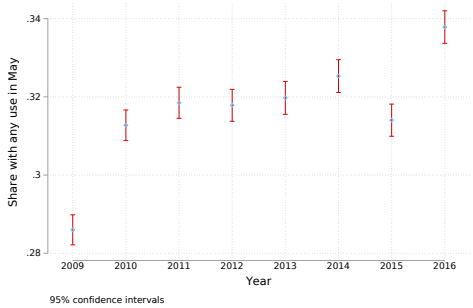
(d) Inpatient Care



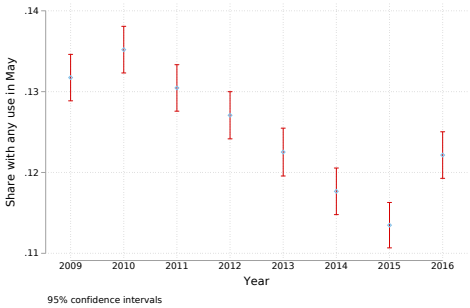
(e) Any Healthcare Use

Note: The figures show the share of 14-year-olds using healthcare each year, by type of care.

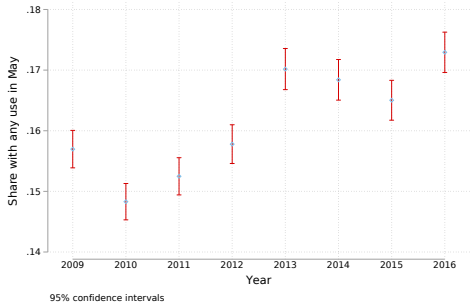
Figure A2.15: Healthcare Use of 14-Year-Olds in May



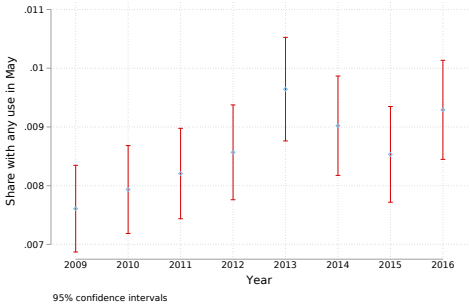
(a) GP Visit



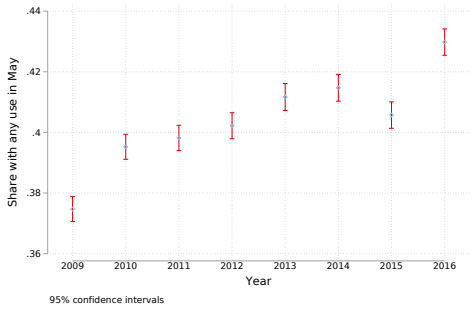
(b) Prescribed Pharmaceuticals



(c) Outpatient Care



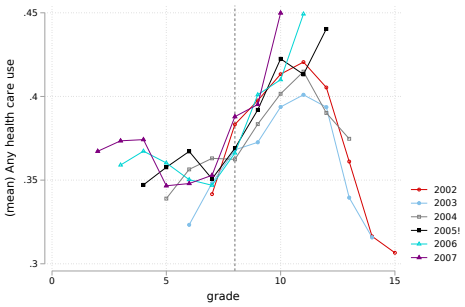
(d) Inpatient Care



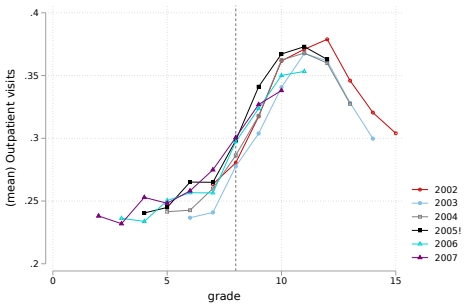
(e) Any Healthcare Use

Note: The figures show the share of 14-year-olds using healthcare each May, by type of care.

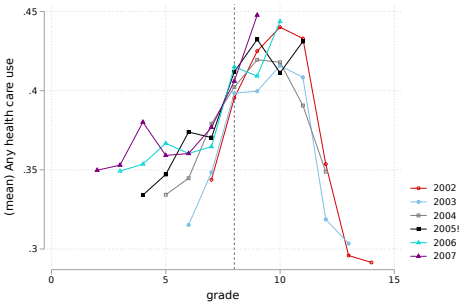
Figure A2.16: Healthcare Use by Cohorts



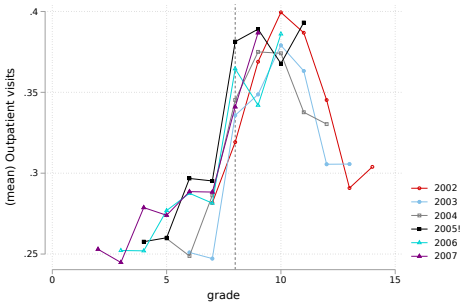
(a) Any Healthcare Use, Year



(b) Outpatient Visit, Year



(c) Any Healthcare Use, May



(d) Outpatient Visits, May

Note: The figures show the healthcare use of student cohorts. The x-axis shows the grade, and each point corresponds to average usage in the corresponding school year (instead of the calendar year). Different colours indicate the different cohorts, labelled with their school starting year (e.g., the cohort starting school in 2005 was in 8th grade in 2013). The vertical line indicates 8th grade. Note that children are assigned to the cohort when they should officially start school, i.e., children starting school earlier or later, or grade-repeaters are not in the same grade as the cohort they are assigned to.

Weather Conditions Around Testing Day

Table A2.13: Weather Conditions Around Testing Day

	Monthly	Daily				Deviation, relative to		
	average [°C]	average [°C]	max [°C]	min [°C]	PPT [mm]	monthly avg	prev. day	next day
27/05/2009	18	19.2	26.5	16.6	1	+1.2	-4.4	+3
26/05/2010	16.6	18.7	25.8	13.3	5.3	+2.1	+0.9	-1.1
25/05/2011	17.4	21.3	27.2	18	-	+3.9	-2.1	+1.3
30/05/2012	18.5	20.8	26.3	14.6	-	+2.3	+3.5	-0.3
29/05/2013	17.2	18.3	24.7	11.2	4.3	+1.1	+3.1	+3.3
28/05/2014	16.4	18.5	23	15.5	-	+2.1	-2.8	+2.4
27/05/2015	17.3	13.6	17.4	11.4	-	-3.7	-3.2	+0.4
25/05/2016	17.2	19.4	25.5	13.8	-	+2.2	+3.1	-1.1
24/05/2017	18.2	18.2	22.7	14.3	-	0	-1.4	+2.3

Note: The table presents temperature and precipitation (rainfall) data for the testing days: monthly average temperature, daily average, maximum and minim temperature in degrees Celsius, and daily precipitation in millimetres. The last three columns show the deviation of the daily average temperature in comparison to the monthly average, the previous day's average and the next day's average temperature. Numbers in red (+) indicate a warmer testing day, while numbers in blue (-) indicate a colder testing day in the respective comparison.

Role of Competition Between Schools

Some results suggest that the introduction of the minimum requirement affected higher-performing schools as well through a trickle-down effect and by the higher-stake nature of the testing. Thus, further analysis could assess the role of competition among schools in the change in behaviour. As a first step, following Cilliers et al. (2021) I assess the importance of local rankings instead of the national ranking.

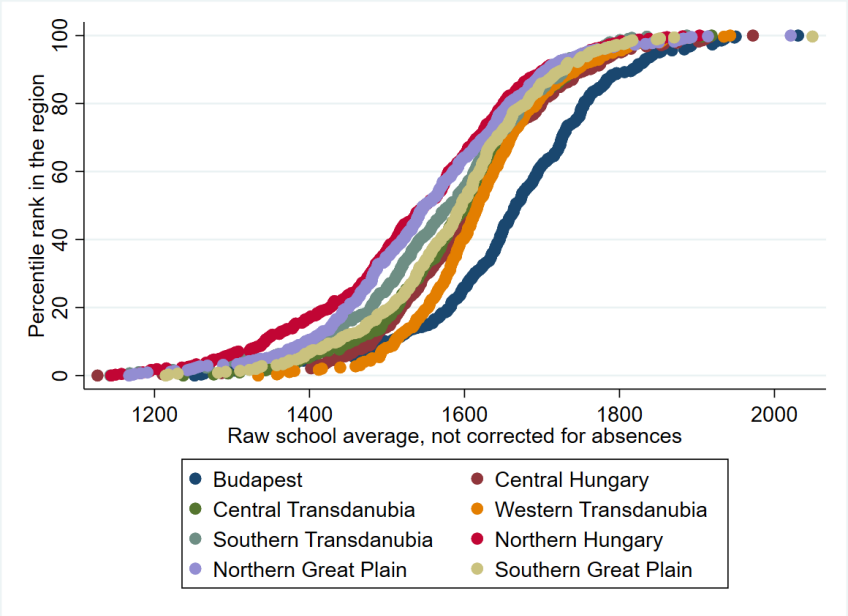
Free school choice resulted in the Hungarian school system's marketization. As seen in Figure A2.1, 35% of students are not enrolled in the closest school to their home. Several rankings are available online for parents who have to choose a school for their children, and every year the list of the 100 best schools is published by a prestigious weekly economic and political magazine. An important component of this evaluation is the schools' performance on the NABC. Since most of the 100 best schools are located in Budapest, and distance from school is still an important determinant of school choice (especially at the primary school level), local rankings are more salient and particularly important in parental decision-making. Most schools do not have a chance to make it to the top 100, but they can aim to be the best in their region, in their county, or in their district.

Local Rankings

Figure A2.17 illustrates why it is important to look at local school markets instead of the whole country. It shows the relationship between average scores and rankings of schools. It can be seen that the same average score ranks a school much lower in Budapest than in any other region - e.g. if a school performs around the national average of 1,600

points, this would put them in the bottom quartile in Budapest, but in the top quartile in Northern Hungary. Differences are even more pronounced when looking at smaller geographical units (counties or districts).

Figure A2.17: Regional Rankings



Note: The figure shows the relationship between schools' average mathematics test scores and their percentile rank in the region, separately for each region. The sample is all primary schools in 2012.

I look at how local rankings matter following Cilliers et al. (2021). In this case R_s of Equation (2.2) corresponds to a categorical variable indicating in which part of the local rankings the school is located. Rankings are based on the average mathematics score of the schools, and local rankings are at the regional, county and district levels.

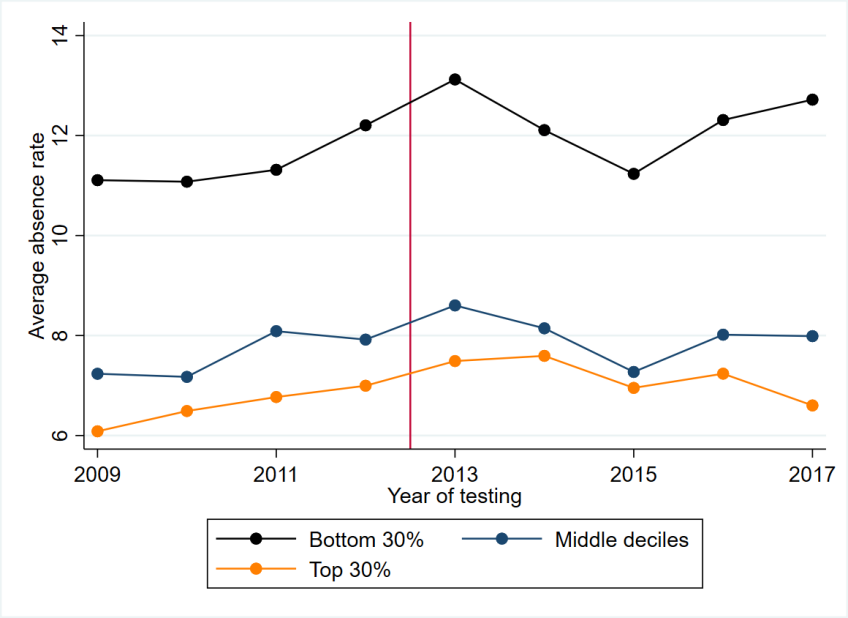
Table A2.14 shows the results of estimating Equation (2.2) using ranking-based definitions of R_s. The results show that part of the differences seen at the national level captures regional differences, but significant differences remain. I find a larger increase in post-policy absence rates in the bottom 30% of schools than in the top 30% of schools at any local market. Figure A2.18 presents the event study figures corresponding to Column 3 of Table A2.14.

Table A2.14: Estimated Post-policy Effects, by Local Rankings

Level	(1) National in 2012	(2) Within-region decile in 2012	(3) Within-county decile in 2012	(4) Within-district decile in 2012
1-3rd deciles # post2012	0.722**	0.496*	0.577**	0.567**
	(0.294)	(0.293)	(0.292)	(0.287)
4-7th deciles # post2012	0.153	-0.0823	-0.0671	0.126
	(0.221)	(0.222)	(0.222)	(0.227)
2012 included	-	-	-	-
Balanced sample	✓	✓	✓	✓
Observations	17,184	17,184	17,184	17,184
Number of schools	2,148	2,148	2,148	2,148
R-squared	0.225	0.224	0.224	0.224

Note: The table presents the estimated effects of the policy on absence rates, by (local) rankings. Rankings are based on the 2012 mathematics results of schools. The baseline is the group of schools in the 8-10th deciles in 2012. Time-variant school-level controls: % Exemption, % Grade repeater, number of 8th graders in school. School and year fixed effects are included. Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Figure A2.18: Absence Rates, by Local (Within-County) Rankings



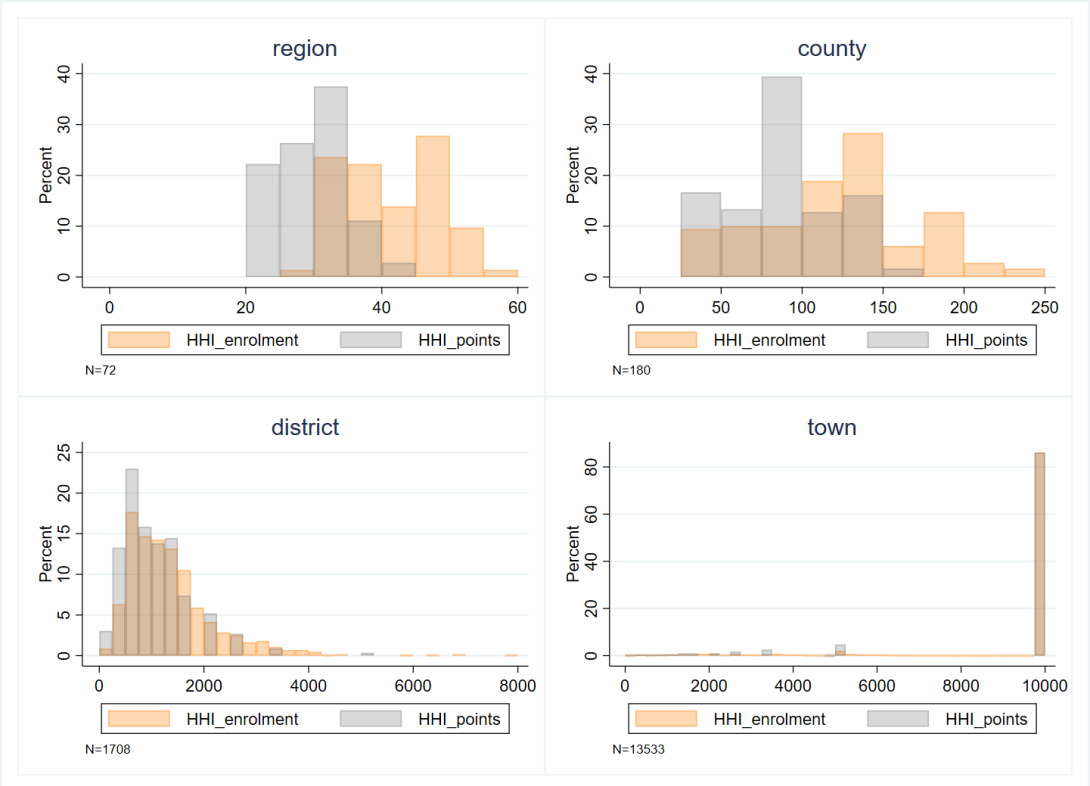
Note: The figure shows absence rates between 2009 and 2017 by local (within-county rankings). Groups of schools were created according to their ranking on the mathematics test within their county of residence in 2012.

Competition Measures

As market concentration measures, I use the Herfindahl-Hirschman Index (HHI) based on schools' share of students and share of the total test scores. I consider geographic areas (region, county, district, town) as the local markets. The upper bound of 10,000 of HHI thus would mean that one single school accounts for either all of the students in the market, or for all the points obtained in the market. This methodology follows Azar et al. (2022) who measure labour market concentration based on firms' share of vacancies, but it is also widely applied in the sports economics literature to measure the competitive balance of leagues (Depken, 1999; Rockerbie and Easton, 2022).

Figure A2.19 shows the distribution of Herfindahl-Hirschman Indices calculated based on the share of students (orange) and the share of points (grey), using different market definitions. It can be seen that as the size of the local market decreases, concentration (naturally) increases.

Figure A2.19: Market Concentration



Note: The figure shows the distribution of Herfindahl-Hirschman Indices calculated based on the share of students (orange) and the share of points (grey), using different market definitions

Table A2.15 shows the results of regressing the logarithms of concentration measures on log absences at the county and district levels. The negative coefficient estimates indicate that higher concentration (less competition) is associated with lower absence rates. However, these estimates are small in magnitude and not statistically significant when fixed effects are included and standard errors are clustered at the market level.

Table A2.15: Effect of Competition

	County-level				District-level			
	(1) log absence	(2) log absence	(3) log absence	(4) log absence	(5) log absence	(6) log absence	(7) log absence	(8) log absence
log(HHI_enrolment)	-0.165***		-0.213		-0.0380***		-0.0316	
	(0.0174)		(0.159)		(0.0140)		(0.0787)	
log(HHI_points)		-0.175***		-0.0907		-0.0211		-0.0331
		(0.0186)		(0.153)		(0.0135)		(0.0874)
'Area' FE			✓	✓			✓	✓
Observations	180	180	180	180	1,704	1,704	1,704	1,704
Number of 'area'			20	20			374	374
R-squared	0.664	0.663	0.513	0.507	0.375	0.373	0.215	0.215

Note: The table presents the estimated effect of competition on absence rates at the county and district level. Robust/Cluster-robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1.

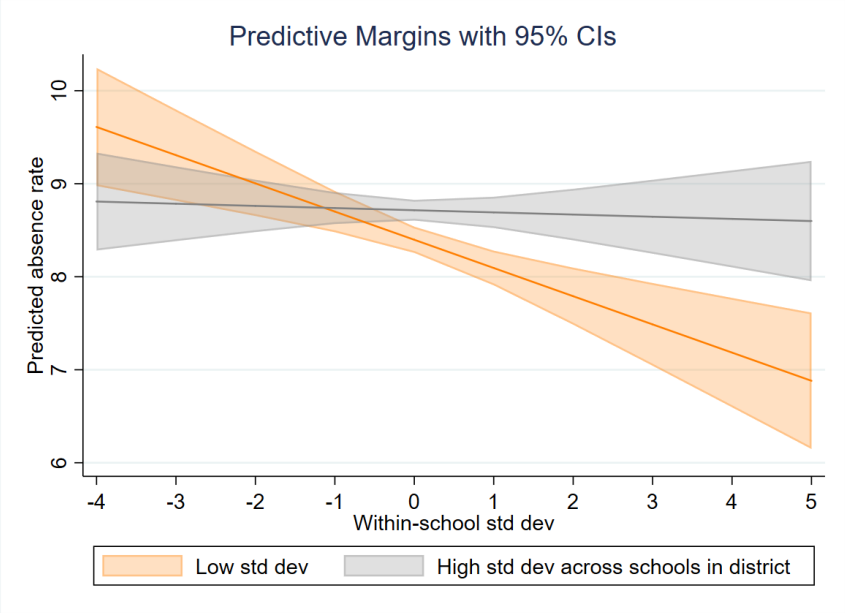
Within-school and Across-school Variation

Both within- and across-school variation matter. The potential gain from sending home one (or more) student(s) depends on how similar or dissimilar are schools on the relevant local market, and how similar are the students within the school. E.g., in the extreme case when students' performance is exactly at the same level, sending home anyone would not change the school's average. On the other hand, if there are outlier students in the school, their absence can alter the school's result substantially.

$$Absrate_{st} = \alpha_0 + \alpha_1 X_{st} + \beta_1 \sigma_{st} + \beta_2 \sigma_{dt} + \beta_3 (\sigma_{st} * \sigma_{dt}) + \phi_s + \delta_d + \tau_t + \epsilon_{st} \qquad (2.4)$$

where σ_{st} and σ_{dt} are school-level and district-level standard errors, respectively.

Figure A2.20: Role of Within- And Across-School Variation



Note: The figure shows predictive absence rates after estimating Equation (2.4). In grey, it is shown how the predicted absence rate of a school depends on the within-school standard deviation when schools within a district are not similar (above median standard deviation across schools). In orange, the same relationship is shown when schools within a district are more homogeneous (below median standard deviation across schools).

Chapter 3

Geographic and Socioeconomic Variation in Healthcare: Evidence from Migration*

3.1 Introduction

Equitable access to health and healthcare is an important policy goal for national governments and international organizations. Nevertheless, inequalities in health are large and persistent even in the most developed countries, including countries with universal health insurance (OECD, 2019). The exact causes of these inequalities vary and include inequities in access, behavioral differences, as well as differences in the utilization of health systems even if access is nominally universal.

While *socioeconomic* inequalities are an important concern, significant *geographic* variation in healthcare use has also been documented in a variety of countries and health insurance programs (Bíró and Prinz, 2020; Finkelstein et al., 2016; Godøy and Huitfeldt, 2020). Such variation may be concerning for policy makers because it may serve as evidence of health inequality, access inequality, or of inefficient program design. Therefore understanding the sources of this variation and separating the role of supply-side factors (e.g. access to physicians and hospitals, physician preferences) and demand-side factors (e.g. patient preferences, differences in the health of residents) is important.

In this paper, we examine the interaction between geographic and socioeconomic inequality in healthcare spending in the context of Hungary, an emerging economy and former socialist country in Eastern Europe with universal health insurance. By exploiting patient migration we disentangle the role of individual (demand-side) and place (supply-

*Joint work with Péter Elek, Anita Győrfi, and Dániel Prinz. A shorter version of the chapter has been published as a CERS Working Paper No. 23/18 (Institute of Economics, Centre for Economic and Regional Studies). Parts of the literature review and the appendix have been included in the published book chapter Bíró et al. (2024).

side) effects in healthcare use, and show how the importance of place effects varies by socioeconomic status (SES). To do so, we leverage high-quality administrative panel data on demographics, healthcare use, incomes, social insurance and welfare benefit take up, and other domains over the 2009-2017 period.

We start by documenting significant geographic variation in healthcare spending. Per capita healthcare spending in the highest-spending district is 1.5 times higher than in the lowest-spending district, and significant variation exists for all of its components: this ratio is 2.4 for spendings in outpatient care, 1.7 for inpatient care, and 2.0 for prescription drug spending.

This wide variation across different areas in healthcare spending may be caused by a number of different factors. Health and the need for treatment may vary, preferences of the residents of different areas could be heterogeneous, access to care may differ, or physicians may have differing practice styles across regions. Therefore after documenting significant cross-sectional variation across districts, we turn to decomposing this variation into place ("supply-side") and patient ("demand side") components. To do so, we follow a "movers" approach that allows us to estimate two-way fixed effects models in which place and patient effects can be separately identified. This approach has been used in labor economics to separate firm and worker effects to understand earnings inequality, and also in prior studies in health economics to study the role of place in geographic variation e.g. Abowd et al., 1999; Card et al., 2013; Finkelstein et al., 2016. The idea behind this approach is that patients who move between different places allow us to identify the effect of these places on healthcare spending, independent from compositional differences and demand-side factors.

We find that there is considerable heterogeneity in the role of place across levels of care. Place effects explain 66% of the variation in outpatient spending and 31% of the variation in drug spending, but almost none of the observed variation in inpatient spending. This may be because outpatient specialist visits are the most discretionary, while inpatient stays are mostly non-discretionary and related to serious illness.

There is also important heterogeneity by socioeconomic status in the role of place in outpatient spending. In our working age sample, the estimated place share is 74% for non-workers and 83% for those in the bottom quartile of the wage distribution, which is significantly higher than the place share of 34-40% for workers above the median wage (where wages are measured two years prior to the move). We find similar patterns among the elderly: place-specific factors explain 84% of the variation for low-SES pensioners but only 66% for high-SES pensioners.

Finally, to understand the mechanisms underlying our results, we examine the correlates of the estimated place effects. We find that outpatient place effects are positively associated with local outpatient care capacity, but the relationship is non-linear, pointing to capacity constraints. Importantly, we find that outpatient capacity influences the

utilization of lower-income individuals more strongly than that of higher-income ones.

The analyses provide four key insights: (1) Place-specific factors play a key role in determining residents' healthcare use. (2) Low-SES individuals are particularly dependent on place. (3) There are asymmetries in the importance of place: moving to a lower-utilization district results in a larger adjustment of utilization. (4) Capacities are strongly associated with place effects, and the relationship is particularly strong for low-SES individuals. In line with the literature, these suggest that care provision to vulnerable groups is particularly affected when capacities are insufficient. This might occur because high-SES patients have more financial resources, better access to information, and higher-quality interaction with physicians which enables them to access the scarce resources.

Our results imply that effective access differs across socioeconomic groups, even in a system of universal healthcare. Increasing capacities might enable better access to care for low-SES individuals and enable providers to spend more time and resources on all patients.

We most directly contribute to the literature using "movers" to understand the role of supply-side and demand-side factors in geographic variation in healthcare (Badinski et al., 2023; Finkelstein et al., 2016; Godøy and Huitfeldt, 2020; Johansson and Svensson, 2022; Moura et al., 2019; Salm and Wübker, 2020; Zeltzer et al., 2021). We make three contributions to this literature. First, our work highlights that the role of supply-side factors is heterogeneous across levels of care and across socioeconomic groups.[1] In particular, we show that place matters the most to low-income individuals. Second, we show that place is likely more important for low-SES groups because they are disproportionately affected by capacity constraints. Third, while the existing literature has focused on advanced economies, to the best of our knowledge we are the first to focus on a former socialist country in Eastern Europe.

More broadly, our work is related to the literature on inequalities in healthcare use. While most studies focus on demand-side reasons for inequality, including financial constraints (Allin and Hurley, 2009), less flexibility at work (Acton, 1975), and informational differences about the benefits of medical care (Cutler and Lleras-Muney, 2010; Glied and Lleras-Muney, 2008), a more recent strand of the literature studies the potential supply-side sources of inequalities in healthcare utilization (Brekke et al., 2018; Currie et al., 2022; Kristiansen and Sheng, 2022; Singh and Venkataramani, 2022; Turner et al., 2022). We make two contributions to this literature. First, we study the interaction of geographic and socioeconomic inequalities. Second, we provide evidence on the importance of capacity constraints as a potential mechanism underlying inequalities.

The remainder of the paper proceeds as follows. Section 3.2 reviews the related lit-

[1]Previous work studied heterogeneities in terms of age (Finkelstein et al., 2016), gender (Moura et al., 2019; Salm and Wübker, 2020), and education (Godøy and Huitfeldt, 2020), while we focus on socioeconomic status defined by income level.

erature. Section 3.3 provides background on the institutional framework of Hungarian healthcare. Section 3.4 describes our data and sample construction. Section 3.5 introduces our empirical framework. Section 3.6 presents our results in the following order: (i) descriptive statistics, (ii) baseline results, (iii) socioeconomic heterogeneities, and (iv) potential mechanisms. Finally, Section 3.7 concludes.

3.2 Literature

Mover Identification

We most directly contribute to the literature that has used "movers" to understand the role of supply-side and demand-side factors in geographic variation in healthcare (Badinski et al., 2023; Finkelstein et al., 2016; Godøy and Huitfeldt, 2020; Johansson and Svensson, 2022; Moura et al., 2019; Salm and Wübker, 2020; Zeltzer et al., 2021).[2] The results of the literature show that the importance of place effects varies by country, by healthcare type, and by socioeconomic status. The place share θ is the lowest with 5-10% in German ambulatory care (Salm and Wübker, 2020) and in drug spending in Sweden (Johansson and Svensson, 2022), while place accounts at most for around half of the variation among US Medicare beneficiaries, in Norway and in the ambulatory care in Israel (Finkelstein et al., 2016; Godøy and Huitfeldt, 2020; Zeltzer et al., 2021). Instead of patient moves, Molitor (2018) exploits cardiologist migration to disentangle the role of physician practice style and place effects. He finds that place effects account for 60-80% of regional variation in physician treatment choices.

We make three contributions to this literature. First, our work highlights that the role of supply-side factors is heterogeneous across both types of care and across socioeconomic groups.[3] In particular, we show that place matters the most to low-income individuals. Second, we show that place is likely more important for them because they are disproportionately affected by capacity constraints. Third, while the existing literature has focused on advanced economies, to the best of our knowledge we are the first to focus on a former socialist country in Eastern Europe.

The mover identification strategy was first used in the labor economics literature to disentangle worker and firm effects (Abowd et al., 2002; Abowd et al., 1999; Card et al., 2013). Since that it has been applied in a number of settings, such as explaining the role of place in the evolution of brand preferences (Bronnenberg et al., 2012), teacher value-added (Chetty et al., 2014), neighbourhood effects and intergenerational mobility Chetty and Hendren (2018), the causes of nutritional inequality (Allcott et al., 2020), voter behaviour (Cantoni and Pons, 2022), or integration of international immigrants (Bailey

[2]For a more detailed overview, see Bíró et al. (2024).

[3]Previous work studied heterogeneities in terms of age (Finkelstein et al., 2016), gender (Moura et al., 2019; Salm and Wübker, 2020), and education (Godøy and Huitfeldt, 2020), while we focus on socioeconomic status defined by income level.

et al., 2022).[4] Some of these studies also investigate potential mechanisms by looking at heterogeneities across age, gender, race or income groups (Cantoni and Pons, 2022; Chetty and Hendren, 2018). Methodological advances in the related literature include identifying the assumptions for causal interpretation (Hull, 2018), or corrections for limited mobility bias (Andrews et al., 2012; Bonhomme et al., 2023, 2019; Borovičková and Shimer, 2017; Kline et al., 2020).[5]

Inequalities in Healthcare Use

More broadly, our work is related to the literature on inequalities in healthcare use. While most studies focus on demand-side reasons for inequality, including financial constraints (Allin and Hurley, 2009), less flexibility at work (Acton, 1975), and informational differences about the benefits of medical care (Cutler and Lleras-Muney, 2010; Glied and Lleras-Muney, 2008), a more recent strand of the literature studies the potential supply-side sources of inequalities in healthcare utilization (Brekke et al., 2018; Chen and Lakdawalla, 2019; Currie et al., 2022; Kristiansen and Sheng, 2022; Martin et al., 2020; Turner et al., 2022). In the Hungarian context, the recent survey evidence by Lucevic et al. (2019), and the descriptive work by Bíró and Prinz (2020) document large socioeconomic differences in unmet healthcare needs, healthcare spending and mortality.

We make two contributions to this literature. First, we study the interaction of geographic and socioeconomic inequalities. Second, we provide evidence on the importance of capacity constraints as a potential mechanism underlying inequalities.

While our second step to further decompose the supply side is relying only on correlational evidence, it is in line with potential mechanisms studied in the related literature. Brekke et al. (2018) find that patients with lower education levels have shorter consultations in primary care, and low-income patients receive fewer medical tests as well. Their suggested mechanism is that consultation quality depends on the patient's communication and cognitive skills.[6] Kristiansen and Sheng (2022) show the importance of the

[4]Other examples using mover identification within the health economics literature can be grouped in smaller sub-groups. First, a sub-strand investigates the role of place in mortality (Deryugina and Molitor, 2020, 2021; Finkelstein et al., 2021). The challenge of this literature comes from the fact that mortality outcome is observed only once for each individual, i.e., no panel structure can be exploited as for healthcare use. Thus, selection on unobservables remains a potential confounder. They aim to correct for the potential correlation between unobserved health capital and destination choice using variation in the observables. Second, another sub-strand exploits doctor switches to disentangle the role of provider practice style and patient characteristics (instead of place and patient effects). They either rely on switches because of patient mobility (Ahammer and Schober, 2020; Koulayev et al., 2017; Skipper and Vejlin, 2015), or patients involuntarily switching providers after practice closures because of physicians' death or move (Fadlon and Van Parys, 2020; Ginja et al., 2022; Kwok, 2019; Simonsen et al., 2021). Identification – as above – relies on sufficient amount of switches.

[5]In case of the mover-based designs the bias in fixed effect estimates attributed to the incidental parameter problem is referred to as limited mobility bias. For more, see Appendix Section A3.1.

[6]The theoretical literature/ex-post rationalizations also tend to rely on assumptions about high-SES patients' better communication skills which results in better signalling of their illness severity (Brekke et al., 2018; Kaarboe and Siciliani, 2023), while Chen and Lakdawalla (2019) assumes doctors care about the utility of the patients, and therefore their income and SES as well, and show that an increase in

physician-patient match by providing evidence of low-SES patients receiving more and better care when treated by physicians coming from similar families. Turner et al. (2022) focus on emergency visits and document longer waiting times, less complex care, lower hospital admission rates, and higher mortality for more deprived patients. They argue that besides the quality of patient-physician interaction, physicians' unconscious bias also plays a role. In a field experiment Angerer et al. (2019) show that discrimination based on socioeconomic status (education degree) is present among Austrian healthcare providers.

As an additional mechanism, our results suggest that capacity constraints disproportionately affect low-SES patients. We argue that the observed asymmetries by the direction of the move are due to the limited patient choice in case of capacity constraints, which particularly affect low-SES people. Patients moving to a lower utilization district cannot keep their initial utilization level, thus, they adjust more to the destination's average healthcare use.

In the literature, there is mixed evidence on the role of capacities. Using the same empirical strategy as ours, Molitor (2018) finds that doctors moving to a less-intensive region adjust less, and concludes that capacity constraints do not play a role. Exploiting an exogenous change in Neonatal Intensive Care Unit (NICU) capacity, Freedman (2016) finds no effect on the sickest infants, but increased use for those who are at the margin of needing intensive care. In the Hungarian context, Elek et al. (2015) showed in a quasi-experimental setup that extending outpatient capacities increases outpatient care utilization.

Regarding the relationship between capacities and inequalities, Singh and Venkataramani (2022) argue that biases (specifically, racial disparities) in provider behaviour may arise when hospitals operate at capacity. Their suggested channels are limited provider bandwidth or reliance on biased algorithms. Relatedly, Bosque-Mercader et al. (2023) find evidence that socioeconomic inequalities arise within hospitals: low-income patients experience longer waiting times and a higher probability of surgery cancellations. They also find that high-income patients are more likely to exit waiting lists voluntarily. They might opt for private care or another public hospital because they can communicate their needs more effectively, and as a consequence, doctors might be able to offer them treatment in other institutions. Kaarboe and Carlsen (2014) however does not find evidence of discrimination in waiting times in Norway. Using Danish data Simonsen et al. (2020) also find inequalities in waiting times only for a limited number of procedures, however, in these cases they show that geographic and institutional differences across hospitals play a crucial role in explaining them.

physician reimbursement results in a larger utilization increase for high-SES patients than for low-SES patients.

3.3 Background

Hungary, a European Union member state with a population of about 9.8 million inhabitants, has a single-payer healthcare system, where services are administered by the National Health Insurance Fund Administration (NHIFA). Primary, specialist outpatient, and inpatient care are all free of charge at the point of use. (However, informal payments[7] were common in the public system and private healthcare has become more important in the study period, especially in outpatient care.) Outpatient care is reimbursed by the NHIFA based on procedure codes associated with visits. Inpatient reimbursements are based on diagnosis-related groups (DRGs). Primary care is financed on a capitation basis. Prescription drugs are subsidized, where subsidy rates range from 25% to 100% and are slightly less than 50% on average.

The country is divided into 197 districts, corresponding to the local administrative unit (LAU) level 1 classification of Eurostat. The average population of districts is approximately 50,000. They are generally composed of a seat town with nearby smaller towns and villages. The capital city of Budapest, with a population of 1.75 million, consists of 23 districts. Specialist outpatient services are available in the vast majority of district seats, and hospitals operate in roughly half of them. The twenty counties (including Budapest) represent the next administrative level, where county seats provide higher-level inpatient services. On the primary care level, there are around 6,600 general practices in the country.[8]

Previous research on Hungary has also documented large geographic and socioeconomic inequalities both in healthcare use and in mortality (Bíró and Prinz, 2020; Lucevic et al., 2019; Orosz, 1990; Szende and Culyer, 2006; Van Doorslaer et al., 2006). It is often argued that the root of inequalities in post-socialist countries is the institutional system that evolved during the socialist times, with a particular example of the above-mentioned gratuity payments that were (and are) common in most Central and Eastern European countries. Van Doorslaer et al. (2006)'s cross-country study finds pro-rich inequality in outpatient specialist visits in OECD-countries, including Hungary.[9] Szende and Culyer (2006) and Lucevic et al. (2019) conduct representative survey studies to assess the burden of informal payments and unmet healthcare needs, respectively, and find large disparities across socioeconomic groups. Using administrative data of full-time workers Bíró and Prinz (2020) provide descriptive evidence of geographic heterogeneity, positive association between labor income a healthcare spending (with further heterogeneities across regions),

[7]Informal payments are unofficial payments to healthcare providers and include gratuities or tips for the staff, bribes, or even payments that are demanded by the staff. Gaal et al. (2006) estimated that the magnitude of informal payments corresponded to 1.5-4.6% of total (official) healthcare costs in Hungary in 2001.

[8]For more details on the healthcare system see Gaál et al. (2011).

[9]The same study (Van Doorslaer et al., 2006), however, does not find inequalities in the number of GP visits in Hungary.

and negative association between labor income and mortality. All studies called for reforms improving access to public healthcare services and specifically targeting low-income groups.

3.4 Data and Sample

3.4.1 Data Sources and Variables

We use an individual-level administrative panel data set that covers monthly healthcare, labor market and demographic information for the years 2009–2017 on a random 50% sample of the 2003 population of Hungary. The healthcare data contains variables that capture the frequency of use, including the number of outpatient visits, inpatient days, and prescriptions. It also contains information on expenditures measured by total reimbursement amounts (and out-of-pocket payments for prescriptions) by type of care. Importantly for our analysis, reimbursement rates do not vary by district or provider for outpatient care, inpatient care, or prescriptions. We do not specifically examine primary care, financed on a capitation basis, due to lack of detailed data.

We break outpatient care use down into six categories by specialty of care, specifically examining internal care, surgery and trauma, gynaecology, rheumatology, cardiology, and laboratory diagnostics. Similarly, we divide pharmaceutical use by category based on the Anatomical Therapeutic Chemical (ATC) classification, and focus in our analyses on antidiabetics, antihypertensives, psycholeptics, psychoanaleptics, antiinfectives, and drugs for obstructive airway disease.

The labor market segment of the dataset contains monthly employment status, occupation classification using the International Standard Classification of Occupations (ISCO), labor market earnings, and information on unemployment, disability, and pension benefits. Finally, the demographic variables include gender, age, and most importantly the district of residence.

We also use district-level indicators on healthcare supply, geography, and broad socioeconomic status from various other databases, including the Pulvita system of the National Directorate General for Hospitals (OKFŐ), data on general practices from the NHIFA, as well as municipal statistics included in the Settlement Statistics Database System (T-STAR) and additional data from the Central Statistical Office. We measure outpatient care supply with per capita outpatient capacity (weekly number of specialist outpatient hours) and inpatient care supply with the per capita number of hospital beds in the district. Distance from the district seat to the county seat captures access to higher-level healthcare and other services as well as employment opportunities. We also use the per capita taxable income of the district to control for broad socioeconomic status.

3.4.2 Movers

We categorize a person as a mover if her district of residence changed exactly once in the period between 2010-2016. There are two types of addresses in Hungary: a permanent address is defined for every citizen at every time, while a small fraction of the population also has a temporary address. We can observe both permanent and temporary addresses in the data. We define movers based on the change of their permanent residence. To improve precision, we also check whether a mover acquired a new temporary residence that coincides with the destination district up to six months before the change of her permanent residence. For movers who moved to their new permanent address after such a change in the temporary address, we shift the time of the move accordingly. This modification applies to around 15% of movers.

Because we want to study moves that plausibly affect the context of healthcare use, such as hospitals and providers accessed, we exclude moves that are within the commuting zones of larger cities. In particular, to exclude mobility within the agglomeration of Budapest and of county seats, we do not examine within-county moves and moves between Budapest and the surrounding Pest County.

To further improve the precision of our identification of moves, we exclude cases for which an individual changes her residence but does not appear to be getting her prescriptions in the area to which she moved. We do so based on a variable which provides information on the county where an individual filled most of her prescriptions in a given quarter (missing if no prescription was filled). In particular, we exclude cases where the modal county of prescriptions coincides with the destination county in fewer than 50% of post-move quarters (defined based on change of residence).

Finally, we restrict the sample to those who were aged between 30 and 80 at the time of the move. The reason for this restriction is that the study-related temporary moves of younger people are less reliably observed in the data, while people aged above 80 years are more likely to live in nursing homes.[10]

Throughout the paper we use data annualized by the time of the move (and not by calendar year). Relative year zero is defined as the first four quarters when the person entirely lives in the destination district according to her place of residence. To verify that individuals whom we categorize as movers actually move, Appendix Figure A3.1 shows separately the annual share of individuals for whom the county where they claimed most of their prescriptions is their origin county and the same share for destination counties. The figure suggests that although there is some discrepancy between the two location indicators, the shares change by about 60% from relative year -1 to 0.

Table 3.1 indicates that movers are younger, use less healthcare and are more likely

[10] According to Monostori and Gresits (2019), less than 3% of the 75-79 age group lived in a nursing home in 2016, while this share increased to 5% and 9%, respectively, for the 80-84 and 85-89 age groups.

Table 3.1: Summary Statistics

	(1) Non-movers	(2) Movers
Male	0.47	0.49
Age	47.7	42.8
Outpatient visits	7.3	6.3
Outpatient spending (HUF)	12,221	10,630
Inpatient days	2.1	2.0
Inpatient spending (HUF)	38,366	28,985
Drug prescriptions	18.1	12.6
Drug spending (HUF)	53,670	36,749
Total spending (HUF)	104,254	76,364
Working	0.45	0.49
White collar job	0.20	0.27
Blue collar job	0.25	0.22
Unemployment benefit	0.02	0.02
Pensioner	0.20	0.12
Number of individuals	3,606,622	62,301

Note: The table shows summary statistics of non-movers and movers. Annual healthcare use measures and labor market participation measures are calculated in 2009, the first year of our data.

to be employed than non-movers (all differences are statistically significant at the 1%-level). Appendix Figure A3.2 displays the evolution of the rate of employment, old-age pensioners, disability pensioners, and unemployment benefit recipients among movers. There is a slight drop in employment (and a corresponding slight, less than 2 percentage points, increase in unemployment) around the time of the move but no such change is seen for old-age and disability pensions.

3.5 Empirical Framework

3.5.1 Fixed Effects Model

To separately identify the individual- and place-specific components of healthcare use, our empirical strategy exploits migration across geographic areas. Following Abowd et al. (1999) and Finkelstein et al. (2016), the baseline model for the healthcare use of individual i at place j and time t takes the following fixed effects specification:

$$y_{ijt} = \alpha_i + \gamma_j + \tau_t + x_{it}\beta + \varepsilon_{it}, \tag{3.1}$$

where α_i is individual (patient) fixed effect, γ_j is place fixed effect and τ_t is time fixed effect. x_{it} is a vector of individual-level time-dependent observable characteristics.

As outcome variables (y_{ijt}) we use a wide range of healthcare use measures. The main results in Section 3.6.2 are presented for frequency and spending measures of outpatient care, inpatient care and prescription drug use. In the Appendix we also present results on subcategories of outpatient spending by specialties (e.g., cardiology, rheumatology, etc.), and on therapeutic classes of prescription drugs (e.g., antidiabetics, antiinfectives, etc.). x_{it} always includes the interaction of gender and five-year age groups, while in some specifications we also control for the (potentially endogenous) labor force status of the individual.

Based on distributional considerations, the literature generally uses log(spending) or, to account for zeros, log(1 + spending) as outcome variables in a (log-)linear setting. Our dependent variables are mainly count data (number of visits or days) or non-negative continuous data with a substantial amount of zeros (healthcare spending), so it is more natural to specify the conditional expectation in a Poisson model:

$$E\left(y_{ijt}\right) = \exp\left(\alpha_i + \gamma_j + \tau_t + x_{it}\beta\right) \tag{3.2}$$

where, for simplicity, the conditions are omitted.

In fact, such a Poisson specification has advantages over the more usual log-linear OLS specification for modeling non-negative continuous data with possibly many zeros see e.g. Correia et al., 2020; Silva and Tenreyro, 2006.[11]

Identification of α_i and γ_j in the above models hinges on the presence of movers. Without movers, average patient characteristics would be inseparable from place fixed effects. Moreover, a sufficient number of moves is needed, i.e., the number of moves between any place-pairs (j' and j'') should increase large as the sample size goes to infinity.

The causal interpretation of the fixed effects results rests on two main assumptions that we discuss in the following paragraphs: exogenous mobility and additive separability.

First, the exogenous mobility assumption states that the residual in Equation (3.1) is unrelated to mobility. This would be the case if individuals were randomly allocated or relocated to different regions, conditional on observables and time-invariant unobservable factors. While the model allows for movers and non-movers to differ not only in the level of healthcare use, but also in the trends around the move, it does not allow for trends to systematically vary with the origin and destination place of movers. For example, if a negative health shock resulted in individuals systematically moving to places with higher average healthcare utilization than their origin place, the estimated place effect would be contaminated by the effect of the health shock.

Godøy and Huitfeldt (2020) – following Card et al. (2013) – identify three potential forms of endogenous mobility: sorting on match effects, drift, and correlated fluctuations

[11] See Badinski et al. (2023) for an application of the Poisson model in the mover-based setting.

in the transitory error. (i) The match effect captures the bias caused by individuals sorting into districts based on the presence of specialized care. If sorting on match effects is dominant, we would expect an increase in healthcare use around after all moves. Even when individuals move to a district with a lower average utilization level than their origin district, their utilization would increase. (ii) Drift describes the trend in an individual's healthcare use over time. If e.g., individuals with gradually deteriorating health are more likely to move to high-utilization districts, the place share will be overestimated. This is because the drift component of the error term would be positively correlated with the place effect and the outcome variable as well. (iii) The transitory error captures shocks or any other fluctuations which might be also correlated with systematic moves. Again, if e.g., individuals experiencing a simultaneous health shock systematically choose higher utilization districts, the place effects will be overestimated. Since an event study representation enables us to assess the validity of this assumption, we discuss it in more detail in Section 3.6.3.

Second, we also assume the fixed effects α_i and γ_j to be additively separable. The outcome variable is typically defined as the logarithm of healthcare spending, thus, the assumed functional form is log additive. Additive separability of the logarithms implies a multiplicative relationship of patient and place effects for the levels of healthcare utilization. This corresponds to the intuition that utilization will vary more for patients who tend to use more healthcare (have a high α_i) because of sickness or their preference, in comparison to patients who rarely use healthcare (have a low α_i).[12]

Turning to the decomposition of inequalities to place and patient shares, the place share is defined as the share of the difference in average usage between places j'' and j' that is attributed to place effects:

$$S_{place}(j'', j') = \frac{\gamma_{j''} - \gamma_{j'}}{\bar{y}_{j''} - \bar{y}_{j'}} \tag{3.3}$$

where $\bar{y}_j$ denotes the average healthcare utilization in place j.

The patient share is the share of the overall difference attributed to patients:

$$S_{patient}(j'', j') = \frac{\bar{\hat{c}}_{j''} - \bar{\hat{c}}_{j'}}{\bar{y}_{j''} - \bar{y}_{j'}} \tag{3.4}$$

where $\bar{\hat{c}}_j$ denotes the average patient compound effect in place j including patient fixed effects α_i, time-varying patient characteristics x_{it} and time effects λ_t.

In practice, the additive decomposition is performed by dividing the sample of ge-

[12] To relax the assumption of additive separability, the literature also follows Card et al. (2013) and estimates a fully saturated model by adding interaction fixed effects for each patient-place pair. Finkelstein et al. (2016) and Godøy and Huitfeldt (2020) both find that the adjusted R^2 is only slightly higher in the saturated model than in the initial model which implies that match effects do not play an important role.

ographic units into a low- and a high-utilization group and (after estimating Equation (3.2)) computing the sample analogues of the place and patient shares $S_{place}(High, Low)$ and $S_{patient}(High, Low)$. High- and low-utilization groups might be defined as the regions with utilization above or below the median, in the top quartile/decile or bottom quartile/decile. Note that as the size of the groups shrinks, precision is decreasing.

3.5.2 Difference-in Differences and Event Study Representation

The above specification can be transformed for movers as follows:

$$E\left(y_{it}\right) = \exp\left(\alpha_i + \gamma_{o(i)} + \tau_t + \mathbb{I}_{\{t \geq t_i^0\}} \times \left(\gamma_{d(i)} - \gamma_{o(i)}\right) + x_{it}\beta\right) =$$

$$\exp\left(\underbrace{\alpha_i + \gamma_{o(i)}}_{\alpha_i'} + \tau_t + \mathbb{I}_{\{t \geq t_i^0\}} \underbrace{\left(\frac{\gamma_{d(i)} - \gamma_{o(i)}}{\log \bar{y}_{d(i)} - \log \bar{y}_{o(i)}}\right)}_{\theta} \underbrace{\left(\log \bar{y}_{d(i)} - \log \bar{y}_{o(i)}\right)}_{\Delta_i} + x_{it}\beta\right) =$$

$$\exp\left(\alpha_i' + \tau_t + \mathbb{I}_{\{t \geq t_i^0\}} \times \theta \times \Delta_i + x_{it}\beta\right) \quad (3.5)$$

where $o(i)$ is the origin and $d(i)$ is the destination district, $\alpha_i' = \alpha_i + \gamma_{o(i)}$ is an individual fixed effect, t_i^0 denotes the time of the move, and the indicator function $\mathbb{I}$ takes value one after the move. The variable $\Delta_i = \log \bar{Y}_{d(i)} - \log \bar{Y}_{o(i)}$ is the difference between the log of the average healthcare utilization in the destination and the origin districts. The parameter of interest, θ shows the average change in healthcare utilization after moves as a share of the difference between the average utilization in the destination and the origin districts. For non-movers, $\mathbb{I}$ is zero for all time periods and the equation becomes $\exp\left(\alpha_i' + \tau_t + x_{it}\beta\right)$.

Parameter θ can be interpreted as the place share that measures the fraction of geographic differences explained by differences in place characteristics. If only place effects matter, individuals will adjust their healthcare use entirely to the destination area's average utilization, and $\theta = 1$. On the other extreme, if only patient characteristics matter, the move will not result in a change in utilization, hence $\theta = 0$.

Equation (3.5) is an individual-level fixed-effects model. As long as the conditional expectation is well-specified, the model can be consistently estimated with the Poisson fixed effects (FE) estimator, free from the incidental parameter problem see e.g. Wooldridge, 2010. Also, equation (3.5) corresponds to difference-in-differences with continuous treatment, and nonlinearities in the θ parameter can be examined by presenting different coefficients for various treatment intensities, e.g. for positive and negative moves (Callaway et al., 2021).[13]

[13]See Callaway et al. (2021) for assumptions with continuous treatment, and Hull (2018) for the limits of causal interpretation in multiple-treatment models. (proposing an estimator to identify mover average treatment effect (MATE)). Recent advances of the difference-in-differences literature could be also incorporated. Some recent studies show that their results are robust to allowing for treatment effect

The fixed effects model (3.5) can be rewritten in an event study framework, where θ_k is estimated separately for each time period (k is the year relative to the move):

$$E\left(Y_{it}\right) = \exp\left(\alpha_i' + \tau_t + \sum_{k=-5}^{k=4} \theta_k \times \mathbb{I}_{\{k=t-t_i^0\}} \times \Delta_i + x_{it}\beta\right). \tag{3.6}$$

The regressions are estimated on the sample when the year relative to the move is between -5 and +4, where θ_{-1} is the reference year parameter and thus set to zero.

Godøy and Huitfeldt (2020) show that if the identifying assumptions of the two-way fixed effects model hold, the estimates of θ_k in equation (3.6) reflect the true place share, and the relationship can be written as:

$$\theta_k = \begin{cases} 0 & k \leq 0 \\ \frac{\gamma_{d(i)} - \gamma_{o(i)}}{\log \bar{y}_{d(i)} - \log \bar{y}_{o(i)}} & k \geq 0 \end{cases}, \tag{3.7}$$

i.e., coefficients corresponding to pre-move years are 0 (no anticipatory adjustment), while in post-move years they are equal to the place share (the magnitude of the adjustment). In the year of the move ($k = 0$), the coefficient should be in this interval.

The event study specification in equation (3.6) also allows us to provide graphical evidence on the identifying assumptions discussed above. We will come back to this in Section 3.6.3, after presenting our main results.

3.5.3 District-Level Correlates of Healthcare Use

After estimating the place-specific component of healthcare utilization, denoted by γ_j in equation (3.2), we turn to analyzing what underlying factors may explain this component. Potential explanatory variables include healthcare supply variables such as the availability of outpatient and inpatient units, and the quality of equipment as well as the specialization and the beliefs (about effective treatments) of the physicians working in these facilities. Also, non-healthcare-specific factors such as long-term local economic, social and geographic conditions may play a role. Importantly, while the approach introduced above allows us to use movers to identify the place-specific component of healthcare separately from individual-specific factors, our analysis of the correlates of this place-specific component uncovers associations, rather than causal effects.

The approach generally followed by the literature e.g. Finkelstein et al., 2016 investigates these relationships via a two-step approach by correlating the estimated place effects with the place-level observables. Here, in a similar (and in special cases identical) one-step

heterogeneity: Johansson and Svensson (2022) applies alternative estimators developed by Chaisemartin and D'haultfoeuille (2020) and de Chaisemartin et al. (2022) for settings with continuous treatment, while Badinski et al. (2023) estimates the event study with the estimators of Callaway and Sant'Anna (2021) and Sun and Abraham (2021).

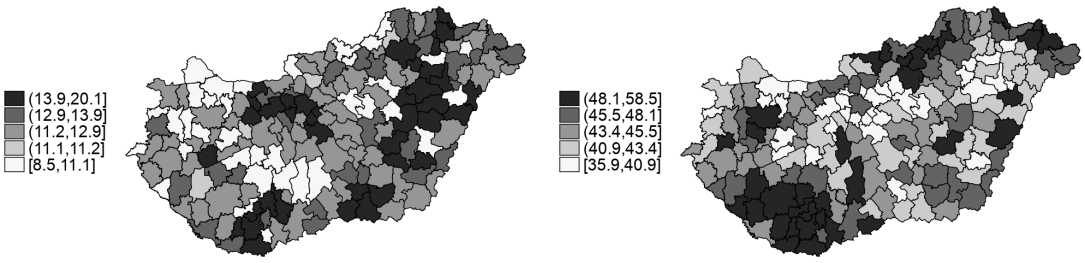
approach, we directly use the movers to estimate panel models of healthcare utilization with individual fixed effects and place-level explanatory variables:

$$E\left(y_{it}\right) = \exp(\alpha_i' + \tau_t + \sum_{k=-5}^{k=4} \mathbb{I}_{\{k=t-t_i^0\}} \times \delta_k + z_{j(i,t),t} \times \eta + x_{it}\beta) \tag{3.8}$$

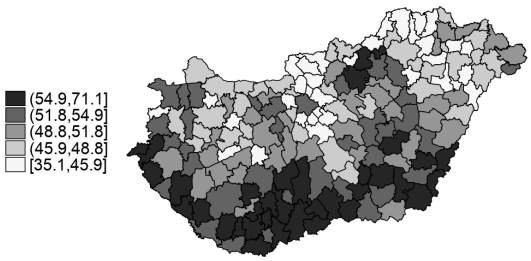
where z_{jt} denotes the observed (and possibly time-varying) place characteristics of district j (number of outpatient hours and hospital beds, distance from county seat, dummy for county seat, and per capita taxable income), and we control for individual, calendar time and event time fixed effects and gender - age group interactions. As Agha et al. (2019) point out, individual fixed effects filter out time-invariant patient characteristics similarly to as in equation (3.5).[14]

Figure 3.1: Geographic Variation in Healthcare Spending

(a) Outpatient Spending (b) Inpatient Spending



(c) Drug Spending



Note: Figure shows average outpatient, inpatient, and prescription drug spending by district in thousand HUF. The 197 districts are divided into quintiles by type of spending. The lower and upper limits of each quintile are displayed in the legend. The sample includes all movers and non-movers ($N = 3,662,646$ individuals).

[14]See also Zeltzer et al. (2021) for a similar solution.

3.6 Results

3.6.1 Descriptive Analysis

Figure 3.1 and Appendix Table A3.1 show the district-level variation of per capita outpatient, inpatient and drug spending and utilization. Each of the three types of healthcare utilization shows significant variation across areas, although the geographic patterns are different. As column (5) of Appendix Table A3.1 shows, total healthcare spending in the highest-spending district is 1.5 times higher than in the lowest-spending district, and significant variation exists for all of its components: this ratio is 2.4 for outpatient spending, 1.7 for inpatient spending, and 2.0 for drug spending. As reimbursement rates are set at a national level, spending variation is driven by the quantity and composition of utilization, rather than geographic variation in reimbursement rates. Utilization varies significantly as well, with the highest-utilization district recording 2.0 times more outpatient visits, 2.6 times more inpatient days, and 1.7 times more prescriptions than the lowest. Total spending is higher by 20%, outpatient spending by 59%, inpatient spending by 25%, and drug spending by 26% in the top quartile of districts than in the bottom quartile (column 6) on average.

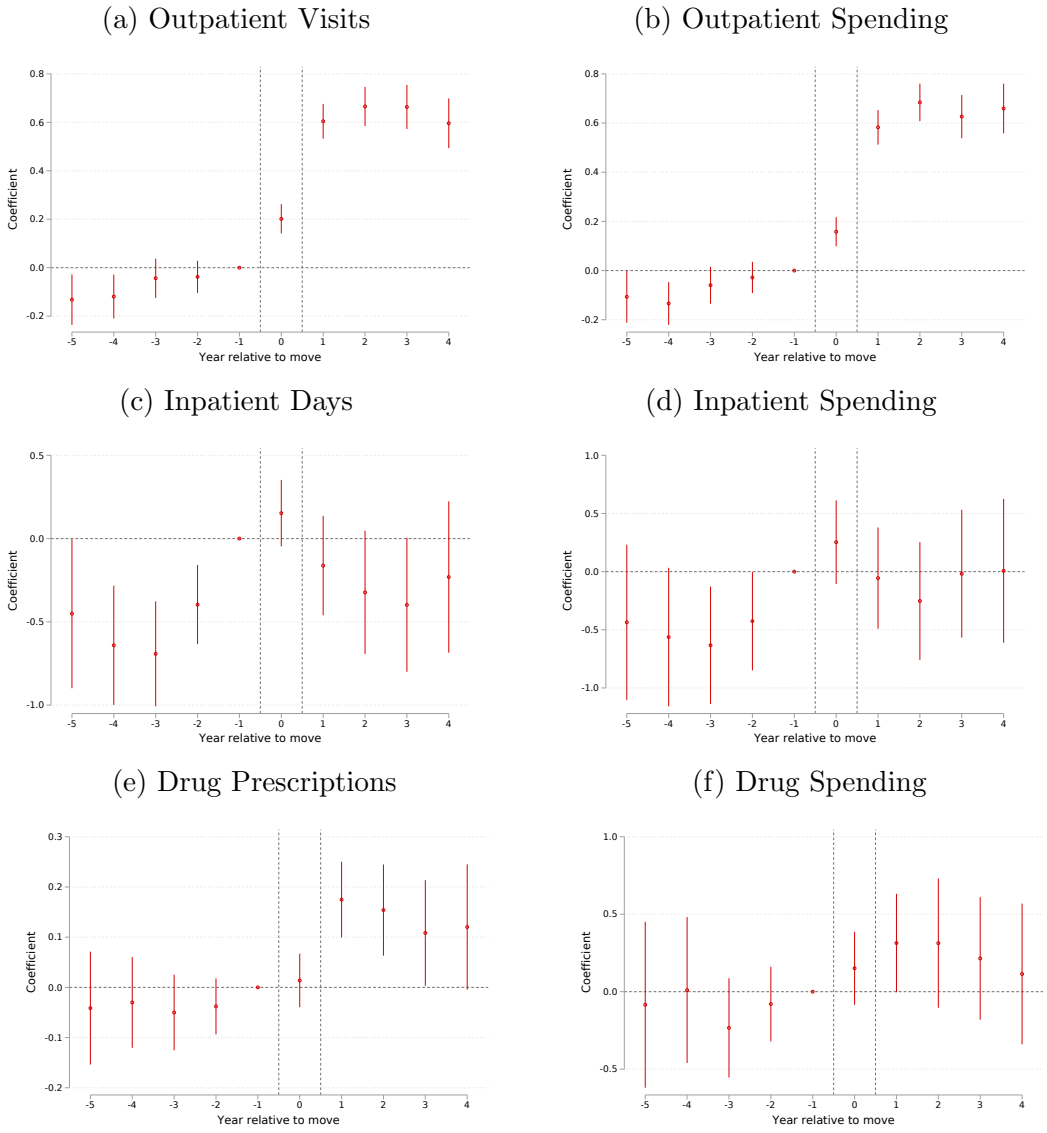
3.6.2 The Role of Place at Different Levels of Care

Our main results are presented in Figure 3.2, Table 3.2, and Table 3.3. Figure 3.2 shows our event study estimates of place effects from estimating equation (3.6). In line with our identifying assumptions, it shows little evidence of pre-trends in any of the measures of utilization before the move. Then after the move, measures of utilization adjust towards the level of utilization in the destination district.

Panels (a) and (b) of Figure 3.2 suggest that outpatient utilization (visits and spending) adjust by more than 60% of the gap between the average utilization in the origin and destination districts, suggesting that the place component of outpatient utilization is explains more than three-fifths of the variation. Based on estimating equation (3.5), column (1) of Table 3.2 shows that pooling over the entire post-move period, the place component of spending is 66% on average. Columns (1) and (2) of Table 3.3 show that place effects account for 57-59% of the difference in outpatient utilization between above- and below-median districts and also between the top and bottom quartiles of districts. The remaining 41-42% of the difference is accounted for by demand-side factors.

Turning to inpatient utilization, panels (c) and (d) of Figure 3.2 suggest that place effects are negligible. There are several reasons why place-specific factors may matter for outpatient but not for inpatient use. Inpatient care is associated with more serious illness. This means that it is likely to be less discretionary or dependent on physician practice

Figure 3.2: Event Study



Note: The figure shows event study estimates of place effects for outpatient, inpatient, and prescription drug utilization. These are the coefficients θ_k from estimating equation (3.6). The bars show 95% confidence intervals. Controls include calendar year fixed effects and gender – age group interactions. The sample includes all movers ($N = 266,290$ individual-years).

styles and more dependent on individual health status. It is also less likely to be subject to capacity or access constraints than outpatient care.

Finally, panels (e) and (f) of Figure 3.2 show our event study estimates of prescription drugs. Panel (e) and column (3) of Table 3.2 suggest that place effects explain approximately 18% of the frequency of utilization, while panel (f) and column (3) of Table 3.2 show a larger, 31% place share for spending. Table 3.3 shows similar estimates for spending from our additive decomposition. The difference in the share of variation explained for quantity and spending is consistent with places influencing the types and consequently the cost of drugs prescribed on top of the quantity prescribed.

Table 3.2: Difference-in-Differences: Average Place Effects

	(1) Outpatient care	(2) Inpatient care	(3) Pharmaceuticals
Frequency	0.659***	0.0136	0.183***
	(0.0316)	(0.148)	(0.0397)
Spending	0.659***	0.252	0.305*
	(0.0298)	(0.191)	(0.170)
Observations	266,290	128,271	257,731

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

Note: The table shows pooled difference-in-differences estimates of place effects for outpatient, inpatient, and prescription drug utilization. These are the coefficients θ from estimating equation (3.5). Controls include calendar year fixed effects and gender – age group interactions. For each utilization type, the first row shows a measure of frequency and the second row shows spending. Frequency measures are outpatient visits, inpatient days, and number of prescriptions. Standard errors are clustered at the individual level. Number of observations are individual-years.

Going beyond the broad categories of outpatient, inpatient, and pharmaceutical utilization, we also examine our results for subcategories of outpatient and prescription drug spending in Appendix Table A3.2, Appendix Figures A3.5, and A3.6. The top panel of Appendix Table A3.2 and Appendix Figure A3.5 suggest that place effects are relatively similar across specialties. The lowest point estimate (58%) is obtained for cardiology, while place effects are the largest for internal care and lab diagnostics (68-69%). This is consistent with lab diagnostics being somewhat more discretionary and subject to wider variation in practice patterns.

The bottom panel of Appendix Table A3.2 and Appendix Figure A3.6 reveal much more significant differences in place effects across drug classes. It appears that place effects are substantial for antiinfectives, which include antibiotics, but are small for other classes. This is consistent with the use of antibiotics often being discretionary and highly dependent on place-specific supply-side factors as documented in other countries as well. At the same time, the use of drugs like antidiabetics and antihypertensives is less likely to respond to place-specific factors in the short term.

We now turn to examining the robustness of our main results to several alternative specifications. We re-estimate our main results from equation (3.5). (1) First, we include non-movers in the estimation sample to increase precision of the coefficient estimates of the control variables. (2) Second, since in Figure A3.2 we observed changes in labor market and income status around the moves, to account for potential endogeneity we include controls for labor market status and income in the estimating equation, (3) Third, we control for differences in the age- and gender-composition of the districts in calculating $\Delta_{i,}$. In this case, the demand side only accounts for other patient characteristics and

Table 3.3: Additive Decomposition

	(1)	(2)	(3)	(4)	(5)	(6)
	Outpatient Spending		Inpatient Spending		Drug Spending	
	Top vs. bottom		Top vs. bottom		Top vs. bottom	
	50%	25%	50%	25%	50%	25%
Mover sample						
Difference in log average utilization	0.28	0.45	0.28	0.46	0.32	0.53
Place component	0.17	0.26	0.10	0.24	0.13	0.16
Patient component	0.12	0.19	0.18	0.22	0.20	0.38
Place share	0.59	0.57	0.35	0.52	0.39	0.30
Patient share	0.41	0.42	0.65	0.48	0.61	0.70
Full sample						
Difference in log average utilization	0.28	0.45	0.14	0.23	0.15	0.24
Place component	0.14	0.23	0.01	0.05	0.05	0.06
Patient component	0.15	0.22	0.13	0.18	0.10	0.18
Place share	0.49	0.50	0.08	0.20	0.35	0.25
Patient share	0.51	0.50	0.92	0.80	0.65	0.75

Note: The table shows additive decomposition estimates of place effects for outpatient, inpatient, and prescription drug utilization. These are based on the coefficients γ_j from estimating equation (3.2). Controls include calendar year fixed effects and gender – age group interactions. For each utilization type, the first column shows the difference between above- and below-median districts and the second column shows the difference between the bottom and top quartiles of districts. The top panel is based on the mover sample, while the bottom panel is based on the full sample

preferences, and consequently, place share is expected to be slightly higher than in the baseline case. (4) Fourth, we consider larger geographical units, and calculate Δ_i as the difference between the log usage of the destination and origin county instead of districts. (5) Fifth, we consider the logarithm of healthcare use as outcome variables using log(y+1) and log(y+0.01) where y is the measure of healthcare use. Finally, we compute standard errors with two-way clustering by individual and place. The first row (I) of Appendix Table A3.3 repeats our baseline results from Table 3.2. The other rows (II-VIII) show results from the alternative specifications. These are very similar to the baseline results.

The results so far offer two key takeaways. First, place-specific factors matter for healthcare use. Using moves across districts we account for individual-specific or demand-side factors, and highlight the importance of place-specific or supply-side factors. In other words, the different composition of individuals living in different areas cannot explain all of the substantial geographic variation in healthcare utilization. Second, the extent to which place matters varies across types of care. Place matters the most for outpatient care, explaining two-thirds of the geographic variation. This is consistent with the idea

that outpatient care is the most likely to be discretionary and subject to practice style variation, as well as capacity constraints and access differences. Place also matters for prescription drugs, explaining about a fifth of the variation in the number of prescriptions and about a third of the variation in spending. Prescription drug spending is likely to be influenced by supply-side factors, such as the practice style of the physicians writing prescriptions in an area, but is not subject to capacity and access constraints in the way outpatient care can be. Finally, place-specific factors do not seem to explain the variation in inpatient utilization. This may be because inpatient stays are mostly non-discretionary but instead result from serious illness. Local, district-specific capacity constraints are likely to also matter less.

3.6.3 Assessing the Identifying Assumptions

In this section we assess the validity of the assumptions regarding exogenous mobility, additive separability, and the representativeness of movers on the whole population.

The above-presented event study framework does not only show the relative importance of place, but also enables us to provide (at least) graphical evidence that endogenous mobility is not a major concern in our analysis. We have already seen that the plots in Figure 3.2 convincingly show the adjustment in postmove years in all cases except for inpatient care. Now let us discuss in detail the possible forms of endogenous mobility introduced above: sorting on match effects, drift, and correlated fluctuations in the transitory error.

First, if sorting on the match component is dominant, we would expect to see more individuals moving to districts with higher utilization than the other way around. Appendix Figure A3.3 shows the distribution of the destination-origin differences in the log average utilization measures (Δ). Since the distribution is symmetric and close to a normal distribution, we can conclude that moving is not unidirectional, similar amount of individuals are moving to districts with higher utilization levels as to lower-utilization districts. Second, the presence of match effects would also imply an increase in utilization (and thus, negative coefficient estimates) for those individuals who move to a lower-spending district. Figure 3.3 shows event study results of outpatient spending for "positive" and "negative" moves separately.[15] The place share is positive for both positive and negative moves, i.e. there is an upward adjustment in spending when the average spending in the destination district is higher than in the origin district, and a downward adjustment when it is the other way around. This further suggests that individuals do not tend to systematically move to higher spending districts, and moving to lower spending districts does not happen because of sorting based on idiosyncratic match components.[16]

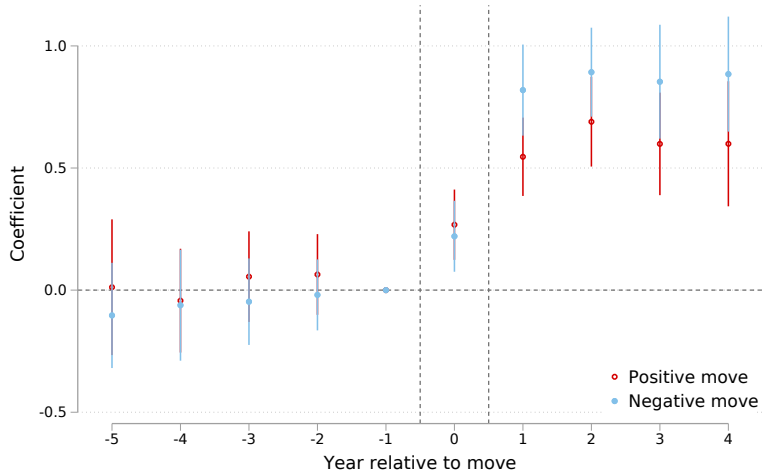
[15] Table A3.6 shows estimates of θ for all main outcome variables, by the direction of move.

[16] Note, however, that the place share is higher for negative than for positive moves. See more below.

Third, a problematic drift component could be directly assessed by looking at the patterns in θ_k. If drift is present, we would expect different pre-move trends in patients' utilization depending on whether their destination district has higher or lower average utilization than the origin district. The evolution of healthcare utilization of "positive" and "negative" movers shown in Appendix Figure A3.4 and also the more formal event study results of Figure 3.2 suggest roughly parallel pre-move trends, hence drift is unlikely to threaten our results.[17] The lack of systematic trend in pre-move utilization suggests that drift in individuals' health is uncorrelated with their mobility pattern.

Fourth, if instead of a systematically declining health status, a sudden health shock induces individuals to move to high-utilization districts (transitory error), we would expect to see an increased utilization level in the first year after the move, especially for individuals moving from low to high utilization districts. Since Figure 3.3 shows no such patterns, we consider this quite unlikely.

Figure 3.3: Event Study: Outpatient Spending by Move Type



Note: The figure shows event study estimates of place effects for outpatient spending by type of move. These are the coefficients θ_k from estimating equation (3.6). The red hollow circles show estimates of place effects for moves from lower- to higher-utilization districts ("positive" moves) and the blue full circles show estimates of place effects for movers from higher- to lower-utilization districts ("negative" moves). The bars show 95% confidence intervals. Controls include calendar year fixed effects and gender – age group interactions. The sample includes all movers ($N = 132,957$ individual-years for "positive" moves and $N = 133,333$ for "negative" moves).

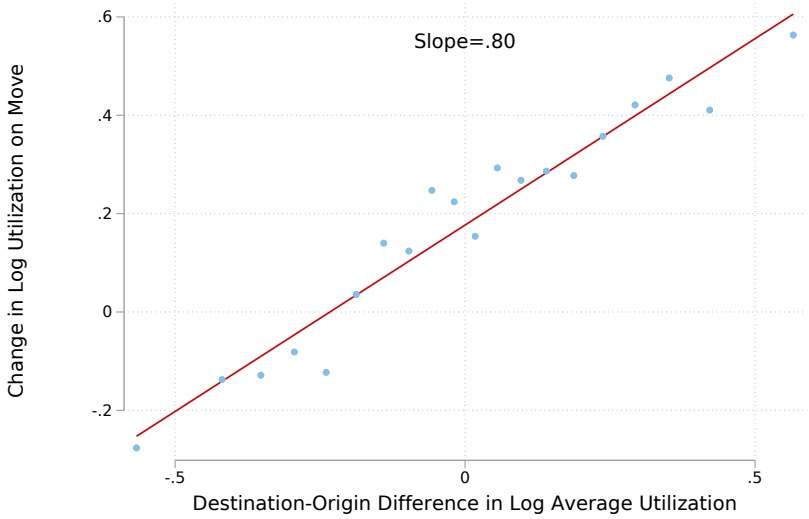
Finally, the event study plots also show that individuals adjust their outpatient care use and pharmaceutical consumption—but not inpatient care use—immediately after

[17] A slight pre-trend may be caused by the measurement error in observing the exact date of the move as suggested by Appendix Figure A3.1.

moving to a new district, without any post-trends. This suggests that habit formation is unlikely to be important in our time frame.

Figure 3.4 provides another graphical representation of the place share. It shows the average change in utilization (defined as the difference between the log average utilization postmove and the log average utilization two to five years premove) depending on the size of the (Δ_i) (the destination-origin difference in utilization). If only place effects would matter, the slope of the fitted line would be 1, while if patient characteristics and preferences would fully determine utilization, the slope would be 0. The slope of 0.8 suggests that the place share is 80%. Although this is even higher than the average place shares suggested by the event study estimations, it further underlines the importance of place.

Figure 3.4: Change in Outpatient Spending by Size of Move



Note: The figure shows the change in log average outpatient spending before and after the move in groups defined by ventiles of the destination - origin difference of log district-level average spending (Δ_i). The x-axis displays the mean of Δ_i for movers in each ventile. The y-axis shows the log average utilization one to four years postmove minus log average utilization two to five years premove, for each ventile. The line of the best fit is obtained from simple OLS estimated on the 20 observations. Sample is all mover-years two to five years premove and two to five years postmove: N=293,219.

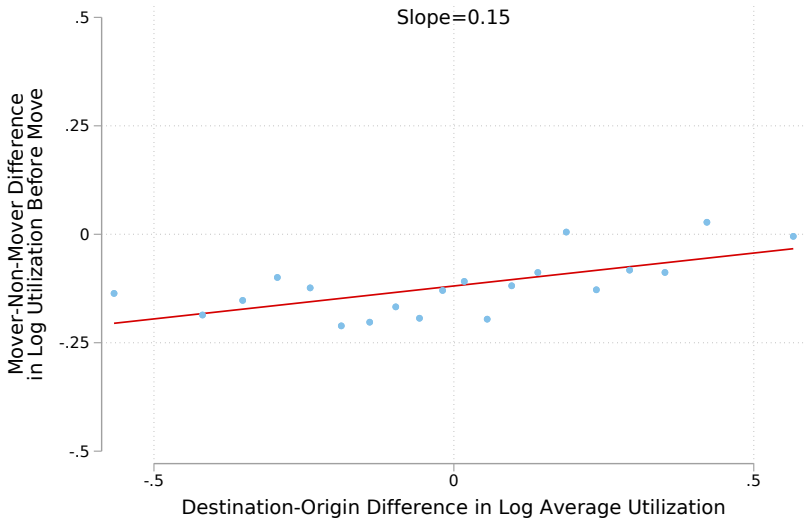
While the absence of sorting on match effects also supports the assumption of additive separability, taking a closer look at Figure 3.3 and the asymmetry of positive and negative moves raises questions on its validity. The log additive functional form would imply that individuals moving from A to B would adjust utilization levels by the same extent in absolute terms as individuals moving from B to A. According to Figure 3.3, pre-trends look similar (there is no anticipatory adjustment in either case), and both types of movers adjust their utilization to the average utilization of their destination district.

However, "negative" moves have a stronger impact on outpatient utilization: when an individual moves to a lower-utilization district, their outpatient spending drops by 86% of the origin-destination gap on average but when they move to a higher-utilization district, their utilization increases by 57% of the gap on average. The same pattern is visible in Appendix Figure A3.8, which displays the change in log outpatient spending before and after the move by ventiles of Δ_i. The figure provides a closer look at the same plot we showed in Figure 3.4, focusing on the potential asymmetry. These suggest that moving to a lower-spending district results in larger adjustments, possibly due to the restricted availability of care there. Such asymmetries are found by Molitor (2018) and Johansson and Svensson (2022) as well.

Since the additive separability assumption would imply no such asymmetries, these results could limit the interpretation of our estimates. Note however that higher place shares for negative moves than for positive moves go against the threat of selective moving. If individuals experiencing a health shock would be more likely to move to higher utilization districts, we would expect higher place shares for positive moves. While the limitations should be further examined, the asymmetry between positive and negative moves still provides a relevant insight into the potential mechanisms behind both the geographic and socioeconomic inequalities. We come back to this in Section 3.6.6.

For the broader validity of the results, it is also required that movers and non-movers have the same γ_j place component in utilization. In Table 3.1 we have already shown that movers and non-movers differ. Now, Figure 3.5 compares the premove utilization of movers to the utilization of matched non-movers. The matched sample is constructed by randomly selecting a nonmover for every mover-year observation that shares the same origin HRR, gender and five-year age bin. The graph plots the difference between movers' and nonmovers' average utilization, depending on the size of the (Δ_i), as in Figure 3.4. The negative values on the y-axis reinforce that movers use less healthcare than nonmovers. The weak, but significant positive relationship (0.15) indicates that moving to a higher utilization district (making a positive move) is associated with relatively higher premove utilization than those moving to a lower-utilization district (negative move). Since movers differ from non-movers along a number of dimensions, including healthcare use, and we have shown significant heterogeneities even within the mover sample, we cannot exclude the possibility that the true place effects are different for non-movers. Note, however, that such differences would not pose a threat to our identification as they are absorbed by the patient fixed effects.

Figure 3.5: Mover-Non-Mover Premove Differences in Log Utilization



Note: The figure shows the difference between movers' and nonmovers' average utilization in groups defined by ventiles of the destination - origin difference of log district-level average spending (Δ_i). The x-axis displays the mean of Δ_i for movers in each ventile, as in Figure 3.4. The y-axis shows the log average utilization of movers in their origin district two to five years premove minus log average utilization of their matched nonmovers in the same years, for each ventile. The matched nonmover sample is obtained by randomly selecting a nonmover for every mover-year observation that shares the same origin HRR, gender and five-year age bin. The line of the best fit is obtained from simple OLS estimated on the 20 observations. Sample is all mover-years two to five years premove: N=130,461.

3.6.4 Socioeconomic Heterogeneity in Outpatient Care Use

The role of place may be different for different groups of individuals. Accordingly, we measure the heterogeneity of θ with respect to socioeconomic characteristics (age group, gender, previous labour market status, income, occupation type). We create two age groups: 40-54 (prime age) and 65-79 years (pensioner) at the time of the move.[18] We further split the younger age group into socioeconomic categories measured two years before the move (workers and non-workers according to their wage quartiles) and the older age group into two categories (whether the sum of pension and labour income, the former dominating in this age group, is below or above the median).

Given the large (average) place shares in outpatient care and the intuition that the measured socioeconomic characteristics play the largest role in outpatient care utilization, we focus our analysis on outpatient spending.[19]

We estimate versions of equation (3.5) by allowing θ to depend on the above socioeconomic characteristics via interactions, and test whether the group-specific place shares

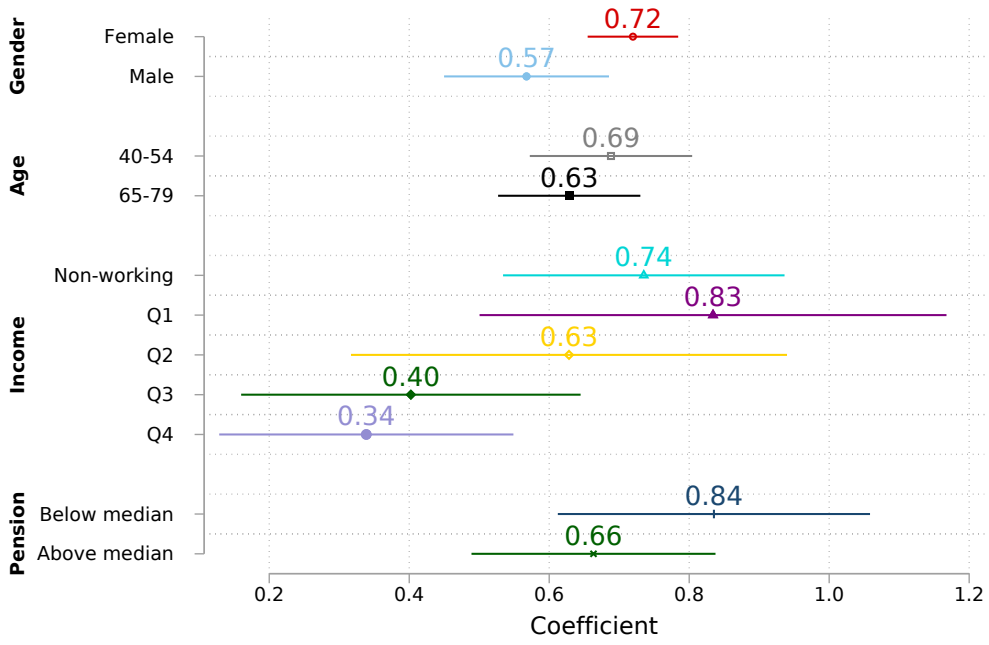
[18] In this analysis we aim to exclude people who move in relation to retirement or childbirth, therefore we focus on these two age categories.

[19] The results for inpatient and drug spending can be found in Appendix Table A3.5.

are significantly different.[20] In the main specification, we calculate the individual Δ_i from district-level averages within groups. This specification takes into account that between-district variation in utilization may differ across groups (although only slightly because of the multiplicative scale). As robustness checks, Appendix Table A3.4 presents results where Δ_i is defined as the difference between the overall raw district averages instead of the group-specific averages. These specifications correspond to heterogeneity in the place effects (group-specific γ_j)[21], and heterogeneity only in the adjustment to those place effects.

Figure 3.6 and Appendix Table A3.4 provide an overview of our estimates of place shares for outpatient spending by socioeconomic groups and Appendix Figure A3.7 shows the corresponding event study plots.

Figure 3.6: Difference-in-Differences: Average Place Effects—Heterogeneity



Note: The figure shows pooled difference-in-differences estimates of place effects for outpatient spending for different subgroups. These are the coefficient θ from estimating equation (3.5), estimated separately for each subgroup. The bars show 95% confidence intervals. Controls include calendar year fixed effects and gender – age group interactions. Wage income categories (non-working, quartiles of positive wage income) are defined for the 40-64 age group. Pension income categories (below or above the median) are defined for the 65-79 age group. Income is measured two years before the move. The sample includes all movers ($N = 135,898$ individual-years for women; $N = 123,448$ individual-years for men; $N = 79,761$ individual-years for the 40-64 age group; and $N = 34,856$ individual-years for the 65-79 age group).

[20] For descriptive statistics on the healthcare use of different groups see Appendix Table A3.1.

[21] For an exploration of possible ways heterogeneities might arise and can be accounted for, see Cantoni and Pons (2022).

Heterogeneity by gender and age is relatively muted. Although the point estimate of the average place share for women (72%) is slightly higher than for men (57%), the differences are not statistically significant. Similarly, the point estimate for the younger age group of 40 to 54 (69%) is slightly higher than for the older age group of 65 to 79 (63%), they are not statistically different from each other.

Heterogeneity is more pronounced across income groups. In the working-age population, the average place share is 74% for individuals who do not work and 83%, 63%, 40%, and 34%, respectively, in the four quartiles of the wage income distribution of those who work. For older individuals, the place share is 84% for those with below-median and 66% for those with above-median pensions. As Appendix Table A3.4 shows, these results are robust to using the alternative definition of Δ_i.

The heterogeneity found in the place shares raises questions about whether different moving patterns or differences in the health status of the groups drive our results. Of course, for healthier patients, there is less room for adjustment to the destination district's utilization level. However, we do not consider this could be the main driver behind our results. First, high-SES individuals, compared to low-SES ones, are only slightly more likely to move to districts with higher average utilization than their origin district (shares of positive moves: 51.4% vs. 48.0%) and they adjust less even if they move to a lower utilization district. Second, we estimate the place shares by pre-move health status (measured by whether drug spending two years before the move was below or above the median calculated by calendar year, age group and gender), and as the bottom panel of Appendix Table A3.4 shows, the place shares are similar (57-63%), meaning that individuals with better and worse health two years before the move will adjust to a similar extent after moving to a different district.

3.6.5 District-Level Correlates of Healthcare Use

Our results so far show that place-specific factors impact healthcare utilization and that these impacts are heterogeneous across groups of individuals. In the final part of the paper, we examine what characteristics of places are correlated with the estimated causal effects of place on utilization as identified by movers across areas. Unlike our main results, this analysis is only correlational but nevertheless can shed light on potentially important mechanisms.

Based on equation (3.8), Table 3.4 shows how district-level variables are associated with the healthcare use of movers controlling for individual fixed effects, calendar time, event time and gender – age group interactions.[22] Outpatient utilization is positively associated with outpatient capacity, and inpatient days (but not inpatient spending) with inpatient capacity (measured as the number of hospital beds). Substitution between the

[22]Summary statistics of the district-level variables are displayed in Appendix Table A3.7.

two types of care may be important since outpatient utilization is negatively associated with the number of hospital beds and inpatient care use is negatively associated with the number of outpatient hours.

Table 3.4: District-Level Correlates of Healthcare Utilization

	(1) Outpatient visits	(2) Outpatient spending	(3) Inpatient days	(4) Inpatient spending	(5) Drug prescriptions	(6) Drug spending
Outpatient hours,	0.079***	0.102***	-0.059**	-0.005	0.019***	0.031
per 100 capita	(0.006)	(0.007)	(0.029)	(0.020)	(0.005)	(0.019)
Hospital beds,	-0.047***	-0.097***	0.129*	0.072	-0.002	0.015
per 100 capita	(0.016)	(0.018)	(0.073)	(0.052)	(0.013)	(0.039)
County seat	-0.172***	-0.226***	0.121	-0.048	-0.042**	-0.064
	(0.025)	(0.028)	(0.115)	(0.080)	(0.019)	(0.057)
Distance from	-0.018***	-0.019***	0.017	-0.016	-0.008**	-0.018*
county seat, 10 km	(0.004)	(0.005)	(0.018)	(0.013)	(0.003)	(0.010)
Log income	-0.005	-0.027	0.169	-0.143	0.050	-0.277**
per capita	(0.040)	(0.048)	(0.167)	(0.141)	(0.033)	(0.113)
Observations	203,910	203,910	100,465	100,437	198,029	198,029

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

Note: The table shows the estimated relationship between healthcare utilization and district-level measures of capacity, access, and income. These are the coefficient η from estimating equation (3.8). Controls include calendar year fixed effects and gender – age group interactions. Number of observations are individual-years.

Distance from the county seat is negatively associated with healthcare use, while the county seat dummy also has a negative coefficient in some specifications because more specialized capacities in the county seat partly serve the population of neighboring rural-type districts as well. Finally, apart from drug spending, the average taxable income of the district as a general socioeconomic indicator does not affect the healthcare use of movers after controlling for the above supply and geographic variables.

Appendix Table A3.8 shows the same associations when instead of estimating equation (3.8) we simply regress the estimated place effects from equation (3.2) on the potential explanatory variables. The magnitudes of the associations between place-specific characteristics and place effects are similar to the results of the one-step procedure presented above.

In Table 3.5 we examine heterogeneity by age and gender in the association between capacity measures and healthcare use, also allowing for non-linearities in the relationship. Column (1) suggests that the partial association between outpatient capacity and outpatient use is stronger at lower levels of capacity, which is consistent with capacity constraints being more binding. Column (2) shows that outpatient capacity is more strongly associated with care use for women, but there is no significant heterogeneity by age. In line with our previous results on capacity constraints being more important for

Table 3.5: Nonlinear and Heterogeneous Effect of Outpatient Capacity on Outpatient Visits of Movers

	(1) 30-79 years	(2) 30-79 years	(3) 40-54 years	(4) 65-79 years
Outpatient hours, per 100 capita	0.133***	0.068***	0.091***	0.067***
	(0.019)	(0.009)	(0.020)	(0.024)
Outpatient hours, per 100 capita^2	-0.0064***			
	(0.0022)			
Interaction with female		0.023***	-0.007	0.030*
		(0.009)	(0.016)	(0.018)
Interaction with (age-40 years)		-0.00027		
		(0.00028)		
Interaction with wage or pension income (million HUF)			-0.0108**	-0.0027
			(0.0049)	(0.0074)
Number of observations	203,910	203,910	49,504	22,688

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

Note: The table shows the estimated relationship between healthcare utilization and district-level measures of capacity, stratified by gender, age, and income. These are the coefficient η from estimating versions of equation (3.8) with quadratic terms or interactions. Controls include calendar year fixed effects and gender – age group interactions. Number of observations are individual-years.

lower-income individuals, column (3) shows that in the 40-54 years old population, the association between outpatient capacity and utilization is stronger for low-wage, working-age individuals. (Column 4 shows that heterogeneity by pension income is not significant in the 65-79 years old group.)

While estimating the effect of outpatient capacity on healthcare utilization, in the above regressions we controlled for individual fixed effects and other healthcare supply, geographic and socioeconomic variables. However, it is theoretically still possible that unobserved variables that change upon moving confound the results. Hence, it is instructive to compare the magnitude of the estimated effects of outpatient capacity with a quasi-experiment, in which new outpatient service locations were established in 2010-2012 in twenty Hungarian districts that had lacked such capacities before (Elek et al., 2015). This development increased the average number of weekly outpatient hours from zero to 1.2 per 100 inhabitants in these districts, while holding fixed all other observable and unobservable characteristics.

Elek et al. (2015) estimated in a difference-in-differences framework that the number of outpatient visits increased on the log scale by 0.217 as a result of the development. A mechanical application of our mover-based results would imply an increase of $0.133 \times 1.2 - 0.0064 \times 1.2^2 = 0.150$ in the quadratic specification (column (1) of Table 3.5). Hence, after taking into account the nonlinear effect of outpatient capacity, the two identification strategies (one based on the changing capacities that movers face upon

moving, the other based on a quasi-experiment of increasing capacities in some given districts) yield surprisingly similar results. This suggests that unobserved variables play a relatively minor role in our mover-based correlational analysis of place effects, and points to the validity of such estimation strategies in explaining the variation in healthcare use.

We note that the development of new outpatient units (the quasi-experiment) increased women's outpatient care utilization more strongly than men's, which is in line with our mover-based results showing women's greater responsiveness to outpatient hours. Also, Elek et al. (2019) estimated a decrease in inpatient care use as a result of the quasi-experiment and hence a substitution between outpatient and inpatient care, which is reflected in the mover-based setting (see the negative estimated effect of outpatient capacity on the number of inpatient days in Table 3.4 and Appendix Table A3.8).

3.6.6 Discussion

In this section we aim to summarize and bind together the results presented above. The several steps of our analysis provide four key insights. (1) Place-specific factors play a key role in determining residents' healthcare use. (2) Low-SES individuals are particularly dependent on place. (3) There are asymmetries in the importance of place: moving to a lower-utilization district results in larger adjustment of utilization. (4) Capacities are strongly associated with place effects, and the relationship is particularly strong for low-SES individuals.

These pieces of evidence point towards the importance of capacity constraints, and these constraints particularly affecting care provision to vulnerable groups. First, our heterogeneity analysis and correlational evidence directly support the argument that low-SES individuals are more affected by supply-side factors, and within that, capacities. Second, we argue that the observed asymmetries by the direction of the move are due to the limited patient choice in lower utilization districts where capacity constraints are more likely. Patients moving to a lower utilization district cannot keep their initial utilization level, thus, they adjust more to the destination's average healthcare use.

The results are in line with a number of demand-side and supply-side reasons of inequalities investigated in the literature. These factors can all be amplified when available capacities are insufficient.

Demand-side sources of inequalities include direct financial constraints Allin and Hurley (2009), less flexibility at work which increases the cost of doctor visits Acton (1975), different time and risk preferences (Cutler et al., 2019; Fuchs, 1982)[23], and informational differences about the benefits of medical care (Cutler and Lleras-Muney, 2010; Glied and Lleras-Muney, 2008).

[23]Fuchs (1982) tests the theory that education would affect health through differences in time and risk preferences, more intuitively patience, however, his findings provide only limited support for the theory. Cutler et al. (2019) also finds that patient preferences are less important than physician beliefs.

It is also documented that access to care differs across socioeconomic groups. Low-SES patients use less healthcare not only because of demand-side reasons, or because within the same district they have access only to less experienced physicians and lower quality of care. Effective access differs even within the same hospital or at the same physician. A more recent strand of the literature studies the potential supply-side sources of inequalities in healthcare utilization, such as the quality of physician-patient communication, unconscious bias or discrimination by providers (Angerer et al., 2019; Brekke et al., 2018; Chen and Lakdawalla, 2019; Currie et al., 2022; Kristiansen and Sheng, 2022; Turner et al., 2022).

While the evidence on the role of capacities is mixed, Singh and Venkataramani (2022) also argue that biases (specifically racial disparities) in provider behaviour may arise when hospitals operate at capacity. Their suggested channels are limited provider bandwidth or reliance on biased algorithms. Using the same empirical strategy as ours, Molitor (2018) also suggests that capacity constraints would imply stronger postmove adjustment in case of 'negative' moves compared to 'positive moves'. He finds that doctors moving to a less-intensive region adjust less, and thus, capacities are unlikely to constrain their treatment decisions. In our setting, however, the observed asymmetries are in line with capacity constraints being binding in low-utilization districts.

It is possible that when there is not enough capacity in a district, high-SES individuals are more able to access those scarce resources. This might occur because of the above-mentioned mechanisms, i.e. having better access to information, or better physician-patient interaction. More direct channels would be more financial resources[24], or more flexibility at work.

This suggests that increasing capacities might enable better access to care for low-ses individuals (extensive margin), but also -in line with other findings in the literature- enables providers to spend more time and resources on any individual patient (intensive margin). When resources are scarce, low-ses groups are hurt disproportionately even in a system of universal health care. Since the latter findings are only correlational, the mechanisms behind the role of capacities in socioeconomic disparities should be further investigated.

In addition, our results are also in line with previous studies using data from Hungary. Using earlier time periods of administrative data Bíró and Prinz (2020) provide descriptive evidence on regional and socioeconomic differences. They suggest that while residents of poorer regions might be in worse health, their access to outpatient care is limited. A number of studies document the prevalence of gratuity payments which can serve as a direct financial channel exacerbating inequalities in access (Gaal et al., 2006;

[24]In our study period gratuity payments were common in Hungary. Although it is a publicly funded system without any co-payments to be made by the patients (except for drug spending), the system of gratuities effectively results in access directly depending on the financial constraints of the patients. For more, see Szende and Culyer (2006).

Orosz, 1990; Szende and Culyer, 2006). As discussed above, Elek et al. (2015) showed in a quasi-experimental setup that extending outpatient capacities increases outpatient care utilization. Other quasi-experiments provide evidence that the number of GP visits and antibiotics consumption decreases in villages with unfilled general practices (Bíró and Elek, 2019), and study outpatient-inpatient substitution patterns using exogenous capacity changes (Elek et al., 2019).

3.7 Conclusion

Substantial geographic variation in healthcare utilization has been documented in a variety of countries and healthcare settings. This paper documents the interaction of this geographic variation with socioeconomic status in the context of Hungary, a healthcare system with universal coverage. Our results show that place matters for healthcare use but it matters differentially for different types of care and different people.

Using movers to decompose utilization into place- and individual-specific components we have demonstrated that place effects explain two-thirds of geographic variation in outpatient spending, but only one-third of prescription drug spending and almost none of inpatient spending. Heterogeneity across income groups is equally pronounced: for working-age individuals who do not work or who belong to the bottom income quarter, place effects explain four-fifth of geographic variation in outpatient care use, while they explain less than two-fifths for individuals with above-median incomes. This suggests that supply-side factors matter more for lower-income individuals.

Our results suggest that capacity constraints may be an important explanation for geographic variation and the documented patterns of heterogeneity. Place effects are larger for more discretionary outpatient care use. They are also larger for lower-income individuals who are presumably more likely to be affected when there is a shortage of physicians or other capacity. Directly assessing the relationship between outpatient capacity and utilization also reveals a positive relationship between these two variables, with magnitudes in line with previous quasi-experimental evidence.

Increasing the capacity of the healthcare system might enable better access to care for low-SES individuals and enable providers to spend more time and resources on all patients. When resources are scarce, low-SES groups are hurt disproportionately even in a system of universal health care. Future work should investigate the causal mechanisms behind the role of healthcare capacity in socioeconomic disparities when access is nominally equal and universal.

References

Abowd, J. M., Creecy, R. H., and Kramarz, F. (2002). Computing Person and Firm Effects Using Linked Longitudinal Employer-Employee Data. *Center for Economic Studies, US Census Bureau*.

Abowd, J. M., Kramarz, F., and Margolis, D. N. (1999). High wage workers and high wage firms. *Econometrica*, *67*(2), 251–333.

Acton, J. P. (1975). Nonmonetary Factors in the Demand for Medical Services: Some Empirical Evidence. *Journal of Political Economy*, *83*(3), 595–614.

Agha, L., Frandsen, B., and Rebitzer, J. B. (2019). Fragmented division of labor and healthcare costs: Evidence from moves across regions. *Journal of Public Economics*, *169*, 144–159.

Ahammer, A., and Schober, T. (2020). Exploring Variations in Health-Care Expenditures-What Is the Role of Practice Styles? *Health Economics*, *29*(6), 683–699.

Allcott, H., Diamond, R., Dubé, J.-P., Handbury, J., Rahkovsky, I., and Schnell, M. (2020). Food deserts and the causes of nutritional inequality. *The Quarterly Journal of Economics*, *134*(4), 1793–1844.

Allin, S., and Hurley, J. (2009). Inequity in Publicly Funded Physician Care: What is the Role of Private Prescription Drug Insurance? *Health Economics*, *18*, 1218–1232.

Andrews, M. J., Gill, L., Schank, T., and Upward, R. (2008). High Wage Workers and Low Wage Firms: Negative Assortative Matching or Limited Mobility Bias? *Journal of the Royal Statistical Society*, *171*(3), 673–697.

Andrews, M. J., Gill, L., Schank, T., and Upward, R. (2012). High Wage Workers Match With High Wage Firms: Clear Evidence of the Effects of Limited Mobility Bias. *Economics Letters*, *117*(3), 824–827.

Angerer, S., Waibel, C., and Stummer, H. (2019). Discrimination in Health Care: A Field Experiment on the Impact of Patients' Socioeconomic Status on Access to Care. *American Journal of Health Economics*, *5*(4), 407–427.

Badinski, I., Finkelstein, A., Gentzkow, M., and Hull, P. (2023). Geographic variation in healthcare utilization: The role of physicians. *NBER Working Paper*, *No. 31749*.

Bailey, M., Johnston, D. M., Koenen, M., Kuchler, T., Russel, D., and Stroebel, J. (2022). The Social Integration of International Migrants: Evidence From the Networks of Syrians in Germany. *NBER Working Paper*, *No. 29925*.

Bíró, A., and Elek, P. (2019). The Effect of Primary Care Availability on Antibiotic Consumption in Hungary: A Population Based Panel Study Using Unfilled General Practices. *BMJ open*, *9*(9), e028233.

Bíró, A., Elek, P., and Kungl, N. (2024). Multi-dimensional Panels in Health Economics with an Application on Antibiotic Consumption. In Matyas, L. (Eds.) *The Econometrics of Multi-dimensional Panels: Theory and Applications* (pp. 479–509). Springer.

Bíró, A., and Prinz, D. (2020). Healthcare Spending Inequality: Evidence from Hungarian Administrative Data. *Health Policy*, *124*(3), 282–290.

Bonhomme, S., Holzheu, K., Lamadon, T., Manresa, E., Mogstad, M., and Setzler, B. (2023). How Much Should We Trust Estimates of Firm Effects and Worker Sorting? *Journal of Labor Economics*, *41*(2), 291–322.

Bonhomme, S., Lamadon, T., and Manresa, E. (2019). A Distributional Framework for Matched Employer Employee Data. *Econometrica*, *87*(3), 699–739.

Borovičková, K., and Shimer, R. (2017). High Wage Workers Work for High Wage Firms. *NBER Working Paper*, *No. 24074*.

Bosque-Mercader, L., Carrilero, N., García-Altés, A., López-Casasnovas, G., and Siciliani, L. (2023). Socioeconomic Inequalities in Waiting Times for Planned and Cancer Surgery: Evidence From Spain. *Health Economics*, *32*(5), 1181–1201.

Brekke, K. R., Holmås, T. H., Monstad, K., and Straume, O. R. (2018). Socio-Economic Status and Physicians' Treatment Decisions. *Health Economics*, *27*(3), e77–e89.

Bronnenberg, B. J., Dubé, J.-P. H., and Gentzkow, M. (2012). The Evolution of Brand Preferences: Evidence from Consumer Migration. *The American Economic Review*, *102*(6), 2472–2508.

Callaway, B., Goodman-Bacon, A., and Sant'Anna, P. H. C. (2021). Difference-in-Differences with a Continuous Treatment.

Callaway, B., and Sant'Anna, P. H. (2021). Difference-In-Differences With Multiple Time Periods. *Journal of econometrics*, *225*(2), 200–230.

Cantoni, E., and Pons, V. (2022). Does Context Outweigh Individual Characteristics in Driving Voting Behavior? Evidence From Relocations Within the United States. *American Economic Review*, *112*(4), 1226–1272.

Card, D., Heining, J., and Kline, P. (2013). Workplace Heterogeneity and the Rise of West German Wage Inequality. *Quarterly Journal of Economics*, *128*(3), 967–1015.

Chaisemartin, C. D., and D'haultfoeuille, X. (2020). Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects. *American Economic Review*, *110*, 2964–2996.

Chen, A., and Lakdawalla, D. N. (2019). Healing the poor: The influence of patient socioeconomic status on physician supply responses. *Journal of Health Economics*, *64*, 43–54.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014). Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates †. *American Economic Review*, *104*(9), 2593–2632.

Chetty, R., and Hendren, N. (2018). The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects. *The Quarterly Journal of Economics*, *133*(3), 1107–1162.

Correia, S., Guimarães, P., and Zylkin, T. (2020). Fast poisson estimation with high-dimensional fixed effects. *Stata Journal*, *20*(1), 95–115.

Crea, G., Galizzi, M. M., Linnosmaa, I., and Miraldo, M. (2019). Physician Altruism and Moral Hazard: (No) Evidence From Finnish National Prescriptions Data. *Journal of Health Economics*, *65*, 153–169.

Currie, J., Kurdyak, P., and Zhang, J. (2022). Socioeconomic status and access to mental health care: The case of psychiatric medications for children in ontario canada. *NBER Working Paper*, *No. 30595*.

Cutler, D., Skinner, J. S., Stern, A. D., and Wennberg, D. (2019). Physician Beliefs and Patient Preferences: A New Look at Regional Variation in Health Care Spending. *American Economic Journal: Economic Policy*, *11*(1), 192–221.

Cutler, D. M., and Lleras-Muney, A. (2010). Understanding differences in health behaviors by education. *Journal of Health Economics*, *29*, 1–28.

de Chaisemartin, C., d'Haultfoeuille, X., Pasquier, F., and Vazquez-Bare, G. (2022). Difference-In-Differences Estimators for Treatments Continuously Distributed at Every Period. *SSRN*.

Deryugina, T., and Molitor, D. (2020). Does When You Die Depend on Where You Live? Evidence from Hurricane Katrina. *The American Economic Review*, *110*(11), 3602–3033.

Deryugina, T., and Molitor, D. (2021). The Causal Effects of Place on Health and Longevity. *Journal of Economic Perspectives*, *35*(4), 147–170.

Dhaene, G., and Jochmans, K. (2015). Split-Panel Jackknife Estimation of Fixed-Effect Models. *The Review of Economic Studies*, *82*(3), 991–1030.

Drenik, A., Jäger, S., Plotkin, P., and Schoefer, B. (2023). Paying Outsourced Labor: Direct Evidence From Linked Temp Agency-Worker-Client Data. *Review of Economics and Statistics*, *105*(1), 206–216.

Elek, P., Molnár, T., and Váradi, B. (2019). The closer the better: Does better access to outpatient care prevent hospitalization? *European Journal of Health Economics*, *20*, 801–817.

Elek, P., Váradi, B., and Varga, M. (2015). Effects of geographical accessibility on the use of outpatient care services: Quasi-experimental evidence from panel count data. *Health Economics*, *24*(9), 1131–1146.

Fadlon, I., and Van Parys, J. (2020). Primary Care Physician Practice Styles and Patient Care: Evidence From Physician Exits in Medicare. *Journal of Health Economics*, *71*, 102304.

Finkelstein, A., Gentzkow, M., and Williams, H. (2016). Sources of geographic variation in health care: Evidence from patient migration. *The Quarterly Journal of Economics*, *131*(4), 1681–1726.

Finkelstein, A., Gentzkow, M., and Williams, H. (2021). Place-Based Drivers of Mortality: Evidence from Migration. *American Economic Review*, *111*(8), 2697–2735.

Freedman, S. (2016). Capacity and Utilization in Health Care: The Effect of Empty Beds on Neonatal Intensive Care Admission. *American Economic Journal: Economic Policy*, *8*(2), 154–185.

Fuchs, V. (1982). Time Preference and Health: An Exploratory Study. In Fuchs, V. (Eds.) *Economic Aspects of Health* (pp. 93–120). University of Chicago Press.

Gaal, P., Evetovits, T., and Mckee, M. (2006). Informal Payment for Health Care: Evidence From Hungary. *Health Policy*, *77*, 86–102.

Gaál, P., Szigeti, S., Csere, M., Gaskins, M., and Panteli, D. (2011). Hungary: Health system review. *Health Systems in Transition*, *13*(5), 1–266.

Ginja, R., Riise, J., Willage, B., and Willén, A. (2022). Does Your Doctor Matter? Doctor Quality and Patient Outcomes. *NHH Dept. of Economics Discussion Paper*, *08/2022*.

Glied, S., and Lleras-Muney, A. (2008). Technological Innovation and Inequality in Health. *Demography*, *45*(3), 741–761.

Godøy, A., and Huitfeldt, I. (2020). Regional variation in health care utilization and mortality. *Journal of Health Economics*, *71*, 102254.

Hull, P. (2018). Estimating Treatment Effects in Mover Designs. (April).

Johansson, N., and Svensson, M. (2022). Regional variation in prescription drug spending: Evidence from regional migrants in Sweden. *Health Economics*, *31*(9), 1862–1877.

Kaarboe, O., and Carlsen, F. (2014). Waiting Times and Socioeconomic Status. Evidence from Norway. *Health Economics*, *23*(1), 93–107.

Kaarboe, O., and Siciliani, L. (2023). Contracts for primary and secondary care physicians and equity-efficiency trade-offs. *Journal of Health Economics*, *87*, 102715.

Kline, P., Saggio, R., and Sølvsten, M. (2020). Leave-Out Estimation of Variance Components. *Econometrica*, *88*(5), 1859–1898.

Koulayev, S., Simeonova, E., and Skipper, N. (2017). Can Physicians Affect Patient Adherence with Medication? *Health Economics*, *26*(6), 779–794.

Kristiansen, I. L., and Sheng, S. Y. (2022). Doctor who? the effect of physician-patient match on the ses-health gradient. *Center for Economic Behavior and Inequality*, *Working Paper 05/22*.

Kwok, J. (2019). How Do Primary Care Physicians Influence Healthcare? Evidence on Practice Styles and Switching Costs From Medicare.

Lucevic, A., Péntek, M., Kringos, D., Klazinga, N., Gulácsi, L., Brito Fernandes, Ó., Boncz, I., and Baji, P. (2019). Unmet Medical Needs in Ambulatory Care in Hungary: Forgone Visits and Medications From a Representative Population Survey. *The European Journal of Health Economics*, *20*, 71–78.

Martin, S., Siciliani, L., and Smith, P. (2020). Socioeconomic inequalities in waiting times for primary care across ten oecd countries. *Social Science and Medicine*, *263*, 113230.

Molitor, D. (2018). The Evolution of Physician Practice Styles: Evidence from Cardiologist Migration. *American Economic Journal: Economic Policy*, *10*(1), 326–56.

Monostori, J., and Gresits, G. (2019). Ageing. In Monostori, J., Őri, P., and Spéder, Z. (Eds.) *Demographic Portrait of Hungary 2018* (pp. 131–149). Hungarian Demographic Research Institute.

Moura, A., Salm, M., Douven, R., and Remmerswaal, M. (2019). Causes of regional variation in Dutch healthcare expenditures: Evidence from movers. *Health Economics*, *28*(9), 1088–1098.

OECD. (2019). Health for Everyone? Social Inequalities in Health and Health Systems.

Orosz, E. (1990). Regional Inequalities in the Hungarian Health System. *Geoforum*, *21*(2), 245–259.

Otero, C., and Munoz, P. (2022). Managers and Public Hospital Performance.

Salm, M., and Wübker, A. (2020). Sources of regional variation in healthcare utilization in Germany. *Journal of Health Economics*, *69*, 102271.

Silva, J. S., and Tenreyro, S. (2006). The log of gravity. *The Review of Economics and Statistics*, *88*(4), 641–658.

Simonsen, M., Skipper, L., Skipper, N., and Thingholm, P. R. (2021). Discontinuity in Care: Practice Closures Among Primary Care Providers and Patient Health Care Utilization. *Journal of Health Economics*, *80*, 102551.

Simonsen, N. F., Oxholm, A. S., Kristensen, S. R., and Siciliani, L. (2020). What Explains Differences in Waiting Times for Health Care Across Socioeconomic Status? *Health Economics*, *29*(12), 1764–1785.

Singh, M., and Venkataramani, A. (2022). Capacity Strain and Racial Disparities in Hospital Mortality. *NBER Working Paper*, *No. 30380*.

Skipper, N., and Vejlin, R. (2015). Determinants of Generic vs. Brand Drug Choice: Evidence From Population-Wide Danish Data. *Social Science and Medicine*, *130*, 204–215.

Sun, L., and Abraham, S. (2021). Estimating Dynamic Treatment Effects in Event Studies With Heterogeneous Treatment Effects. *Journal of Econometrics*, *225*(2), 175–199.

Szende, A., and Culyer, A. J. (2006). The Inequity of Informal Payments for Health Care: The Case of Hungary. *Health Policy*, *75*(3), 262–271.

Turner, A. J., Francetic, I., Watkinson, R., Gillibrand, S., and Sutton, M. (2022). Socioeconomic inequality in access to timely and appropriate care in emergency departments. *Journal of Health Economics*, *85*, 102668.

Van Doorslaer, E., Masseria, C., Koolman, X., et al. (2006). Inequalities in Access to Medical Care by Income in Developed Countries. *Canadian Medical Association Journal*, *174*(2), 177–183.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT Press.

Zeltzer, D., Einav, L., Chasid, A., and Balicer, R. D. (2021). Supply-side variation in the use of emergency departments. *Journal of Health Economics*, *78*, 102453.

Appendix

A3.1 Limited Mobility Bias in the Health Economics Literature

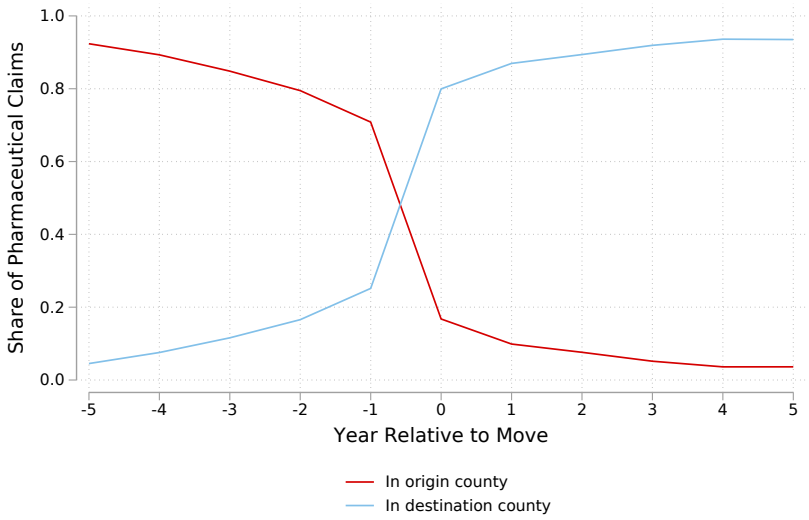
In case of the mover-based designs the bias in fixed effect estimates attributed to the incidental parameter problem is referred to as limited mobility bias. This arises specifically because a large number of fixed effects have to be identified from the potentially limited number of movers. Because of the bias in the fixed effects, the variance of place or physician effects will be upward biased, and consequently, the match component (covariance) will be downward biased.

The "limited mobility bias" term was introduced in the labor economics literature after discrepancies in the results on assortative matching emerged. While not finding positive correlation between the worker and firm fixed effects was initially considered as evidence against positive assortative matching, later it was shown that it is rather caused by the bias in the estimations Andrews et al. (2008). Since then several bias-correction methods were proposed to correct for the limited mobility of workers Andrews et al. (2012), Bonhomme et al. (2023, 2019), Borovičková and Shimer (2017), and Kline et al. (2020).

Because of the differences between the settings in the labor economics and health economics literature, it is less common to address the question when disentangling patient and place effects. First, the number of geographic units considered in the above cited health economics studies is generally lower than the number of firms considered when disentangling worker and firm effects. Consequently, the average number of movers across regions is higher than the average number of workers who switch workplaces within a certain group of firms. Thus, the health economics literature generally argues that the number of movers is sufficient for identification. The question is addressed by only a few studies (Crea et al., 2019; Godøy and Huitfeldt, 2020; Kwok, 2019; Otero and Munoz, 2022; Simonsen et al., 2021).

Godøy and Huitfeldt (2020) use a split-sample jackknife approach following Dhaene and Jochmans (2015) to account for potential limited mobility bias in their variance decomposition exercise. They find that the unadjusted estimate of the place share is 32%, while the bias-corrected estimate is 26%, but the confidence intervals overlap. Simonsen et al. (2021) also use a split-sample correction procedure following Drenik et al. (2023). Kwok (2019) applies a fixed effects method for bias correction developed by Kline et al. (2020) to obtain unbiased estimates of the variance of primary care physician and patient fixed effects and their covariance. The method allows for heteroskedastic errors and autocorrelation within a patient-physician match. Finally, Otero and Munoz (2022) use the homoskedasticity-based correction of Andrews et al. (2008), and find evidence of negative assortative matching.

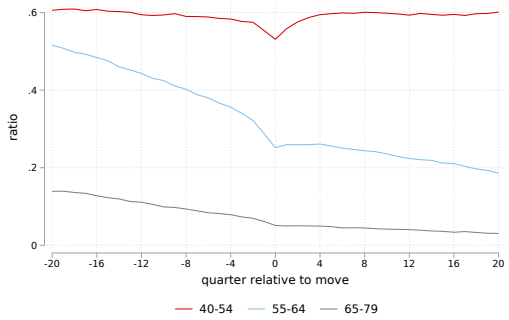
Figure A3.1: Evolution of Share of Pharmaceutical Claims in Origin and Destination County



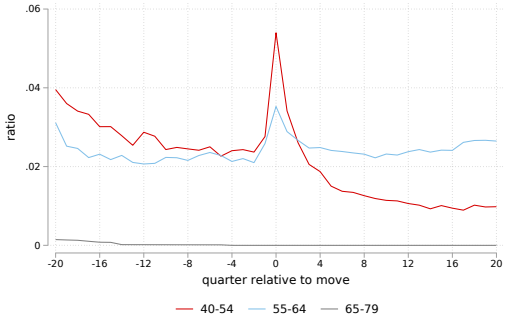
Note: The figure shows the evolution of the share of movers for whom the county where they claimed most of their prescriptions is their origin county and their destination county. Quarterly data are annualized by year relative to the move. Relative year zero is defined as the first four quarters when the individual lived in the destination district according to the place of residence. The sample includes all movers ($N = 64,590$ individuals).

Figure A3.2: Evolution of Labor Market Outcomes

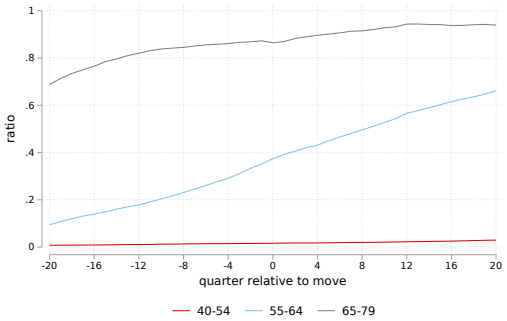
(a) Employment



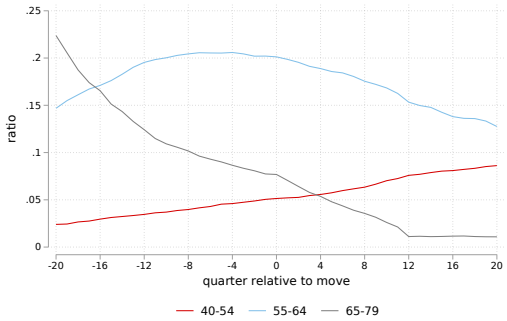
(b) Unemployment Benefit



(c) Old-Age Pension



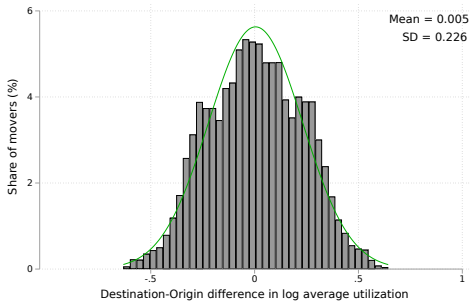
(d) Disability Pension



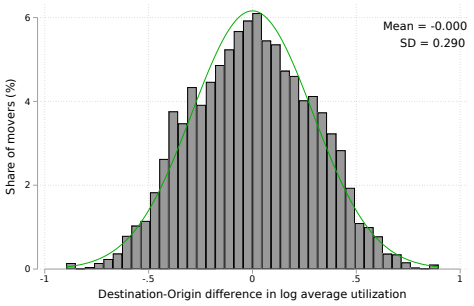
Note: The figure shows the evolution of labor market outcomes, including the probabilities of being employed, receiving unemployment benefits, receiving an old-age pension, and receiving a disability pension among movers by age group. The sample includes all movers ($N = 64,590$ individuals).

Figure A3.3: Distribution of Destination-Origin Difference in Log Utilization

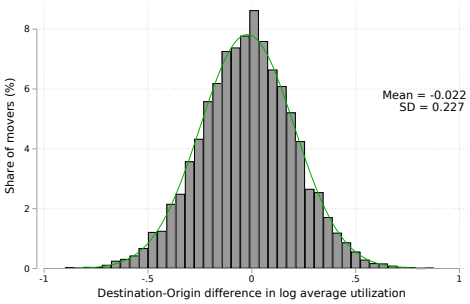
(a) Outpatient Visits



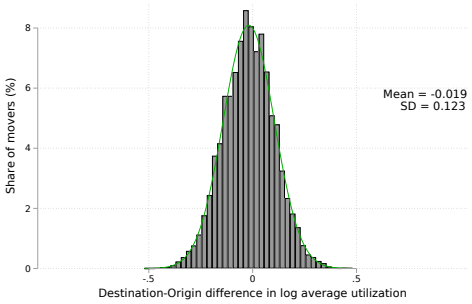
(b) Outpatient Spending



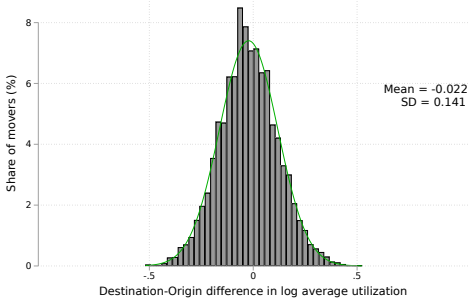
(c) Inpatient Days



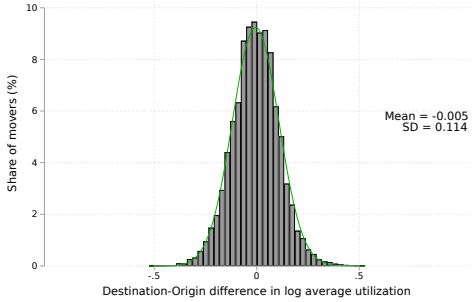
(d) Inpatient Spending



(e) Drug Prescriptions



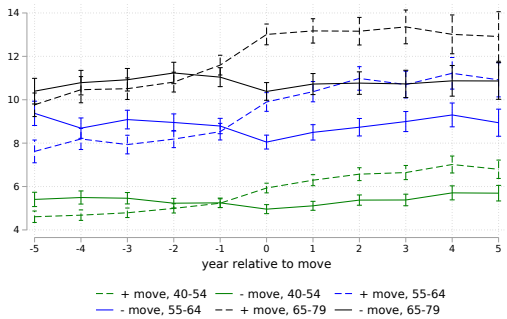
(f) Drug Spending



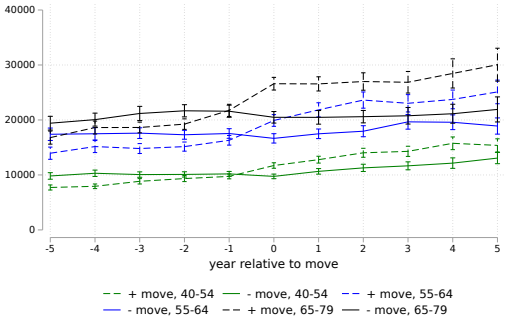
Note: The figure shows the distributions of the logarithmic difference between the average outpatient, inpatient, and prescription drug utilization of a mover's origin district and destination district. The sample includes all movers ($N = 64,590$ individuals).

Figure A3.4: Evolution of Healthcare Utilization of Movers

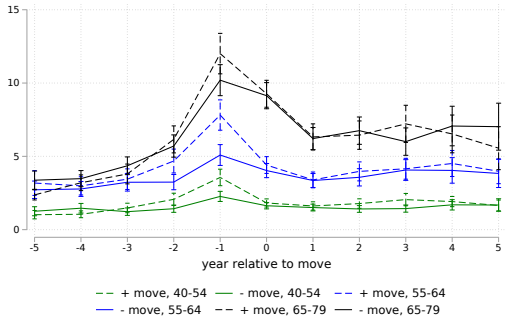
(a) Outpatient Visits



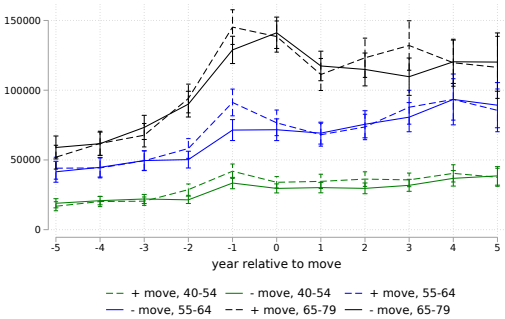
(b) Outpatient Spending



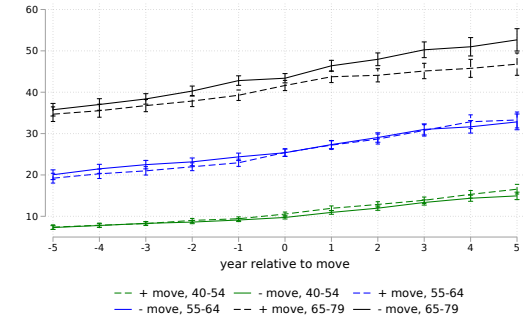
(c) Inpatient Days



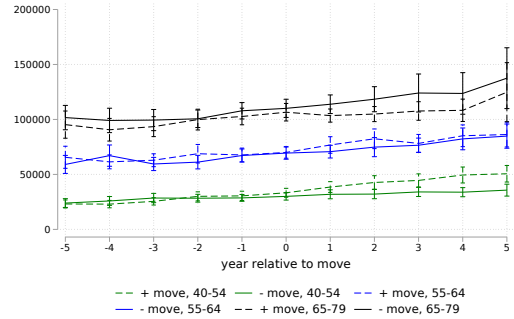
(d) Inpatient Spending



(e) Drug Prescriptions

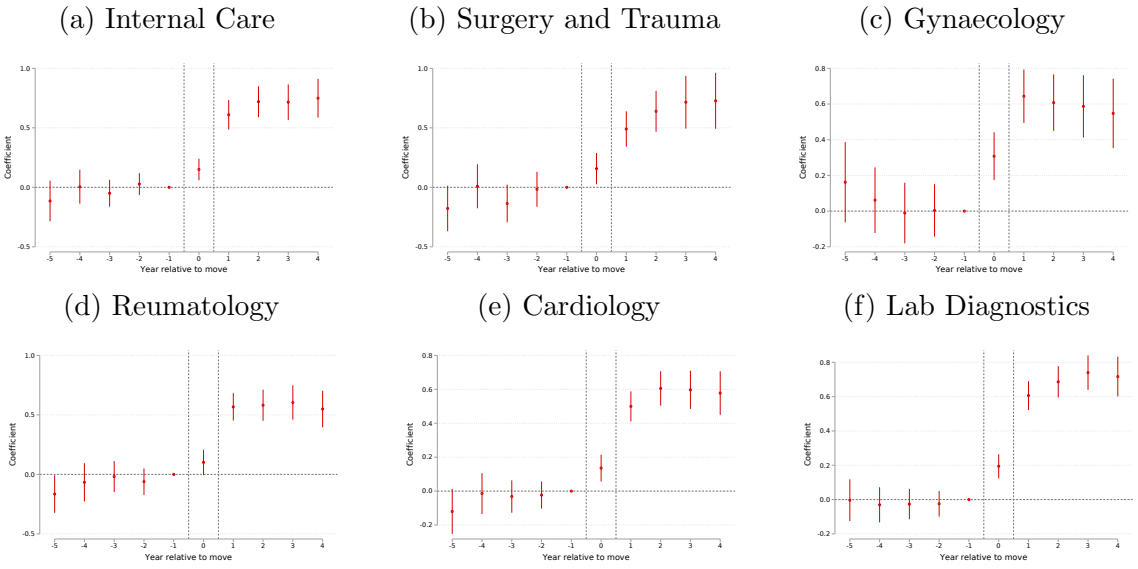


(f) Drug Spending



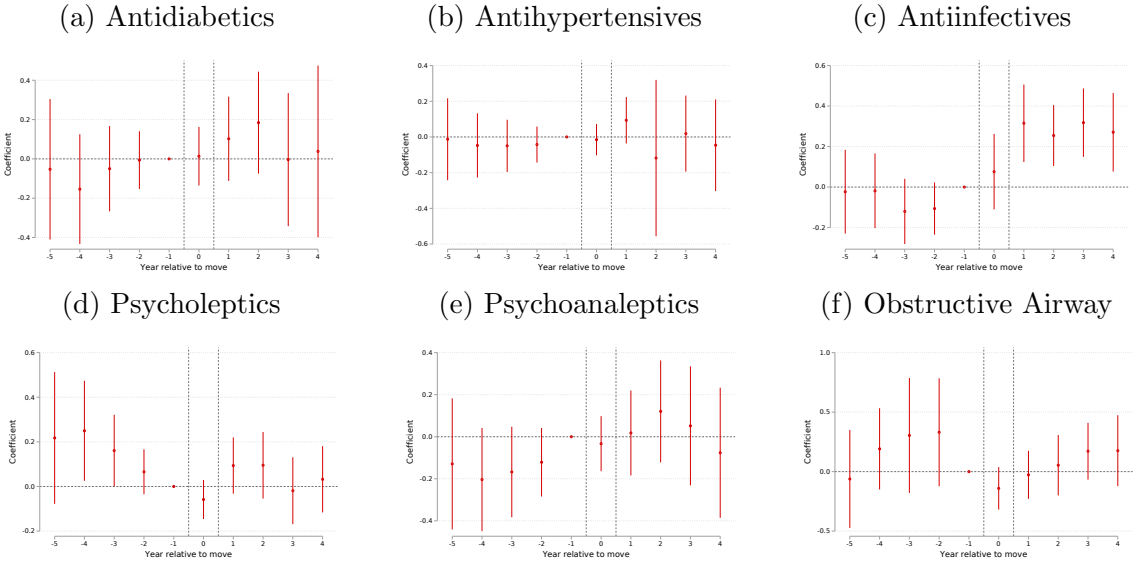
Note: The figures show healthcare utilization of movers of three age groups split by the direction of the move (positive or negative difference between the average utilization of the destination and origin district). 95% confidence intervals for the means are shown.

Figure A3.5: Event Study: Outpatient Specialties



Note: The figure shows event study estimates of place effects for outpatient spending by specialties. These are the coefficients θ_k from estimating equation (3.6). The bars show 95% confidence intervals. Controls include calendar year fixed effects and gender – age group interactions. The sample includes all movers ($N = 266,290$ individual-years).

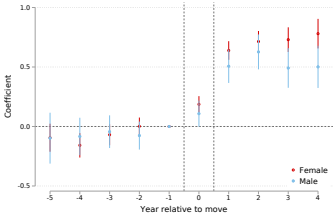
Figure A3.6: Event Study: Therapeutic Classes of Drugs



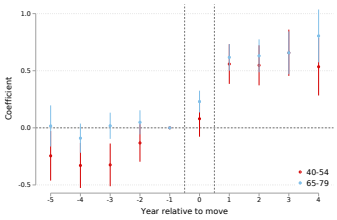
Note: The figure shows event study estimates of place effects for prescription drug spending by therapeutic class. These are the coefficients θ_k from estimating equation (3.6). The bars show 95% confidence intervals. Controls include calendar year fixed effects and gender – age group interactions. The sample includes all movers ($N = 266,290$ individual-years).

Figure A3.7: Event Study: Heterogeneity

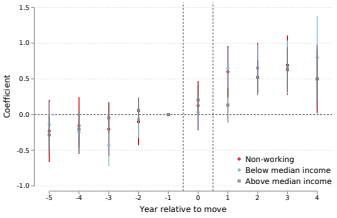
(a) By Gender



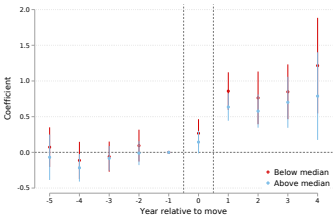
(b) By Age Group



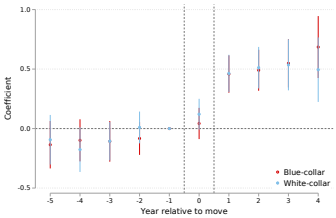
(c) By SES (40-54)



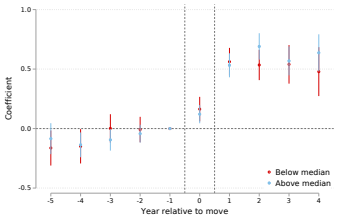
(d) By SES (65-79)



(e) By Job Type

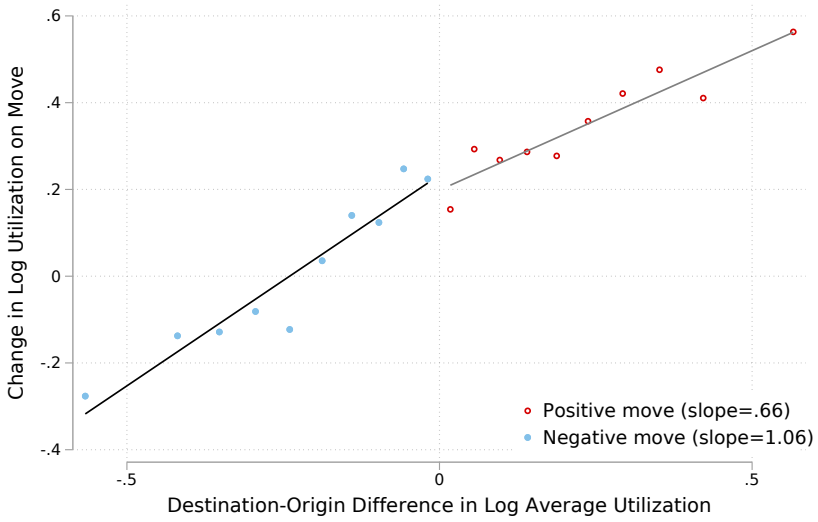


(f) By Health Status



Note: The figure shows event study estimates of place effects for outpatient spending by subgroup. These are the coefficients θ_k from estimating equation (3.6). The bars show 95% confidence intervals. Controls include calendar year fixed effects and gender – age group interactions. Wage income categories (non-working, quartiles of positive wage income) are defined for the 40-64 age group. Pension income categories (below or above the median) are defined for the 65-79 age group. Income is measured two years before the move. Job type is defined by International Standard Classification of Occupations (ISCO) code. Health status is measured by drug spending (below or above the median calculated by calendar year, age group and gender). The sample includes all movers ($N = 266,290$ individual-years).

Figure A3.8: Change in Outpatient Spending by Size of Move, Positive and Negative Moves



Note: The figure shows the change in log average outpatient spending before and after the move in groups defined by ventiles of the destination - origin difference of log district-level average spending (Δ_i). Colors indicate negative ($\Delta_i < 0$) and positive moves ($\Delta_i > 0$). The x-axis displays the mean of Δ_i for movers in each ventile. The y-axis shows the log average utilization one to four years postmove minus log average utilization two to five years premove, for each ventile. The line of the best fit is obtained from simple OLS estimated separately for negative and positive moves.

Table A3.1: Summary Statistics and Regional Variation of Healthcare Use

	(1) Mean	(2) S.D.	(3) Lowest spending district	(4) Highest spending district	(5) Difference max-min (%)	(6) Difference top-bottom quartile (%)
Total spending	120.2	353.6	99.1	151.3	53	20
Outpatient spending	15.0	30.2	9.3	22.7	145	59
Inpatient spending	47.9	220.4	38.3	64.2	68	25
Drug spending	57.3	230.7	41.6	81.2	95	26
Outpatient visits	7.8	12.5	5.3	10.3	95	47
Inpatient days	2.3	12.2	1.6	4.1	163	55
Prescriptions	20.9	30.8	16.3	27.4	68	30
Total spending, by group						
Female	126.8	343.4	103.6	157.1	52	22
Male	112.7	364.7	92.7	149.0	61	22
Age groups						
40-54	80.7	307.0	59.6	119.8	101	33
65-80	228.3	445.4	175.4	279.3	59	18
40-54 years						
Non-working	119.0	399.9	67.7	190.6	182	58
Below median wage	73.7	288.5	55.7	104.1	87	27
Above median wage	55.1	217.2	39.9	74.9	88	30
65-80 years						
Below median pension	215.0	421.5	173.7	258.6	49	19
Above median pension	233.5	466.2	174.2	285.1	64	19
Job types						
Blue-collar	68.4	253.1	53.1	111.5	110	30
White-collar	70.6	276.8	52.3	97.5	87	24

Note: The table shows individual-level summary statistics (mean and standard deviation) and measures of regional variation for healthcare spending (thousand HUF / year) and use (as frequency variables). Column (3) and (4) show usage in the highest and lowest spending districts and Column (5) shows the percentage difference between the two. Column (6) shows the percentage difference between average usage in the top quartile of districts and the bottom quartile. The bottom part of the table shows differences in total spending by groups.

Table A3.2: Difference-in-Differences: Average Place Effects—Outpatient Specialties and Therapeutic Categories

	(1)	(2)	(3)	(4)	(5)	(6)
	Outpatient specialties					
	Internal care	Surgery, trauma	Gynecology	Reumatology	Cardiology	Lab diagnostics
Spending	0.687***	0.634***	0.663***	0.611***	0.573***	0.682***
	(0.0561)	(0.0643)	(0.0406)	(0.0490)	(0.0391)	(0.0367)

			Drug categories			
	Anti-diabetics	Antihyper-tensives	Anti-infectives	Psycho-leptics	Psycho-analeptics	Obstructive airway
Spending	0.126	0.031	0.353***	0.019	0.111	-0.125
	(0.124)	(0.121)	(0.057)	(0.064)	(0.105)	(0.152)

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

Note: The table shows pooled difference-in-differences estimates of place effects for outpatient specialties and therapeutic classes. These are the coefficient θ from estimating equation (3.5). Controls include calendar year fixed effects and gender – age group interactions.

Table A3.3: Difference-in-Differences: Average Place Effects—Robustness

	(1) Outpatient visits	(2) Outpatient spending	(3) Inpatient days	(4) Inpatient spending	(5) Drug prescriptions	(6) Drug spending
(I) Baseline	0.659***	0.659***	0.0136	0.252	0.183***	0.305*
	(0.0316)	(0.0298)	(0.148)	(0.191)	(0.0397)	(0.170)
(II) Full sample	0.669***	0.655***	0.349**	0.473**	0.238***	0.504***
	(0.0353)	(0.0317)	(0.158)	(0.206)	(0.0431)	(0.185)
(III) With controls	0.655***	0.656***	-0.00143	0.229	0.189***	0.307*
	(0.0316)	(0.0299)	(0.147)	(0.191)	(0.0397)	(0.170)
(IV) Adjusted	0.670***	0.663***	0.0294	0.218	0.189***	0.334*
	(0.0320)	(0.0301)	(0.154)	(0.197)	(0.0436)	(0.194)
(V) County-level	0.670***	0.674***	-0.215	0.156	0.116**	0.443*
	(0.0377)	(0.0361)	(0.232)	(0.285)	(0.0501)	(0.228)
(VI) Log(y+1)	0.657***	0.678***	0.125**	0.0976	0.0294	-0.0205
	(0.0295)	(0.0355)	(0.0632)	(0.0707)	(0.0363)	(0.0559)
(VII) Log(y+0.01)	0.683***	0.674***	0.103	0.0975	-0.0250	-0.0295
	(0.0362)	(0.0379)	(0.0659)	(0.0709)	(0.0488)	(0.0593)
(VIII) Two-way SE	0.660***	0.660***	0.0157	0.244	0.183***	0.309*
	(0.0326)	(0.0318)	(0.148)	(0.195)	(0.0444)	(0.161)

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

Note: The table shows pooled difference-in-differences estimates of place effects for outpatient, inpatient, and prescription drug utilization. These are the coefficient θ from estimating equation (3.5). Controls include calendar year fixed effects and gender – age group interactions in all rows. Row (I) replicates our baseline specification from Table 3.2 estimated on movers. Row (II) re-estimates the same specification on the full sample. Row (III) re-estimates the same specification but also includes controls for employment and income. Row (IV) re-estimates the same specification using Δ_i calculated as the destination-origin difference of the log average usages, controlling for gender and age. Row (V) re-estimates the same specification using Δ_i calculated as the difference between the log usage in the destination and origin counties instead of districts. Rows (VI) and (VII) estimate log specifications instead of Poisson. Row (VIII) re-estimates the baseline, clustering standard errors by individual and place. For each utilization type, the first column shows a measure of frequency and the second column shows spending.

Table A3.4: Difference-in-Differences: Average Place Effects—Heterogeneity for Outpatient Spending

	(1)	(2)	(3)	(4)	(5)	(6)
		Baseline		Δ_i not group-specific		
	Place share	(S.E.)	Difference	Place share	(S.E.)	Difference
Gender						
Female	0.720***	(0.0330)		0.719***	(0.0333)	
Male	0.568***	(0.0601)	-0.152**	0.551***	(0.0577)	-0.168**
Age group						
40-54 years	0.688***	(0.0592)		0.645***	(0.0532)	
65-79 years	0.629***	(0.0519)	-0.0604	0.784***	(0.0645)	0.139*
SES, working age						
Non-working	0.735***	(0.103)		0.768***	(0.0906)	
q1	0.834***	(0.170)	0.100	0.715***	(0.163)	-0.0582
q2	0.628***	(0.159)	-0.0662	0.614***	(0.157)	-0.121
q3	0.402***	(0.124)	-0.347**	0.491***	(0.147)	-0.296*
q4	0.339***	(0.107)	-0.400***	0.419***	(0.110)	-0.363**
SES, pensioners						
Below median pension	0.835***	(0.114)		1.008***	(0.143)	
Above median pension	0.663***	(0.0889)	-0.168	0.762***	(0.101)	-0.242
Job type						
Blue-collar	0.549***	(0.0618)		0.637***	(0.0696)	
White-collar	0.528***	(0.0628)	-0.0161	0.506***	(0.0562)	-0.125
Health status						
Below median	0.572***	(0.0465)		0.659***	(0.0536)	
Above median	0.633***	(0.0409)	0.0642	0.648***	(0.0411)	-0.00840

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

Note: The table shows pooled difference-in-differences estimates of place effects for outpatient spending by subgroup. These are the coefficient θ from estimating equation (3.5). Controls include calendar year fixed effects and gender – age group interactions. Wage income categories (non-working, quartiles of positive wage income) are defined for the 40-54 age group. Pension income categories (below or above the median) are defined for the 65-79 age group. Income is measured two years before the move. Job type is defined by International Standard Classification of Occupations (ISCO) code. Columns (4), (5), and (6) replicate the same results using the aggregate, rather than group-specific utilization difference between the origin and destination district.

Table A3.5: Difference-in-Differences: Average Place Effects—Heterogeneity for Inpatient and Drug Spending

	(1)	(2)	(3)	(4)	(5)	(6)
	Inpatient spending			Drug spending		
	Place share	(S.E.)	Difference	Place share	(S.E.)	Difference
Gender						
Female	0.305	(0.222)		0.467***	(0.172)	
Male	0.214	(0.309)	-0.0903	0.0522	(0.256)	-0.417
Age group						
40-54 years	-0.117	(0.245)		0.0874	(0.183)	
65-79 years	0.725*	(0.391)	0.850*	0.186	(0.276)	0.0942
SES, working age						
Non-working	-0.330	(0.254)		-0.0278	(0.0934)	
q1	0.544	(0.627)	0.955	0.485	(0.524)	0.446
q2	1.718*	(1.007)	1.849*	0.293	(0.313)	0.376
q3	-0.825	(0.844)	-0.447	-0.00660	(0.352)	0.0795
q4	1.540**	(0.745)	1.625*	-0.243	(0.333)	-0.122
SES, pensioners						
Below median pension	0.244	(0.624)		0.0451	(0.682)	
Above median pension	1.159**	(0.582)	0.902	0.182	(0.390)	0.146
Job type						
Blue-collar	0.649	(0.453)		-0.0385	(0.215)	
White-collar	0.953**	(0.456)	0.351	0.621**	(0.311)	0.637
Health status						
Below median	-0.212	(0.262)		-0.0336	(0.506)	
Above median	0.784***	(0.256)	0.966***	0.294	(0.222)	0.386

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

Note: The table shows pooled difference-in-differences estimates of place effects for inpatient and prescription drug spending by subgroup. These are the coefficient θ from estimating equation (3.5). Controls include calendar year fixed effects and gender – age group interactions. Wage income categories (non-working, quartiles of positive wage income) are defined for the 40-64 age group. Pension income categories (below or above the median) are defined for the 65-79 age group. Income is measured two years before the move. Job type is defined by International Standard Classification of Occupations (ISCO) code.

Table A3.6: Difference-in-Differences: Average Place Effects—Positive and Negative Moves

	(1) Outpatient visits	(2) Outpatient spending	(3) Inpatient days	(4) Inpatient spending	(5) Drug prescriptions	(6) Drug spending
Baseline	0.659***	0.659***	0.0136	0.252	0.183***	0.305*
	(0.0316)	(0.0298)	(0.148)	(0.191)	(0.0397)	(0.170)
Positive move	0.399***	0.573***	0.171	0.0419	0.221**	0.385
	(0.0777)	(0.0687)	(0.404)	(0.562)	(0.106)	(0.352)
Negative move	0.779***	0.862***	0.250	0.648*	0.0934	0.607
	(0.0794)	(0.0791)	(0.276)	(0.380)	(0.0796)	(0.371)

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

Note: The table shows pooled difference-in-differences estimates of place effects for outpatient, inpatient, and prescription drug utilization. These are the coefficient θ from estimating equation (3.5). Controls include calendar year fixed effects and gender – age group interactions in all rows. The first row replicates our baseline specification from Table 3.2. The second row re-estimates the same specification on the sample of moves from lower-utilization to higher-utilization districts ("positive" moves). The third row re-estimates the same specification on the sample of moves from higher-utilization to lower-utilization districts ("negative" moves). For each utilization type, the first column shows a measure of frequency and the second column shows spending. Frequency measures are outpatient visits, inpatient days, and number of prescriptions.

Table A3.7: Summary Statistics for District-Level Variables

	(1) Mean	(2) S.D.	(3) Lowest	(4) Highest
Outpatient hours, per 100 capita	1.643	1.172	0.000	6.674
Hospital beds, per 100 capita	0.464	0.559	0.000	3.100
County seat	0.103	0.305	0.000	1.000
Distance from country seat, 10 km	3.155	1.968	0.000	9.901
Log income, per capita	13.88	0.186	13.52	14.38

Note: The table shows summary statistics of district-level variables between 2009-2017 (excluding the districts of Budapest).

Table A3.8: Regressions of Place Effects on District-Level Variables

	(1) Outpatient visits	(2) Outpatient spending	(3) Inpatient days	(4) Inpatient spending	(5) Drug prescr.	(6) Drug spending
Outpatient hours,	0.075***	0.092***	-0.089**	0.003	0.017**	0.053**
per 100 capita	(0.012)	(0.014)	(0.039)	(0.031)	(0.008)	(0.021)
Hospital beds,	-0.043	-0.097***	0.185**	0.052	-0.002	0.003
per 100 capita	(0.028)	(0.032)	(0.090)	(0.073)	(0.019)	(0.049)
County seat	-0.155***	-0.197***	0.201	0.024	-0.025	-0.133*
	(0.042)	(0.049)	(0.138)	(0.112)	(0.029)	(0.075)
Distance from	-0.013*	-0.017*	0.020	-0.010	-0.003	-0.015
county seat, 10 km	(0.007)	(0.008)	(0.024)	(0.020)	(0.005)	(0.013)
Log income	0.010	-0.009	-0.029	-0.058	-0.012	-0.233**
per capita	(0.063)	(0.073)	(0.206)	(0.167)	(0.043)	(0.112)

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

Note: The table shows the weighted least squares regression results (weighted by the population of the districts) of the estimated place effects on district-level variables. Districts of Budapest are excluded because of the irrelevance of the district-level supply variables there.
Place effect were estimated using Equation (3.2). Transitory years -1 and 0 were excluded.
Number of districts: N =174.

Abstract

This thesis consists of three chapters applying applied microeconomics tools in the fields of education and health economics, utilizing unique administrative datasets.

The first two chapters investigate fraudulent behaviour in standardized student assessments using Hungarian data. In the first chapter I study the prevalence of test score manipulation and find no evidence of systematic manipulation. I argue this is because of the unique testing environment where the low-stakes testing is paired with strict quality assurance. This means that compared to previously studied tests there is not only less incentive to cheat, but it is also more costly. I employ three methods: the algorithm by Jacob and Levitt (2003), a clustering technique and a simpler screening based on summary statistics. In addition, I shed light on the limitations of the methods with a simulation exercise.

In the second chapter, I provide suggestive evidence that a less costly form of manipulation, namely test pool manipulation, is present in the testing. I exploit a policy change which introduced higher stakes for schools and employ a difference-in-differences estimation strategy. First, in line with the incentives of the new policy, I find that post-policy absence rates increased particularly in schools "at risk" of not meeting the minimum requirement. Finally, using multiple imputation I aim to quantify schools' gains from these manipulations and absences. I find that although the variation in absences is large, schools do not benefit substantially from it.

The third chapter, co-authored with Péter Elek, Anita Győrfi and Dániel Prinz, focuses on geographic and socioeconomic variations in healthcare use. Exploiting migration across regions in Hungary we show that place-specific factors account for 66% and 31% of the variation in outpatient and drug spending, respectively, but play no role in inpatient care use. Notably, place effects explain 80% of outpatient spending variation for non-employed working-age individuals and those below the first quartile of the wage distribution, but less than 40% for individuals with above-median wage incomes. We also find a positive association between place effects and outpatient capacity, especially for low-income individuals. These results suggest that even in a system with universal coverage, access to healthcare can significantly vary, particularly for vulnerable groups.

Zusammenfassung

Diese Dissertation besteht aus drei Kapiteln, in denen angewandte mikroökonomische Methoden in den Bereichen Bildung- und Gesundheitsökonomie angewendet werden.

Die ersten beiden Kapitel untersuchen betrügerisches Verhalten bei standardisierten Schülerbewertungen anhand ungarischer Daten. Im ersten Kapitel untersuche ich die Verbreitung der Manipulation von Testergebnissen und finde keine Evidenz von systematischer, gezielter Verfälschung der Antworten. Ich argumentiere, dass dies auf die einzigartige Testumgebung zurückzuführen ist, in der niedrige Anreize zur Täuschung mit strengen Qualitätskontrollen kombiniert werden. Ich wende drei Methoden an: den Algorithmus von Jacob und Levitt (2003), eine Clustering-Technik und ein Screening auf Basis von zusammenfassender Statistiken. Mit Simulationen beleuchte ich zusätzlich die Beschränkungen dieser Methoden.

Im zweiten Kapitel liefere ich Hinweise auf Manipulationen des Testpools, z.B. durch Verhindern der Test-Teilnahme von leistungsschwachen Schülern. Ich nutze eine politische Maßnahme, die höhere Einsätze für Schulen eingeführt hat, und wende eine Differenz-von-Differenzen-Schätzung an. Im Einklang mit den Anreizen der neuen Politik zeige ich, dass die Abwesenheitsraten nach der Einführung in Schulen, die 'gefährdet' waren, die Mindestanforderungen nicht erfüllen zu können, stärker anstiegen. Schließlich versuche ich mit multipler Imputation die Gewinne der Schulen aus diesen Manipulationen zu quantifizieren. Ich stelle fest, dass, obwohl die Variation bei den Abwesenheiten groß ist, die Schulen nicht wesentlich davon profitieren.

Das dritte Kapitel, Zusammenarbeit mit Péter Elek, Anita Győrfi und Dániel Prinz, konzentriert sich auf geografische und sozioökonomische Variationen bei der Inanspruchnahme von Gesundheitsversorgung. Durch die Verfolgung von Bewegungen zwischen Regionen in Ungarn zeigen wir, dass ortsbezogene Faktoren für 66% bzw. 31% der Variation bei den Ausgaben für ambulante Behandlungen und Arzneimittel verantwortlich sind, während diese Faktoren bei der stationären Behandlung keine Rolle spielen. Wir zeigen, dass es große Heterogenitäten in der ambulanten Versorgung gibt. Der Anteil von ortsbezogenen Faktoren ist in Gruppen mit niedrigem Einkommen am höchsten, und gleichzeitig zeigt sich hier auch eine besonders starke positive Assoziation zwischen diesen Faktoren und ambulanten Kapazitäten. Diese Ergebnisse legen nahe, dass der Zugang zur Gesundheitsversorgung selbst in einem Kontext mit allgemeinen Krankenversicherung variiert.