Evaluations of education policies in Indonesia

Berkhout, E.

Evaluations of Education Policies in Indonesia

Emilie Berkhout

Many children in low- and middle-income countries are not receiving adequate education, despite attending school. This dissertation contributes to the understanding of how education policies affect student academic performance by investigating this issue in the context of Indonesia. It provides evaluations of various policies that were implemented to enhance the quality and fairness of the Indonesian education system.

Emilie Berkhout holds a BSc degree in Economics and Business from the University of Amsterdam and a MSc degree in Development Economics from the Vrije Universiteit Amsterdam. She wrote her PhD dissertation at the School of Economics of the University of Amsterdam. She worked as a researcher at the Amsterdam Institute for Global Health and Development and as a consultant for the Research on Improving Systems of Education Programme.

Evaluations of Education Policies in Indonesia

Emilie Berkhout

Evaluations of Education Policies in Indonesia

Emilie Berkhout

ISBN: 978-94-6473-097-5

Cover: *Tony Liong*, Batik Parang motif from Yogyakarta. The motif resembles waves crashing against rocks, conveying a message of resilience. It symbolizes never to give up when facing challenges in life and to keep striving for our ideals and personal growth.

Printing of this thesis was financially supported by the RISE Programme.

Evaluations of Education Policies in Indonesia

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. P.P.C.C. Verbeek
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op vrijdag 7 juli 2023, te 16.00 uur

door Emilie Maria Berkhout
geboren te Heemskerk

Promotiecommissie

Promotores:	prof. dr. M.P. Pradhan	Universiteit van Amsterdam
	prof. dr. H. Oosterbeek	Universiteit van Amsterdam
Overige leden:	prof. dr. E.J.S. Plug	Universiteit van Amsterdam
	prof. dr. M. de Haan	Universiteit van Amsterdam
	dr. G. Sorrenti	Universiteit van Amsterdam
	prof. dr. J.J. de Laat	Utrecht University
	dr. A. Singh	Stockholm School of Economics

Faculteit Economie en Bedrijfskunde

Acknowledgements

While I am thankful for the support of many individuals, it is my advisor, Professor Menno Pradhan, to whom my research career owes a significant debt of gratitude. Upon completing his Development Economics course during my undergraduate program, he offered me a research assistant position at the Amsterdam Institute for International Development (AIID). This organization later merged with the Amsterdam Institute for Global Health and Development (AIGHD), where I have been able to continue my work. After collaborating on various impact evaluations, he encouraged me to pursue a PhD. Menno's mentorship has motivated me to strive for excellence and challenge myself.

I would also like to express my deep appreciation to my second advisor Professor Hessel Oosterbeek for his guidance throughout my doctoral studies. His insightful feedback and constructive criticism have helped me to refine my research questions and analyses, and have contributed significantly to the quality of my dissertation. I feel fortunate to have had the opportunity to work with such exceptional advisors, and I am thankful for their support and encouragement throughout this journey.

Being able to write my dissertation as part of the Research on Improving Systems of Education (RISE) Programme has been an invaluable experience. The RISE Programme gave me access to a network of outstanding researchers and practitioners. In particular, I would like to express my thanks to Professor Lant Pritchett, the research director of RISE, for instilling confidence in me as a researcher and helping me disseminate my research through presentations at RISE conferences and publications in the RISE working paper series. I am also thankful to Professor Jishnu Das and other anonymous RISE researchers for providing thorough feedback on my papers, which substantially improved their quality. Additionally, I am grateful to my co-author, Rahmawati, for sharing her expertise on exam cheating and the integrity index.

The RISE Indonesia team played an indispensable role in my dissertation journey. I am fortunate to have had my co-author, Amanda Beatty, as additional mentor. Amanda's unwavering support helped me persevere through the arduous writing process. In addition, the researchers at the SMERU Research Institute made my PhD journey a lot more enjoyable. My dissertation benefited enormously from excellent contributions of Daniel, Goldy, Bima, Asri, and Arya. Moreover, my visits to Indonesia were always filled with fun experiences thanks to Risa, Florischa, Shintia, and Delbert, who kindly showed me around Jakarta and Yogyakarta. I will forever cherish their friendships.

My personal growth as a researcher has been greatly influenced by AIGHD, both through their organizational support and the emotional support provided by my former colleagues. The guidance of Melinda Vigh†, Youdi Schipper and Professor Chris Elbers during impact evaluations prior to my dissertation research helped me prepare for the rigorous research required in a PhD program. In addition, my colleagues helped me to stay focused and motivated during the ups and downs of the research process. Special thanks to Gerton, who has been a constant source of moral support throughout my employment at AIID and AIGHD, Marije, who encouraged me to pursue a PhD, and my office mates Marjan, Heleen, Roxanne and Puck, who lend me a listening ear when I faced challenges and provided me with numerous laughter-filled breaks.

Being part of a research group with outstanding economists from both the University of Amsterdam and the Vrije Universiteit Amsterdam was a privilege. I learned a lot from the labor seminars at the Tinbergen Institute and our work-in-progress meetings. The feedback I received during these meetings greatly enhanced this dissertation. Furthermore, I would like to thank my Doctoral Committee for their time and effort in reading my dissertation and posing challenging questions during the defense.

To all of my fellow PhD students, thank you for making my research journey a stimulating and enjoyable experience. I am proud of the strong sense of community that we have built together, for which I want to express a special thanks to Eva, Raviar, Lisa and the members of the Faculty PhD Council. I am also grateful for my office mates at the UvA throughout the years, Luisa, Maddalena, Simon, Julius, Stanislav, Leonard, Timo and Andrej. Your encouragement, valuable feedback and interesting discussions made my research experience more fulfilling.

Most of all, I would like to express my deepest appreciation to my family and friends who have been my support system throughout my academic journey. To my parents, Hans and Gina, and my brother Mark, thank you for always believing in me and encouraging me to pursue my ambitions. To my sister Sophie, thank you for being my rock and for helping me whenever I was struggling with coding. To my girls, “het fort”, thank you for being my cheerleaders, providing much-needed laughs during stressful times. Last but not least, I want to express my gratitude to my dearest Romain. Thank you for supporting me in every way possible, even when it meant scraping data that ultimately went unused.

Acknowledgements of financial support

The studies in this dissertation were funded by the United Kingdom's Foreign, Commonwealth, and Development Office (FCDO), the Australian Government's Department of Foreign Affairs and Trade (DFAT), and the Bill and Melinda Gates Foundation under the RISE Program.

Contents

Chapter 1

Introduction

Improving education has been a central part of economic growth strategies for developing countries. As a result, developing countries have dramatically increased school enrollment over the past decades. The share of children at primary school age who were out-of-school decreased from 43 percent to 19 percent in low-income countries between 2000 and 2020. At the lower secondary level, the share of out-of-school children in low-income countries decreased from 52 percent to 32 percent during the same period. Low-income countries closed more than half of the gap in enrollment rates with high-income countries (UNESCO, 2022).

Despite this schooling expansion, the share of the population obtaining basic skills remains low. Many children leave primary school not knowing how to read or do simple calculations. Using data across nearly 50 developing countries, Pritchett and Sandefur (2020) calculated that at least 40 percent of women would still be illiterate even if all women would complete primary education. In mathematics and science, 62 percent of the world's secondary school students do not acquire basic skills (Gust et al., 2022). Yet, cognitive skills predict individual earnings and economic growth better than years of schooling (Hanushek, 2013; Hanushek and Woessmann, 2008). As the underperformance of education systems in developing countries widens socioeconomic inequality, the World Bank declared a global learning crisis (World Bank, 2018b). Correspondingly, the United Nations (UN) updated their Millenium Development Goal of "achieving universal primary education" (MDG 2) for 2015 to "ensuring equitable quality education for all children around the world" as one of the Sustainable Development Goals (SDG 4) for 2030 (United Nations General Assembly, 2015).

How can governments achieve equitable quality education? This thesis contributes to that discussion by evaluating education policies in the context of Indonesia. Indonesia is a lower-middle income country that has been committed to improving its education system in order to sustain its economic growth. The primary school completion rate has been near-universal since 1988, and lower secondary school completion increased from about 60 percent in 2000 to nearly 90 percent in 2020 (UNESCO, 2021). However, many Indonesian children still lack foundational skills. In 2018, only 30 percent of 15-year-old students was proficient in reading and 28 percent was proficient in mathematics (OECD, 2019). The thesis contains three studies that examine effects on student academic performance of different policies that aimed to improve equitable quality education.

Chapter 2 reports the development of learning outcomes during a period in which the Indonesian Government made large changes to the way they financed and managed their education system. It decentralized managerial and financial responsibilities of education to districts to reduce administrative bottlenecks and to increase responsiveness to local needs in 2001. Moreover, the Government has more than doubled its education expenses since 2000, and has allocated 20 percent of its budget to education spending since 2009 (World Bank, 2013). A large part of those expenses is spent on teachers, as the Government hired more civil servant teachers and doubled the salaries of certified teachers (De Ree et al., 2018). At the same time, the 2005 Teachers and Lecturers Law increased standards for teachers. The share of teachers with a bachelor's degree increased from 37 to 90 percent between 2003 and 2016 (World Bank, 2018a).

To examine how learning outcomes developed during that period, we generated "learning profiles". These learning profiles show how numeracy skills accumulate as children progress through their schooling. We use a household survey that is representative for 83 percent of the Indonesian population. The survey asked numeracy questions to children between 7 and 18 years old in 2000, 2007 and 2014. Children out of school were also tested. A general concern with studying learning outcomes during a period of increased enrollment is that lower-achieving students select into the education system and drive average scores down. We can rule out that such selection drives our results.

The learning profiles revealed two problems in the Indonesian education system. First, many children fall behind curriculum expectations in early grades, but they still move forward in their schooling. As skills are cumulative, we see that only few students eventually learn more complicated skills such as calculating with fractions. Second, despite massive investments in education, numeracy skills declined between 2000 and 2014. The decline in skills was larger in higher grades as more students fell behind with each grade. This decline was substantial; the average grade 6 student in 2014 performed at the same level as the average grade 4 student in 2000. The results suggest that the curriculum is too ambitious.

Chapter 3 studies one aspect of the Indonesian education system that may have hindered learning improvements: widespread cheating. Even though many students lacked basic skills, national exam results showed graduation rates close to 100 percent. An algorithm that detects suspicious answer patterns, developed by the Government, found sufficient evidence for cheating practices in at least one third of all junior secondary schools. Cheating can hold back improvements in learning outcomes for two reasons. First, cheating distorts the information about education quality for policymakers. National exams are generally the only information source on learning outcomes that they have for their policy decisions. Second, if students and teachers know that they can achieve high scores by cheating, they may put less effort into learning.

To fight cheating, the Indonesian Government implemented computer-based testing (CBT) on the grade 9 national exam in 2015. The main cheating method of students and teachers was to get a hold of answer sheets. CBT makes answer sheets useless because it generates an almost unique test version for each student using an item bank on a server. To estimate the impact of this policy, we exploit the staggered implementation of CBT and use administrative data on school-average exam results between 2015 and 2019. We estimate a two-way fixed effect model based on the latest difference-in-difference literature.

We show that computers successfully prevented cheating. Average grade 9 exam scores dropped substantially after implementation of CBT, exposing large-scale cheating practices. To confirm that our findings are indeed due to a decrease in cheating, we show that exam scores dropped most for schools with suspicious answer patterns in previous years. The drop was similar for schools with and without a computer lab, suggesting that the decline in scores was not driven by a lack of computer skills. The intervention was particularly effective because it also decreased cheating in surrounding schools that still took the exam on paper. This suggests that cheating became more difficult, possibly because it became less accepted. Hence, the intervention improved the reliability of the exam to inform policymakers on learning outcomes. As high stakes on the exam remained for students and schools, we expected them to put more effort into learning. However, we do not find improvements in computer-based exam scores within three years of implementation.

Chapter 4 focuses on equity in access to quality education. Oftentimes, school quality varies widely, even within the same region. Parent preferences and school admission policies determine which children have access to the better schools. These school admission policies can have important implications for inequality later in life if the school someone attends affects learning outcomes. Specifically, there is an ongoing debate about whether secondary schools should be allowed to select students based on previous test scores. On the one hand, such a system is considered fair because it rewards hard work and fuels talent. On the other hand, it may increase inequality because test scores are not just a product of effort, but also of a students' background. Selective admissions generally give wealthier students access to the best schools.

Due to concerns about inequality, Indonesia moved away from test score selection for their public junior secondary schools. Public secondary schools perform better than private schools and are free of charge, but they can only accommodate about 60 percent of students. While the highest-scoring students enrolled in the public schools under the old policy, students living in the closest neighborhoods to each public school could enroll under the new policy.

We evaluate this policy change in the context of Yogyakarta, where some of the best performing public junior secondary schools in the country are located. As there is little neighborhood segregation in Yogyakarta, the new policy change gave many low-achieving students access to public schools and displaced high-achieving students to private schools. We study whether low-achieving students benefited from enrollment in public schools, whether displaced high-achieving students learned less in private schools, and how a change in peers affected students who stayed in the same schools. We identify students whose access changed and students whose access stayed the same by simulating public school access under each policy for two student cohorts admitted before and after the policy change.

We find that giving low-achieving students access to high-quality schools does reduce learning inequality, but mostly at the expense of high-achieving students. Low-achieving students who gained access to public schools saw modest benefits, while high-achieving students who lost access to public schools saw a twice-as-large learning decline. Moreover, students who remained in public schools learned less with lower-scoring peers, while students who remained in private schools did not benefit from higher-scoring peers. Survey results suggest that teachers simplified their instructions to accommodate for the lower-achieving students, suggesting that high-achieving students were stimulated less than before.

In summary, this thesis investigates how some of Indonesia's efforts to improve equitable quality education contributed to learning outcomes. First, a combination of doubled education spending, decentralized management, more higher-educated teachers and increased secondary enrollment did not improve average numeracy skills in the country. Numeracy skills even declined during that period. Second, high stakes on the exam did not increase exam scores within three years after cheating was no longer possible. Because initially cheating schools were located in rural and poorer areas, this may point to constraints on school resources to improve learning outcomes. Third, more equitable school admissions led to somewhat lower average learning outcomes as high-achieving students lost more learning than low-achieving students gained.

The results suggest that, even for a country with a sufficient education budget, achieving equitable quality education is a challenge. The findings highlight two general lessons.

First, as long as there are many low-quality schools, equitable quality education cannot be achieved. When education is compulsory, and seats in quality schools are limited, some students would have to enroll in low-quality schools. Some regions may not have quality schools at all. Thus, to achieve quality education for all, the focus should be on improving the schools in the bottom of the quality distribution. Second, we need to better understand how policies translate into classroom practices, and what kind of support teachers need. The results raise several questions, such as how teachers deal with students who are behind the curriculum, and why exam scores did not increase after teachers and students could not cheat anymore. In addition, the results show that policy changes can lead to behavioral responses in school management, teachers and students, which in turn affect school performance. Understanding classroom practices and these behavioral responses would help with designing education policies that make all actors aligned towards, and equipped for, improving learning for all.

Chapter 2

Schooling Progress, Learning Reversal: Indonesia's Learning Profiles Between 2000 and 2014[1]

2.1 Introduction

Over the past twenty years, Indonesia has made dramatic progress in improving junior and senior secondary enrollment. While the country had achieved universal primary enrollment in 1988 (Government of Indonesia, 1998), between 2000 and 2014, the timeframe of this study, it saw a 17 percentage point improvement in junior secondary enrollment, to 77 percent, and a 20 percentage point improvement in senior secondary enrollment, to 59 percent (Statistics Indonesia, 2020).

Simultaneous with extending years of schooling for millions of children, the country also made massive investments in education with the stated goal of improving quality. In 2002, the 1945 Constitution was amended to require that 20 percent of the budget be allocated to education spending. In 2005, the government passed the Teachers and Lecturers Law, which required higher qualification standards for new and existing teachers and effectively doubled civil servant teacher salaries (UU No. 14, 2005). Indonesia's move to decentralization in 2001 also extended to education policy such that its approximately 500 districts could make decisions on education delivery and adjust policy to local context and needs (UU No. 22, 1999).

[1] This chapter is joint work with Amanda Beatty, Luhur Bima, Menno Pradhan, and Daniel Suryadarma, and is published in the International Journal of Educational Development (Beatty et al., 2021). The authors thank Jishnu Das, Michelle Kaffenberger, Lant Pritchett, Anu Rangarajan, Niken Rarasati, Shintia Revina, Andrew Rosser, and Abhijeet Singh for helpful comments and suggestions. We are grateful to Thomas Coen for his valuable comments and work on an earlier version of this chapter. Alia An Nadhiva provided excellent research assistance.

Despite reforms that provided more educational resources, raised standards, and increased school access, the country continues to face learning challenges. In 2018 Indonesia scored 379 out of 500 on the mathematics portion of the Programme for International Student Assessment (PISA); a score of 379 is 7th from the lowest score among the nearly 80 countries or states taking the test (OECD, 2019). PISA defines Level 2 as "achieving at least a minimum proficiency level," and the Sustainable Development Goals (SDG) use PISA "Level 2" as a metric for SDG Target 4.1 (UNESCO Institute for Statistics, 2018). Fewer than 1 in 3 students in Indonesia were able to perform at Level 2 or above in mathematics (OECD, 2019). Indonesia demonstrated similar results in the Trends in International Mathematics and Science Study (TIMSS) in 2015, in which only 50 percent of 4th graders met the lowest benchmark defined as having "some basic mathematical knowledge." Another 27 percent of students did not even meet the lowest benchmark, and no students met the highest benchmark (Mullis et al., 2016). Looking at Indonesia's historic performance on these assessments in mathematics, it has largely stayed the same over time for PISA (OECD, 2019) and fallen for TIMSS since 2003 (Mullis et al., 2012, 2008, 2004).

This chapter takes a deeper look at the contrast between the positive trends in enrollment and the more negative or static international assessment findings on learning. It is unclear what is driving these two outcomes. It could be that newer learners entering the system (i.e., possibly students from households with less educational exposure, facing greater challenges staying in school, or keeping up with the instructional pace) bring down average learning. It could also be that learning at least did not go up because the system's quality deteriorated; or the answer could be a combination of these hypotheses. We further explore this contrast using a unique longitudinal household-level dataset, the Indonesian Family Life Survey (IFLS). The IFLS includes not only variables on household characteristics but also mathematics assessments for children age 7 and up in 2000 and 2014. We use the testing data to develop a set of mathematics learning profiles that show learning by age and grade-level; and we assess how learning varies by background characteristics and over time. We are able to examine the trends in learning for in-school and out-of-school children, in contrast to international assessments, which only assess in-school children. Moreover, we can assess how learning changed as enrollment rose in Indonesia.

To better understand how learning changed in the face of this improvement in enrollment, we first answer the following questions: What did children in school know compared to curriculum expectations? How much did in-school children learn as they progressed through school? These two questions allow us to frame children's basic numeracy competencies within the context of what the education system expects children to know by a particular grade and examine if schooling is delivering more learning with each additional year. We answer these questions using just the IFLS 2014 for children across all schooling-relevant ages. Then we ask: Did learning change over time? Specifically, we compare learning profiles of all children and of enrolled children between 2000 and 2014. This is one of two studies that analyses learning accumulation in Indonesia over time. Afkar et al. (2018) looked at mathematics learning for in-school children between 2011 and 2012; we utilize data for all school-age children from 2000, 2007, and 2014.

We finally answer the question: Did different subgroups demonstrate different learning profiles? We pursue this analysis in order to understand if one group is driving our findings and examine if different groups disproportionally benefit or lose from education system changes during this timeframe. We look at separate effects for children in different wealth groups, males and females, children whose mothers have different education levels, and different provinces.

2.2 Background

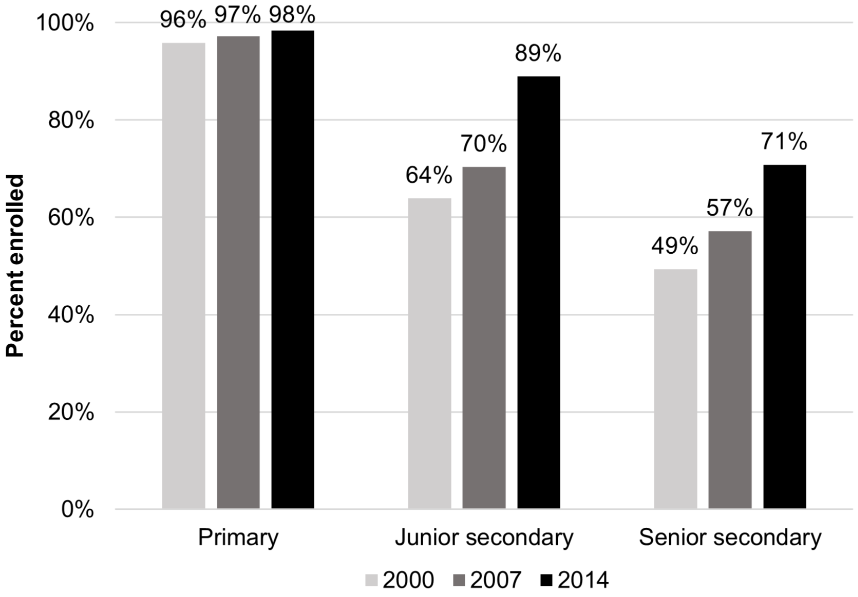
2.2.1 Changes to Indonesia's Educational Landscape Between 2000 and 2014

In this section we offer context to our research questions regarding whether, for whom, and why learning may have changed from 2000 to 2014. We describe changes to the education landscape during that timeframe, including the shift towards decentralization, rising enrollment, increased education spending, lower teacher-student ratios, improved teacher qualifications, curriculum changes that focus less time on mathematics, and eliminating class grades as a criterion for graduation.

Indonesia generally, and its education system specifically, went through dramatic changes starting in 1999 when the country transitioned to democracy, which included a shift towards decentralization, offering more financial and political autonomy to its now 514 districts. In 2003, the government solidified this initiative in education by granting more autonomy to districts to manage education (UU No. 20, 2003). Since 2003, civil servant teachers have been hired by the central Ministry of Education and Culture (MoEC), which also sets the curriculum, upper-grade assessments, and accredits schools; but districts distribute and manage teachers, hire and fire non-civil servant teachers, allocate funding to schools, manage school infrastructure, and carry out a range of other functions. This move towards decentralization meant that the country saw more geographic variation in education delivery than it had in previous decades.

Enrollment had already begun to rise at the primary level (grades 1 to 6) before 1999 as primary school attendance had been compulsory since 1984 (UU No. 20, 2003), and enrollment was near universal since 1988 (Government of Indonesia, 1998). Junior secondary (grades 7-9), which became compulsory in 2003, and senior secondary (grades 10-12) enrollment saw significant growth during our study period, 2000-2014. The IFLS 2014 data shows that junior secondary enrollment increased by 25 percentage points, from 64 percent to 89 percent; and senior secondary enrollment increased by 22 percentage points, rising from 49 percent in 2000 to 71 percent in 2014 (Figure 2.1).[2] (The IFLS dataset is described in detail in Section 2.3.) These figures were 79 percent for junior secondary and 61 percent for senior secondary nationally in 2019 (Statistics Indonesia, 2020).[3]

Figure 2.1: Educational Enrollment by Year and School Level



Source: IFLS 3 (2000), IFLS 4 (2007) and IFLS 5 (2014)
Note: The figure shows the total of net enrollment and completion rates. Net enrollment and completion rates are calculated as a percentage of respondents who are within the anticipated age range and who (1) ever enrolled in the specified school level and are still enrolled, or (2) ever enrolled in the specified school level and finished that school level: 7- to 12-year-olds for primary school, 13- to 15-year-olds for junior secondary school, and 16- to 18-year-olds for senior secondary school.

Not surprisingly, attainment for people ages 20 to 30 also reflect these enrollment trends. Between 1993 and 2014, average years of schooling increased from 7.1 years to 10.5 years (authors' analysis of IFLS). In 2014, according to the IFLS, 95 percent had

[2] All analyses in this chapter focuses on all school types combined. This includes secular public schools, religious public schools, and secular and religious private schools.

[3] The discrepancy between the IFLS and the national statistics likely reflects the fact that the IFLS is representative of 83 percent of the population and the omitted 17 percent represents mainly very remote areas.

completed primary school; this attainment went up slightly between 2000 and 2014, from 91 percent. In 2014, this figure was 82 percent for junior secondary and 57 percent for senior secondary, up from 64 percent and 38 percent respectively in 2000. There was also little within-school-level drop-out among 20 to 30-year olds. Almost 95 percent of students who enrolled at any level of schooling completed it.

Government spending on education grew significantly over our study period. In 2002, the government amended the 1945 Constitution to require that 20 percent of the budget be allocated to education spending. Indonesia achieved this goal in 2009, nearly doubling spending on education over just five years (World Bank, 2013). By 2014, spending per year reached over 300 trillion Rupiah or nearly US$21 billion (World Bank, 2018a). A large share of the increased funding for education was spent on employing more teachers and driving down class sizes. The student-teacher ratio was 22-1 in 1999; and even in the midst of increasing enrollment was 16-1 by 2010, one of the lowest ratios in the region (UNESCO Institute for Statistics, 2018). A larger education budget was also spent on increasing pay for teachers as stipulated in the 2005 Teachers and Lecturers Law, although research demonstrated that this did not affect learning (De Ree et al., 2018).

Teachers became on average more highly educated over this timeframe. Between 2003 and 2016, due to changes to teacher certification requirements resulting from the 2005 Teachers and Lecturers Law, the share of teachers with a bachelor's degree rose from 37 to 90 percent (World Bank, 2018a). There is evidence that teachers' education may not explain much variation in teacher effectiveness in developed countries (Hanushek et al., 2005); in Indonesia, teachers with bachelor's degrees perform only slightly better on a series of math, science, and Indonesian test questions than teachers with less education (De Ree, 2016).

While we might not expect spending or improved teacher qualifications to improve learning, we would not expect those improvements to have a negative effect. We now discuss several changes – children's exposure to mathematics content and national examination incentives – that could have negatively affected learning over the study period. Curriculum changes reduced the number of hours of math instruction per week. The 1994 curriculum mandated 10 hours a week of math instruction for grades 1-3 and eight hours a week for grades 4-6. In 2004, the curriculum required teachers in grades 1-3 to teach math "thematically," which meant that teachers were to cover all academic subjects related to a theme or topic; and lowered math instruction limits to five hours per week for grades 4-6 (Sugiarti, 2014). Shifting to thematic lessons was an adjustment for teachers who received little training or guidance in implementing this approach. The curriculum change could have prompted teachers to cover less material, but it is also possible that teachers found it challenging to teach with less structured guidance.

The 2003 National Education System Law changed the significance of leaving exams. Prior to 2003, a student's graduation from 6th, 9th or 12th grade was based on yearly grades and national exam results. After 2003, the country took a lower stakes approach of basing promotion on a combination of teacher discretion and the leaving exams. Districts also took over responsibility for the grade 6 leaving exam, so the content varied by district, although MoEC's testing center still had responsibility for overseeing the junior secondary and senior secondary leaving exams. In 2014, grade 6 and 9 exam scores still had stakes in some areas as they could have been used for admission to junior secondary and senior secondary schools, and admission to some schools was highly competitive.

2.2.2 Learning Profiles Literature

We generate learning profiles to examine learning across grades and over time. A learning profile is a plot of skills, knowledge, or subject-matter competence across multiple grades or ages, among in-school and/or out-of-school children. It represents the skill or knowledge that a cohort of children accumulates during schooling (Kaffenberger, 2019). Kaffenberger (2019) identifies three main categories of learning profiles: contemporaneous cross-section (knowledge across a cross-section of respondents in different grades and ages), adult retrospective (knowledge of a cross-section of adults who have completed schooling), and true panel (knowledge of the same respondents over time). This study uses IFLS to generate contemporaneous cross-section and true panel profiles.

The majority of studies that employ learning profiles use contemporaneous cross-section. Assessments by organizations such as the ASER (Annual Status of Education Report) Centre, Uwezo, and USAID, which created the EGRA/MA (Early Grade Reading Assessment and Early Grade Math Assessment), generated some of the first examples of learning profiles in developing countries. For example, Jones et al. (2014) used Uwezo data to show that in Kenya, Tanzania, and Uganda more than half of 10-year-olds and one-third of 13-year-olds could not recognize a single written word or recognize numbers. Spaull and Kotze (2015) showed that the poor-wealthy gap in Grade 3 was three grade levels. Pritchett and Beatty (2015) used ASER data to illustrate the concept of learning profiles and incongruence between curriculum pace and actual student learning.

Less common are adult retrospective and panel profiles. Kaffenberger and Pritchett (2020) created learning profiles across ten countries using Financial Inclusion Insight data with young adults ages 18 to 37, and Pritchett and Sandefur (2020) used DHS literacy data from women aged 25 to 34 in 51 countries. The child-level longitudinal study, Young Lives, utilizes similar questions across four countries – Ethiopia, India, Peru, and Vietnam – and has in several papers demonstrated vast differences in learning gains over time across countries using panel learning profiles (Singh, 2020a; Rolleston and James, 2015; Rolleston, 2014). The LEAPS program in Punjab, Pakistan followed the same children over four rounds or years of schooling, highlighting learning changes as children transitioned from public to private school and vice versa (Bau et al., 2021; Andrabi et al., 2008).

Afkar et al. (2018) produced the first study of learning profiles and the first panel profiles in Indonesia. They examined changes in math learning for 40,000 children in 360 primary and junior secondary schools over two sequential years (2011 and 2012), using anchor items that were similar across grades. They find that approximately 40 percent of students did not master basic numeracy questions after three years in school and that in many schools, learning did not keep up with curriculum expectations.

While profiles naturally differ across countries, a common theme across the papers cited above and others is that profiles are shallow in many developing countries, meaning students learn little as they progress through school. This finding is consistent with the "learning crisis" message from the 2018 World Bank World Development Report. Afkar et al. (2018) illustrate the shallow learning profile. They find that the same number of students who can recognize numbers by the end of grade 2 can do one-digit multiplication by the end of grade 3, indicating that only those who can recognize numbers are the ones who will learn one-digit multiplication, i.e., those who are behind do not catch up.

Another common finding across papers is that in countries with shallow learning profiles, much of the potential gains in learning are through improvements in the quality of learning per grade rather than the expansion in schooling. For example, Singh (2020a) uses panel profiles, also with Young Lives data, to make comparisons of different countries with differential schooling productivity and shows that the effect of another grade of schooling in Vietnam is 0.25 to 0.40 standard deviations higher than in other countries. Exposing students to a more productive schooling environment like that in Vietnam closes nearly all of the cross- country achievement gap for students in Peru and India and 60 percent of the students in Ethiopia. Similarly, in a context in which even the advantaged have shallow learning profiles, Akmal and Pritchett (2021) generate simulations using ASER and Uwezo data to show that even helping poor students achieve the attainment profiles of the rich doesn't necessarily generate large learning gains. In India, Pakistan, and Uganda, just 60 percent of poor students would be numerate and able to read a simple story if they achieved the attainment levels of the rich.

2.3 Data

We construct learning profiles using three waves of the IFLS, collected in 2000 (IFLS 3), 2007 (IFLS 4), and 2014 (IFLS 5) (Strauss et al., 2016, 2009, 2004; Frankenberg et al., 1995). The IFLS is a panel survey, started in 1993, that follows the same households and their offspring (if household members form a new household) at each survey round. The over 30,000 respondents live in 13 of 27 provinces, and the survey is representative of 83 percent of the Indonesian population. The IFLS randomly selected enumeration areas in each province from a nationally representative sampling frame used in the 1993 SUSENAS, a socioeconomic survey designed by the Indonesian Central Bureau of Statistics.[4] Within each EA, households were randomly selected from the 1993 SUSENAS listings (Frankenberg et al., 1995). The 2000 and 2014 waves serve as the primary source for analysis presented in this chapter; we also use the 2007 data for panel analysis in Section 2.5.2.

While the IFLS was primarily designed to measure demographic changes, it includes a multiple-choice numeracy test with nine items shown in Table 2.1. Different age groups took one of two versions of the test with different levels of difficulty. Test 1 is the first four items and Test 2 is the latter four items in Table 2.1. The one overlapping question (56/84) is shaded in grey and was included in both versions. All items are multiple choice with four answer options, except for the first three questions, which had three answer options. Table 2.1 shows which respondent groups took which test items in which years. For the analysis presented in this chapter, we mainly use results from respondents between ages 7 to 18 because the analysis primarily focuses on school-age children.

The mathematics test was first included in the IFLS in 2000. Children aged 7 to 14 took Test 1 while 15 to 18-year-old adolescents took Test 2. In the 2007 and 2014 IFLS, adolescents 15 years old or above were asked to take Test 1 again if they also took it seven years earlier when they were between 7 and 14 years old. Therefore, of the respondents 15 years old and above, a large percentage took all ten items across the two versions in the same IFLS year (88 percent in 2007 and 71 percent in 2014). (These students took the overlapping item twice, so we characterize this as ten items total.) Table 2.1 also shows our mapping of the items to the skill or concept that a child should have mastered by a certain grade according to the 2006 and 2013 national curriculum standards (Kementerian Pendidikan dan Kebudayaan, 2013; Badan Standar Nasional Pendidikan, 2006).

[4] The IFLS over-sampled rural enumeration areas and enumeration areas in smaller provinces to facilitate urban-rural and Javanese-non-Javanese comparisons. We use sampling weights to correct for this.

Table 2.1: IFLS's Numeracy Questions, Expected Grade Mastery According to the Curriculum, and Ages in Which Children Were Tested in Which IFLS Year

Numeracy skill	Test question	Expected grade level mastery	Ages tested		
			2000	2007	2014
2-digit subtraction	49-23	1	All 7-14	All 7-14, 88% of 15-18	All 7-14, 71% of 15-18
3-digit addition and subtraction	267+112-189	2			
1-digit addition and multiplication	(8+9)*3	3			
Subtracting fractions	1/3-1/6	4			
2-digit division	56/84	4	All 7-14, All 15-18	All 7-14, All 15-18	All 7-14, All 15-18
Order of operations	(412+213)/(243-118)	3	All 15-18	All 15-18	All 15-18
Decimals	0.76-0.4-0.23	4			
Calculating interest (Percent 1)	Ali put 75,000 rupiah in his savings account. If he receives 5% interest a year, how much interest does Ali receive on his savings after one year?	5			
Calculating percent (Percent 2)	If 65% of people smoke, and the current population is 160 million, how many people do not smoke?	5			

Note: Data source is IFLS 2000, 2007, 2014, and Badan Standar Nasional Pendidikan, 2006 and Kementerian Pendidikan dan Kebudayaan, 2013. We examined the 2006 and 2013 curricula to determine the grade in which the numeracy skill was covered in the curriculum; and to examine if there were changes with curricula reforms. In the IFLS data, Test 1 is referred to as EK 1 while Test 2 is referred to as EK 2.

Table 2.2 shows the sample size for the numeracy test in each survey wave. We excluded from the analysis those individuals for whom the complete numeracy test is missing because they refused, could not be contacted, did not have enough time, or any other reason unrelated to competencies (5.5 percent of the sample). We also excluded those individuals for whom educational attainment is missing (0.1 percent of the sample for whom we have a numeracy score).

Table 2.2: Numeracy Question Sample Sizes, Ages 7–18

	2000	2007	2014
Respondents interviewed (attempted + did not attempt numeracy test)	9,579	9,517	11,362
Respondents who answered at least one numeracy question	9,208	9,162	10.697
Percent of respondents who answered at least one numeracy question for whom we imputed at least one item*	21.5	16.7	14.7

Source: IFLS 3, 2000, IFLS 4, 2007, and IFLS 5, 2014
Note: Table includes in- and out-of-school children. In our analysis we also include students above 18 years old who are still enrolled in senior secondary school. This amounts to 84 students in 2000, 80 in 2007 and 63 in 2014. These individuals are excluded from the table as they are over 18.
* Imputation methods discussed in Section 2.4.

2.4 Methods

As discussed above in Section 2.3, there are two versions of the numeracy test—an easy version (Test 1) and a more difficult version (Test 2). We applied a test equating procedure using Item Response Theory (IRT) to generate a measure of numeracy skills that is comparable between the two versions of the test and adjusts for question difficulty. To link the test versions, we employed a horizontal test equating procedure using the group of respondents that answered both versions, called anchor respondents.

Responses from the anchor respondents generated the difficulty level and discrimination power of each of the ten items.[5] As mentioned above, there is one overlapping item in Test 1 and Test 2: 56/84. While the question is the same in both versions, the notation was slightly different ($\frac{56}{84}$). We chose to treat the overlapping question as separate questions in each version because one-third of the respondents that answered both versions gave two different answers.

To estimate each respondent's numeracy score using IRT, we use a three-parameter logistic model. Three parameters, item discrimination power, item difficulty, and a guessing parameter, are used to determine the fourth parameter, which is student ability. The difficulty parameter relates to the ability of an individual, such that if the difficulty parameter is equal to the ability parameter, the individual is equally likely to answer correctly or incorrectly. The discrimination parameter reflects how fast the probability of success changes with ability near the item difficulty. The higher the discrimination parameter, the better the item can differentiate high ability students with those with low ability. Putting these parameters in a formula, the probability of person j providing a positive answer to item i is given by

[5] Note that there are no anchor groups in the 2000 survey. The numeracy score is based on the anchor respondents in 2007 and 2014. Technically, we assume that the relative difficulty levels and discrimination power of the items remained the same over time and is the same across the country.

$$\Pr(Y_{ij} = 1|\theta_j) = c_i + (1 - c_i)\,\frac{\exp(\alpha_i\ (\theta_j - b_i))}{1 + \exp(\alpha_i\ (\theta_j - b_i))} \qquad \theta_j \sim N(0,1) \tag{2.1}$$

where α_i represents the discrimination of item i, b_i represents the difficulty of item i, c_i represents the guessing correction called the pseudo guessing parameter and θ_j is the latent trait (or ability) of person j (StataCorp, 2017). We present the results for θ and weigh them using sampling weights. We present Bayesian Markov chain Monte Carlo estimates of the latent ability θ.[6]

The ability parameter reflects the respondent's numeracy skill level. Even though the limited number and scope of the items pose constraints to our numeracy skill measure, tests of psychometric properties of the measure show that the test items are adequate for the numeracy comparisons we make.[7] We standardize the numeracy skill measure using the mean and standard deviation of grade 1 students in the 2000 sample and rescale the measure to have a mean of 0 and a standard deviation of 100 for grade 1 students in 2000. This way, our measure shows the improvement in learning relative to grade 1 in terms of grade 1 standard deviations.[8] Throughout the rest of the chapter, we call this the "standardized numeracy score".

The numeracy test responses contain missing values, and we find that missing data patterns are systematic. We find that the share of missing values generally increases as the question difficulty increases, measured by the grade in which the items are expected to be mastered according to the curriculum, and that the highest share of missing values is concentrated among the youngest respondents (see Table A2.2). This provides evidence that the missing value patterns are associated with lower skills, so we infer that respondents likely left these questions blank because they didn't know the answer. Because leaving these values out of our analysis would bias the results, we impute the missing items as if the respondent gave an incorrect answer. Table 2.2 shows the percentage of observations that we imputed with an incorrect answer. We impute at least one item response on the test for 22 percent of the 2000 sample and 15 percent of the 2014 sample. As a robustness check, we also perform our analysis without imputed values and by imputing missing values with random guessing and find that the learning profiles are

[6] We use the openIRT Stata program developed by Tristan Zajonc. Maximum likelihood estimates of latent ability are similar and available upon request.

[7] We check the validity of the score with factor and infit and outfit analysis, and we examine the reliability using Cronbach's alpha and the IRT discrimination coefficients. In addition, we run tests on the IRT assumptions of unidimensionality, no differential item functioning, and conditional local independence.

[8] Standardizing using the grade 1 mean and standard deviation could result in unrealistically large difference in learning across grades, because we might expect the grade 1 standard deviation to be relatively small as the test is actually too difficult for these students. However, our results look similar when we use the grade 5 standard deviation for the standardization. For ease of interpretation (improvements relative to grade 1), we use the standardization using the grade 1 mean and standard deviation.

steeper when imputing with wrong answers, because ignoring missing values or imputing with random guessing inflates scores of children in lower grades who had the most missing values. However, it does not alter our conclusions about differences in learning between subgroups and learning over time (see Appendix A2.2).

For individual items shown in Figure 2.2, we correct the percent correct for guessing such that, in expectation, a zero is given for those who randomly guessed and a 1 is given for those who knew the correct answer. As the test items are multiple choice, respondents could correctly answer a question by chance alone. We use the following method by Afkar et al. (2018). If α is the fraction that knows the answer and y is the fraction that answered correctly, then:

$$y = (1 - \alpha) \times \frac{1}{K} + \alpha \times 1 \tag{2.2}$$

for K answer options. Those who guess have a probability of $\frac{1}{K}$ to answer correctly, while those who know the answers have a probability of one. We present the results for α and weight them using sampling weights.

In Section 2.5.3, we show the standardized numeracy score by gender, region (province), mother's education level and wealth quintile. For the differences by wealth, we generate an asset index using Principal Component Analysis (PCA) at the household level (Filmer and Pritchett, 2001).[9] For differences by region, we show the average difference in learning between 2000 and 2014 for the 13 provinces included in the IFLS.[10] The IFLS data is representative at the provincial level (Frankenberg et al., 1995). We estimate the following regression model using Ordinary Least Squares to measure the change in the standardized numeracy score between 2000 and 2014 within each of the provinces

$$Y_{ipg} = \beta_1 + \beta_2 W_{ipg} + \sum_{p=1}^{13} \beta_{2,p} \times P_{ig,p} + \sum_{p=1}^{13} \beta_{3,p} \times W_{ig,p} \times P_{ig,p} + \gamma_g + \epsilon_{ipg} \tag{2.3}$$

where Y is the standardized numeracy score for student i from province p in grade g. W is a dummy variable for the 2014 IFLS wave, P are dummy variables for the 13 provinces, γ_g are grade fixed effects, and ϵ is an error term.

[9] The included assets are a house, land, other buildings, poultry, livestock or fish pond, vehicles (cars, boats, bicycles, motorbikes), household appliances (radio, television, fridge, etc.), savings or certificate of deposit or stocks, credits (money owed to the household), jewellery, and household furniture and utensils.

[10] These are North Sumatra, West Sumatra, South Sumatra, Lampung, Jakarta, West Java, Central Java, Yogyakarta, East Java, Bali, West Nusa Tenggara, South Kalimantan and South Sulawesi.

2.5 Learning Outcomes Results

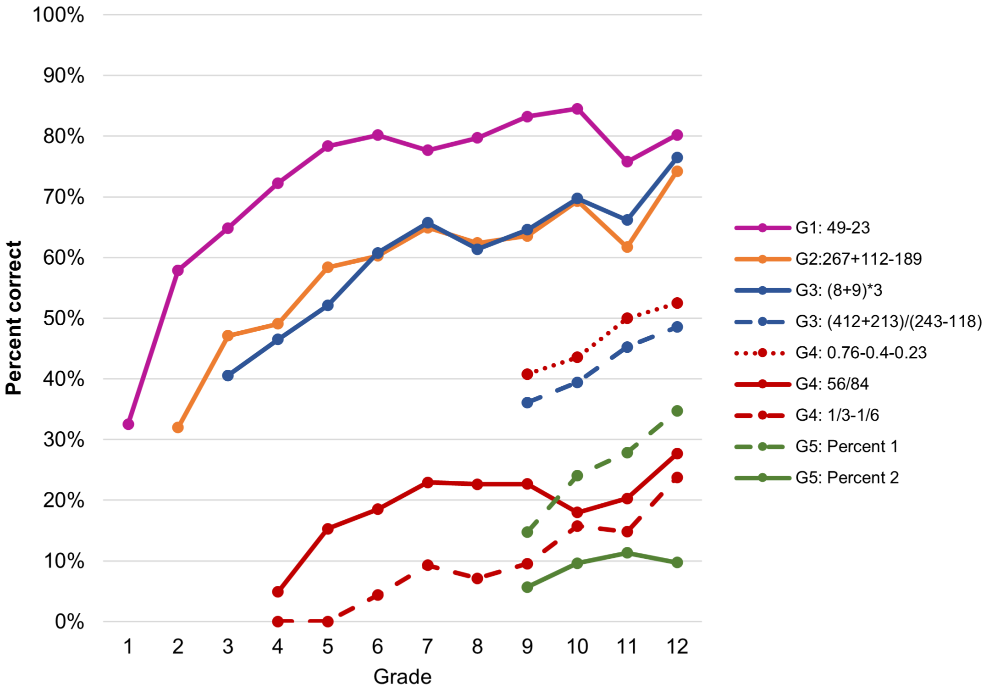
In this section we shed light on mathematics learning gains, using questions from the IFLS that were asked of respondents in both 2000 and 2014.

2.5.1 What Did Children in School Know in 2014 Compared to Curriculum Expectations? How Much Did In-School Children Learn From One Grade to the Next?

Our first finding is that learning levels were low in 2014 and by extension, children did not keep up with curriculum expectations. Figure 2.2 shows descriptive learning profiles for the 2014 IFLS questions for each grade, by item, mentioning what grade level the item content is covered in the curriculum. Just 67 percent of students in grade 3 could answer the simplest grade 1 question, 49-23, correctly. This low level of learning is even more pronounced for more "difficult" questions, such as those requiring calculating fractions or percent. Only 36 percent of 12th graders could correctly answer a word problem on calculating percent (Percent 1 in Figure 2.2) and no 5th graders could answer 1/3-1/6, a grade 4 question, correctly.

Second, children learned little as they progressed through school. There was particularly little improvement in most numeracy skills after primary school (grade 6). For example, using the grade 1-level question, 49-23, which just 65 percent of grade 3 students could answer, we find that this mastery improved by approximately 15 percentage points by 6th grade but there was no improvement between grades 7 and 12. The solid-line grade 1-3 items shown in Figure 2.2 start with around 30-40 percent of students correctly answering the problem in the relevant grade level. In subsequent grades in primary school, the share of students correctly answering the question grew by approximately just 5 to 10 percentage points per grade; this share fell to 1 percentage point per grade in junior secondary school. For the items only asked of students in grades 9 to 12, the share of students answering correctly generally only improved by 1 to 4 percentage points per grade, with the exception of the percent problem regarding interest (Percent 1 in Figure 2.2) for which we see up to a 5 percentage point improvement per grade in the share of students answering correctly in grades 9 to 12.

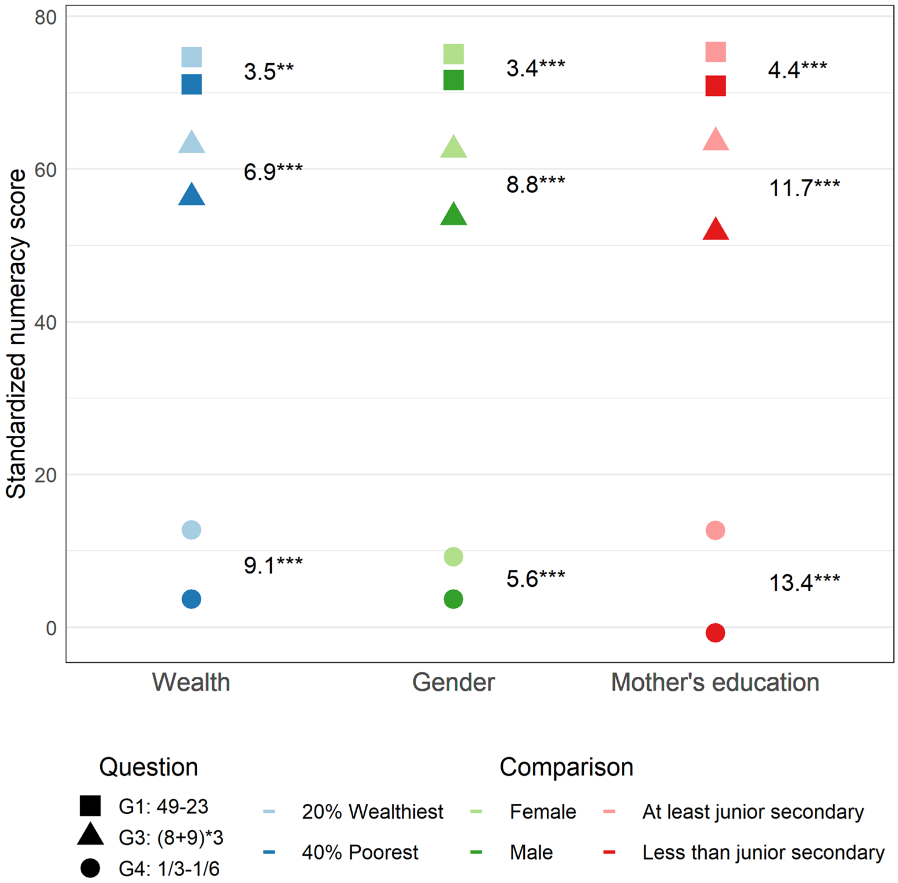
Figure 2.2: Learning by Grade Level and Item, Enrolled Students in 2014



Source: IFLS 5 (2014)
Note: Results show the percent who answered each question correct among currently enrolled students. The sample sizes for each grade change depending on the number of children in that grade and what questions students should have mastered according to the curriculum per Table 2.1. Some results are presented beginning with students who enrolled in 9th grade as harder item-level questions were only asked among an older age group (15 years and older). Grade-level 1, 2 and one level 3 ((8+9)*3) questions have three answers; all remaining questions have four answers. The questions for Percent 1 and Percent 2 are in Table 2.1. Results are adjusted for guessing as described in Section 2.4.

Looking at subgroup differences for these items, we find that differences grow with question difficulty, as shown in Figure 2.3. While there was hardly any difference (3 percentage points) between the wealthiest 20 percent and the poorest 40 percent of the population in the grade 1 level question $(49 - 23)$, this difference was 9 percentage points with a grade 4 level question $(1/3 - 1/6)$. We find the largest difference between students whose mothers completed at least junior secondary school and students whose mothers completed less than junior secondary school. Students with mothers with higher attainment were 13 percentage points more likely to correctly answer the grade 4 question, while almost none of the students whose mothers completed less than junior secondary school could answer that question. For the hardest question, the smallest subgroup gap is that between males and females, yet there is still a 5 percentage point difference. All differences are statistically significant.

Figure 2.3: Subgroup Differences for Three Questions, Enrolled Students in 2014

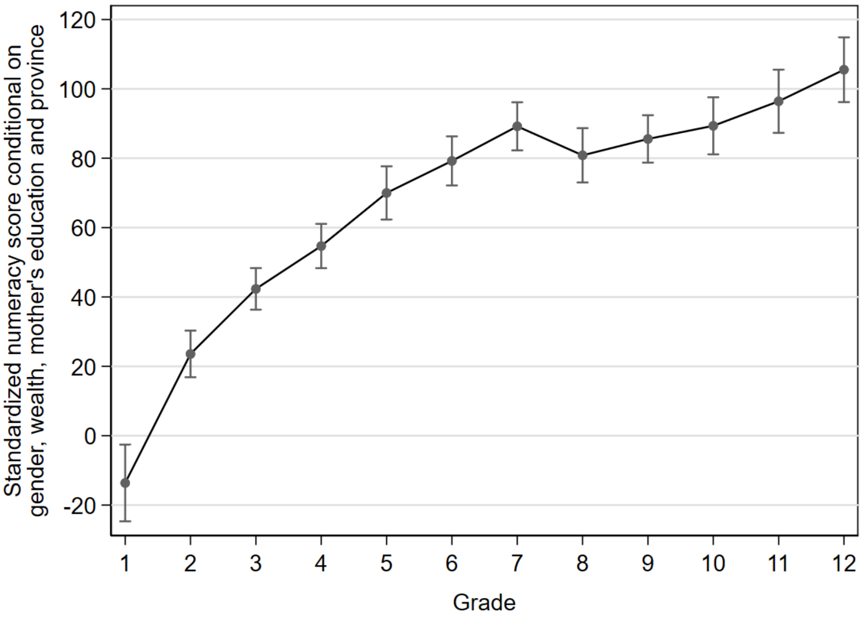


Source: IFLS 5 (2014)
Note: Results show the subgroup standardized numeracy score of the three different items and the subgroup difference among currently enrolled (40 percent poorest, males, and students with mothers who completed less than junior secondary school). The sample sizes for question change depending on the number of children enrolled in grades in which students should have mastered the question according to the curriculum per Table 2.1. For example, the students included in bars for the G4 question are enrolled in grade 4 to 12. Results are adjusted for guessing as described in Section 2.4. * p<0.10 ** p<0.05 *** p<0.01

In addition to looking at performance on each individual question by current grade level, we use IRT to develop a numeracy score that incorporates responses to all questions and adjusts for question difficulty, as discussed in Section 2.4. Recall that we normalize the scores to have a mean of 0 and a standard deviation of 100 for grade 1 students in the year 2000 to get to the standardized numeracy score. Figure 2.4 shows the score gains from an additional year of schooling from grades 2 to 12, relative to grade 1, using data from 2014. We control for gender, whether the child's mother completed junior secondary school, wealth quintile, and province. The controls do not alter these results much (see Figure 2.5 for the 2014 learning profile without controls), so differences in student composition across the grades in terms of these background characteristics do not explain the differences in the standardized numeracy score across grades.

We find that the standardized numeracy score improves by 119 points between grade 1 and grade 12 – over a full standard deviation gain throughout a child's entire schooling. Putting this result in context, if we consider what type of trajectory we would expect of a student meeting grade-level expectations, a grade 5 student who was able to correctly answer the relatively easy version of the test (five items that are at grade levels 1-4) correctly would have a score of 238, or more than a 2 standard deviation improvement. In this case, the improvement of 88 points from grades 1 to 5 is only a third of the improvement in the score that we would expect if all students learned these basic skills. Given that these items reflect content covered in grades 1 to 5, it is not surprising that most learning takes place during primary school. Between grades 2 and 7, there is an approximate 15-point improvement per grade, or almost a fifth of a standard deviation per grade, compared to an approximate 6-point improvement per grade in grades 8 to 12.

Figure 2.4: Change in Standardized Numeracy Score Due to an Additional Year of Schooling Controlling for Gender, Mother's Education, Wealth Quintile, and Province



Source: IFLS 5 (2014)
Note: Point estimates and 95 percent confidence interval for progress in numeracy score relative to grade 1. Standard errors are corrected for clustering at the enumeration are level and observations are weighted using survey weights. The controls are gender, a dummy indicating if the child's mother's completed at least junior secondary school, wealth quintile (poorest 40 percent, middle 40 percent or wealthiest 20 percent) and province dummies. Results are adjusted for guessing as described in Section 2.4.

2.5.2 Did Learning Change Over Time?

Because IFLS asked the same questions across survey rounds, it allows us to observe changes in learning between 2000 and 2014. When we apply survey weights, our results for the full sample of respondents between 7 and 18 years old are representative for that population. Table A2.1 shows the balance of the weighted sample between 2000 and 2014. The survey population changed minimally between 2000 and 2014. There were no or very small differences in the gender ratio, age, or distribution of the sample across provinces over time; the main difference was that the population stayed in school longer and was somewhat wealthier.

Figure 2.5 shows the IRT results for enrolled students and all (in-school and out-of-school) students. The solid lines show the enrolled students' performance using the standardized numeracy score performance by grade and year. There are negative values in 2014 because we show learning levels relative to the 2000 grade 1 mean, which is standardized to be 0. It does not mean that there is negative learning, but means that the 2014 grade 1 students performed less well on the test than the 2000 grade 1 students. The striking finding in Figure 2.5 is that the slopes in 2000 and 2014 are nearly identical, with learning levels slightly higher in 2000. This difference between 2000 and 2014 is statistically significant, as shown in Table A2.3.[11] Describing this another way, a grade 6 student in 2014 performed at the same level as a grade 4 student in 2000.

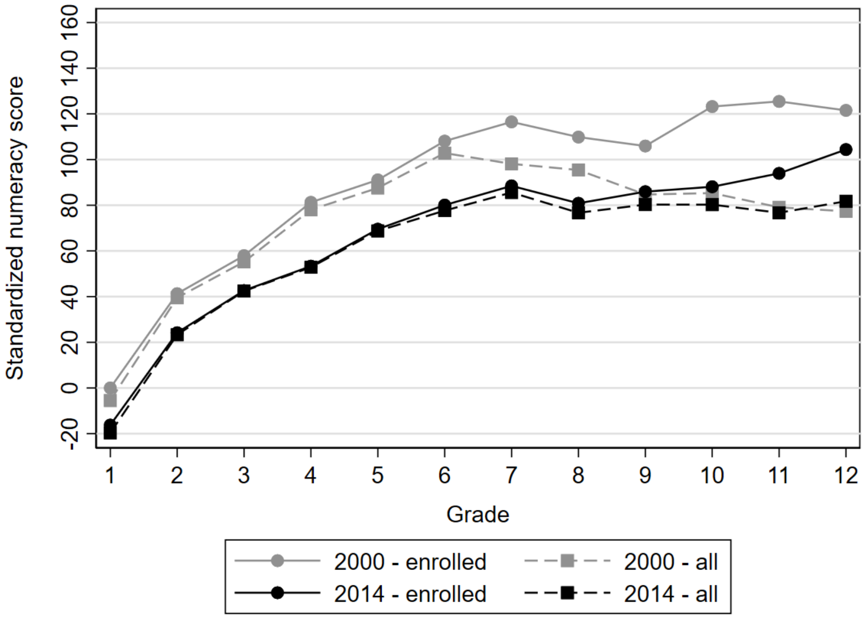
The dotted lines in Figure 2.5 show performance for all children, including out-of-school children, using the standardized numeracy score performance by grade (or the grade they would have been in for their age) and year. We include unenrolled children in this analysis to help answer the question of whether the results we see could be driven by a change in enrollment over time. Enrollment increased between 2000 and 2014, and it increased most for relatively poor children whose mothers completed less than nine years of schooling (not shown). Therefore, the composition of enrolled students is different in 2014 than in 2000, and one might hypothesize that the decline in learning between 2000 and 2014 is at least partly explained by this composition effect.

The enrollment rate for primary school, i.e. grade 1 to 6, has been nearly universal since before 2000, so the lower numeracy score in 2014 cannot be driven by selection. We can see this in Figure 2.5 because the dotted and flat lines for both years are nearly identical for grades 1 to 6. For the secondary schools, as shown in Figure 2.1, junior secondary school (grades 7-9) enrollment increased by 20 percentage points (from 70 percent to 90 percent) during this time frame; and senior secondary school (grades 10-12) enrollment increased by 24 percentage points, rising from 47 percent in 2000 to 71 percent in 2014. Figure 2.5 reflects this trend as the 2014 dotted and straight lines are nearly identical through grade 9 whereas the 2000 lines diverge more beyond grade 6.

[11] As a robustness check, we checked whether this result is driven by differential item functioning between the years. This is not the case. Results are available upon request.

Figure 2.5 shows that learning declined for all children, including enrolled students, between 2000 and 2014, indicating that this difference is not driven by a change in the student composition due to increased enrollment because there is a consistent difference in learning between the years when we include all children. The difference between 2000 and 2014 is also not driven by our imputation method. Figure A2.1 shows that we also find a decline in learning if we do not impute or if we consider missing answers as random guessing.

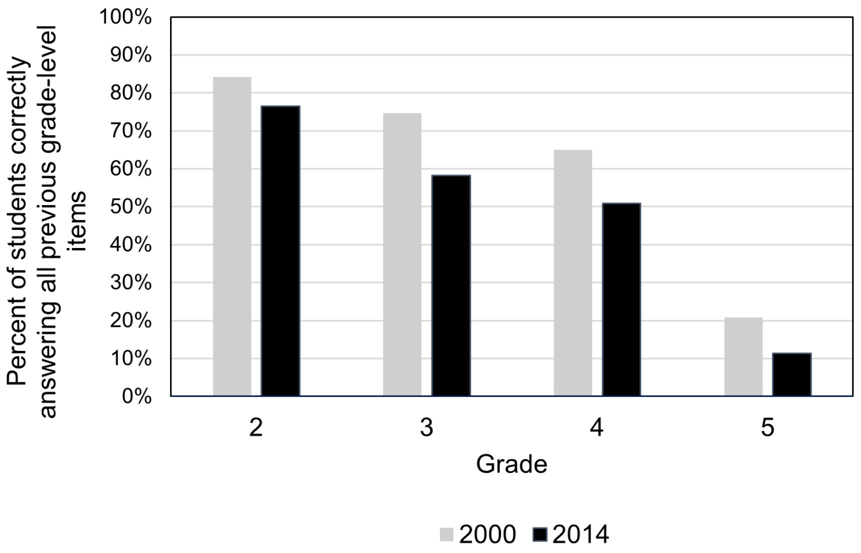
Figure 2.5: Standardized Numeracy Score in 2000 and 2014 by Grade Level Completed (for Enrolled Children) or Grade Level They Would Have Completed (for All Enrolled and Unenrolled Children)



Source: IFLS 3 (2000) and IFLS 5 (2014)
Note: Results are adjusted for guessing as described in Section 2.4.

Another way of examining the change in learning over time is to simply look at the share of students answering all relevant grade-level questions correctly. Figure 2.6 shows that this share is lower for students in every grade in 2014 compared to 2000. For example, we expect that a 4th grader would be able to answer questions for grade 3 and below. In 2000, the share of students who could do this was 65 percent; by 2014, 51 percent of 4th graders answered all grade 1, 2, and 3 level questions correctly. Figure 2.6 also demonstrates that the decline is not due to a single item since we see this trend across items; and the results are consistent across grade levels.

Figure 2.6: Percent of Students Who Answered Items Appropriate to Their Grades in 2000 and 2014



Source: IFLS 3 (2000) and IFLS 5 (2014)
Note: Expected grade-level mastery is described in Table 2.1. Figure shows percentage of students enrolled in each grade that correctly answered all items with an expected grade-level mastery below their enrolled grade. Results are not adjusted for guessing as this analysis involves combining items at the respondent level rather than looking at group means that reflect the percent correct of specific items.

Above we considered whether learning improved over time for different cohorts of students. Because IFLS is a panel dataset, we can also examine changes in learning among the same respondents in the 2000, 2007, and 2014 surveys, i.e., we can construct a panel learning profile.[12] In Table 2.3, we look at learning among children who were enrolled in grades 1 to 5 in either 2000 or 2007, who were also tested seven years later. The "gain" columns show the change in the standardized numeracy score over seven years of schooling for those individuals who were part of the panel, i.e., for whom the survey followed over time. For example, those students who were in grade 1 in 2000 gained 62 points between 2000 and 2007.

Consistent with Figure 2.5, we first find that on average children progressing through grades 1 to 8 between 2000 and 2007 learned more than the children progressing through the same grades between 2007 and 2014. Learning went down over time. The average gain over seven years for the 2000 cohort was 86 points, whereas this gain was 55 points or half a standard deviation, for the 2007 cohort. The smallest gains were for the older children, i.e., the children in more advanced grades than grades in which much of the material tested would have been taught.

[12] We do not consider the 2007 survey in any other analysis in this chapter since 2007 is more of a midterm result and does not add to existing information about the learning decline other than to confirm it.

We find that the panel results shown in Table 2.3 are much lower than the non-panel results shown in Figure 2.4, meaning that this causal learning profile is flatter than the contemporaneous cross-section profile we show in Figure 2.4. For several cohorts, the change in learning for the non-panel students is double that of the panel students. This indicates that the actual changes in learning were even lower than those shown using the descriptive profile. Because the contemporaneous cross-section profiles are declining, it is logical that the panel profiles demonstrate even lower learning gains.

Table 2.3: Change in Mean Standardized Numeracy Score Between 2000, 2007 and 2014, Among Panel Respondents

"Baseline" grade	"Endline" grade	Gain in numeracy score	
		2000-2007	2007-2014
1	8	86.1	54.5
2	9	57.9	47.6
3	10	55.0	29.6
4	11	39.1	18.4
5	12	43.1	15.4

Source: IFLS 3, 2000, IFLS 4, 2007, and IFLS 5, 2014
Note: Baseline is the year 2000 in column 3 and the year 2007 in column 4, while the endline is the year 2007 in column 3 and the year 2014 in column 4. Results are adjusted for guessing as described in Section 2.4.

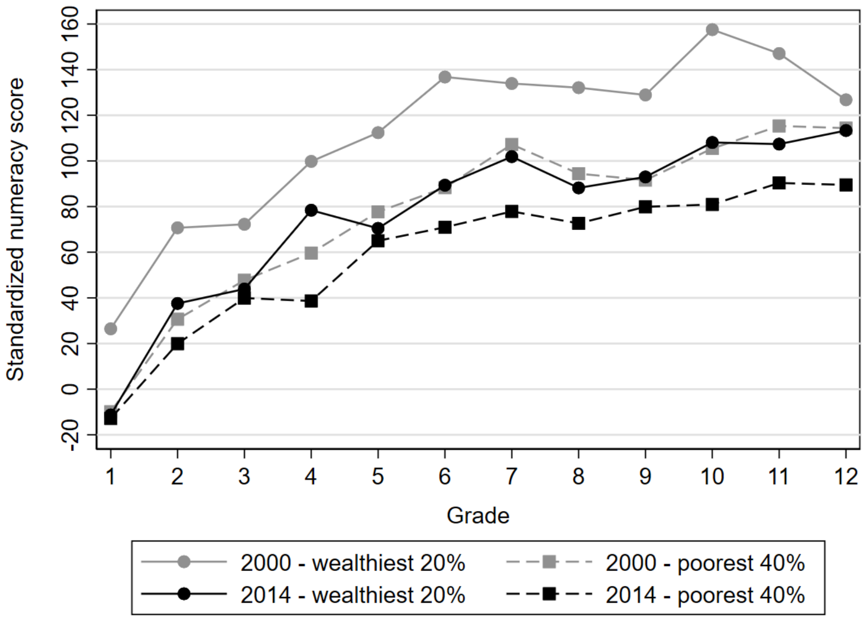
2.5.3 Did Different Subgroups Demonstrate Different Learning Profiles?

In addition to looking at learning progress for all children together, we investigate how learning varied across different groups of children, specifically how it varied by gender, wealth quintile, mother's education level, and province. We also compare differences in learning over time with changes in enrollment between subgroups in order to explore whether the decline in learning could have been due to changing enrollment. We show these results for enrolled students only as the primary focus of this analysis is what children are learning from the education system. Our findings do not differ significantly when we include out-of-school children. For the analysis in this section, we calculate the subgroup differences by regressing the numeracy score on the subgroup and grade dummies (Table A2.4). Column 1 in Table A2.3 presents the result of a regression of the standardized numeracy score on each of the subgroups and grade dummy variables in 2014 to show the coefficients and significance levels of the differences in that year.

In Figure 2.4, we showed that the standardized numeracy score declined overall between 2000 and 2014. We ask whether this decline was different for different subgroups looking first at the difference between the wealthiest 20 percent and the poorest 40 percent as shown in Figure 2.7. We determined these wealth categories within each year. The rich-poor gap declined markedly between 2000 and 2014. The mean rich-poor gap per grade was 37 points (about a third of a standard deviation) in 2000 and it went down to 17 points in 2014. As to be expected given the Figure 2.4 results, learning declined for both groups. This decline was greater for the wealthier group (Table A2.4). The mean 2000 to 2014 decline per grade was 36 points for the rich and 16 points for the poor (Table A2.4). The results for the rich in 2014 were very similar to the poor in 2000.

We posit that the 2000 to 2014 decline is a learning effect rather than an enrollment effect due to changes in student composition because the wealthier group saw a smaller change in enrollment than the poor group, and yet learning still went down for the wealthiest students. Between 2000 and 2014, enrollment rose for the wealthiest 20 percent by 8 percentage points in junior secondary school and 13 percentage points in senior secondary school while these figures were 27 and 30 percentage points respectively for the poorest 40 percent. If we consider results for all children (not shown), including unenrolled children, we find a similar pattern.

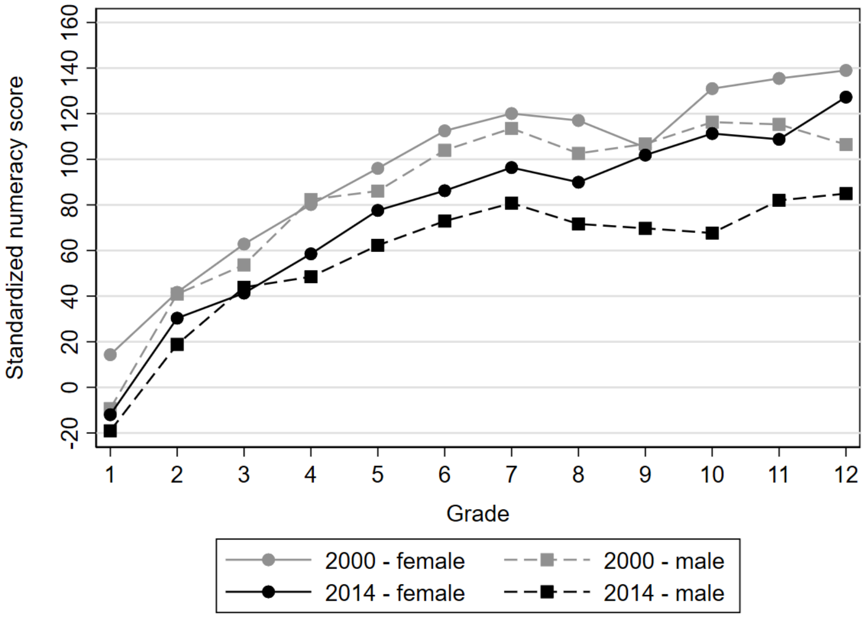
Figure 2.7: Standardized Numeracy Score for Poorest 40 Percent and Wealthiest 20 Percent in 2000 and 2014



Source: IFLS 3 (2000) and IFLS 5 (2014)
Note: Results are adjusted for guessing as described in Section 2.4.

Figure 2.8 shows similar results by gender. We see that scores declined for both females and males from 2000 to 2014, but that males saw a larger drop, and that the male-female gap also widened between 2000 and 2014. The average male-female difference in each grade was 10 points in 2000, and this rose to 18 points in 2014 (with females consistently scoring higher). The average decline in scores in each grade from 2000 to 2014 was 20 points for females and 27 points for males (Table A2.4). This was especially high for males after grade 6, where the 2000 to 2014 difference was 34 points. We do not find a gender difference in attainment over time for primary or junior secondary school. The senior secondary graduation rate difference by gender declined over time; by 2014 the male senior secondary graduation rate was four percentage points higher than that for girls. Thus this gender difference in learning was unlikely due to gender differences in enrollment. Enrollment went up by 14 percentage points for males and 20 percentage points for females in junior secondary school over this timeframe; it rose by 23 percentage points for both genders for senior secondary.

Figure 2.8: Standardized Numeracy Score for Females and Males in 2000 and 2014

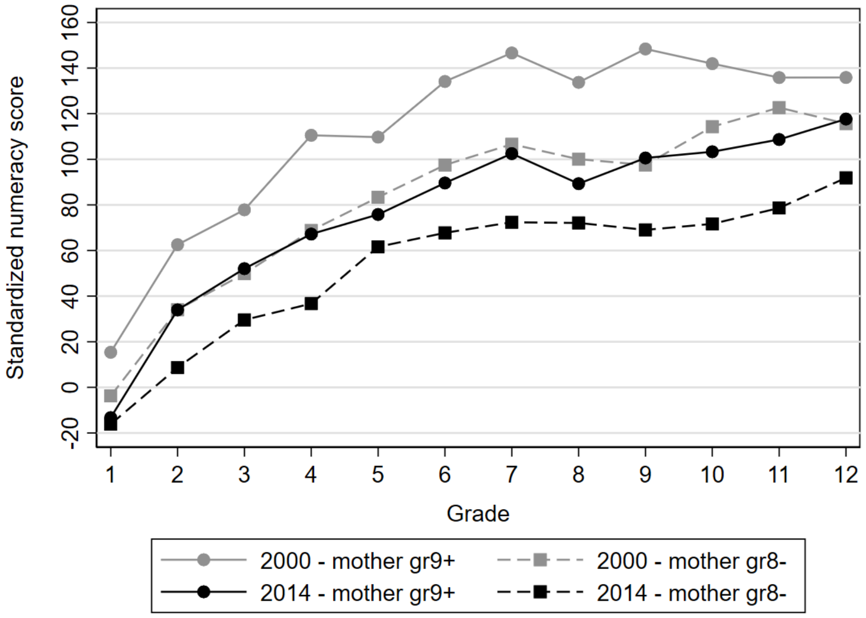


Source: IFLS 3 (2000) and IFLS 5 (2014)
Note: Results are adjusted for guessing as described in Section 2.4.

Given that mothers' education is a strong predictor of educational outcomes (see for example Suryadarma et al., 2006), we also consider how results differ for children whose mothers have different levels of schooling (Figure 2.9). We use junior secondary school as a cut-off such that we look at differences between children whose mothers completed junior secondary school (grade 9) or above compared to children whose mothers completed less than junior secondary school (grade 8 or below). Consistent with the other figures, we find a decline in learning for both groups over time. The decline is slightly larger for children with mothers with more schooling. Between 2000 and 2014, mean learning within each grade decreased by 36 points for students with mothers who completed at least junior secondary school while it decreased by 28 points for students with mothers with less schooling (Table A2.4). The gap between students with mothers who completed at least junior secondary school and students whose mothers completed less schooling decreased from 31 points in 2000 to 24 points in 2014. Interestingly, learning levels among students with mothers with less schooling in 2014 were nearly identical to students with mothers with more schooling in 2000.

As shown in Section 2.2.1, average years of schooling rose during the 14-year study period, so the share of mothers with a junior secondary degree or above also rose, from 24 percent of students in 2000 to 53 percent in 2014 (Table A2.1). Among children with a mother with a junior secondary degree or above, in 2000, 98 percent of their children were enrolled in junior secondary school (and 93 percent in senior secondary); which confirms that the decline in learning is not due to enrollment changes, at least for this group.

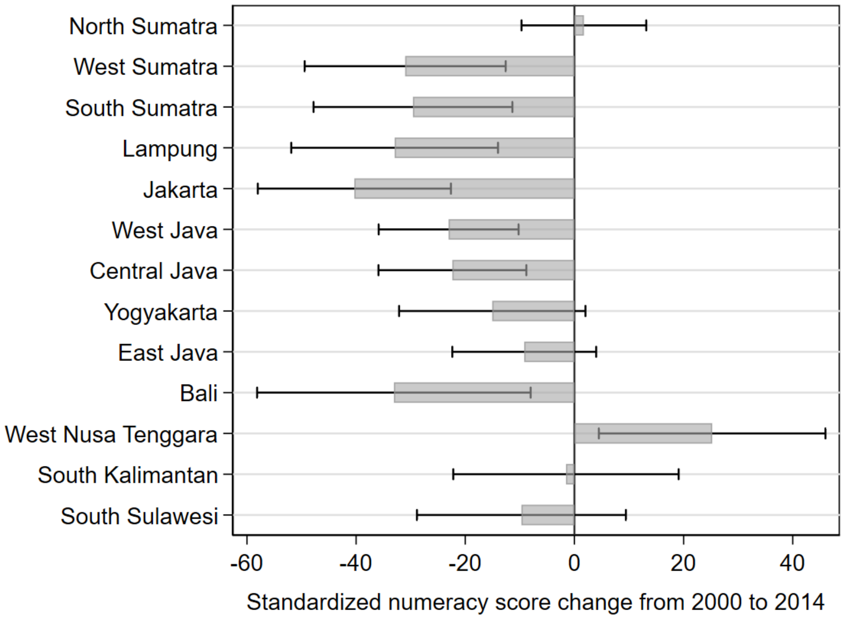
Figure 2.9: Standardized Numeracy Score for Children Whose Mothers Completed Grade 9 and Above and Whose Mothers Completed Grade 8 or Below in 2000 and 2014



Source: IFLS 3 (2000) and IFLS 5 (2014)
Note: Results are adjusted for guessing as described in Section 2.4.

Because Indonesia is incredibly diverse and we might expect a diversity in learning outcomes in different parts of the country, we also consider regional differences, shown in Figure 2.10. IFLS includes 13 out of 27 provinces and is representative at the province level for the provinces surveyed. Figure 2.10 shows the change in standardized numeracy test score results for all available provinces. We present the coefficients β_3 as estimated using Equation 2.3 in Section 2.4 for all the 13 provinces that are represented in the IFLS survey. These are the coefficients of the interaction terms between the dummy variable for the 2014 IFLS wave and each of the provinces, showing the difference in the standardized numeracy score between 2000 and 2014 within each province. Not surprisingly, there was a great diversity in mean standardized numeracy scores in 2000. They ranged from 19 points in West Nusa Tenggara to 119 points in West Sumatra, with a mean of 82 points across provinces. We find that scores declined in all but three provinces. Only one province, West Nusa Tenggara, which had the lowest baseline score, saw a positive and significant difference; declines were significant for 7 out of 13 provinces. In Jakarta, which started with an average score of 109 in 2000, the average score declined up to 40 points, or a bit over a third of a standard deviation. Again, we find a larger decline for groups with initially higher scores. The provinces with a significant decline in the numeracy score had an average standardized numeracy score in 2000 of 92; the provinces with no change had an average initial score of 76.

Figure 2.10: Difference in Average Standardized Numeracy Score for Students Enrolled in Grade 1 to 12 From 2000 and 2014, by Province



Source: IFLS 3 (2000) and IFLS 5 (2014)
Note: Bars present the coefficients and black lines indicate the 95 percent confidence interval of separate regressions for each province of the standardized numeracy score on an indicator for 2014 and grade fixed effects, applying survey weights (β_3 in Equation 2.3). The standard errors are corrected for clustering at the enumeration area level. Results are adjusted for guessing as described in Section 2.4.

2.6 Discussion and Conclusion

Between 2000 and 2014, Indonesia witnessed major progress in junior and senior secondary enrollment, as shown in Figure 2.1: a growth of 20 percentage points in junior secondary school and 24 percentage points in senior secondary school. Average years of schooling completed among 18 to 24-year-old respondents went up by 1.4 years over this 14-year time frame. We find that simultaneous to this progress, learning levels remained low. For example, looking at the simplest question in our study, a grade 1 question, 49-23, 65 percent of students in grade 3 in 2014 were able to answer it correctly. None of the 5th graders answered a more difficult question, 1/3-1/6, a grade 4 question, correctly. We find that the disparity between subgroups in terms of ability grew as the questions grew in difficulty.

In a study that tested children in nine grades at two points in time, in 2011 and 2012, Afkar et al. (2018) also find similarly low levels of learning in Indonesia. Just 57 percent of children could correctly answer a one-digit multiplication question by the end of grade 3; 50 percent could order four-digit numbers from big to small by the end of grade 2; and 60 percent could recognize two-digit numbers by the end grade 2. PISA and TIMMS results also reinforce this finding of similarly low learning levels (OECD, 2019; Mullis et al., 2016).

We further show that learning declined over 14 years. This decline amounted to approximately one-fourth of a standard deviation based on a scale normalized to grade 1 learning levels in 2000. This decline was the equivalent of two grades of learning; the average grade 6 student in 2014 performed at the same level as the average grade 4 student in 2000. Comparing these results to international assessments, Indonesia's TIMSS scores declined for grade 8 mathematics between 2003 and 2011 (Luschei, 2017). In PISA, mathematics scores over a similar timeframe (2003 to 2018) improved by just a few points on average over the six PISA tests that Indonesia participated in (OECD, 2019).

A critical outstanding question is why learning declined. There are several reasons we reject the hypothesis that it declined due to the changes in enrollment. First, we see a decline in learning at the primary level while primary school enrollment was basically universal by 1988. If there was a compositional effect at higher grades, we would expect to see differences in the decline in these grades compared to primary – which we don't.

Second, looking at the entire population (in- and out-of-school children) across all ages, we still see a decline, as shown in Figure 2.5; so there wasn't a selection effect. The decline for the children in school is greater in magnitude than the improvement in learning for the children who entered school and wouldn't have otherwise. Taking all 18-year-old respondents in 2014, using 2014 enrollment levels but the 2000 learning profile, we would expect them to have an average standardized numeracy score of 100; but instead they have an average score of 73 due to the declining learning profile. It is possible that learning for in-school children declined due to increased enrollment because more students stressed the system (and thus lowered quality for all) or due to peer effects from new learners who were not in school in 2000. However, our finding that learning also declined at the primary level where enrolment did not change between 2000 and 2014 makes the case against system stress or negative peer effects, unless those challenges were unique to grades 7 to 12.

Third, learning declined for nearly all subgroups, even those that had high levels of enrollment in 2000. For example, learning actually declined more for the wealthiest 20 percent than for the poorest 40 percent and for children with mothers with more education than for children with less education, despite the fact that enrollment changed less for these subgroups. Between 2000 and 2014, enrollment rose for the wealthiest 20 percent by 8 percentage points in junior secondary school and 13 percentage points in senior secondary school. Ninety-eight percent of children with a mother with a junior secondary degree (93 percent for senior secondary) were already enrolled in junior secondary in 2000 and enrollment for this group did not change much by 2014.

The learning decline is especially surprising given all the education system upgrades that took place over this timeframe. These include nationwide decentralization in 2001 to allow districts more flexibility with introducing innovative education policies and adjusting policy to reflect local context; the 2002 amendment to the Constitution that required 20 percent of the budget be devoted to education expenditures—resulting in a threefold increase in real education budget; and the 2005 teacher certification policy as a way to improve teacher quality. The increased budget allowed for a decline in the student teacher ratio during this period and one aspect of teacher quality, the share of teachers with bachelor's degree, rose from 37 to 90 percent (World Bank, 2018b).

However, many of these policies were not directly targeted at learning or specifically at improving foundational skills like the numeracy questions analyzed in this chapter. Given the mixed evidence of the impact of spending on learning, it is not guaranteed that the 2002 budget requirements on education spending would have had an impact on learning (World Bank, 2018b; Vegas and Coffin, 2015). Indeed, a study examining the impacts of the teacher compensation component of the teacher certification law of 2005 showed that it had no impact on learning (De Ree et al., 2018). Districts could use greater education policy autonomy to achieve goals that are not necessarily aligned with improving student learning, such as satisfying certain constituent demands for job opportunities within the school system.

What then could have caused the learning decline? In the absence of a causal study, we only have several conjectures. First, as mentioned in Section 2.2.1, children's exposure to math changed over this timeframe. The 1994 curriculum mandated 10 hours a week of math instruction for grades 1 to 3 and eight hours a week for grades 4 to 6. In 2004, the curriculum was to be taught "thematically" for grades 1 and 3 and instruction time went down to five hours per week for grades 4 to 6. Of course it is possible that thematic teaching was a more efficient and holistic way of learning; but cutting math instruction time in half could potentially have an effect on learning.

Second, related to dosage or exposure to material, grade repetition went down by 38 percent (from 17 percent in 2000 to 11 percent in 2014), indicating that perhaps students who might have needed more support by repeating a grade would have been able to in 2000 but not in 2014 (authors' analysis with IFLS, not shown). By 2014 fewer children were behind grade level and more children were either at the appropriate grade level for age or ahead (meaning young for their grade) compared to 2000. For the richest 20 percent, the percent of students repeating a grade dropped from 14 to 6 percent, and for the poorest 20 percent, this only declined from 19 to 17 percent. Thus it is possible the decline in grade repetition for the rich contributed to the learning decline, although we would not expect this to have a very large overall effect given that the decline across all groups was 6 percentage points.

Third, class grades became less important which could have affected student incentives to learn. Prior to 2003, a student graduated from 6th, 9th or 12th grade based on yearly grades and national exam results. After 2003, grades were less important as graduation was determined by a combination of teacher discretion and national exam results. During this timeframe, districts took over responsibility for the grade 6 leaving exam, so the content varied by district. Thus the weight of exams in graduation could have affected incentives for learning during the 2000 to 2014 timeframe.

Consistent with many studies outside of Indonesia, importantly the World Bank's World Development Report 2018 (World Bank, 2018b), this study makes it clear that rising enrollment does not necessarily translate to improved test performance. Indonesia took costly measures to address education challenges over the 2000 to 2014 timeframe and yet not only did learning not improve but it declined. This study shows that policy should more carefully explore and target the major barriers to learning, which appear not to be financing, teacher qualifications, or teacher-student ratios; they could be the duration of exposure to mathematics or incentives to learn, but more study is needed to uncover the primary barriers to improving learning. Moreover, this study emphasizes the importance of comparable, low stakes exams that ask similar questions over time for monitoring purposes. We hope that this study will encourage more government-supported outcomes monitoring, a key starting point to any strategy that seeks to transform education systems and prioritize learning.

Appendix

A2.1 Balance Between the 2000 and 2014 Sample

Table A2.1 shows the difference in characteristics between the sample included in the 2000 IFLS sample and the 2014 IFLS sample. Applying the sampling weights, the samples are representative for the population between 7 and 18 years old in the 13 provinces in each of these years. Since the population can change over time, we do not expect the samples to be the same. The sample in 2014 is slightly younger than the one in 2000 (0.2 years), they completed half a year of schooling more, and 30 percentage points more mothers completed at least junior secondary school. The population also improved their wealth with 0.2 standard deviation. The gender ratio and the distribution of the sample across the provinces remained virtually the same. Note that we standardize the asset index and determine the wealth quantiles separately in each year at the household level. Since there can be multiple respondents in one household, the fraction can be slightly different at the individual level.

Table A2.1: Balance Between the IFLS Sample in 2000 and 2014

	2000	2014	Difference
Age in years	12.41	12.23	-0.18***
	(3.49)	(3.28)	(0.05)
Fraction male	0.52	0.52	-0.00
	(0.50)	(0.50)	(0.01)
Completed years of schooling	4.99	5.50	0.51***
	(3.31)	(3.22)	(0.07)
Fraction of mothers that completed at least junior secondary school	0.24	0.53	0.29***
	(0.43)	(0.50)	(0.01)
Standardized asset index	0.06	0.23	0.17***
	(0.96)	(0.83)	(0.03)
Fraction living in [. . .]			
North Sumatra	0.06	0.07	0.01***
	(0.23)	(0.25)	(0.00)
West Sumatra	0.04	0.04	-0.00
	(0.19)	(0.19)	(0.00)
South Sumatra	0.04	0.04	-0.00
	(0.20)	(0.20)	(0.00)
Lampung	0.04	0.03	-0.01
	(0.20)	(0.18)	(0.00)
Jakarta	0.05	0.05	0.00
	(0.21)	(0.22)	(0.00)
West Java	0.28	0.24	-0.04***
	(0.45)	(0.43)	(0.01)
Central Java	0.15	0.17	0.02**
	(0.36)	(0.37)	(0.01)
Yogyakarta	0.05	0.04	-0.01*
	(0.22)	(0.21)	(0.00)
East Java	0.19	0.19	0.00
	(0.40)	(0.40)	(0.01)
Bali	0.02	0.02	0.00
	(0.12)	(0.13)	(0.00)
West Nusa Tenggara	0.03	0.03	0.00*
	(0.16)	(0.17)	(0.00)
South Kalimantan	0.03	0.03	0.00
	(0.17)	(0.17)	(0.00)
South Sulawesi	0.04	0.04	0.00
	(0.19)	(0.20)	(0.00)

Source: IFLS 3, 2000, and IFLS 5, 2014

Note: Table includes all respondents between 7 and 18 years old, and respondents older than 18 years that are still enrolled in senior secondary school. Values are weighted using the sampling weights. Standard errors in parentheses and corrected for clustering at the EA level. * $p<0.10$ ** $p<0.05$ *** $p<0.01$

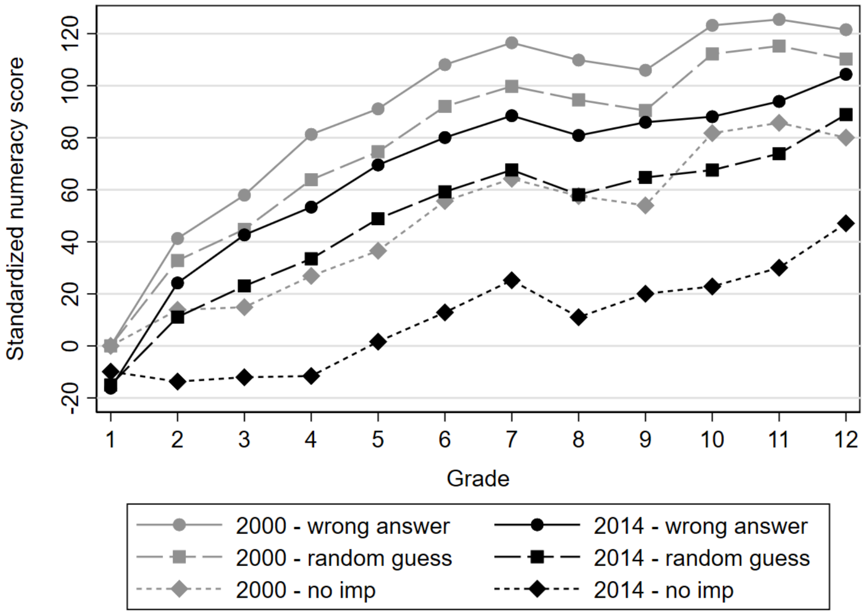
A2.2 Different Imputation Methods as Robustness Checks

We conduct several tests to assess the robustness of our findings to different imputation specifications. Our primary results are presented using imputations of wrong answers for (partial) missing cases, meaning that we assume that a student did not know the answer to the question if he or she left the field blank. We think that the latter is likely to be the case, because there are more missing values amongst younger kids and more difficult items (see Table A2.2). In Figure A2.1, we present our primary approach (impute with wrong answers or 0), the standardized numeracy score when not imputing missing values and the standardized numeracy score if we would impute with random guessing. Children that did not know the answer to the question could make a guess instead of leaving the field blank. When a question has 4 answer options, we impute 25 percent of the missing values randomly with a correct answer.

Overall, we find that results from our primary approach are similar to results without conducting any imputation and to results when imputing missing values with random guessing. The other imputation methods result in a somewhat flatter learning profile, but in all cases most learning takes place between grade 1 and 6 and the learning profile declines between 2000 and 2014.

The learning profile from our primary imputation approach is steeper, because ignoring missing values and imputation with random guessing inflate scores of children in lower grades. We standardize such that the grade 1 mean is 0 and the grade 1 standard deviation is 100. There are more missing answers for students in lower grades, especially for grade 1 students, so if we assume that all missing values are wrong answers, it makes sense that we find more learning over grades than with the other methods.

Figure A2.1: Results When Using Different Imputation Methods



Source: IFLS 3 (2000) and IFLS 5 (2014)
Note: Results are adjusted for guessing as described in Section 2.4.

Table A2.2: Fraction Missing by Item and Age

Item / Age	7	8	9	10	11	12	13	14	15	16	17	18
G1: 49-23	0.19	0.09	0.05	0.04	0.03	0.02	0.02	0.02	0.01	0.01	0.02	0.02
G2:267+112-189	0.26	0.13	0.07	0.05	0.04	0.03	0.03	0.04	0.02	0.02	0.03	0.02
G3: (8+9)*3	0.33	0.17	0.10	0.07	0.05	0.04	0.04	0.04	0.03	0.02	0.04	0.03
G3: (412+213)/(243-118)									0.07	0.07	0.08	0.08
G4: 56/84	0.45	0.34	0.21	0.16	0.10	0.08	0.06	0.06	0.05	0.04	0.07	0.05
G4: 1/3-1/6	0.45	0.33	0.19	0.14	0.10	0.07	0.05	0.05	0.04	0.03	0.05	0.04
G4: 0.76-0.4-0.23									0.07	0.07	0.09	0.09
G5: Percent 1									0.08	0.08	0.09	0.09
G5: Percent 2									0.08	0.07	0.08	0.08

Source: IFLS 3, 2000, and IFLS 5, 2014

A2.3 Regression Analysis of Subgroup Differences and Differences over Time

As part of our subgroup analysis, we also use regression analysis to examine what factors might explain learning differences among children in the same grade. Table A2.3 shows that the differences between the subgroups and over time that we described in Section 2.5.3 are significant.

We test the significance of the sub-group differences and differences over time using three regressions.

First, we test the significance of the differences between the subgroups in the 2014 IFLS wave by regressing the standardized numeracy score on subgroup indicators, controlling for grade in which the student is enrolled and weighting the observations using the sampling weights, as shown in Equation 2.4 for individual i from province p and grade g,

$$Y_{i,p,g} = \beta_1 + \beta_2 MALE_{i,p,g} + \beta_3 SES_{i,p,g} + \beta_4 MOTHEDUC_{i,p,g} + \phi_p + \gamma_g + \epsilon_{i,p,g} \tag{2.4}$$

in which Y is the standardized numeracy score that follows from IRT. MALE, SES and MOTHEDUC are dummy variables indicating the subgroups, ϕ_p are province fixed effects, γ_g are grade fixed effects and ϵ is an error term. We estimate the model using Ordinary Least Squares (OLS) and the standard errors are corrected for clustering at the enumeration area level.

Second, we test the significance of the difference in the standardized numeracy score over time by including the 2000 IFLS wave and by adding a dummy for the 2014 IFLS wave to Equation 2.4. This way, we test whether the difference over time is significant while controlling for background characteristics and grade as shown in Equation 2.5 for individual i in IFLS wave w and grade g,

$$\begin{aligned} Y_{i,p,w,g} = \beta_1 + \beta_2 MALE_{i,p,w,g} + \beta_3 SES_{i,p,w,g} + \beta_4 MOTHEDUC_{i,p,w,g} \\ + \beta_5 W_{i,p,g} + \phi_p + \gamma_g + \epsilon_{i,p,w,g} \end{aligned} \tag{2.5}$$

in which W is a dummy variable for the 2014 IFLS wave.

Table A2.3 shows the results of the regression analysis. All subgroup differences in the standardized numeracy score are statistically significant in the 2014 sample, except for the difference between the 40% poorest and 40% middle SES students. The differences by gender and mother's education are the largest, where girls and students with mother's that completed at least junior secondary school scored about a fifth of a standard deviation higher on the numeracy test. The decline in the standardized numeracy score of enrolled students between 2000 and 2014 is 29 points and statistically significant, even when controlling for the background characteristics of students.

Table A2.3: Subgroup Differences in Standardized Numeracy Score in 2014 and the Difference in the Standardized Numeracy Score Between 2000 and 2014

	(1)	(2)	(3)
	Standardized Numeracy Score		
	Subgroup Comparison 2014	Difference over Time	Difference over Time with Controls
Male	-18.72***		-14.37***
	(2.29)		(1.60)
Poorest 40%	-	-	-
Middle 40%	5.00*		6.94***
	(2.61)		(2.24)
Wealthiest 20%	7.77***		16.07***
	(2.96)		(2.69)
Mother completed at least	21.27***		22.62***
junior secondary school	(2.71)		(2.15)
Year 2014		-23.70***	-29.46***
		(2.16)	(2.29)
Constant	-19.92***	2.75	-6.10
	(7.66)	(4.84)	(6.36)
Province Fixed Effects	Yes	No	Yes
Grade Fixed Effects	Yes	Yes	Yes
Years Included	2014	2000 and 2014	2000 and 2014
Observations	9133	16873	15993

Source: IFLS 3, 2000, and IFLS 5, 2014
Note: Standard errors in parentheses and corrected for clustering at the EA level. * p<0.10 ** p<0.05 *** p<0.01

Third, we test the significance of the difference in the standardized numeracy score over time for each of the subgroups by estimating the following equation for each of the subgroups separately,

$$Y_{i,w,g} = \beta_1 + \beta_2 W_{i,g} + \gamma_g + \epsilon_{i,w,g} \tag{2.6}$$

for student i from IFLS wave w in grade g. Again W is a dummy variable for the 2014 IFLS wave and we include grade fixed effects γ_g. Note that the grade fixed effects are allowed to differ between the subgroups. Also note that we estimate the same model for each of the provinces, for which we show the results in Figure 2.10 in Section 2.5.3.

The results in Table A2.4 show that the standardized numeracy score significantly declined for all subgroups. It declined more for boys, for wealthier students and for students whose mothers completed at least junior secondary school. With almost two fifths of a standard deviation, the standardized numeracy score declined most for the wealthiest 20% and for students whose mothers completed at least junior secondary school.

Table A2.4: Subgroup Differences in the Change in the Standardized Numeracy Score Between 2000 and 2014

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Standardized Numeracy Score						
	By Gender		By Wealth			By Mother's Education	
	Female	Male	Bottom 40%	Middle 40%	Top 20%	Less than JSS	At least JSS
Year 2014	-19.97***	-27.11***	-16.21***	-21.47***	-35.77***	-28.02***	-36.01***
	(2.72)	(2.68)	(3.32)	(3.16)	(3.88)	(2.83)	(3.22)
Constant	11.97*	-2.97	-5.31	-0.27	25.60***	0.07	19.81***
	(6.29)	(6.07)	(7.25)	(6.15)	(8.69)	(5.79)	(6.54)
Grade Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	8258	8615	6707	6433	3662	8918	7421

Source: IFLS 3, 2000, and IFLS 5, 2014

Note: Models include enrolled students in grade 1 to 12 in 2000 or 2014. Standard errors in parentheses and corrected for clustering at the EA level. JSS stands for junior secondary school. * p<0.10 ** p<0.05 *** p<0.01

Chapter 3

Using Technology to Prevent Fraud in High Stakes National School Examinations: Evidence from Indonesia[13]

3.1 Introduction

Cheating happens in high stakes school examinations, from the "cheating mafia" in India (Anderman, 2015) to fraudulent practices in prestigious high schools in the United States (Safi, 2018). Systemic cheating is difficult to eliminate, as all stakeholders benefit. Students and teachers do not have to exert much effort to attain higher grades. Since official grades are what teachers and bureaucrats are held accountable for, they may prefer to allow cheating. Honest test takers lose, as their results reflect poorly relative to cheating students and teachers. In addition, when honest test takers are the minority, the costs to report cheating practices are high and the chance that the authorities seriously attempt to reduce cheating is low (Borcan et al., 2017). Therefore, honest students and teachers may be compelled to also cheat. The resulting equilibrium is systemic and sustained cheating practices.

[13] This chapter is joint work with Menno Pradhan, Rahmawati, Daniel Suryadarma, and Arya Swarnata. An earlier version of this chapter was circulated under the title "From Cheating to Learning: An Evaluation of Fraud Prevention on National Exams in Indonesia". The authors thank the Ministry of Education and Culture for sharing their data. We would like to acknowledge Hessel Oosterbeek, Amanda Beatty, Lant Pritchett, our colleagues at the University of Amsterdam and the Vrije Universiteit Amsterdam, and the audience at the RISE Annual Conference 2019 and the IRSA Conference 2019 for their valuable feedback. We also thank Brian Arieska Pranata for his insights into cheating practices in Indonesia.

We evaluate the Indonesian government's flagship policy to eliminate cheating in national school examinations: computer-based testing (CBT). Cheating in the paper-based national exams in Indonesia was widespread. It has been reported in the popular press (e.g. Economist, 2011; Sundaryani, 2015; Jong, 2015), but there have been hardly any instances where it was prosecuted. Reported cheating ranged from students copying each other's answers to teachers and principals providing answer keys to students prior to or on the exam day. Anecdotes of teachers correcting students' answers before grading also exist.

In 2015, the extent of cheating became apparent when the central government began measuring and disseminating a school "integrity index" (Rahmawati and Asrijanty, 2016). The integrity index identifies cheating through suspicious answer patterns, a method that has been validated in schools in Chicago (Jacob and Levitt, 2003) and has been used in multiple studies to measure cheating in exams (Battistin et al., 2017; Martinelli et al., 2018). In Italy, a similar method was used to sanction schools for cheating on a standardized national test in primary and high schools (Lucifora and Tonello, 2020). Classrooms with identical answer strings or counter-intuitive performance on items of certain difficulty levels, such as scoring high on difficult items while incorrectly answering easier items, are given a lower integrity index. The index was shared with district governments, who are responsible for ensuring a fair examination in Indonesia's decentralized setting. The results revealed widespread cheating. One-third of the schools were flagged by the Ministry of Education as suspicious, compared to 5 percent of the classrooms in Chicago (Jacob and Levitt, 2003) and 5 percent in Italy (Angrist et al., 2017). In Mexico 7 percent of high school exams were flagged as suspicious (Martinelli et al., 2018), which increased to 32 percent after two years of monetary incentives based on test scores for students and teachers.

The Ministry of Education introduced CBT in 2015 with the explicit and singular aim to eliminate cheating (Rahmawati and Asrijanty, 2016). With CBT, the test items are drawn directly from a server, so test versions vary across students and across classrooms. Cheating is virtually impossible as teachers and students do not know the questions beforehand, and the variations of test versions are in the thousands. The different test items also void the students' ability to work together during the exam. In addition, teachers cannot change students' answers, because the computer program grades the exam. Although this mode of CBT is not new (for instance, see Wang et al. (2008) for CBT use in the United States), implementing CBT on a national scale in a developing country is exceptional. The program started with 40 junior secondary schools in 2015. By 2019, 78 percent of Indonesia's junior secondary schools (43,841 schools with 3,554,556 exam takers) participated in CBT. It is also resource intensive and technically complicated.

Evaluating its impact would provide insights for other countries that may be interested in adopting the approach to improve the integrity and measurement of national school examinations.

We implement a difference-in-difference analysis for each cohort of schools that switched to CBT between 2017 and 2019 using publicly available data at the school level on the average exam score, the variance, the number of students taking the national exam and the integrity index. By construction, the integrity score is only available for schools using the paper-based test.[14] We use the Callaway-Sant'Anna approach (Callaway and Sant'Anna, 2020), taking heterogeneous treatment effects and treatment anticipation into account.

We find that school level exam scores decreased on average by 5.4 points (0.4 standard deviation) in the first year of participation in CBT. We also find that the negative impact on the mathematics score was larger than on the Indonesian or English score. To confirm that this effect is due to a reduction in cheating rather than a change in the test taking mode, we split the sample by high and low integrity (defining low integrity as being flagged by the Ministry) and availability of a computer lab in 2015 (used as a proxy for familiarity with working on computers). We find a much larger drop in scores for low integrity schools (-8.8 points) than for high integrity schools (-2.1 points), and availability of computers made little difference to the estimates, indicating the the drop in test score was indeed because of reduced cheating.

We also find that the standard deviation in test scores within schools increased with 0.5 points as a result of CBT compared to a within-school standard deviation of 5.5 in comparison schools, suggesting that the CBT method was better able to distinguish between high and low performing students. Finally, we show that the correlations between exam scores and district characteristics are more in line with expectations based on the literature after schools adopt CBT. These findings indicate that the exam results under CBT provide a better signal of true learning levels.

In districts where CBT was implemented at a faster pace, the integrity index of schools that were still doing paper-based tests rose faster and their test scores declined more. The estimates indicate that the integrity index of schools still taking the paper-based test increases by 1 point (on a 1-100 scale) as CBT implementation among schools in the same district expands by 10 percentage points. These findings suggest that the roll-out of CBT affected local cheating practices, creating a small spillover effect on comparison schools. This spillover effect could arise from the fact that exams are proctored by teachers from other schools in the same district. Teachers from schools that switched to CBT may have become stricter when proctoring schools that conduct paper-based exams to ensure a fair

[14] The algorithm checks whether students copy answers, which is impossible for CBT exams as all students receive different questions.

competition. Another reason could be that honest students and teachers in paper-based exam schools no longer feel high pressure to cheat as scores in nearby CBT schools have dropped. Finally, with more schools participating in CBT, exam answer keys are harder to acquire. Data availability limits our scope to test which of these potential explanations actually take place.

We perform two robustness checks to correct for these spillover effects. We estimate our model holding the integrity index of the comparison schools constant and we allow the trend of the comparison group to vary with the share of schools that implement CBT in the district. As expected, this increases the impact estimates of CBT somewhat, particularly for the later years.

We do not find evidence for improvements in exam scores within three years of implementation, which would point to a shift in focus from cheating to learning. With the correction for spillovers, none of the cohorts indicate that the impact of CBT diminished over time. Note that the longest trends of three years is for the 2017 cohorts, where cheating was relatively low. Unfortunately, the government cancelled national exams in 2020 and 2021 due to the COVID-19 pandemic, preventing us from estimating longer-term effects of CBT on test scores.

The switch to CBT is highly cost-effective. The annual cost of the national exam administration declined from about 9.2 million dollars to 2.4 million dollars, because printing and distributing the exams on paper was no longer necessary (Siddiq, 2018). Additional cost savings are also enjoyed by universities and employers, which can rely on national exams as an accurate measure of learning achievement. Although the intervention requires a significant upfront investment in computers, internet, and servers, these costs are mostly fixed. Moreover, the computers could also support teaching and learning activities outside of exam periods.

This study makes the following contributions. First, it adds to a small literature on the effects of programs aimed at reducing cheating in schools. These programs include cameras in classrooms in Romania (Borcan et al., 2017), random assignment of external monitors in Italy (Bertoni et al., 2013; Lucifora and Tonello, 2020), centralization of grading in New York (Dee et al., 2019) and tablet-based testing in India (Singh, 2020b). All these studies found that the programs reduced cheating and, in turn, test scores. The main difference between those studies and ours is the scale of the effort to reduce cheating. We examine a national-level program that affects around 5 million students. We show that it is possible for a government of a developing country with widespread cheating to substantially reduce cheating in a high stakes national exam with the utilization of technology.

Second, we provide suggestive evidence that the introduction of CBT changed the norms with respect to cheating. As more schools could no longer cheat, other schools followed suit by also reducing cheating, either through peer pressure or voluntarily. These indirect effects are in accordance with findings by Bertoni et al. (2013), who find that external monitors in one classroom also reduced cheating in other classrooms in the same school without an external monitor. Moreover, our finding that computer-based exam scores did not improve over time when cheating is no longer possible relates to a literature on effects of high stakes exams. Contrary to this chapter, these studies generally find that the introduction of high stakes testing improves learning outcomes in contexts with little cheating (e.g., Jacob, 2005; Angrist and Lavy, 2009). Perhaps, schools need more time or school resources in initially cheating schools are insufficient to achieve improved student learning. Our findings also speak to the broader literature on group norms and enforcement (Feldman, 1984; Galbiati et al., 2021).

The rest of the chapter proceeds as follows. In the next section we provide background information on the Indonesian national examination. Section 3.3 describes the data and Section 3.4 explains our empirical strategy. We report on the impact of CBT on exam scores in Section 3.5 and we discuss the results in the final section.

3.2 The Indonesian National Examination

The Indonesian education system implements national examination at the end of junior and senior secondary school (grade 9 and 12, respectively). Students take multiple choice exams in Indonesian, English, mathematics and science. Graduation has been independent from the national exam since 2015.[15] However, these exams remain high-stakes. The national exam score is used to determine admission into higher education levels.[16] This is especially true for the grade 9 exam, which we focus on in this chapter. Admission into senior secondary schools is highly influenced by the grade 9 exam, as the majority of seats in senior secondary schools are allocated based on grade 9 exam scores (Berkhout et al., 2022).

High exam scores are not only important for students, but also for schools and district governments. The score contributes substantially to school and local government achievement indicators (Economist, 2011). Although there is no legislation for holding

[15] In 2010, students had to score higher than 55 out of 100 on average across four subjects to graduate. Between 2011 and 2014, schools gained more autonomy in the graduation of their students when a composite score of the national exam and school exams determined graduation.

[16] Although this is true until 2019, admission into higher education levels has not been determined by the national exam since 2020. The 2020 exam was cancelled due to the COVID-19 pandemic and the national exam was replaced with a low stakes competency assessment in 2021.

schools accountable on their exam scores, local governments consider performance on the national exam as a matter of prestige. They put pressure on school principals and teachers to achieve high grades.

As argued by Neal (2013), using one assessment system to measure student achievement and school quality creates incentives to cheat for both the students and the educators. Anecdotal evidence indicates that cheating in national exams was indeed widespread in Indonesia (Economist, 2011; Jong, 2015). Students copied each other's answers or used answer sheets, which they illegally bought or received from the teacher. Not only did teachers allow these cheating practices to take place, they were active participants. The exam answer sheets were collected and scanned at the provincial level and graded centrally by the Ministry of Education (MoE), but the teacher could still interfere with the answer sheets beforehand, for example by correcting the wrong answers before they were sent to the provincial office.

Prior to 2015, the Government of Indonesia (GoI) tried to prevent cheating in the national exam by increasing the number of unique booklets in an exam room from two to five in 2011, and from five to 20 in 2013. However, students and teachers still managed to cheat. Therefore, since 2015, the GoI took additional measures with the aim to identify and reduce cheating.

First, the Center for Assessment and Learning (Pusmenjar or *Pusat Asesmen dan Pembelajaran*) of MoE develops an algorithm that generates a score to identify cheating at the school level, based on methods developed in the education literature (Hanson et al., 1987; Widiatmo, 2006; Van Der Linden and Sotaridona, 2006). The algorithm detects suspicious response patterns across students in the same schools and districts (Rahmawati and Asrijanty, 2016). It combines two cheating detection methods: (i) answer copying detection, where identical patterns of wrong and correct answers within a classroom or school are seen as an indication of answer copying and therefore increase suspicion of cheating; (ii) aberrant response detection, where unexpected patterns, for example consistently answering easier items incorrectly while getting more difficult items correctly, are seen as an indication of of cheating. The second method is performed because identical wrong answers could also result from teachers incorrectly teaching the concept that the item tests, which is not an indication of cheating. In addition, Pusmenjar adds more qualitative checking. First, Pusmenjar checks school exam results in previous years. A school that achieved uniformly correct answers would be suspected of cheating if it had a track record of poor performance. Second, the school-level integrity index is validated against a qualitative measure of school quality determined by respective provincial governments and against school accreditation reports. The methods produce an index which estimates the probability that a school cheated in the exam. An integrity index, which is the complement of the probability to cheat, is then calculated for each school.

The integrity index measures cheating on a continuous scale between 0 and 100, where a lower score means that there is more evidence for cheating. Pusmenjar considers an integrity index below 70 as low integrity, between 70 and 80 as fair integrity and above 80 as high integrity.[17] The integrity index is robust to type 1 errors, but it is prone to type 2 errors. This means that when the score is low, there is compelling evidence for cheating. At the same time, exam scores of schools with high integrity could still include cheating (Rahmawati and Asrijanty, 2016).

The GoI shares the results of the integrity index with district governments to signal that they do not only care about high grades on the national exam, but also about how the exam scores are achieved. However, the GoI does not implement sanctions based on the integrity index.

Second, the GoI implemented of computer-based testing. Students receive the exam items directly from a server with an item bank containing 30,000 items per subject each year. The system draws items from this bank, then gives them to the students in random order. Randomization happens both horizontally (i.e. different items across forms) and vertically (i.e. different order of items), such that each student in the exam room has a unique test version.

CBT prevents cheating in a number of ways. The test versions vary across students, classrooms and schools. This makes copying answers ineffective for students. In addition, neither teachers, school principals nor students have access to the test beforehand and answer sheets of the paper-based exams are useless. Finally, grading is done automatically as soon as a student completes an exam and encrypted student responses are sent directly to the central server of the MoE, so modification of the student responses by other parties is impossible.

Some parts of the test procedure remain the same as with paper-based testing (PBT). The paper-based and computer-based exams test the same competencies and are the same across Indonesia. The items for each of the 20 paper-based test versions are taken from the same item bank as the computer-based test versions.[18] In addition, both paper-based and computer-based exams are monitored by teachers from other schools in the district, who are randomly assigned by the district government. The teacher is not allowed to be in the classroom with his or her own students during the exam.

[17] These threshold values are based on the correlation between a change in the integrity index and a change in exam scores between 2015 and 2016. Pusmenjar found for schools with an integrity index above 80 in both years that exam scores do not vary much over time, while the exam scores of schools that started with an integrity index below 70 in 2015 and had an integrity index above 80 in 2016 declined substantially (Rahmawati and Asrijanty, 2016).

[18] Five test versions are shuffled in four different ways to create 20 test versions per classroom.

The CBT is rolled out in phases, starting from 2015. In that year 40 junior secondary schools switched from PBT to CBT. Implementation then ramped up. A total of 43,841 junior secondary schools (78%) implemented CBT in 2019 (see Table 3.1). Schools - in some cases, districts - self-selected into the CBT program. Only in 30 out of 514 districts did all public schools switch to CBT in the same year, showing that CBT implementation was rarely clustered at the district level. After receiving an application to participate in CBT, the relevant district government determines if these schools can take the exam on computers.[19] Schools are also allowed to use computers in neighboring schools. Once a school switches to CBT, their integrity index is not calculated anymore.[20]

Table 3.1 shows the average exam score, integrity index and access to electricity, internet and computers in 2015 for all junior secondary schools, grouped by the year in which the schools switched to CBT. In the first two years, only a small percentage of schools took the CBT. From 2017, large groups of schools switched. The table confirms that adopters in the first two years are significantly different from schools that adopted CBT later or those did not adopt CBT until the end of our study period. The late CBT adopters had fewer computers, lacking electricity, and low access to internet in 2015. Schools that switched later also had lower average exam scores and integrity in 2015. In 2019, only less than a quarter of junior secondary schools (10,750 schools) still took exams on paper.

[19] The district government checks if the schools have a sufficient number of computers and stable electricity supply. Schools with computers, but without a stable internet connection can download the exams and conduct the exams offline. The questions are only revealed once the students commence the exam.

[20] It is impossible to calculate the integrity index in a comparable way as it is partly based on how often students copy each other's answers. Yet in CBT, every student receives different questions.

Table 3.1: Staggered Adoption of CBT

	(1)	(2)	(3)	(4)	(5)	(6)
	No CBT	2015	2016	2017	2018	2019
	Mean	Difference between cohort [...] and no CBT group				
Exam Score	58.68	19.71***	10.97***	3.38***	3.62***	0.87
	[12.35]	(1.45)	(1.75)	(0.97)	(0.96)	(0.80)
Integrity	65.62	34.38***	11.87***	9.57***	1.75	2.61**
	[17.91]	(1.04)	(1.76)	(1.24)	(1.32)	(1.14)
Exam Participants	57.44	172.16***	101.91***	66.52***	22.31***	6.60**
	[67.54]	(14.32)	(15.29)	(4.52)	(4.24)	(3.11)
Student-Teacher Ratio[1]	13.20	3.61***	3.55***	2.73***	0.87*	-0.16
	[8.29]	(0.68)	(0.43)	(0.40)	(0.45)	(0.39)
Share teachers with 4-year degree[1]	0.82	0.09***	0.05***	0.06***	0.05***	0.04***
	[0.22]	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
Public school	0.71	0.14**	-0.31***	-0.32***	-0.42***	-0.35***
	[0.46]	(0.06)	(0.10)	(0.02)	(0.02)	(0.02)
Rural[1]	0.89	-0.69***	-0.71***	-0.41***	-0.14***	-0.09***
	[0.31]	(0.08)	(0.08)	(0.05)	(0.02)	(0.03)
Electricity[1]	0.86	0.14***	0.14***	0.14***	0.12***	0.10***
	[0.34]	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Internet[1]	0.65	0.25***	0.32***	0.28***	0.26***	0.21***
	[0.48]	(0.05)	(0.02)	(0.02)	(0.02)	(0.01)
Computers[1]	0.19	0.74***	0.64***	0.50***	0.29***	0.17***
	[0.39]	(0.06)	(0.02)	(0.02)	(0.02)	(0.02)
Observations	10,705	40	856	9,377	16,183	12,963
Cumulative (%)	21.4	0.1	1.8	20.5	52.8	78.6

Note: The table includes 50,124 panel schools. Standard deviations are provided between brackets and standard errors between parentheses.

[1] We only have this information for schools that fall under the Ministry of Education. These are 33,331 schools in total, or from the first to the last row: 39, 766, 7,335, 8,004, 7,296 and 9,891 schools. * p<0.10 ** p<0.05 *** p<0.01

3.3 Data and Descriptive Statistics

We use publicly available administrative data from Pusmenjar. The data source is called Pamer (*Pengoperasian Aplikasi Laporan Pemanfaatan Hasil Ujian Nasional*) and it reports the national examination results. The dataset contains exam score means in mathematics, Indonesian, English and science, the number of students taking the exam, the standard deviations and the integrity index at the school level. We have access to mean exam scores from 2010 to 2019, standard deviations from 2010 to 2018, and the integrity index from 2015 to 2018.[21] In addition, we know which schools switched to CBT between 2015 and 2018. The exam scores are between zero and 100, and the final

[21] A CD with the national examination data, including the integrity index, can be requested from the Ministry of Education. Exam score data between 2015 and 2019, but not the integrity index, can also be accessed at https://hasilun.pusmenjar.kemdikbud.go.id/.

exam score is the average score in mathematics, Indonesian, English and science. We use this average score as our main outcome variable throughout the paper. We complement this data with information on school resources in 2015 from datasets called *Dapodik* and *Sekolah Kita*.

The exam information is available for all junior secondary schools in Indonesia, both public and private. Private school students also take the national exam because it determines continuation to senior secondary school. Our sample consists of 56,500 schools. For our analysis, we focus on 50,124 schools that participated in the national exam each year between 2015 and 2019.[22] The data on school resources are only available for 34,412 junior secondary schools that fall under MoE. We do not have school resource information for religion-based schools under the Ministry of Religious Affairs.

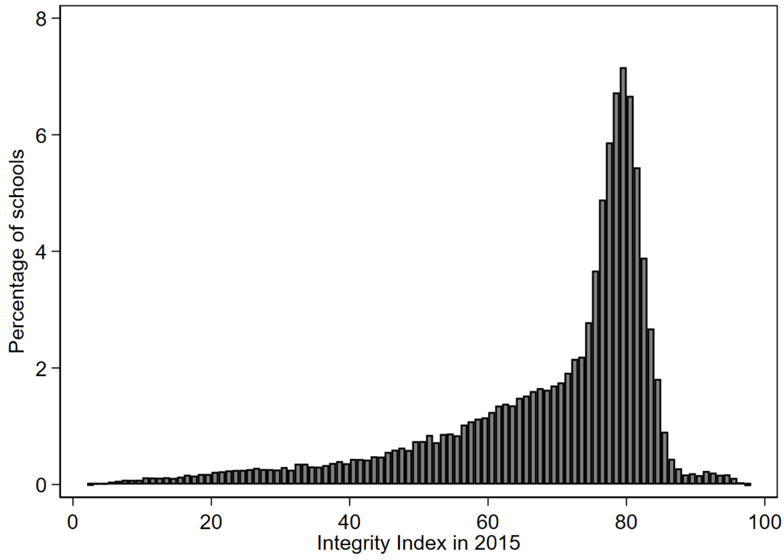
Figure 3.1 presents the distribution of the integrity index in 2015 and box plots of the 2015 exam scores of schools grouped by their integrity. Figure 3.1a confirms that cheating was widespread when CBT was introduced. Only 24 percent of schools achieved high integrity above 80 in 2015. Moreover, a third of the schools scored below 70, which Pusmenjar uses as a threshold for sufficient evidence for cheating. This is more than in Italy and Chicago, where a similar algorithm flagged the exams of about 5 percent of classrooms as compromised (Angrist et al., 2017; Battistin et al., 2017), but less than in Andhra Pradesh, India, where a similar algorithm flagged 38 to 43 percent of classrooms (Singh, 2020b). The integrity index was relatively constant over time, so we interpret this as a school characteristic: schools with lower integrity indices are more likely to cheat on the exam in any year. For 74 percent of schools, the difference between the 2015 and 2016 integrity index was less than 10 points. We did not consider later years to check the stability of the integrity index, because CBT seemed to affect the integrity index of surrounding schools through spillover effects (as discussed in Section 3.5.3). In Table A3.2 we show differences between schools with an integrity index above and below 70. Low integrity schools are generally smaller schools in rural areas, but teacher qualifications and the share public schools are similar between the groups.

The box plots of the exam scores in Figure 3.1b show that the lower the integrity index, the higher the paper-based exam scores.[23] In addition, it shows that a high school average exam score does not automatically translate into a high integrity index, meaning that the integrity index can distinguish between high scoring schools that do and do not cheat.

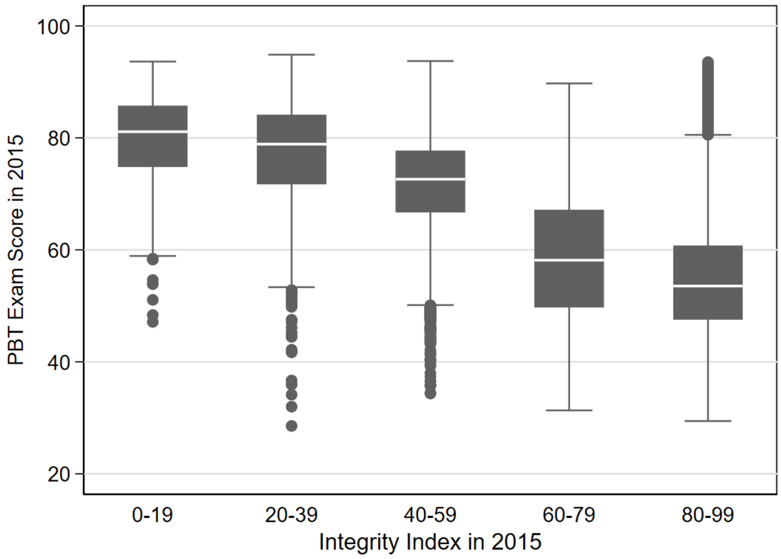
[22] There are 188 panel schools that switched to CBT but switched back to PBT before 2019. We leave these schools out of the analysis.

[23] The pairwise correlation coefficient of the integrity index and exam scores in 2015 is -0.6 and is statistically significant at the 1 percent level.

Figure 3.1: Integrity Index Distribution and Correlation with Exam Scores



(a) Distribution of Integrity Index

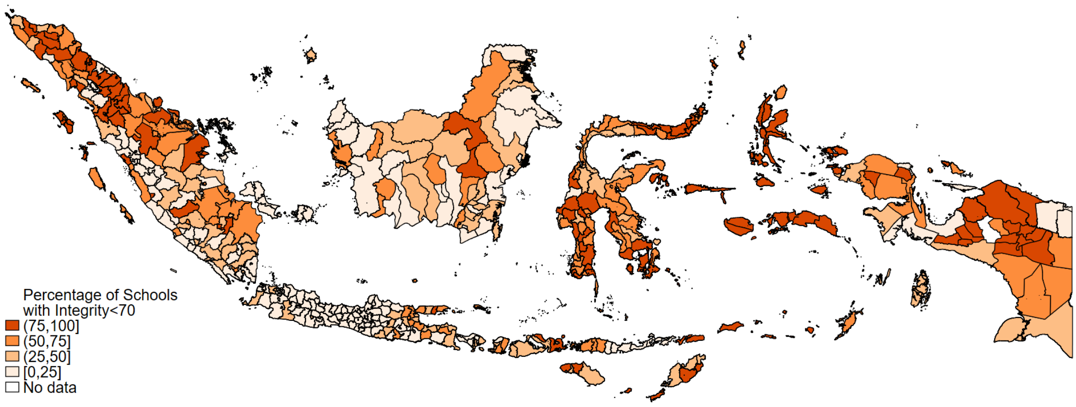


(b) Exam Score Distribution by Integrity

Note: Figures include 44,186 schools for which the 2015 integrity index is non-missing. Panel (a) has a bandwidth of 1. Panel (b) shows the median, the 25th and the 75th percentile, the upper and lower adjacent values and outliers.

There is a strong regional dimension to cheating in Indonesia. Figure 3.2 shows the percentage of schools that had an integrity index below 75 in 2015 by district. Districts with many low integrity schools were often located next to each other. The regional concentration of cheating was also apparent in Italy, where most cheating took place in the southern provinces (Angrist et al., 2017).

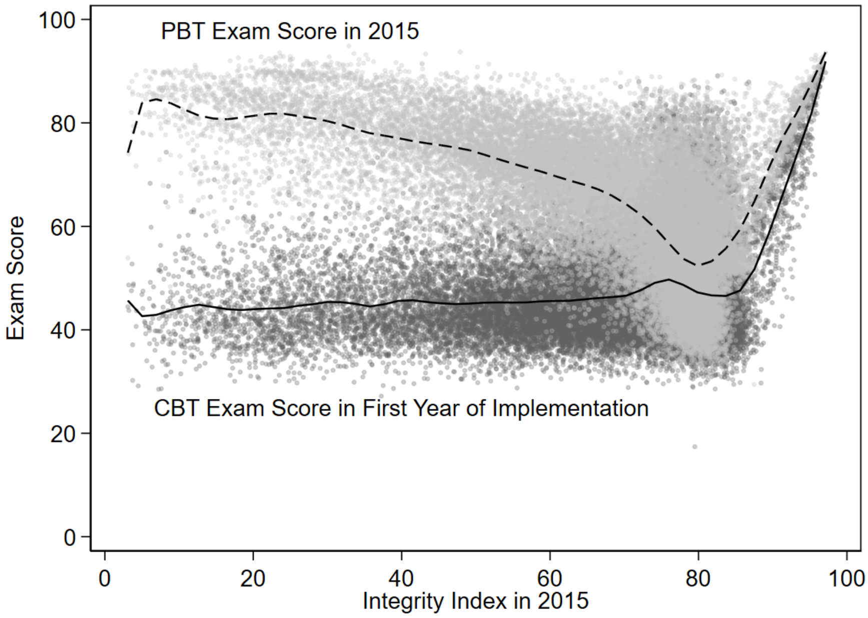
Figure 3.2: Regional Variation in Integrity in 2015



Note: Data shown at the district level.

To get an idea of how CBT affected the exam scores, we plot the 2015 exam score and the exam score in the first year of CBT implementation as a function of the integrity score in 2015 (see Figure 3.3). The dashed line indicates that in 2015, high exam scores could be obtained either through cheating or in an honest way. After switching to CBT, however, the schools that did so by cheating saw their exam score drop substantially. For schools with an integrity score below 70, the exam score dropped by 27 points on average. For honest schools we observe a much more modest drop. Note that these differences cannot be interpreted causally, because it does not correct for the general trend in exam scores over time. For instance, part of the decline in scores could be driven by changes in the difficulty of the exam. We correct for the exam score trend in the impact analysis in Section 3.5.

Figure 3.3: Difference between 2015 Paper-Based and First Computer-Based Exam Scores of Treatment Schools by Integrity



Note: The lines represent smoothed results of a local polynomial regression. The figure includes 34,783 out of 39,379 treatment schools for which the 2015 integrity index is non-missing. The CBT score polynomial regression result combines the exam scores of all treatment schools in the first year of CBT implementation, which is between 2016 and 2019. 95% confidence interval in grey.

To assess whether the CBT exam scores capture true achievement better than PBT exam scores, we correlate CBT and PBT scores with district and school indicators for which we have a strong prior on how they are related to learning outcomes, based on the literature. For eight out of ten indicators reported in Table 3.2, the correlations for the CBT exam scores are more closely aligned with expectations than for the PBT score. Only for average district years of schooling, and "rural school" level, the correlation for PBT is more aligned with expectations. In both cases, the difference in correlation is statistically insignificant.

Table 3.2: Correlations Between Exam Scores and District and School-level Indicators by Test Taking Method

	PBT	CBT	Difference
			p-value
District Indicators			
Average Years of Schooling	1.17	0.85	0.61
	(0.05)***	(0.03)***	
Share of Population that went to Preschool	5.54	7.69	0.68
	(0.49)***	(0.30)***	
Share of Population that is Literate	-12.09	16.03	0.07
	(1.33)***	(0.67)***	
Net Enrolment in Junior Secondary School	11.08	14.22	0.76
	(1.18)***	(0.69)***	
Log Expenditure per Capita	2.18	5.29	0.20
	(0.25)***	(0.19)***	
Share of Population that is Poor	4.39	-23.59	0.04
	(1.37)***	(0.83)***	
Share of Population with Internet Access	4.91	13.17	0.07
	(0.48)***	(0.34)***	
School Indicators			
Share Teachers with 4-year Degree	1.83	4.10	0.05
	(0.51)***	(0.37)***	
Rural School	-4.58	-3.08	0.20
	(0.18)***	(0.14)***	
Internet Access	0.44	1.35	0.11
	(0.22)**	(0.13)***	
Year Fixed Effects	Yes	Yes	
Observations	39,379	39,379	

Note: District indicators come from Kemendikbud 2018 and school indicators from Dapodik 2015 (except for the number of exam participants which we have for each year). Table includes PBT scores in 2015 and the CBT scores in the first year of implementation for the 2016, 2017, 2018 and 2019 CBT cohorts, as in Figure 3.3. Each correlation coefficient is estimated separately because of strong correlations between indicators and is corrected for time trends in exam scores. Standard errors are corrected for clustering at the district level. * p<0.10 ** p<0.05 *** p<0.01

The implementation of CBT was accompanied by a stark reversal of the rankings across schools and districts. Table 3.3 presents the rank correlations of the average exam score between 2015 and other years (4 years before and 4 years after 2015) at the school and district level. The table includes 226 out of 514 districts in which all schools implemented CBT by 2019, so the rank correlation between 2015 and 2019 provides an indication of how different school and district ranks were with and without cheating.[24] The rank correlations between 2015 and earlier years show whether these ranks also differed across years when cheating was still possible, and the rank correlations between 2015 and 2018 show the gradual change in ranks as more schools in the districts switched to CBT.

[24] We performed the same exercise on the full sample and found similar but less distinct patterns, see Table A3.1 in the Appendix.

The first column looks at the school percentile correlation across all 226 districts. In years before the start of the CBT program in 2015, the rank correlation varied between 0.45 and 0.61 with higher rank correlations closer to the base year. In the years after the start of the CBT program, the same pattern is observed but the rank correlation dropped to 0.18 in 4 years. Column 2 presents the average rank correlations of schools within districts. Interestingly, the opposite pattern arises. Average rank correlations after 2015 are somewhat higher than they were before CBT started. On the other hand, the rank correlations across districts dropped sharply after implementation of CBT. While the rank correlation was in the range of 0.53 to 0.63 before the start of CBT, it turned even negative in years thereafter. The evidence shows that the loss in rank correlation is mostly resulting from rank reversals across districts, and less so from rank reversal of schools within districts. This is in accordance with the findings in Figure 3.2, that shows that cheating in concentrated at the regional level.

Table 3.3: Rank Correlation over Time for Districts with Full CBT Implementation by 2019

	(1) School Percentile	(2) School Rank Within District	(3) District Rank
2011	0.45	0.63	0.53
2012	0.50	0.66	0.59
2013	0.59	0.74	0.58
2014	0.61	0.76	0.65
2015	1	1	1
2016	0.67	0.82	0.67
2017	0.48	0.75	0.51
2018	0.23	0.65	0.06
2019	0.18	0.68	-0.1
Observations	24,028	24,028	226

Note: Table shows the Pearson pairwise correlation coefficient of the rank in each year with the rank in 2015. It includes 24,028 panel schools from 226 (out of 514) districts in which all schools implemented CBT by 2019. None of these schools implemented CBT in 2015, 3 percent in 2016, 35 percent in 2017 and 80 percent in 2018. There are between 11 and 952 schools in a district (228 on average).

3.4 Empirical Strategy

To measure the impact of CBT on test scores, we conduct a series of difference-in-difference (DiD) estimations. We estimate the impact separately for each group of schools that switched to CBT in a different year, which we call treatment cohorts. We perform a separate DiD estimation for each CBT cohort because recent studies found that treatment effect estimators from two-way fixed effects models (period and group fixed effects) are biased when there are heterogeneous and dynamic treatment effects. These studies show that the coefficients are weighted sums of the average treatment effects across cohorts and the weights of some cohorts could be negative (Goodman-Bacon, 2021; Callaway and Sant'Anna, 2020; De Chaisemartin and D'Haultfoeuille, 2020; Sun and Abraham, 2020). Because of the integrity differences between the CBT cohorts, we expect heterogeneous treatment effects and estimate the average treatment effect for each cohort separately.

To estimate a combined effect across cohorts, we use the identification strategy of Callaway and Sant'Anna (2020) for staggered DiD models with variation in treatment timing. From the growing literature on DiD models with two-way fixed effects, the Callaway and Sant'Anna estimators suit our study best. Their method allows for dynamic effects when the treatment is binary and the design is staggered, and it allows for conditional parallel trends (De Chaisemartin and D'Haultfœuille, 2022). We expect dynamic effects over time as schools might try to improve their exam scores again after experiencing a drop. Like all recently suggested estimators for DiD models with two-way fixed effects, they take out wrong comparison schools from the control group. A standard two-way fixed effect model includes all schools that did not switch treatment status in the comparison group, including schools that remained treated. In case of dynamic effects, this creates a bias in the estimates as the common trend assumption does not hold. We use the not-yet-treated schools as the comparison group. Due to selection into CBT implementation, these schools were more likely to be similar than the never treated schools that had not switched to CBT yet by 2019. The results can be interpreted as a sample-weighted average treatment effect on the treated schools.

We allow for one period of anticipation of the treatment, meaning that we use the second-to-last year before treatment implementation as the base period to estimate the treatment effects. Anticipatory behavior could occur if treatment schools knew in advance the year they would switch to CBT. The treatment schools might want to limit the drop in exam scores once they implement CBT by already reducing cheating practices in year prior to implementation. As shown later in the results, we indeed find evidence for small anticipation effects on the last year before switching to CBT.

We are interested in the effect of CBT on school mean exam scores and the spread of the exam scores within schools.[25] We estimate the following model for each of the treatment cohorts separately using data between 2015 and 2019,

[25] Since the exam has not been a graduation requirement since 2015, we cannot analyze passing rates.

$$Y_{st} = \alpha_t + \alpha_s + \sum_{e=-K}^{-3} \delta_e \cdot D_{st}^e + \delta_{-1} \cdot D_{st}^{-1} + \sum_{e=0}^{L} \beta_e \cdot D_{st}^e + \epsilon_{st} \tag{3.1}$$

where Y is the average exam score or the standard deviation of student exam scores within school s at time t, α_t and α_s are time and school fixed effects[26], respectively, D_{st}^e are indicators for a school s being e periods away from initial treatment at time t, and ϵ_{st} is the error term. K and L are the earliest and latest period in the data available for a specific cohort, which are 2010 and 2019, respectively. Since we allowed for one period of anticipation, the coefficients are relative to period $e = -2$, or the second to last period before treatment. The parameters of interest are β_e, measuring the effect of participating in CBT at period e. In the years that the treatment schools implement CBT, we expect β to be negative for the school mean exam scores and positive for the within school standard deviation of the exam scores. The standard errors are computed using multiplier bootstrap and are corrected for clustering at the district level. We also present aggregated treatment effects, which are sample-weighted average treatment effects across cohorts in each period relative to period $e = -2$.

The model is estimated on a balanced sampled of schools for which we have complete exam data for each year between 2010 and 2019.[27] We exclude schools that switched to CBT in 2015 or 2016 from our analysis (2 percent of treated schools). These schools had significantly different characteristics from the relevant comparison group, so the common trend assumption is unlikely to hold. The 2015 cohort was a pilot cohort of only 40 schools. The 2016 cohort has a much higher average integrity score and higher average exam scores than the comparison group (about 10 points, see Table 3.1). Hence, it has little common support with the comparison group. In Figure A3.1 we show the propensity to be treated for each cohort, based on a logit model of CBT implementation on the exam scores between 2010 and 2015 and the integrity score in 2015. We find that there is substantial overlap for schools that switched to CBT in 2017 and later, but there is a lack of common support for the 2016 cohort. Therefore, we do not report impact estimates for the 2016 cohort.[28] We also removed 188 schools from our sample that switched back from CBT to PBT before 2019 as De Chaisemartin and D'Haultfoeuille (2020) point out that the weighted average of the cohort-specific treatment effects is only valid when treated

[26] Note that any time-invariant differences between districts are also taken up by these school fixed effects. Figure 3.2 suggests that this may be relevant as cheating was regionally clustered.

[27] In Figure A3.2, we show that results are similar when we estimate the model on the unbalanced panel, that is, when we drop the restriction of complete exam data for all years.

[28] Note that we do not apply propensity score weighting in our analysis. Unweighted trends in exam scores before the intervention were similar between the treatments groups and the comparison groups, so adding weights is not necessary.

groups remain treated after their first year of treatment implementation. The analysis sample contains 39,420 schools out of 56,242 schools that took the exam in 2019.

The causal interpretation of our results depends on two important assumptions. First, we assume that the average student ability within schools is stable over time. Each year a different group of students took the exam. We are only able to attribute a difference in exam scores over time at the school level to CBT if the underlying ability of the students remained the same. The assumption would be violated if students changed schools because of CBT. We argue that this is unlikely. Students enrolled three years before they took the exam, so they could not anticipate whether their school would opt into CBT. To assess whether students did change schools due to CBT, we estimate the impact of CBT on the number of exam participants of each school and find no effect (see Table A3.3). We discuss these results further in Section 3.5.1.

The second assumption is the common trend assumption. We assume that the trend in exam scores of the treatment and comparison group would have been the same if the exams would have remained on paper. The Callaway and Sant'Anna method tests for parallel trends prior to the treatment by computing a pseudo-ATT in each of the pre-treatment periods. The pseudo-ATT is computed by comparing the change in outcomes for a particular treatment cohort relative to the comparison group in two consecutive years, as if treatment had occurred in that period. Hence, the presented results do not include the δ_e coefficients but they include the coefficients of the pseudo-ATT estimates. In the next section, we show that trends were indeed parallel for at least six years prior to CBT implementation.

We present heterogeneous treatment effects by subject, the schools' integrity level in 2015, and by whether the school had access to computers in 2015.[29] With this heterogeneity analysis we test whether the effects we observe from CBT are indeed resulting from a reduction in cheating, and not from other factors associated with the method of exam taking. If the effect is due to a reduction in cheating, it should be larger for schools with a low integrity score. On the other hand, if it is due to students being unfamiliar with working on computers, the effect should be smaller for schools that already had computers in 2015 and should be similar across subjects.

Spillovers could arise if the roll-out of CBT in a district results in a norm change with respect to the acceptability of cheating in the national exam. For example, exam supervision is organized by district governments, which allocate teachers from different schools as proctors to supervise exams. If these proctors came from schools that switched to CBT, they may have been stricter than usual because their school had no option to

[29] This analysis includes 30,198 out of 50,124 schools for which both the integrity score and school resource information are available.

cheat anymore. Allowing the other school to cheat would lead to unfair competition between schools. In addition, the distribution of answer sheets among students and teachers may have been disrupted because the answer sheets are of no use to the ones that took the exam on computers. This way, the probability that a student that took the exam on paper acquiring an answer sheet becomes smaller as more schools switch to CBT. To investigate whether the spillover hypothesis is true, we estimate Equation 3.2 using data from comparison schools only

$$Y_{sdt} = \alpha_t + \alpha_s + \delta_1 \cdot \overline{D}_{dt} + \epsilon_{sdt} \tag{3.2}$$

where Y is the mean exam score or the integrity score in the years 2015 to 2019 of the schools that had not implemented CBT yet by 2019. $\overline{D}_{dt}$ is the fraction of schools that implemented CBT in district d in year t.

We perform two robustness checks that adjust the main estimates for these potential spillover effects. First, we correct the comparison group trend in exam scores for the decline in cheating using the integrity index. This robustness check corrects the estimates for a change in cheating practices in comparison schools directly. However, since we do not have access to the 2019 integrity score, we can only apply this correction for the cohorts that switched to CBT in 2017 and 2018. We correct the main estimation as specified in Equation 3.1 by holding the integrity index of the comparison schools constant. We estimate the following equation

$$Y_{st} = \alpha_t + \alpha_s + \sum_{e=-K}^{-3} \delta_e \cdot D^e_{st} + \delta_{-1} \cdot D^{-1}_{st} + \sum_{e=0}^{L} \beta_e \cdot D^e_{st} + \theta_1 \cdot (1 - D^L_{st}) \times I_{st} + \epsilon_{st} \tag{3.3}$$

which is the same as Equation 3.1, but with the addition of integrity index I_{st} interacted with a dummy variable that indicates comparison schools, which did not implement CBT in the last period ($D^L_{st} = 0$).

Second, we apply a similar correction but now using the share of schools in the district that switched to CBT to capture the spillover effects. We can apply this robustness check to all treatment cohorts and all years. We estimate the following model, allowing the test scores of comparison schools to vary with the share of schools that switched to CBT,

$$Y_{sdt} = \alpha_t + \alpha_s + \sum_{e=-K}^{-3} \delta_e \cdot D^e_{sdt} + \delta_{-1} \cdot D^{-1}_{sdt} + \sum_{e=0}^{L} \beta_e \cdot D^e_{sdt} + \delta_1 \cdot (1 - D^L_{sdt}) \times \overline{D}_{dt} + \epsilon_{sdt} \tag{3.4}$$

which is the same equation as Equation 3.3, but we replace I_{sdt} with $\overline{D}_{dt}$. We basically combine Equation 3.1 and Equation 3.2 in Equation 3.4 because we only allow the comparison group trend to vary with the share of CBT in the district. Conditioning on the integrity index or CBT implementation in the district, we expect the negative treatment effect to be larger because we hypothesize that comparison schools in districts with a higher fraction of treated schools have a more downward trend in exam scores.

One could be concerned about reverse causality between the share of schools implementing CBT in the district and cheating in PBT schools. Although we assume that a higher share of CBT in the district generated spillover effects on PBT schools, it is also possible that schools are less likely to switch to CBT when there is more cheating in PBT schools. We argue that this is not an issue for the interpretation of our estimates because we study changes over time within the same schools and because the decision to switch to CBT was made before taking the exam, so before the decision of PBT schools to cheat in that particular year.

3.5 Results

3.5.1 School Average Exam Scores

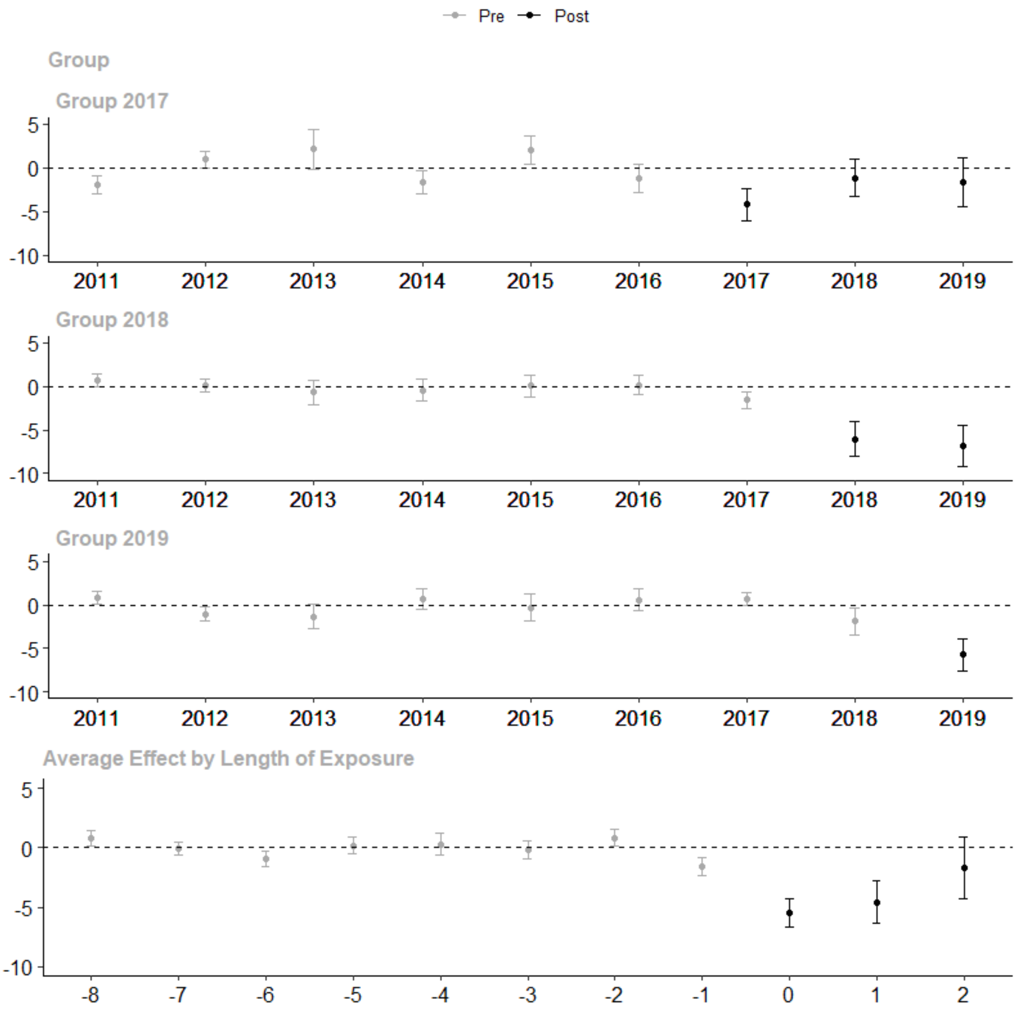
CBT resulted in a drop of 5.4 points in school average exam scores in the first year of implementation. In Figure 3.4 we plot the estimated treatment effect in each year for each cohort (the detailed regression results can be found in Table A3.4 in the Appendix). The effect is larger for the 2018 and 2019 cohorts (6.0 and 5.7 points respectively) than for the 2017 cohort (4.2 points). This makes sense because the integrity index of the 2017 cohort is higher than that of the other cohorts (see Table 3.1). The combined effects, presented in the last row of Figure 3.4 are the sample weighted averages of the cohort effects. Note that for the extreme periods, not all cohorts feed into the estimates. For instance, only the 2017 cohort feeds into the 2 year average effect. For this reason, one should be careful in comparing the first two and last two year average estimates to the other average estimates.

Table A3.5 presents the impact estimation results on exam scores in terms of standard deviations. We used the within-school standard deviation of the test scores, the school level mean exam scores and the number of students that took the exam to calculate the student level mean and standard deviation of the comparison group exam scores in each year and used these to standardize the exam scores.[30] Average school level exam scores drop with 0.4 standard deviation in the first year of CBT implementation (see Table A3.5 in the Appendix).

[30] We do not have access to the within school standard deviation of exam scores in 2019, so for that year we assume that the ratio between the sum of squares across groups and the sum of squares within groups is the same as in 2018.

The pre-intervention trend estimates confirm that the effect arises in the year of opting in. We only find some differences in the pre-trends between the comparison group and the 2017 cohort. The combined estimates also show a small anticipation effect of 1.6 points in the year prior to CBT implementation.

Figure 3.4: Impact Estimation Result on School Exam Scores



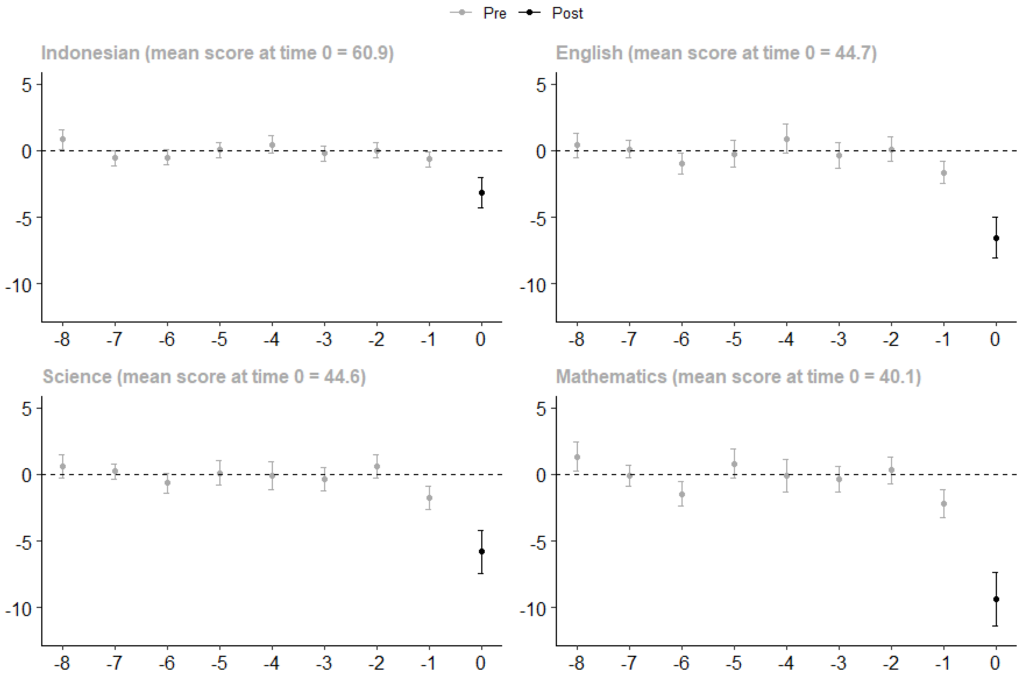
Note: The 2017 cohort includes 8,418 panel schools, the 2018 cohort 13,101 schools and the 2019 cohort 10,052 schools. Plot of post CBT point estimates of β_e in Equation 3.1 with 95% confidence interval, estimated separately for each cohort. The pre-CBT estimates are pseudo-ATT estimates for each pair of subsequent years. Standard errors are corrected for clustering at the district level. The 'average effect by length of exposure' figure shows the sample-weighted average effect across cohorts. Detailed results are available in Table A3.4.

To confirm that the effect is due to a decline in cheating rather than a lack of computer skills, we present average heterogeneous effects by subject, integrity score and availability of a computer lab. We show first year average effects [31] by subject in Figure 3.5. Effects are larger in subjects in which the average exam score in the first year after switching to

[31] We only show first year average effects as this is the only period for which all cohorts contribute to the estimated effect.

CBT were lower. This suggests that there was more cheating on subjects that students found more difficult. The difference in effects across subjects shows that the decline in exam scores was not only due to a lack of computer skills. If that were the case, we would expect the effect to be similar across subjects.

Figure 3.5: Impact Estimation Result on School Exam Scores by Subject



Note: Plot of post-CBT point estimates of β_e in Equation 3.1 with 95% confidence interval, estimated separately for each cohort. The pre-CBT estimates are pseudo-ATT estimates for each pair of subsequent years. Figure shows sample-weighted average effects across cohorts that switched to CBT in 2017, 2018 or 2019. Standard errors are corrected for clustering at the district level. Detailed results are available in Table A3.6.

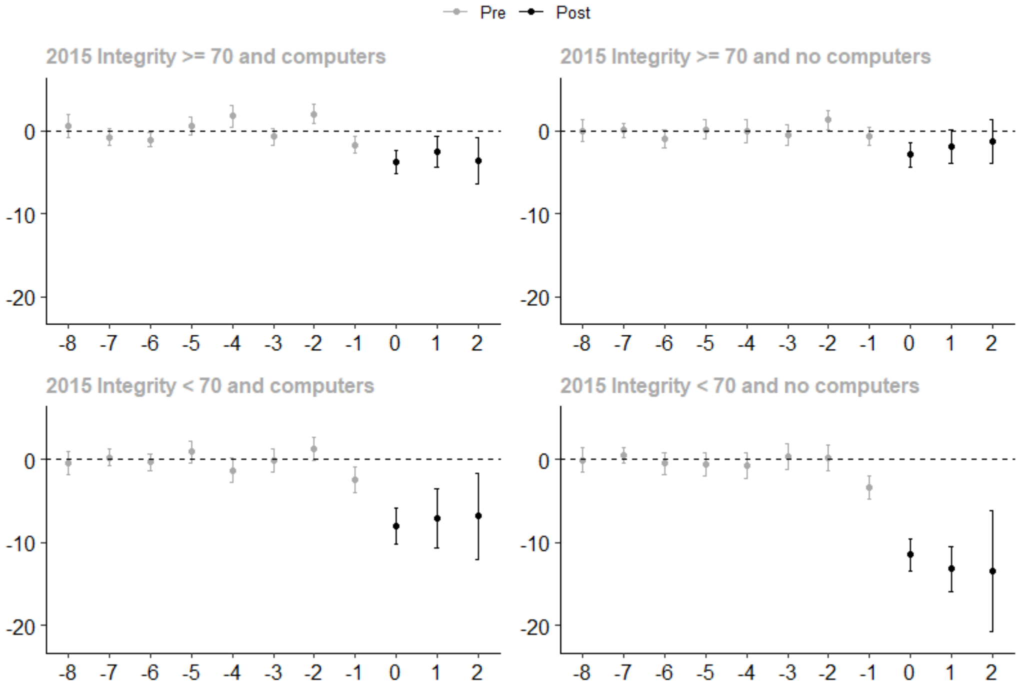
Schools with low integrity and those without computers in 2015 were more affected by the switch to CBT. The effect of integrity is much larger than the effect of having computers, indicating that the effect of the CBT mainly operated through a reduction in cheating rather than through the change in test taking mode (from paper to computers).[32] Figure 3.6 plots the estimates separately for schools with an integrity index below 70 and above 70, and with and without computers in 2015. We focus on the average estimates reported in Table A3.7 in the Appendix. For low integrity schools, CBT resulted in a 8.0-point drop in exam scores while for high integrity schools the drop was 3.7 points. Not having computers resulted in a 3.5-point larger drop for low integrity schools but had no

[32] Evidence for limited effects from the test taking mode were also found in the US (Wang et al., 2008) and India (Singh, 2020b), where computer-based testing yielded similar results as compared to paper-based testing when there was no scope for cheating in either.

significant effect for high integrity schools indicating that for the latter group, familiarity with computers did not drive the small decrease in exam scores.[33]

Figure 3.6 also shows the average effects for the second and third year after adoption. Note that these, contrary to the overall average shows in Figure 3.4, do not indicate the impact declined over time. This suggests that trend observed in in Figure 3.4 was largely driven by different cohorts contributing to different year estimates. Cheating in the 2017 cohort was much less, hence the impact of CBT lower and as this cohort contributed relatively more to later year estimates. Conditioning on computers and integrity results in more homogeneous cohorts and average estimates that are more stable over time.

Figure 3.6: Impact Estimation Result on School Exam Scores by Baseline Integrity and Computer Ownership



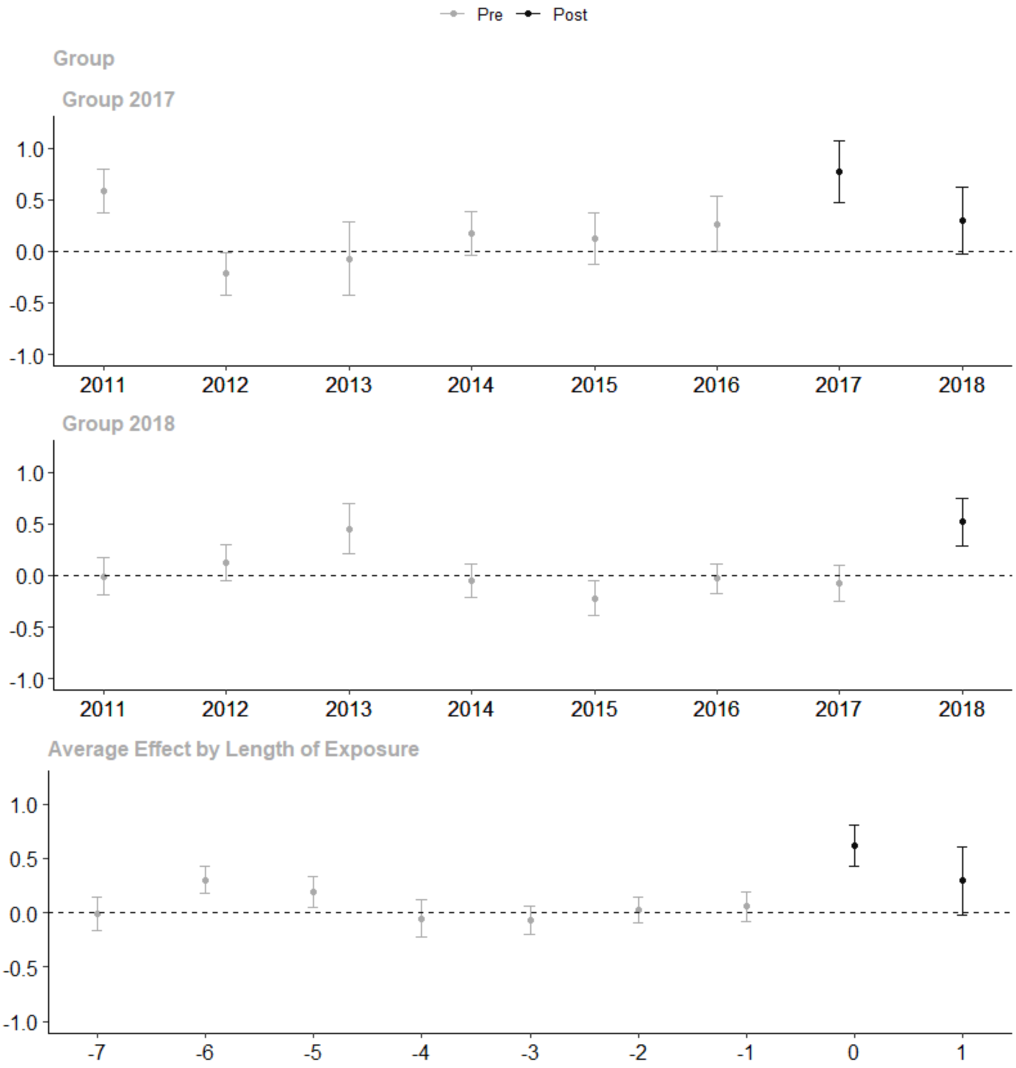
Note: Plot of post-CBT point estimates of β_e in Equation 3.1 with 95% confidence interval, estimated separately for each cohort and integrity and computer ownership category. The pre-CBT estimates are pseudo-ATT estimates for each pair of subsequent years. The figure includes 30,198 schools for which the integrity index and computer information is available in 2015. Figure shows sample-weighted average effect across cohorts that switched to CBT in 2017, 2018 or 2019. The integrity categories are based on the integrity index in 2015. Standard errors are corrected for clustering at the district level. Detailed results are available in Table A3.7.

[33] The drop in exam scores of high integrity schools could be due to the integrity index being conservative. The creators of the index are confident that the exams include cheating when the integrity index is lower than 70, but they are not certain if the exams of schools with an integrity index above 70 do not include cheating (Rahmawati and Asrijanty, 2016). Our results suggest that there were some schools with integrity above 70 that cheated but were not detected by the algorithm.

3.5.2 Variance of the Exam Scores Within Schools

As expected, the standard deviation of (raw) exam scores within schools increased with 0.8 and 0.5 for the 2017 and 2018 cohorts, respectively, when these schools switched to CBT (Figure 3.7). The detailed regression results can be found in Table A3.8 in the Appendix. As described before, it is likely that lower performing students benefited more from cheating before CBT. The disappearance of the treatment effect on the standard deviation in the second year of CBT implementation suggests that lower performing students improved their test scores more than higher performing students.

Figure 3.7: Impact Estimation Result on Standard Deviation of Exam Scores Within Schools



Note: Plot of point estimates of β_y in Equation 3.1 with 95% confidence interval, estimated separately for each cohort. Standard errors are corrected for clustering at the district level. The 'average effect by length of exposure' figure shows the sample-weighted average effect across cohorts. The year 2019 is not included in the figure because the within-school standard deviation of the exam scores is not available in that year. Detailed results are available in Table A3.8.

3.5.3 Local Spillovers of CBT

To investigate the spillovers, we first look at the correlation between a change in the fraction of schools in the district that implemented CBT and a change in exam scores and the integrity index of the comparison schools, for which we estimate Equation 3.2. The results are shown in Table 3.4. We look at the district level, because education policy is determined at that level and proctors are assigned to schools within the district. Recall that it was not whole districts that switched, but schools within districts opted in. The average district has 97 junior secondary schools.

The more schools in a district switched to CBT, the lower the exam scores of the comparison schools and the higher their integrity. Only the exam scores of schools with integrity below 70 significantly decreased as more schools in the district switched to CBT, suggesting that the exam score difference was due to a reduction in cheating practices.The results suggest that the local rollout of the CBT program led to a change in norms with respect to the acceptance of cheating in schools still using the paper based test.

Table 3.4: Correlation between a Change in the Fraction CBT in District and a Change in the Exam Scores and Integrity Index of comparison schools

	(1)	(2)	(3)	(4)
		Integrity < 70	Integrity >= 70	
	Exam Score	Exam Score	Exam Score	Integrity Index
Fraction CBT in the district	-1.65	-9.78	0.83	10.34
(excluding the observed school)	(1.77)	(2.77)**	(1.51)	(2.87)***
School Fixed Effects	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes
Observations	53,525	21,360	25,655	40,352
Number of Schools	10,705	4272	5131	10,671

Note: Model specified in Equation 3.2, estimated on the comparison schools. Standard errors between parentheses and corrected for clustering at the district level. Each regression includes year and school fixed effects. Column 2 and 3 have less observations than column 1 due to missing values of the integrity index. Column 4 has less observations because the integrity index is unavailable in 2019. * p<0.10 ** p<0.05 *** p<0.01

3.5.4 Robustness checks

The spillover affects reported in the previous section affect the interpretation of the results reported in Section 3.5.1 and 3.5.2. In Table 3.5 we present the results for the impact of CBT on exam scores (previously reported in Figure 3.4 and Table A3.4) with and without the correction for spillover effects as discussed in Equation 3.3 and 3.4. The correction does not affect the pretrend coefficients as presented in Figure 3.4 because those are estimated separately as pseudo-ATT's in each two subsequent years, and we do not have integrity scores before 2015. Therefore, we only present the impact estimates in Table 3.5.

As expected, the impact estimates increase in size when we correct for spillover effects. Correcting for the decline in cheating amongst comparison schools using the integrity index makes the biggest difference to our estimates. Because a larger faction of schools switched to CBT as time passes, the later years are affected more. As a result, the declining trend in impact of CBT which was observed for the 2017 cohort is no longer visible with the correction for spillovers.

Overall, the robustness checks make little difference because even though we find evidence for spillover effects, the correlation between the fraction of schools in the district that implement CBT and the average PBT exam score is small and insignificant when estimated on the full sample of schools (column 1 of Table 3.4). In addition, the coefficients shown in Table 3.4 should be interpreted as the difference in the average exam score or integrity index when CBT implementation among other schools in the district increases from 0 percent of schools to 100 percent. The yearly increase in CBT implementation is smaller than that. On average, 1.4 percent of schools in each district implemented CBT in 2016, 19 percent in 2017, 46 percent in 2018 and 70 percent in 2019. This change is too small to generate spillover effects that substantially affect our estimates. Hence, our impact estimates are robust against controlling for the spillover effects.

Table 3.5: Results Corrected for Spillover Effects on Comparison Schools

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dependent Variable: Exam Score	2017 Cohort			2018 Cohort			2019 Cohort	
	Main	Corrected		Main	Corrected		Main	Corrected
CBT 0	-4.19	-4.23	-4.41	-6.04	-6.30	-6.42	-5.69	-8.74
	(0.67)***	(0.88)***	(0.88)***	(0.68)***	(1.02)***	(1.00)***	(0.69)***	(1.15)***
CBT 1	-1.17	-1.42	-1.91	-6.78	-9.44			
	(0.81)	(0.99)	(0.99)*	(0.76)***	(1.31)***			
CBT 2	-1.67	-3.75						
	(1.01)	(1.19)***						
Share CBT in District ×(1−D^L)		-4.67			-6.03			-7.68
		(2.13)**			(2.19)***			(2.35)***
Integrity Index ×(1−D^L)			-0.28			-0.28		
			(0.01)***			(0.01)***		
School Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Number of Schools	15,468	15,468	15,468	20,151	20,151	20,151	17,102	17,102

Note: Model specified in Equation 3.3 and 3.4. Main results are also presented in Figure 3.4. Table includes panel schools that participated in the exam each year between 2010 and 2019. Standard errors between parentheses and corrected for clustering at the district level. Each regression includes year and school fixed effects. The integrity index is unavailable in 2019. * p<0.10 ** p<0.05 *** p<0.01

3.6 Discussion and Conclusion

Cheating is costly to society. Student performance in examinations are used by policymakers to evaluate school quality and by employers in reviewing applicants, yet cheating distorts the signal. Anecdotal evidence about widespread cheating in Indonesia's national examinations had existed for decades. In order to measure the actual magnitude of the problem, the MoE developed an "integrity index" that used answer patterns to detect cheating. In 2015, 33 percent of the junior secondary schools had an integrity index below 70, a threshold that indicates strong evidence of cheating.

To prevent cheating, the MoE decided to use CBT for the national exam starting in 2015. The implementation was gradually phased in, allowing us to estimate the impacts of switching to CBT using difference in differences (Callaway and Sant'Anna, 2020). We present effects for the 2017, 2018 and 2019 cohorts, which together represent 77 percent of all junior secondary schools.

We find that CBT caused a substantial decline in scores. Exam scores decreased by 0.54 points on a 1-10 scale, equivalent to 0.4 standard deviations. For schools for which the integrity index indicated a high likelihood of cheating, the drop is in the 0.80-1.15 range. By comparing the treatment effects on high and low integrity schools and schools with and without computers in 2015, we confirm that the decline in exam scores was mainly driven by a reduction in cheating. If it had been the test taking technology that caused the drop in scores, we would expect to see smaller effects in schools which had access to computers. We observe no such effect. Further, we observe stronger impacts for more difficult subjects for which the payoff of cheating was higher.

We also find that the phase in of CBT at the district level reduced cheating in schools still conducting the paper based exam. While we cannot test for the mechanisms behind this finding, we believe it is indicative of a change in norms. When the schools that adopt CBT have to play by the rules, they may assert pressure on other schools in the same district to do so as well. It may also reflect a change in the logistics of cheating. With more schools switching to CBT, the demand and supply of answer keys are lower. However, the spillover effects are small. Correcting for them does not substantially alter the conclusions with respect to the immediate effect of CBT on test scores.

We find that the impacts are persistent over time. Because the MoE cancelled the national exams for 2020 and 2021, we could only analyze the impacts for a maximum of three years. There is no clear upwards or downward trend in the point estimates, and they are usually statically indistinguishable. For most groups, except of the schools with high integrity and access to computers, the effects remain significantly different from zero. On the one hand, this is positive news in the sense that it indicates that stakeholders were not able to develop alternative ways to cheat. While cheating practices on paper-based exams continue to be discussed at length in newspaper articles, there have been few reported cases of cheating in the computer-based exams (Biantoro and Arfianti, 2019).[34] On the other hand, one would expect that the reduction in cheating opportunities would encourage more thorough preparation for exams, which in turn would lead to higher exam scores in schools that switched to CBT. Unfortunately, this learning effect did not yet materialize after three years of implementation.

[34] There was one teacher that managed to connect his computer with those of the students such that he could control their computers from a distance (Abdi, 2019) and there were some students who took photos of the computer screen during the exam to share questions with others (Alfons, 2019).

Appendix

A3.1 Tables

Table A3.1: Rank Correlation over Time for All Districts

	(1) School Percentile	(2) School Rank Within District	(3) District Rank
2011	0.37	0.63	0.42
2012	0.43	0.64	0.50
2013	0.54	0.73	0.60
2014	0.62	0.77	0.69
2015	1	1	1
2016	0.65	0.81	0.71
2017	0.50	0.76	0.61
2018	0.31	0.69	0.32
2019	0.24	0.68	0.21
Observations	50,084	50,084	514

Note: Table shows the Pearson pairwise correlation coefficient of the rank in each year with the rank in 2015. It includes 50,084 panel schools from 514 districts, only excluding 40 schools that switched to CBT in 2015. None of the schools in the table implemented CBT in 2015, 2 percent in 2016, 20 percent in 2017, 53 percent in 2018 and 79 percent in 2019. There are between 6 and 952 schools in a district (107 on average).

Table A3.2: Difference Between Schools with an Integrity Score Above and Below 70

	(1)	(2)	(3)
2015 Variables	Integrity < 70	Integrity >= 70	Difference
Exam Score	71.28	55.53	-15.75*
	(9.42)	(10.83)	[0.66]
Number of Exam Participants	77.19	105.97	28.77*
	(76.34)	(94.42)	[2.63]
Student-Teacher Ratio	13.92	15.63	1.71*
	(7.22)	(7.94)	[0.33]
Share of teachers with 4-year degree	0.84	0.87	0.03*
	(0.18)	(0.16)	[0.01]
Share of teachers that are civil servant	0.50	0.46	-0.03
	(0.34)	(0.37)	[0.02]
Public School	0.41	0.43	0.02
	(0.49)	(0.49)	[0.02]
Rural	0.75	0.68	-0.07*
	(0.44)	(0.47)	[0.04]
Electricity	0.94	0.99	0.05*
	(0.24)	(0.10)	[0.01]
Internet	0.83	0.89	0.06*
	(0.37)	(0.31)	[0.01]
Computer Lab	0.39	0.56	0.17*
	(0.49)	(0.50)	[0.02]
Observations	16,439	27,747	50,124

Note: Table includes panel schools that participated in the exam each year between 2015 and 2019. Standard deviations between parentheses and standard errors between brackets, corrected for clustering that the district level. * p<0.05

Table A3.3: Impact Estimation Result for Exam Participants

Dependent Variable:	(1)	(2)	(3)	(4)
Exam Participants	2017 Cohort	2018 Cohort	2019 Cohort	Combined
CBT -7			-0.485	-0.485
			(0.300)	(0.285)
CBT -6		-0.458	0.108	-0.211
		(0.347)	(0.313)	(0.249)
CBT -5	-0.029	1.023	0.215	0.488
	(0.410)	(0.356)	(0.332)	(0.208)
CBT -4	-1.236	1.252	-1.457	-0.272
	(0.511)	(0.407)*	(0.330)*	(0.226)
CBT -3	-0.907	0.044	-0.147	-0.265
	(0.595)	(0.350)	(0.333)	(0.274)
CBT -2	4.083	-0.763	0.399	0.874
	(0.476)*	(0.353)	(0.258)	(0.207)*
CBT -1	1.153	-0.806	-0.877	-0.320
	(0.449)	(0.362)	(0.503)	(0.267)
CBT 0	-0.311	-0.593	-1.142	-0.698
	(0.747)	(0.482)	(0.577)	(0.363)
CBT 1	-0.166	-1.840		-1.196
	(0.87)	(0.658)		(0.563)
CBT 2	-1.810			-1.810
	(1.040)			(1.200)
School Fixed Effects	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes
Number of Schools	15,468	20,151	17,102	39,420
Comparison Mean t=0	69.717	71.680	70.074	

Note: Standard errors between parentheses and corrected for clustering that the district level. The "combined" columns show the sample-weighted average effect across cohorts. * p<0.05

Table A3.4: Impact Estimation Result for Raw Exam Scores

Dependent Variable:	(1)	(2)	(3)	(4)
Exam Score	2017 Cohort	2018 Cohort	2019 Cohort	Combined
CBT -8			0.786	0.786
			(0.260)	(0.234)*
CBT -7		0.671	-1.040	-0.072
		(0.257)	(0.267)*	(0.183)
CBT -6	-1.964	0.087	-1.337	-0.913
	(0.353)*	(0.258)	(0.480)	(0.230)*
CBT -5	0.930	-0.676	0.664	0.179
	(0.347)	(0.471)	(0.362)	(0.269)
CBT -4	2.121	-0.449	-0.322	0.277
	(0.696)	(0.371)	(0.453)	(0.306)
CBT -3	-1.640	0.074	0.575	-0.224
	(0.465)*	(0.426)	(0.391)	(0.263)
CBT -2	1.996	0.152	0.704	0.819
	(0.509)*	(0.389)	(0.269)	(0.237)*
CBT -1	-1.183	-1.556	-1.908	-1.568
	(0.588)	(0.315)*	(0.491)*	(0.264)*
CBT 0	-4.193	-6.040	-5.692	-5.437
	(0.671)*	(0.675)*	(0.690)*	(0.402)*
CBT 1	-1.172	-6.782		-4.587
	(0.809)	(0.755)*		(0.682)*
CBT 2	-1.670			-1.670
	(1.005)			(1.009)
School Fixed Effects	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes
Number of Schools	15,468	20,151	17,102	39,420
Comparison Mean t=0	52.303	49.989	50.799	

Note: Standard errors between parentheses and corrected for clustering that the district level. The "combined" columns show the sample-weighted average effect across cohorts. * p<0.05

Table A3.5: Impact Estimation Result for Standardized Exam Scores

Dependent Variable:	(1)	(2)	(3)	(4)
Standardized Exam Score	2017 Cohort	2018 Cohort	2019 Cohort	Combined
CBT -7			-0.094	-0.094
			(0.024)*	(0.025)*
CBT -6		-0.016	-0.105	-0.055
		(0.024)	(0.039)	(0.022)
CBT -5	0.094	-0.070	0.061	0.015
	(0.030)*	(0.039)	(0.029)	(0.023)
CBT -4	0.175	-0.048	-0.022	0.020
	(0.062)	(0.031)	(0.031)	(0.025)
CBT -3	-0.131	0.003	0.042	-0.021
	(0.035)*	(0.031)	(0.027)	(0.019)
CBT -2	0.145	0.011	0.050	0.059
	(0.038)*	(0.028)	(0.019)	(0.019)*
CBT -1	-0.086	-0.104	-0.150	-0.114
	(0.039)	(0.024)	(0.040)*	(0.019)*
CBT 0	-0.324	-0.452	-0.454	-0.418
	(0.047)*	(0.047)	(0.052)*	(0.033)*
CBT 1	-0.077	-0.515		-0.344
	(0.059)	(0.057)*		(0.049)*
CBT 2	-0.107			-0.107
	(0.076)			(0.068)
School Fixed Effects	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes
Number of Schools	15,468	20,151	17,102	39,420
Comparison Mean t=0	-0.030	-0.030	-0.025	

Note: Standard errors between parentheses and corrected for clustering that the district level. Outcome is standardized using the student-level comparison group mean and standard deviation in each year. The “combined” columns show the sample-weighted average effect across cohorts. * p<0.05

Table A3.6: Impact Estimation Result for Raw Exam Scores by Subject

Dependent Variable:	(1)	(2)	(3)	(4)
Exam Score	Indonesian	English	Science	Mathematics
CBT -8	0.827	0.398	0.582	1.339
	(0.251)*	(0.324)	(0.292)	(0.394)*
CBT -7	-0.557	0.117	0.224	-0.072
	(0.200)	(0.251)	(0.210)	(0.271)
CBT -6	-0.497	-0.994	-0.667	-1.495
	(0.199)	(0.283)*	(0.242)	(0.305)*
CBT -5	0.041	-0.233	0.092	0.817
	(0.207)	(0.344)	(0.335)	(0.392)
CBT -4	0.435	0.876	-0.106	-0.098
	(0.255)	(0.357)	(0.356)	(0.424)
CBT -3	-0.217	-0.273	-0.249	-0.156
	(0.193)	(0.311)	(0.305)	(0.347)
CBT -2	0.282	0.859	1.087	1.051
	(0.160)	(0.272)*	(0.321)*	(0.346)*
CBT -1	-0.668	-1.654	-1.772	-2.181
	(0.194)*	(0.291)*	(0.322)*	(0.351)*
CBT 0	-2.526	-5.751	-5.126	-8.344
	(0.341)*	(0.495)*	(0.452)*	(0.608)*
CBT 1	-3.831	-3.861	-3.832	-6.823
	(0.517)*	(0.706)*	(0.778)*	(0.940)*
CBT 2	-2.882	0.550	-0.701	-3.646
	(0.769)*	(1.133)	(1.103)	(1.392)
School Fixed Effects	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes
Number of Schools	39,420	39,420	39,420	39,420
Treatment Mean t=0	60.9	44.7	44.6	40.1

Note: We show mean values of the treatment group here to indicate that most cheating happened in subjects that students struggled with most. Therefore, we show scores that do not include cheating. Standard errors between parentheses and corrected for clustering that the district level. Table shows the sample-weighted average effect across cohorts. * $p<0.05$

Table A3.7: Heterogeneous Impact Estimation Result for Raw Exam Scores

Dependent Variable:	(1)	(2)	(3)	(4)
Exam Score	Integrity >= 70		Integrity < 70	
	No Computers	Computers	No Computers	Computers
CBT -8	0.015	0.609	-0.052	-0.366
	(0.42)	(0.491)	(0.515)	(0.521)
CBT -7	0.060	-0.756	0.545	0.257
	(0.302)	(0.390)	(0.350)	(0.364)
CBT -6	-0.911	-1.053	-0.470	-0.321
	(0.367)	(0.328)*	(0.460)	(0.410)
CBT -5	0.151	0.645	-0.552	0.953
	(0.453)	(0.408)	(0.550)	(0.489)
CBT -4	-0.005	1.775	-0.666	-1.303
	(0.492)	(0.456)*	(0.538)	(0.536)
CBT -3	-0.542	-0.717	0.365	-0.077
	(0.437)	(0.337)	(0.532)	(0.507)
CBT -2	1.315	2.063	0.206	1.320
	(0.417)*	(0.438)*	(0.562)	(0.495)
CBT -1	-0.667	-1.691	-3.298	-2.438
	(0.386)	(0.355)*	(0.538)*	(0.663)*
CBT 0	-2.839	-3.740	-11.503	-8.042
	(0.508)*	(0.480)*	(0.700)*	(0.835)*
CBT 1	-1.954	-2.502	-13.208	-7.052
	(0.719)	(0.658)*	(1.091)*	(1.306)*
CBT 2	-1.331	-3.574	-13.450	-6.830
	(0.929)	(0.967)*	(2.533)*	(1.982)*
School Fixed Effects	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes
Number of Schools	7,433	9,365	4,782	3,005

Note: Table only includes schools for which the integrity index and computer information is available in 2015. Standard errors between parentheses and corrected for clustering that the district level. Table shows the sample-weighted average effect across cohorts. * p<0.05

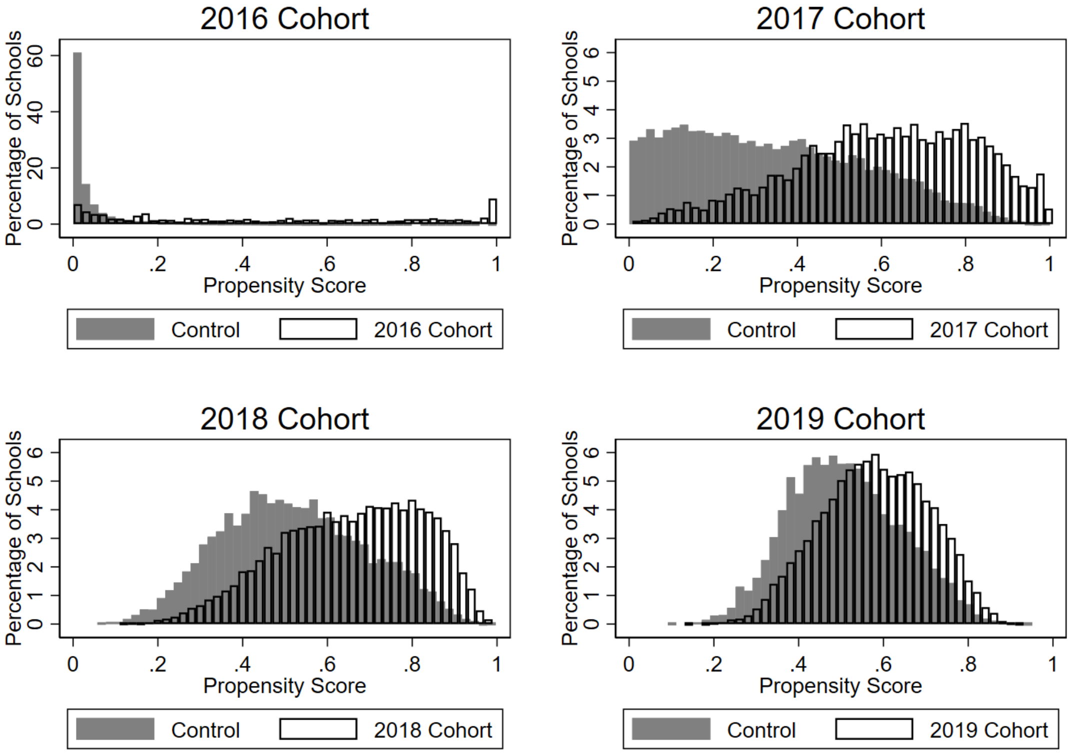
Table A3.8: Impact Estimation Result for Exam Score Standard Deviation Within Schools

Dependent Variable:	(1)	(2)	(3)
S.D Within School	2017 Cohort	2018 Cohort	Combined
CBT -7		-0.010	-0.010
		(0.064)	(0.063)
CBT -6	0.586	0.120	0.302
	(0.078)*	(0.054)	(0.049)*
CBT -5	-0.220	0.454	0.190
	(0.067)*	(0.078)*	(0.059)*
CBT -4	-0.072	-0.047	-0.057
	(0.124)	(0.049)	(0.065)
CBT -3	0.173	-0.221	-0.067
	(0.078)	(0.061)*	(0.051)
CBT -2	0.119	-0.030	0.028
	(0.084)	(0.049)	(0.042)
CBT -1	0.265	-0.076	0.058
	(0.093)*	(0.060)	(0.047)
CBT 0	0.770	0.520	0.618
	(0.114)*	(0.081)*	(0.070)*
CBT 1	0.292		0.292
	(0.122)		(0.113)
School Fixed Effects	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes
Number of Schools	20,082	26,888	70,638
Comparison Mean t=0	5.355	5.841	

Note: Within-school standard deviation is not available for the 2019 exam. Standard errors between parentheses and corrected for clustering that the district level. The “combined” columns show the sample-weighted average effect across cohorts. * p<0.05

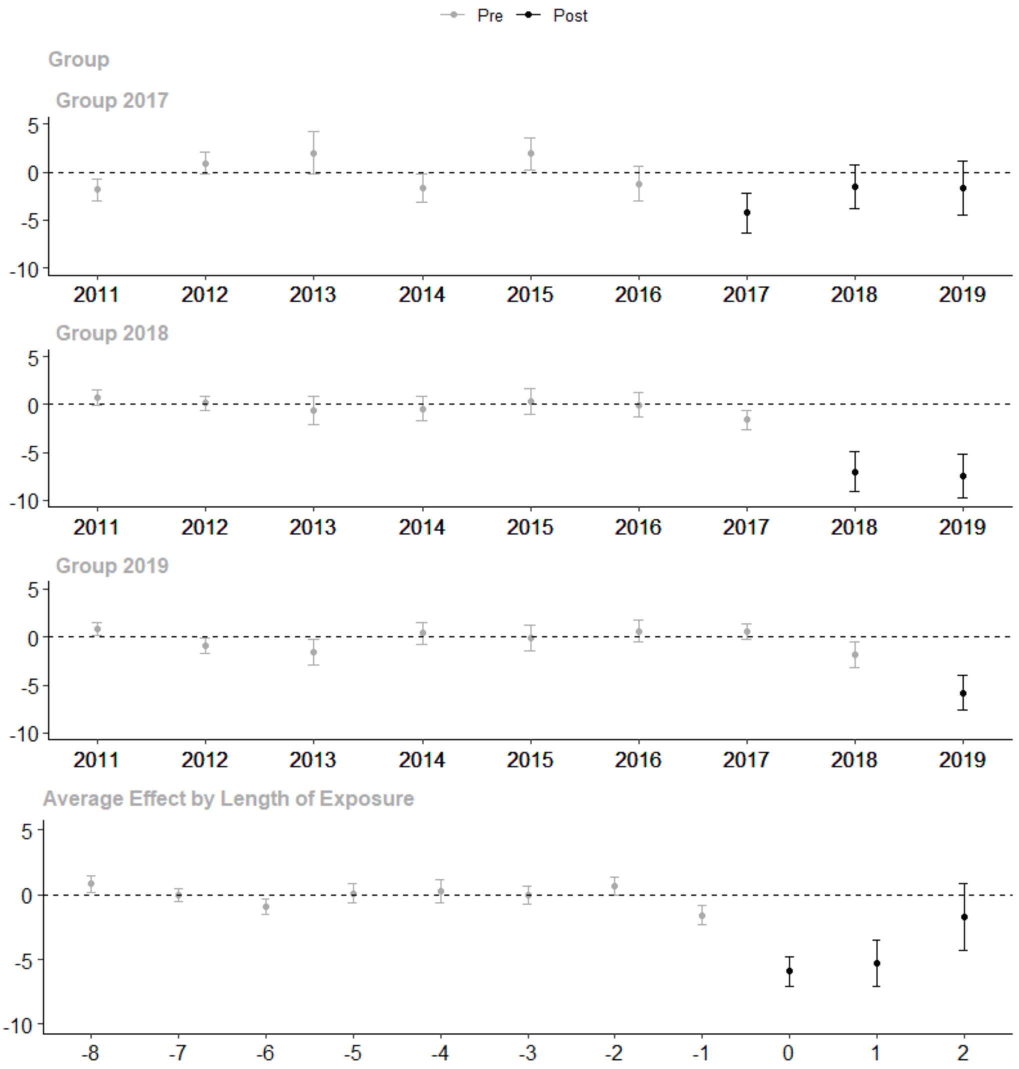
A3.2 Figures

Figure A3.1: Common Support Between comparison schools and Each Treatment Cohort



Note: Propensity score is estimated using the school average exam score in each year from 2010 to 2015 and the integrity index in 2015. Size of bins is 0.02. Y-axis scale of the first histogram that compares the comparison schools to the schools that switched to CBT in 2016 deviates from the scale of the other figures.

Figure A3.2: Impact Results Using Unbalanced Panel Schools



Note: Plot of post CBT point estimates of β_e in Equation 3.1 with 95% confidence interval, estimated separately for each cohort. The pre-CBT estimates are pseudo-ATT estimates for each pair of subsequent years. Standard errors are corrected for clustering at the district level. The "average effect by length of exposure" figure shows the sample-weighted average effect across cohorts.

Chapter 4

Who Benefits and Loses from Making Top Schools Less Selective? Evidence From a Large Change in Student Composition in Indonesian Schools[35]

4.1 Introduction

Education systems that allow high-quality schools to select students based on test scores have come under pressure to become less selective. Although test score selection is a common practice in many countries (OECD, 2020), it is criticized for creating a more favorable learning environment for high-achieving students. High-achieving students can enroll in better schools with better-performing peers than low-achieving students. Selective schools could therefore widen the achievement gap, and in turn increase earnings inequality (Burgess et al., 2020).

Understanding the learning impacts of making high-quality schools less selective is complex. First, as seats are generally fixed, admitting more low-achieving students means that some high-achieving students should be displaced to lower-quality schools. The

[35] This chapter is joint work with Goldy Dharmawan, Amanda Beatty, Daniel Suryadarma, and Menno Pradhan. An earlier version of this chapter was circulated under the title “Who Benefits and Loses from Large Changes to Student Composition? Assessing Impacts of Lowering School Admissions Standards in Indonesia”. We are grateful to Budi Santosa Asrori, Fajar Afrian, and other officials in the Yogyakarta education agency for their support. We thank the SMERU Research Institute for collecting the data, specifically Florischa Tresnatri, Risa Nihayah, Sirojuddin Arif and Wisnu Adi Wijoyo. We greatly appreciate feedback from Evan Borkum, Jishnu Das, Philip Gleason, Hessel Oosterbeek, Lant Pritchett, Shintia Revina, Sudarno Sumarto, several anonymous RISE reviewers, and seminar participants.

learning effects on these groups do not have to cancel each other out, as school effects may be heterogeneous by student achievement and may depend on peer composition. Second, compositional changes may affect learning of incumbent students who stay in the same schools through teacher responses and peer effects. In this chapter, we examine impacts of school integration on student learning throughout an education system. We ask whether low-achieving students benefit from enrollment in selective schools, how these benefits compare to losses for high-scoring students displaced to non-selective schools, and whether a change in peer composition leads to adjustments in teaching strategies and changes in learning for incumbent students.

We study a reform that radically changed the student composition in public and private junior secondary schools in Yogyakarta, Indonesia. Public schools are generally preferred over private schools, because they are free, have more experienced teachers and score higher on the grade 9 exam. The public schools in Yogyakarta are even some of the best performing schools in the country. Yet, they can only serve 60 percent of the student population. The reform altered the primary admission criterion to these oversubscribed public schools, from primary school leaving exam scores to the distance between the students' neighborhood and the school. It was therefore called "the zoning policy". The zoning policy led to a large influx of lower-scoring students in public schools and displacement of many high-scoring students to private schools. Entry scores of incoming students dropped by 0.4 standard deviations (s.d.) in public schools, and increased by 0.4 s.d. in private schools.

We exploit the admission rules to identify students whose public school access changed or remained the same. Both policies rank students on observable characteristics, allowing us to predict public school access under each policy scenario for every student. Our approach detects students who have a high probability of admission to at least one public school if they would apply.[36] That way, we define four subgroups. First, we call students with a low probability of public school access under the merit policy and a high probability under the zoning policy students who "gained access". Those students with a high probability under the merit policy but a low probability under the zoning policy "lost access". Students with a high probability under both policies are called "always access", and those with a low probability under both policies are "never access". Actual school enrollment by public school access group largely changed as predicted. Public school enrollment among "gained access" students increased from 26 to 85 percent[37], while it de-

[36] Note that the policy only affected students who were about the enroll in junior secondary school. Students who were already enrolled could not switch schools.

[37] Some of these students could still enroll in public schools before the reform, because students that opt out of their accessible public school seat open up seats for students with no predicted public school access.

creased from 72 to 27 percent for “lost access” students. Public school enrollment stayed the same for “always access” (87 percent) and “never access” students (16 percent).

To examine learning impacts, we compare test score value-added between the first student cohort admitted based on residence and the last cohort admitted based on merit. We rely on test score and survey data we collected when the students were in grade 8, combined with entry scores from administrative data from the Yogyakarta Education Agency. We estimate a value-added model that compares grade 8 test scores between cohorts, conditional on their grade 6 exam scores and other background characteristics. We do this for all students, and separately for each of the four groups with the same predicted public school access. That way, we identify reduced-form effects from enrollment in a school of different quality with different peers (for “gained access” and “lost access” students), and effects from a different peer group only (for “always access” and “never access” students).

We find that the zoning policy change slightly decreased average learning. “Lost access” students saw a large loss in learning (-0.23 s.d.), while the increase in learning for “gained access” students was smaller and statistically insignificant (0.12 s.d.). “Always access” students learned significantly less with lower-scoring peers (-0.13 s.d.), but “never access” students did not benefit from higher-scoring peers (-0.03 s.d.). Hence, learning losses were larger for high-achieving students than learning gains for low-achieving students.

Using student and teacher survey results, we find suggestive evidence that teachers adjusted their instruction level downward when they had lower-achieving students, but not upward when they had higher-achieving students. Public school teachers seemed to have shifted their attention to the new lower-scoring students as “always access” students found the instruction level easier than before the reform. The perceived difficulty of the instruction level did not change for “never access” students, suggesting that private school teachers did not adjust their teaching to the new higher-scoring students. “Lost access” students found the instruction level easier in private schools than in public schools. Our findings suggest that a lower instruction level played a role in the learning losses of high-scoring students. Heterogeneous classrooms likely made it difficult for teachers to cater to all students’ individual needs.

We also find that “gained access” students reduced their private investments in education, whereas we do not observe any increase in private investments for “lost access” students. The take-up of school-based tutoring classes among “gained access” students halved. Perhaps their parents considered school resources and educational investments as substitutes (see also Das et al., 2013; Pop-Eleches and Urquiola, 2013), or they could not afford the classes. In addition, fewer of them aspired to go to university. The decrease in private investments and aspirations may have limited their learning improvements.

Our findings demonstrate that school integration can affect learning outcomes of students other than those who gain access to better schools when the change in student composition is large. The policy decreased learning inequality, but this was mostly at the expense of high-achieving students. The findings suggest that high-achieving students benefit more from being grouped with higher-scoring peers than low-achieving students. At least in the short run, there seems to be a trade-off between learning inequality and average learning.

This chapter contributes to three strains of literature by showing that admission policies lead to behavioral responses by teachers and students and affect students throughout the system. First, we contribute directly to a small literature on system-wide learning effects of admission policies (Dalla-Zuanna et al., 2022; Muralidharan and Sundararaman, 2015).[38] These studies find no effects from a change in student composition. In Dalla-Zuanna et al. (2022), this may be because the change in student composition was smaller as they focus on academic high schools only. In Muralidharan and Sundararaman (2015) teachers did not adapt their methods, possibly because their experiment only affected one student cohort (Singh, 2015).[39] Second, we contribute to the peer effect literature by showing that a negative shock to peers can decrease learning outcomes. Previous studies that use experimental variation in peers of high-achieving students find no or modest learning losses from lower-scoring peers (Rao, 2019; Imberman et al., 2012; Angrist and Lang, 2004), although a recent paper finds that improving peer test scores can increase learning outcomes of high-achieving students (Berlinski et al., 2022).[40] Finally, this chapter is related to the literature on heterogeneous effects from selective schools relative to non-selective schools. These papers mostly rely on school lotteries or admission cutoffs in regression discontinuity designs. Their estimates capture a composite effect of school quality and peer composition, and they find mixed results.[41] Although their results provide important information on school effects under a specific admission policy, our results

[38] Black et al. (2020) study the impacts of making college less selective on students who gained and lost access, but not on incumbent students. They study college graduation and earning instead of learning outcomes, and find larger benefits for students who gained access.

[39] Other related papers study the expansion of seats in elite schools (Guyon et al., 2012) and a policy change from ability grouping to mixing in schools (Chin and Kwon, 2019). They are not able, however, to separate effects for new and incumbent students in selective schools.

[40] Non-experimental studies that exploit cohort-to-cohort variation in classroom composition often find benefits of higher-scoring peers. However, they suffer from the typical reflection problem, i.e. it is impossible to distinguish the effect of peers on the individual from the effect of the individual on peers if both are determined simultaneously (Paloyo, 2020).

[41] Low-achieving students sometimes benefit more than high-achieving students (Jackson et al., 2020; Shi, 2020), sometimes do not benefit while high-achieving students do (Pop-Eleches and Urquiola, 2013), and sometimes even learn less in selective schools (Oosterbeek et al., 2020; Abdulkadiroğlu et al., 2018).

imply that they are conditional on that student allocation and cannot be generalized to different admission policies.

The rest of the chapter proceeds as follows: the next section gives the framework based on which we develop hypotheses for the learning impacts. Section 4.3 describes the policy reform and the context in which it took place. Section 4.4 describes our data. Section 4.5 explains how we exploit the reform to identify students whose access changed or remained the same and explains our empirical strategy. The impact results are presented in Section 4.6. In Section 4.7 we discuss the results.

4.2 Framework

To understand the mechanisms through which less selective school admissions can affect learning outcomes, we discuss a theoretical model and the related literature. We consider a simple model that assumes that student test scores depend on students' initial ability, the instruction level, school resources, student effort and peers. It builds on the model developed by Duflo et al. (2011) to study tracking in Kenyan primary schools.[42] The model for student i in school s is given by

$$y_{is} = x_{is} + f(\overline{x}_{-is}) + h(x_s^* - x_{is}) + e_{is} + \delta_s + u_{is} \tag{4.1}$$

where x_{is} is the student's initial test score, $\overline{x}_{-is}$ is the average score of other students in the school, x_s^* is the teacher's target level of instruction, e_{is} is the student's own study effort, δ_s captures the quality of school resources, such as the quality of the teachers and available funds, and u_{is} is an error term. $h(x_s^* - x_{is})$ is a decreasing function in the absolute value of the difference between the student's score and the target instruction level. $f(\overline{x}_{-is})$ captures direct peer effects.

Based on this model, we develop hypotheses on mechanisms for learning effects of making high-quality schools less selective, and on average learning effects. We consider a context with two types of schools, high quality and low quality, and two types of students, high-scoring and low-scoring. We take school capacity and class size as fixed. Before the policy change, high quality schools select the highest-scoring students and the lowest-scoring students enroll in the low quality schools. After the policy change, part of the

[42] We added school effects to capture learning effects for students who would enroll in different schools under each policy. We also added student effort as students may respond to different school resources as explained later. We removed teacher effort, written in the Duflo et al. (2011) model as $g(e)$ interacted with $h()$. Duflo et al. (2011) test for changes in teacher effort by comparing teachers who face different incentives (contract teachers and civil servant teachers). We do not test for teacher effort, but from interviews with the principals and teachers we know that they put much effort into improving the performance of low-scoring students.

low-scoring students are admitted into the high-quality schools, displacing high-scoring students to low-quality schools.

Accordingly, we define four groups of students that are differently affected by such a policy: high-scoring students who remain in high quality schools ("always access"), high-scoring students who were displaced to low quality schools ("lost access"), low-scoring students who gained access to high quality schools ("gained access") and low-scoring students who remained in low quality schools ("never access"). For "gained access" and "lost access" students, both the quality of the school they enrolled in and their peer group differs between policies. For "always access" and "never access", only their peer group changes. High-scoring students get a lower-scoring peer group, while low-scoring students get a higher-scoring peer group. These changes by group are shown in the first two rows of Table 4.1.

Table 4.1: Hypotheses on Average Learning Effects on Full Sample and by Subgroup

		Changed Access		Same Access	
		"Lost Access"	"Gained Access"	"Always Access"	"Never Access"
School quality	All	High to Low	Low to High	High	Low
Test performance	All	High	Low	High	Low
School resource effect	0	−	+	0	0
Direct peer effect	0	−	+	−	+
Instruction level effect	−	−	−	−	−
Order of effects by size	(−)	4 (−)	1 (+)	3 (−)	2 (?)

Note: Table shows expected average learning effects when making high-quality schools less selective. The bottom row shows the rank of the expected effects by size, ranking from the larger positive effect to the largest negative effect. The expected direction of the effect is shown between parentheses.

We assume that all students benefit from better school resources. Positive learning effects from school resources are commonly found in the literature. Increased schools funds improved school inputs and student learning outcomes in the United States (Jackson, 2020) and in developing country contexts (Das et al., 2013; Ferraz et al., 2012; Reinikka and Svensson, 2005). The model captures the difference in school quality by a change in δ_s as students enroll in different schools. We assume that school resources stay constant within schools after the policy change, at least in the short-run, e.g. that teachers stay in the same schools.

We assume that the change in peer composition affects learning in two ways: directly through peer interactions, $f(\overline{x}_{is})$, and indirectly by affecting teacher practices, x_s^* (consistent with Sacerdote (2011)). In more heterogeneous classrooms, it may be harder for teachers to adjust their instruction level to the students' needs. A high instruction level may set the pace too fast for students at a low ability level, making them at risk to fall behind (Bau, 2022; Duflo et al., 2011). A low instruction level may be too slow for students with a high ability level, such that they do not achieve their full potential.[43] As making high-quality schools less selective increases heterogeneity in the student composition, we assume a negative instruction level effect for all student groups (row 3 in Table 4.1).

Taking the direct and indirect peer effects together, we expect the learning losses to high-scoring students to be larger than learning gains for low-scoring students. We expect high-scoring students to learn less because of lower-scoring peers and an easier instruction level, while the benefits of higher-scoring peers for low-scoring students are limited due to a more difficult instruction level. The learning effect on low-scoring students depends on the relative magnitude of direct peer effects and the instruction level effect.

These hypotheses are in accordance with findings in the peer effect literature. The extensive peer effect literature generally rejects a linear-in-means model in which peers affect all students in the same way (Sacerdote, 2014). High-scoring students seem to benefit more from higher-scoring peers than low-scoring students (e.g., Antonovics et al., 2022; Hill et al., 2022; Burke and Sass, 2013; Sacerdote, 2011).[44] Booij et al. (2017) show that these nonlinearities may come from the spread of peer scores. Low-ability students would benefit from being grouped with similar students, but conditional on being in a heterogeneous classroom, they do benefit from higher-ability peers.[45] We assume that the nonlinearities from the spread of test scores run through an indirect effect from the instruction level.

Some studies find that differences in school quality and peers may also affect the students' own investments in their learning. This is captured in the model in terms of effort e_{is}. Schools of perceived higher quality can prompt parents to lower their own educational investments for their children (Das et al., 2013; Pop-Eleches and Urquiola, 2013). If school resources and parent educational investments are substitutes, the school resource effect on "lost access" and "gained access" students could be small. In addition, having a lower position relative to peers may decrease confidence and demotivate low-scoring students (Denning et al., 2021; Barrow et al., 2020; Pop-Eleches and Urquiola, 2013; Cullen et al., 2006). We do not include these mechanisms in the table because there

[43] If teachers do not adjust their instruction level, only students whose access changed would be exposed to a different level. Students whose access remained the same would only experience direct peer effects.

[44] An exception is Mendolia et al. (2018), who find stronger peer effects for students in the bottom of the test score distribution.

[45] The peer effects literature mostly exploits small year-to-year variation in peer composition and student fixed-effects models (see Paloyo (2020) for a recent overview), and more recently network models (as summarized in Bramoullé et al. (2020)), which provide limited insight into impacts from larger-scale peer changes that generate changes in teacher behavior.

is little evidence on the relevance of these effects for learning outcomes.[46] Nevertheless, we argue that these adjustments in student effort may reduce school resource and direct peer effects.

The bottom row summarizes the expected average learning effects for each of the student groups. We can only say something about relative effect sizes based on the hypotheses, so the bottom row ranks the expected effects by size and shows the expected direction of the effect between parentheses. We expect to find the largest decline in learning for "lost access" students, and a smaller decline in learning for "always access" students. If the instruction level effect is smaller than the direct peer effect and the school resource effect together, learning would increase for "gained access" students. The learning effect on "never access" students depends on the relative size of the direct peer effect and the instruction level effect. Due to a large decline in learning for high-scoring students and a smaller increase in learning for low-scoring students, we expect less selective school admissions to decrease learning inequality.

Across all students, the model predicts that average learning levels would be lower after making high-quality schools less selective due to more heterogeneous classrooms. In a closed system, the changes in peers and school resources sum up to zero. The overall student composition is the same under both policies, so better peers for one group of students means worse peers for another group. In addition, when the number of seats in high quality schools is constant, admitting one group of students means rejecting another group. If direct peer effects and school resource effects are homogeneous across the test score distribution, as assumed in the table, they would cancel out on average.

In this chapter, we test the bottom row of the table by studying average learning impacts on all students and on each of these groups. We cannot test for the mechanisms directly, so we use intermediate outcome variables from teacher and student surveys to study which mechanisms may have been at play. We examine the change in school quality and peers, and we use proxies for changes in the instruction level and students' own educational investments. The intermediate outcome variables are discussed further in Section 4.4.

4.3 The Policy

We study a unique natural experiment in Yogyakarta, a city on the island of Java in Indonesia. In 2018, the city changed admissions to its selective public junior secondary

[46] Denning et al. (2021) found that the effect from a lower class rank reduced the gains from attending a school with a two standard deviations better performance by 39 percent. However, this may also be explained by a more difficult instruction level with higher-scoring peers. Therefore, we do not try to separate the class rank and instruction level effects for low-scoring students.

schools (grades 7 to 9) from merit-based selection to mostly residence-based selection. Yogyakarta's public schools are free, and they are some of the most reputable schools in the country. Half of the 16 public junior secondary schools scored among the top 1 percent nationally on the 2019 grade 9 leaving exam, and all public school scored among the top 10 percent (Rahmawati, 2019). Students not admitted to these public schools enroll in one of the 41 private schools. Income-eligible students can apply for a publicly-funded voucher that covers 60 to 100 percent of typical private school tuition.[47] Any child in a household eligible for *Kartu Menuju Sejahtera* (KMS), a comprehensive poverty assistance program, is eligible for a voucher, i.e., there is no oversubscription.

Public schools in Yogyakarta have the capacity to serve only 60 percent of students enrolled in junior secondary schools and are highly selective. All students apply to public schools as they are perceived to be of higher quality than the 41 private schools.[48] In the year before the reform, public schools scored on average 23 out of 100 points higher on the grade 9 leaving exam than private schools in the city. Public schools are also generally better resourced. Teachers in public schools have more years of experience and received on average more than double the salary of teachers in private schools in 2018 (see Table A4.10). Our own estimates of school value-added also confirm that most private schools performed worse than public schools (Figure A4.2, we explain how we estimated school value-added in Section 4.4). Public schools traditionally admitted students based on test scores, such that average entry scores, i.e. the grade 6 exam score, among public junior secondary school students were 1.2 s.d. higher than those of private school students (Table A4.1). The so-called "zoning policy" happened after a push from the Central Government for more equality in access to education quality (Kemdikbud, 2017). The policy was expected to improve equality in access to public schools because these schools were mainly located in neighborhoods with relatively low-scoring and poor students (Figure A4.1).

Table 4.2 depicts the percent of seats allocated to different admissions categories under the merit and zoning policies. Under the merit policy, all admissions categories used a grade 6 exam score ranking, but at least 55 percent of seats were reserved for Yogyakarta residents, 25 percent for poor Yogyakarta residents and, at the school's discretion, up to 20 percent for non-Yogyakarta residents. Whether a students was poor was determined by household participation in the KMS program. The second column shows how seats were allocated under the new zoning policy. The most significant change was that now 75 percent of the public junior secondary seats were reserved for applicants who lived closest to the school, i.e., applicants were ranked by distance from their neighborhood to

[47] Private school students receive Rp 2,000,000 or approximately US$140 per semester.

[48] Yogyakarta has a gross enrollment rate of over 100 in junior secondary schools. Nearly all residents attend school until grade 9 or higher as schooling has been compulsory through grade 9 throughout the country since 1994 (Pusat Data dan Teknologi Informasi Kemdikbud, 2019).

a school with closest students ranking highest and admitted using this rank. The zoning policy still reserved some seats for students with the highest grade 6 exam scores, but this declined from 55 to 15 percent of seats at a school for Yogyakarta residents and from 20 to 5 percent of seats for residents outside Yogyakarta. This decline in seats from residents outside of Yogyakarta effectively also meant that public school seats expanded for city residents. The remaining 5 percent of seats were allocated to students who moved to Yogyakarta due to natural disasters in their place of origin or their parent's job assignment to Yogyakarta, who were admitted using the grade 6 exam score rank. The admission system is choice-based. Under both policies, students listed the public schools according to their preference in their application, and were allocated using a Deferred Acceptance (DA) algorithm (for an explanation of this mechanism, see Roth (2008)).[49]

Table 4.2: Allocation of Seats Within Each School Under the Merit and Zoning Policies

Student Category	Selection Criterion	Share of seats	
		Merit Policy	Zoning Policy
Yogyakarta residents	Grade 6 exam score	55	15
Poor Yogyakarta residents	Grade 6 exam score	25	0
Yogyakarta residents	Proximity to school	0	75
Non-Yogyakarta residents	Grade 6 exam score	20	5
Relocated students	Grade 6 exam score	0	5

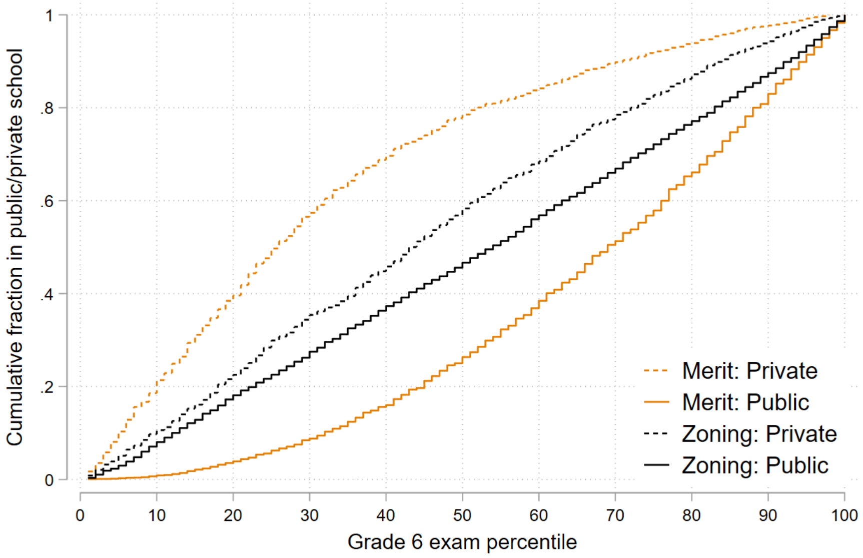
Note: The students' grade 6 exam score is an unweighted average of a student's mathematics, Indonesian, and science scores. Proximity is measured by the difference between the student's neighborhood and each school. Poverty status is determined by participation in the KMS program.

The zoning policy resulted in a massive change in student composition in public and private schools compared to the merit policy. The distributions of entry scores (grade 6 exam scores) of incoming students among public and private schools became much more similar due to the zoning policy, as shown in Figure 4.1. Before the zoning policy, only about 25 percent of public schools students scored below the median, whereas 80 percent of private school students did. After the policy, this gap declined substantially. Almost 50 percent of public schools students had below-median scores compared to 60 percent of private school students. On average, grade 6 exam scores of incoming students were 13 percentiles lower in public schools, equivalent to about 0.4 standard deviation (Table

[49] Under the zoning policy, students were allowed to list all 16 public schools. Under the merit policy, they were only allowed to list their top three preferences. This means that the mechanism was not strategy proof under the merit policy. Students are permitted to enter into multiple admissions categories since these are implemented sequentially. For example, before zoning, a student could enter the grade 6 exam score category (which is first); and if not selected, enter the KMS-participant category. Different school rankings are permitted across categories, meaning that a student could place a school in a different preference rank for each category.

A4.1), and 13 percentiles higher in private schools after the policy change.[50] However, we point out that the zoning policy has targeting shortcomings. Naturally, students and schools are not distributed equally throughout the city such that everyone has equal access to schools of equal quality. In Yogyakarta, some students were not accepted to any public school because they lived too far from any school. Moreover, location-based admissions policies have a history of creating inequities in other countries (Black, 1999). Thus, we evaluate the learning impacts of this policy considering the stated goal of fairness and expanding access, even if we as researchers might have recommended a different policy to achieve these goals.

Figure 4.1: Cumulative Distribution of Grade 6 Exam Scores by Cohort and School Type



Source: Grade 6 exam score 2017 and 2018 and primary data collected in 2019.
Note: Percentiles are calculated based on full population of UASDA takers, within each cohort.

The reform was an exogenous shock to the student cohort that graduated from primary school in 2018. The policy was announced after the grade 6 leaving exam, and just one month before school registration. Students could not influence their chances of public school enrollment by changing address because it considered the registered address of one year before registration. Accordingly, we do not see an increase in the share of students that moved house in grade 6 (Table A4.1). The policy also did not affect student effort on

[50] The Dissimilarity Index across schools, where 1 means perfect segregation and 0 means no segregation, declined from 0.51 to 0.27 when categorizing students into having scored above or below the grade 6 exam median. However, in terms of wealth, the Index barely changed from 0.34 to 0.33 when categorizing students based on eligibility for school vouchers, probably because 25 percent of seats were reserved for poor students under the merit policy. In addition, while grade 6 exam scores scores and KMS participation are somewhat correlated, this correlation is modest; the correlation coefficient is -0.18 for mathematics and -0.24 for Indonesian.

the exam because it was announced after the exam. If it did, we would expect students living close to public schools to exert less effort in the zoning cohort because of admission based on distance to the school. Table A4.2 shows that students living further from public schools did not score significantly higher on the grade 6 exam in any cohort.

The assignment rules of each policy and the exogenous timing of the zoning policy allow us to sort students into groups as defined in Table 4.1: students whose public school access would be the same under both policies ("always access" or "never access") and students who would only have access to public schools under one of the policies ("gained access" and "lost access"). Because the assignment rules are based on observable student characteristics, we can use them to measure the *opportunity* to enroll in a public school for the same student under each policy scenario. Students with high grade 6 exam scores under the merit policy, or with a public school nearby their house under the zoning policy, were essentially offered a public school seat. They could reject the offer if they preferred one of the private schools over the offered public school seat. The exogenous timing of the policy change ensured that grade 6 exam scores and residential location were comparable between pre- and post-reform cohorts, such that students with the same scores and neighborhood were similar. The simulation is explained further in Section 4.5.1.

The zoning reform was the only policy change that took place in junior secondary schools in 2018 in Yogyakarta, so we attribute differences in learning between pre- and post-reform student cohorts to the policy. There were no changes in public school-level funding since budgets are determined on a per-pupil basis, and there was nearly no change in the total number of seats in public schools. The zoning policy did, however, decrease the number of public school seats that could be allocated to students from outside Yogyakarta from 20 to 5 percent, effectively increasing the number of public school seats for students residing in Yogyakarta (Table 4.2). We explain in Section 4.4 how this affects the interpretation of our results. There were no changes to teaching staff in public schools, other than regular teacher retirement. In private schools, teachers turnover was larger due to changes in enrollment (see also Appendix A4.4). We confirm in Section 4.6.3 that our results for students whose access remained the same were not driven by a change in the quality of their teachers.

The policy was partly reverted after only one year of zoning due to parental pressure to allocate more seats based solely on test scores. We call the new policy the "mixed policy", which decreased the share of seats allocated based on proximity. In Appendix A4.3 we analyze learning effects of this policy, although we only have test score data after six months of enrollment. The reversal of the policy prevents us from studying longer-term effects.

4.4 Data and Descriptive Statistics

We combine test score and survey data with administrative data from the Yogyakarta education agency on grade 6 exam scores, public school admissions, KMS participation and house location. We administered a student learning assessment (SLA) (Rarasati et al., 2020) in mathematics and Indonesian for this study in grades 7 and 8 in 2019 and 2020.[51] These include the first student cohort under the zoning policy and the last student cohort under the merit policy.

Our sample covers all 16 public junior secondary schools and 30 out of 41 private schools in Yogyakarta. We used stratified random sampling to create a sample representative of all private junior secondary schools in Yogyakarta.[52] This way, we tested 78 percent of all private school students in the merit cohort and 77 percent in the zoning cohort. Throughout the paper, we apply sampling weights to correct for under-sampling of private school students. The weight for public school students is one, while the weight for private school students is the inverse of the number of sampled private school students divided by the total number of private school students in each cohort. Appendix A4.2 provides more details on our sample.

Our main outcome variable of interest comes from the grade 8 SLA. We standardize the average percent correct on the mathematics and Indonesian SLA to have a mean of zero and a standard deviation of one in the merit cohort.[53] These test scores were collected after 18 months of enrollment, which we believe to be long enough for learning effects from school resources and peers to emerge. However, schools might need more time to adjust to a substantial change in student composition. Teachers in Yogyakarta were not given any training or support before the zoning policy was enacted. Teachers could become better at teaching more heterogeneous classrooms over time. Unfortunately, we are unable to study longer-term effects due to the (partial) reversal of the policy.

To measure learning over time, we use the students' grade 6 leaving exam scores in mathematics and Indonesian as our baseline test score measure.[54] This exam is called *Ujian Akhir Sekolah Daerah* or UASDA and is taken by all students who attend public and private primary schools in Yogyakarta.[55] Exam items differ across years, so to correct for potential differences in test difficulty over time, we standardize the score within each cohort using the mean and standard deviation of the entire population of students who

[51] All testing was completed before schools closed due to the pandemic in March 2020.

[52] We stratified the schools using four geographical strata, two in the north and two in the south. Then we randomly sampled schools within each of the geographical strata.

[53] Results are similar when we estimate learning effects for each subject separately (not shown).

[54] The exam also includes a test in science. We leave this out of our analysis of learning effects as we do not have test scores for science in later grades.

[55] Note that primary school enrollment is universal in Yogyakarta.

took the test. We also show percentile scores. We assume that the tests rank students similarly across years.

The grade 6 exam score is missing for students who graduated from primary schools outside Yogyakarta. As shown in Table 4.2, the Yogyakarta education system is open to students from surrounding districts.[56] Since specific seats in public schools were reserved for students from outside the city, we simply take these students out of our analysis and only consider public school seats for Yogyakarta residents. Therefore, we interpret the results in this chapter as pertaining to students from Yogyakarta.

The missing baseline scores do have consequences for the interpretation of our average results. A larger share of public schools seats was reserved for Yogyakarta residents after the zoning policy (Table 4.2), increasing public school enrollment in our sample from 58.5 to 65.3 percent (shown later in Table 4.3). If public schools produced better learning outcomes, we would expect the increase in seats to improve learning on average. We would underestimate a negative average learning effect of the zoning policy as predicted in Table 4.1. We take this into account when interpreting our results in Section 4.6.

We use the test score data to report differences in the school environment in terms of school quality and peer performance.[57] Our measure for peer performance is the mean grade 6 exam score in the school or classroom, leaving the observed student out. Schools could determine their own policy on student allocation across classrooms, so the peer group at the school level could be different from the classroom peer group. Therefore, we report both.[58] We also report the standard deviation of the peer grade 6 exam score to capture the level of heterogeneity in the student composition.

We define school quality as school value-added before the reform, which is a common method to measure school effectiveness (Angrist et al., 2022). We use an Ordinary Least Squares (OLS) value-added model as specified in Equation 4.2 for student i in school s in the merit cohort,

$$Y_{i,s}^{2} = \alpha_1 Y_{i,s}^{1} + \alpha_2 X_{i,s} + \rho_s + \epsilon_{i,s} \tag{4.2}$$

[56] Grade 6 exams differ across districts, so we did not attempt to collect this data from surrounding districts.

[57] We also considered class size as a mechanism that can affect learning outcomes. We found no change in class size for students whose access remained the same. In public schools, student had 5 more class mates on average than private school students (Table A4.1). Although evidence exists for class size effects, these effects were generally small (Hanushek, 2020). We therefore do not consider this a relevant mechanism in our study.

[58] Note that this measure only includes Yogyakarta residents, so these peer scores do not reflect the full group of peers. Although we do not use peer scores for our learning impact estimates, we do use it to give an impression of the magnitude of the change in peers.

where ρ_s are school indicators that capture the average school value-added in the merit (baseline) cohort.[59] We use these ρ_s estimates to study differences in school value-added for different student types and between cohorts, as explained further in Section 4.5.2. The ρ_s estimates likely contain sampling error, especially the estimates for small schools with few student observations. We follow Bau and Das (2020) by assuming that the sampling error is random and zero on average. Under that assumption, we can include the school value-added estimates as a dependent variable in our analysis that compares school quality between cohorts (see Equation 4.3 in Section 4.5.2).[60] Figure A4.2 plots the estimates of ρ_s for each school and confirms that most public schools produced higher value-added than private schools.[61] In addition to school value-added, we also look at school quality in terms of the average grade 9 exam scores of the schools, which is called the *Ujian Nasional* (UN).

Our measures of school quality do not only capture differences across schools in terms of resources, but also in terms of peer composition. Although school value-added aims to correct for student selection into schools by conditioning on baseline test scores, this does not correct for any behavioral responses to student composition, as discussed in Section 4.2. Yet, the measures are still informative for the magnitude of the expected change in school performance for students who gained or lost access to public schools (Deming et al., 2014).

In addition to testing data, we collected survey data from students, teachers, and school principals in 2019 and 2020. We administered a short survey to all students enrolled in public and sampled private schools about their school preferences, background characteristics and experiences in school. We interviewed teachers to ask about their background characteristics and teaching practice, such as how they might have adjusted lessons or teaching due to the policy changes. We interviewed principals about school facilities and school responses to the policy changes. The results of the teacher and principal survey are presented in Appendix A4.4, and we refer to those results where appropriate.

[59] Angrist et al. (2020) show that the reliability of school value-added estimates can be improved in case of centralized school assignment by adding controls for the probability of assignment to different schools, which they estimate using school preferences and selection criteria. We cannot apply this method because we do not have rank-ordered preference lists for students in private schools.

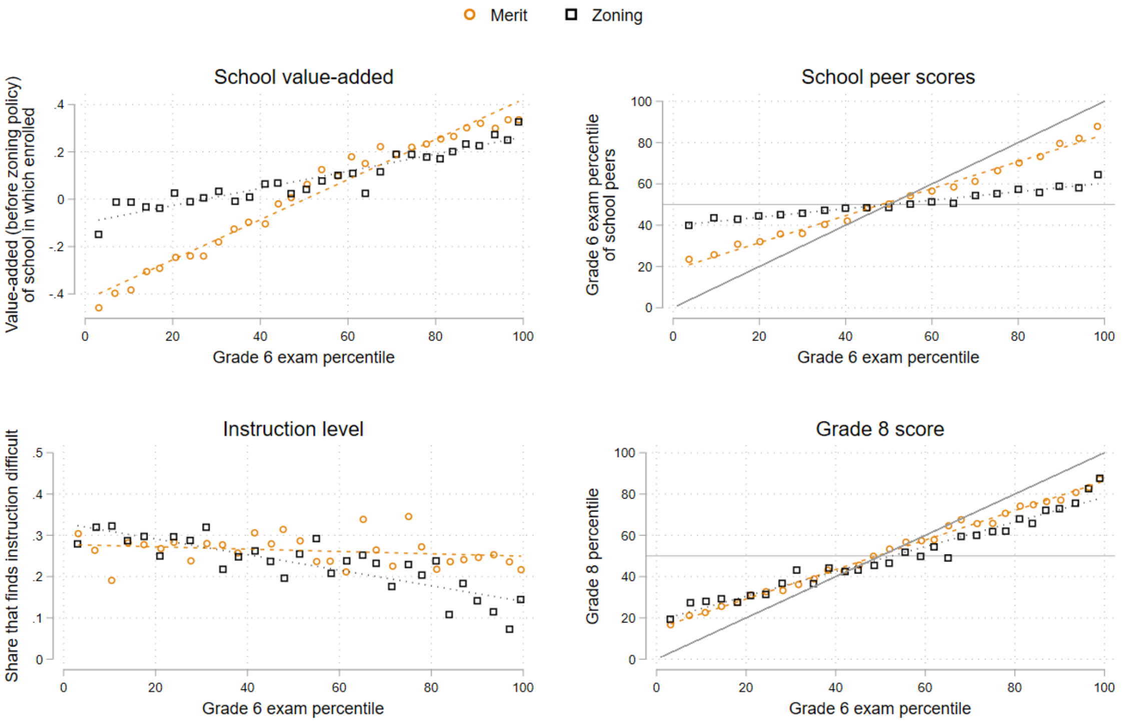
[60] However, including school value-added estimates as predictors on the right side of a regression would lead to attenuation bias (Angrist et al., 2022; Bau and Das, 2020). In that case, the value-added estimates should be corrected for sampling error using the empirical Bayes approach. This approach weights the estimates to shrink them towards the mean school value-added in proportion to their sampling error, as explained in more detail in Angrist et al. (2022). We follow Bau and Das (2020) by not applying this correction on value-added estimates used on the left side of a regression since the OLS estimates are unbiased.

[61] Applying empirical Bayes shrinkage barely affects this ranking: one public school moves up one position in the ranking (not shown).

We use the student survey to study the instruction level and student effort as mechanisms of learning effects. Teacher practices in the classroom are hard to measure, so we use the perceived difficulty level of teacher instructions. We asked students whether they perceived the lessons to be easy, moderate or difficult. Since only about 2 percent of students found the instruction level easy, we present the share who found it difficult. We also use indirect proxies to study student effort. We asked students whether they took tutoring classes and whether they aspire to go to university. Tutoring classes are the primary form of outside educational support in Indonesia. Parents must pay for tutoring – there are no scholarships. Generally, school-based tutoring was more expensive in public schools than in private schools. We see aspiration to go to university as an indicator of student motivation.

Figure 4.2 illustrates how our measures of the school environment differed between cohorts across the grade 6 exam score distribution. We present school value-added, peer scores and the perceived difficulty of the instruction level. The top left figure shows that the zoning policy substantially decreased inequality in access to school quality. Students scoring in the bottom quartile of the grade 6 exam on average enrolled in schools with 0.2 to 0.3 s.d. higher average value-added when allocated according to the zoning policy rather than the merit policy. Students in the top quartile on average enrolled in schools with about 0.1 s.d. lower value-added. The top right figure shows that students of different ability levels were mixed more after the policy. The 45-degree line represents perfect ability grouping across schools, while the horizontal line represents perfect random allocation of students across schools. It shows that the merit policy sorted students across schools, while the zoning policy resulted in a student allocation that was closer to random. The results for the instruction level in the bottom left figure are somewhat noisy, but they roughly show that low-scoring students did not find the instruction level more difficult, while high-scoring students did find it easier. This suggests that teachers shifted their attention towards the low-scoring students.

Figure 4.2: School Quality, Peer Quality, Perceived Instruction Level Difficulty and Grade 8 Test Scores by Grade 6 Exam Percentile



Note: Grade 6 percentile scores are calculated within each cohort, while grade 8 percentiles are calculated on the two cohorts combined. School value-added is measured in terms of standard deviations. Average score of school peers excludes the observed student, and is calculated using students residing in Yogyakarta. Perfect test score selection would result in the 45-degree line, where students are enrolled in schools with students with the same grade 6 score, while perfect random selection would result in the horizontal line at the median student. Figures are conditional on gender, age at the time of UASDA exam, an asset index, an indicator for the mother having completed tertiary education and kelurahan (and indicators for missing values in these controls).

The bottom right figure also shows that grade 8 test scores were different after the reform. Despite improvements in their school environment, below-median-scoring students did not see improvements in their test scores (on average 1.6 percentiles higher in the zoning cohort, equivalent to 0.06 s.d.). At the same time, test scores for above-median-scoring students were lower under the zoning policy even though the decline in school quality was small for them (on average 6.0 percentiles lower in the zoning cohort, equivalent to -0.20 s.d.). Although the zoning policy seems to have decreased learning inequality, this was mainly at the expense of higher-scoring students. In the next sections, we split the cohorts into students whose public school access changed and students whose access remained the same, to study to what extent the learning outcomes were affected by enrollment in a different school or by a change in peers.

4.5 Empirical Strategy

We first discuss how we identify students who remained in public schools or private schools, and students who gained access to public schools or lost access. After that, we describe the empirical model that we use to estimate effects for each of these groups.

4.5.1 Identification of Students Whose Access Changed and Remained the Same

To identify students whose access changed or remained the same, we simulate public school access under the merit and zoning policy scenarios for each student in the two cohorts. We simulate the allocation of students to public school seats based on the selection criteria as shown earlier in Table 4.2. For public school access under the merit policy, we determine an admission cutoff in grade 6 exam scores for regular seats and seats for poor students separately. We retrieve the cutoff value by allocating the merit cohort students who scored highest on the grade 6 exam to public school seats, without making any assumptions on preferences among public schools. We then apply this cutoff to the zoning cohort to find students who would have had public school access under the previous policy.

To determine public school access under the zoning policy, we allocate 15 percent of seats in each school to top-scoring students, and the rest of available seats to students with the smallest distance between their neighborhood of residence and that specific school. For students who were in the "catchment area" of multiple public schools, we assume that they prefer the closest public school to their house.[62] We allocate seats using the DA mechanism.

Using the simulated public school access, we define four subgroups. Students with a grade 6 exam score above the cutoff who live within the catchment area of a public school have access to public schools under both policies ("always access") and those with a grade 6 exam score below the cutoff who live outside of the catchment areas have access under none of the policies ("never access"). Students with a grade 6 exam score below the cutoff who live in a catchment area gained access ("gained access"), while those with a grade 6 exam score above the cutoff who live outside the catchment areas lost access ("lost access"). Figure A4.3 provides a schematic illustration of the simulation.

Our simulation measures which students would be offered a public school seat if all students would apply to all the public schools. We interpret the simulated public school access as the treatment assignment that allows us to conduct an intent-to-treat analysis,

[62] These are 215 students in the zoning cohort and 191 students in the merit cohort, out of 2,322 seats allocated based on distance.

as explained further in Section 4.5.2. However, there may be non-compliance with this treatment assignment as some students may prefer a private school over the public school to which they have access. Table 4.3 shows that the simulation performs well at predicting actual public school enrollment, suggesting that most students preferred their allocated public school seat over a private school seat. The table presents the share of students allocated to each subgroup and actual public school enrollment for each group, comparing the last merit-based student cohort to the zoning cohort. The simulation allocates 46 percent of students to the "always access" group, 25 percent to "gained access", 15 percent to "lost access", and 14 percent to "never access". The "gained access" group is larger than the "lost access" group due to the increase in the share of seats for students from Yogyakarta. Actual public school enrollment among the "always access" group is around 86 percent for both cohorts. About 16 percent of "never access" students could still enroll in public schools. For the "gained access" and "lost access" groups, we find a 47 to 60 percentage point difference in public school enrollment between policies.

Table 4.3: Share of Students in Each Simulated Public School Access Subgroup, and Public School Enrollment by Subgroup

		Changed Access		Same Access	
	All	Lost Access	Gained Access	Always Access	Never Access
Share of students in each group					
Zoning cohort	100	15.1	25.5	45.5	13.9
Merit cohort	100	19.4	24.1	45.1	11.4
Public school enrollment by group					
Zoning cohort	65.3	27.0	85.1	87.1	16.2
Merit cohort	58.5	71.6	25.5	85.3	16.4

Note: "Merit" indicates the last student cohort admitted under the merit policy, and "zoning" indicates the first cohort under the zoning policy. The table includes students with non-missing UASDA and SLA scores, although the simulation is performed on all students with non-missing UASDA scores. Numbers are corrected for under-sampling of private schools using sampling weights.

We could improve the accuracy of our predictions of public school enrollment if we knew which public schools the students prefer over private schools, and the order of their public school preferences. Unfortunately, we only have rank-ordered preference lists for students enrolled in public schools and not for the rejected students who had to enroll in private schools. Moreover, students could only list three public schools under the merit policy, so these lists are incomplete and may include strategizing. This means that the lists do not represent true preferences and would not be the same under the zoning policy as that policy requires a different strategy to secure a seat in a public school.

We argue that the two cohorts represent counterfactual groups under different policies. Balance tests between cohorts for the full sample and each of the groups are shown in

Table 4.4. We test for cohort differences in terms of characteristics including their grade 6 exam score, gender, age, socio-economic status and mother's education. We use questions from the student survey about assets to generate an asset index as a proxy for wealth or income.[63] For each characteristic, we present the mean value of the merit cohort, and the difference between the merit cohort and the zoning cohort.

Cohorts are balanced on average, and within each of the defined groups. Hence, the simulation identifies similar students in terms of these characteristics across cohorts. We only find a slightly lower grade 6 exam score for the zoning cohort in the "lost access" group. We perform robustness checks in Section 4.6.3 to confirm that this difference does not drive our impact results.

The mean values under the merit policy show that the zoning policy gave low-scoring and relatively poor students access to public schools, while rejecting high-scoring, wealthy students. "Gained access" students scored in the 21st percentile of the grade 6 exam, and their mean standardized asset index was -0.37 s.d. "Lost access" students scored in the 66th percentile, and had an asset index of 0.22 s.d.. "Always access" students had the highest baseline scores (73rd percentile), because part of the seats under the zoning policy were still allocated to students with the highest grade 6 exam scores. "Never access" students also had low scores (23rd percentile), but were somewhat wealthier than "gained access" students with a mean asset index of 0.

[63] Using principal component analysis, we converted the asset data into independent components; the component that explained the largest amount of variance of the original data was used as the asset index (Filmer and Pritchett, 2001).

Table 4.4: Balance Between Cohorts, on Average and Within Subgroups

	(1)	(2)	(3)	(4)	(5)	(6)
	Grade 6 percentile	Male	Age	Eligible for voucher	Asset Index	Mother completed tertiary education
All						
Zoning cohort difference	-1.60	-0.00	0.02	-0.01	-0.01	0.03
	(2.94)	(0.01)	(0.02)	(0.02)	(0.04)	(0.02)
Merit cohort mean	51.94	0.49	12.46	0.36	0.03	0.46
Observations	7510	7510	7468	7506	7505	5824
Gained Access						
Zoning cohort difference	1.08	0.01	0.00	-0.03	0.02	0.01
	(1.29)	(0.02)	(0.04)	(0.04)	(0.11)	(0.04)
Merit cohort mean	20.75	0.53	12.61	0.47	-0.37	0.27
Observations	1854	1854	1843	1853	1853	1306
Lost Access						
Zoning cohort difference	-3.58*	-0.02	-0.00	-0.02	0.07	0.10**
	(1.85)	(0.02)	(0.03)	(0.04)	(0.10)	(0.05)
Merit cohort mean	66.40	0.45	12.37	0.27	0.22	0.56
Observations	1300	1300	1291	1300	1299	1052
Always Access						
Zoning cohort difference	-1.67	0.01	0.05**	0.03	-0.06	0.03
	(2.16)	(0.02)	(0.02)	(0.03)	(0.05)	(0.02)
Merit cohort mean	73.21	0.44	12.37	0.33	0.19	0.53
Observations	3409	3409	3391	3406	3406	2834
Never Access						
Zoning cohort difference	0.11	-0.03	0.00	-0.04	-0.01	-0.03
	(1.15)	(0.02)	(0.04)	(0.04)	(0.09)	(0.04)
Pre-zoning	22.90	0.60	12.58	0.36	0.00	0.40
Observations	947	947	943	947	947	632

Note: Students were eligible for school vouchers if their household participated in the KMS program. Missing KMS participation was imputed using the asset index. Age was measured at the time of the grade 6 exam. Standard errors in parentheses and corrected for clustering at the school level. Numbers are corrected for under-sampling of private schools using sampling weights. * p<0.10 ** p<0.05 *** p<0.01

4.5.2 Empirical Model

On each of these groups and on the full sample, we estimate a model that compares the mean grade 8 test score between the merit and zoning cohort, conditional on students' grade 6 exam scores and other background characteristics. This is a value-added model as in Andrabi et al. (2011), specified in Equation 4.3 for student i

$$Y_i^2 = \beta_0 + \beta_1 T_i + \beta_2 Y_i^1 + \beta_3 X_i + \epsilon_i \tag{4.3}$$

where Y^2 is the standardized grade 8 score, and Y^1 is the standardized grade 6 exam score. T is a dummy variable indicating the treated zoning cohort. X is a vector of control variables for gender, age at the time of the grade 6 exam, the asset index, and whether the mother completed tertiary education. X also includes kelurahan (neighborhood) indicators that capture any cohort-invariant characteristics of a student's kelurahan. We interact each control variable with an indicator for missing values of that variable, such that all students with a non-missing grade 8 test scores and grade 6 exam scores are included in the model. ϵ is the residual. Standard errors are corrected for clustering at the school level.

Our coefficient of interest is β_1, which measures the difference in learning between the merit and first zoning cohorts for students from the same kelurahan and with the same grade 6 exam rank.[64] We can interpret this coefficient as the causal effect of the zoning policy on learning if learning between grade 6 and 8, as measured by the UASDA and SLA tests, would have been the same across cohorts in the absence of the reform. We already discussed in Section 4.3 that there were no other policy changes that could have affected value-added. We also showed that cohorts were similar in various characteristics that are generally correlated with learning in Table 4.4. Therefore, without any changes to school resources between cohorts, we argue that a change in test score value-added between two subsequent student cohorts in the absence of the zoning policy would be negligible. We show that our results are robust against several other specifications in Section 4.6.3.

This model measures reduced form estimates. Our measure for public school access does not perfectly predict actual public school enrollment (Table 4.3). The reduced form estimates are still policy relevant because they estimate the effect of giving low-scoring students the opportunity to enroll in better-resourced schools. In a school choice system, policymakers can only control who they give the opportunity to enroll; they cannot control actual enrollment decisions based on the students' preferences. Hence, our estimates inform policymakers about average learning effects from such a policy, taking into account enrollment decisions and other behavioral responses. We do not attempt to estimate the Average Treatment Effect on the Treated (ATT) of public school enrollment for "gained access" and "lost access" students. The exclusion restriction does not hold because the policy also affected students who did not enroll in a different school type (non-compliers within those groups) through a change in peers. We show this in the next section (Table 4.8).[65]

[64] Note that for the interpretation of these coefficients, it does not matter that the UASDA (grade 6) and SLA (grade 8) tests are different, as long as the UASDA score captures the contribution of previous inputs and unobservable resources (Singh, 2015; Andrabi et al., 2011). The coefficient for β_2 is 0.6, see Table 4.6, showing that the UASDA score is strongly correlated with the SLA score.

[65] Our simulated public school access could serve as an instrument for actual public school enrollment to calculate the Wald estimator for students whose access changed. Dalla-Zuanna et al. (2022) perform such an Instrumental Variables (IV) analysis on similar students groups as in our study, but they did not find evidence for peer effects on high school and university completion.

4.6 Results

We present the learning results for the full sample and for each of the four public school access groups. We also show how the school environment changed in terms of school quality and peers.

4.6.1 Average Learning Effects

On average, students enrolled in schools of somewhat better quality after the reform and classroom heterogeneity increased (Table 4.5). The increase in school value-added and the school-average grade 9 exam score is due to the larger number of public school seats for students in our sample (as explained in Section 4.4). Therefore, 6 percentage point more students enrolled in public schools on average. Changes in peers add up to zero, because the standardized grade 6 exam score distribution was the same between cohorts. However, heterogeneity in classroom composition increased by 0.17 s.d.. If school resources generate positive learning effects, the increase in public school enrollment could offset the negative instruction level effect from more heterogeneous classrooms, as explained in Section 4.2.

Table 4.5: Change in School Environment for Full Sample

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Public school	School value-added	School grade 9 exam score	School peers		Classroom peers	
				mean	*s.d.*	*mean*	*s.d.*
Zoning Cohort	0.06***	0.06***	2.12***	-0.05	0.27***	-0.04	0.17***
	(0.02)	(0.02)	(0.61)	(0.06)	(0.04)	(0.05)	(0.04)
Merit cohort mean	0.59	0.01	70.13	0.06	0.58	0.06	0.51
Observations			7510				

Note: Test scores are the average of test scores in mathematics and Indonesian. Observations for teacher experience are smaller due to missing information for these students' teachers. Standard errors in parentheses and corrected for clustering at the school level. Each model controls for the standardized UASDA score, gender, age at the time of UASDA exam, an asset index, an indicator for the mother having completed tertiary education and kelurahan (and indicators for missing values in these controls). Numbers are corrected for under-sampling of private schools using sampling weights. * p<0.10 ** p<0.05 *** p<0.01

We present the impact results on student outcomes in Table 4.6. Despite the increase in public school seats, learning declined by 0.08 s.d. on average (not significant). Under the assumption that school resources benefit all students in the same way, and ignoring effects from student composition, we would expect an improvement in learning equal to the increase in average school value-added of 0.06 s.d. This suggests that, on average, schools were hardly able to maintain the same learning production as they were prior to the policy change, potentially due to more heterogeneous classrooms. In addition, we find a small decline in the share of students who find the instruction level difficult (-3 percentage points, p.p.), in the share that aspires to go to university (-4 p.p.) and in the share that takes tutoring lessons at school (-9 p.p.) or private (-3 p.p.).

The results also demonstrate that the grade 6 exam score is a good predictor of the grade 8 score, despite these being different tests. The coefficient of the standardized grade 6 exam score is around 0.6 s.d., which is comparably high; Andrabi et al. (2011) found test score persistence, or coefficients for lagged test scores, of between 0.2 and 0.5 s.d..

Table 4.6: Effect on Test Scores for Full Sample

	(1)	(2)	(3)	(4)	(5)
	Grade 8 score	Finds instruction level difficult	Aspires to go to university	Takes tutoring classes	
				At school	*Private*
Zoning cohort	-0.08	-0.03*	-0.04***	-0.09*	-0.03**
	(0.05)	(0.02)	(0.01)	(0.05)	(0.01)
Grade 6 score	0.64***	-0.02**	0.10***	0.01	0.05***
	(0.02)	(0.01)	(0.01)	(0.02)	(0.01)
Merit cohort mean	-0.05	0.26	0.80	0.42	0.36
Observations	7510	7264	7464	7439	7430

Note: Test scores are the average of test scores in mathematics and Indonesian. Standard errors in parentheses and corrected for clustering at the school level. Each model controls for the standardized UASDA score, gender, age at the time of UASDA exam, an asset index, an indicator for the mother having completed tertiary education and kelurahan (and indicators for missing values in these controls). Numbers are corrected for under-sampling of private schools using sampling weights. * p<0.10 ** p<0.05 *** p<0.01

4.6.2 Heterogeneous Effects by Public School Access

The change in the school environment for each of the public school access groups is as we would expect, see Table 4.7. Like we showed before in Table 4.3, public school enrollment changed substantially for "gained access" and "lost access" students, while it stayed the same for "always access" and "never access" students. The public schools that the "gained access" students enrolled in produced 0.38 s.d. higher value-added than the private schools that their merit-based comparison group enrolled in. This equivalent figure is 0.22 s.d. lower value-added for the "lost access" group. These changes in school quality are also reflected in the school average grade 9 exam scores. The "gained access" students enrolled in schools that scored 12 points higher on average, while the "lost access" students enrolled in schools that scored 6 points lower.

The improvement in school quality for the "gained access" group is larger than the decline in school quality for the "lost access" group because of differences in their school choices. Their school choices are presented in Figure A4.4. "Lost access" students enrolled in relatively higher-ranked private schools in terms of value-added than "gained access" students did before the zoning policy.[66] Many of them enrolled in one of nine private schools that are ranked higher than the lowest ranked public school.[67] Therefore, in the absence of peer effects, we would expect to find a larger learning benefit to "gained access" students than a loss to "lost access" students.

The peer group changed for all four groups, but the change in peer scores is smaller at the classroom level than at the school level. At the school level, the peers of the "gained access" and "never access" groups scored 0.51 and 0.29 s.d. higher in the first zoning cohort than in the merit cohort, respectively. However, at the classroom level, their peers scored 0.31 s.d. and 0.21 s.d. higher, respectively. The new peers of the "lost access" and "always access" groups scored 0.29 and 0.36 s.d. lower at the school level, but 0.16 and 0.24 s.d. at the classroom level, respectively. This suggests that some public and private schools started grouping students by test scores (i.e. tracking) in response to zoning. In Appendix A4.5, we show that schools with a more heterogeneous student composition were more likely to track. Peer effects are generally stronger at the classroom level than at the grade level (Paloyo, 2020; Burke and Sass, 2013), so tracking could limit these effects. Despite tracking, heterogeneity in classroom composition significantly increased for all groups.

[66] Note that private schools were free to implement their own admission policy, so they may have selected based on test scores.

[67] If the number of private school seats is fixed, this suggests that "never access" students had to enroll in different private schools. Figure A4.4 indeed shows small differences in the distribution of "never access" students across private schools between the merit and zoning cohort. In Section 4.6.3 we show that our results are not driven by the differences in school by including school fixed effects.

Table 4.7: Change in School Environment by Public School Access Subgroup

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
		School	School grade 9				
	Public school	value-added	exam score	School peers		Classroom peers	
				mean	*s.d.*	*mean*	*s.d.*
Gained Access							
Zoning cohort	0.56***	0.38***	11.92***	0.51***	0.23***	0.31***	0.15***
	(0.06)	(0.07)	(1.79)	(0.11)	(0.02)	(0.11)	(0.05)
Merit cohort mean	0.26	-0.30	59.51	-0.67	0.63	-0.73	0.55
Observations	1854						
Lost Access							
Zoning cohort	-0.45***	-0.22***	-6.00***	-0.29***	0.20***	-0.16*	0.13***
	(0.06)	(0.07)	(1.75)	(0.10)	(0.05)	(0.08)	(0.04)
Merit cohort mean	0.72	0.14	74.45	0.33	0.59	0.38	0.52
Observations	1300						
Always Access							
Zoning cohort	0.00	0.02	0.67	-0.36***	0.36***	-0.24***	0.22***
	(0.02)	(0.01)	(0.53)	(0.07)	(0.05)	(0.06)	(0.05)
Merit cohort mean	0.85	0.21	77.62	0.58	0.51	0.62	0.46
Observations	3409						
Never Access							
Zoning cohort	-0.04	0.01	0.64	0.29***	0.12***	0.21***	0.12***
	(0.03)	(0.03)	(0.81)	(0.07)	(0.02)	(0.06)	(0.03)
Merit cohort mean	0.16	-0.27	60.33	-0.62	0.66	-0.68	0.56
Observations	947						

Note: Test scores are the average of test scores in mathematics and Indonesian. Observations for teacher experience are smaller due to missing information for these students' teachers. Standard errors in parentheses and corrected for clustering at the school level. Each model controls for the standardized UASDA score, gender, age at the time of UASDA exam, an asset index, an indicator for the mother having completed tertiary education and kelurahan (and indicators for missing values in these controls). Numbers are corrected for under-sampling of private schools using sampling weights. * p<0.10 ** p<0.05 *** p<0.01

We present effects on grade 8 test scores and other student outcomes in Table 4.8. Learning significantly declined for high-scoring students.[68] We find 0.23 s.d. lower grade 8 test scores for "lost access" students, and 0.13 s.d. lower test scores for "always access" students. However, learning improvements for low-scoring students were smaller and not statistically significant. Learning improved by 0.12 s.d. for "gained access" students (insignificant). We even find a negative coefficient for "never access" students (-0.03 s.d.), even though they got higher-scoring peers.[69]

The relative size and direction of these effects are in accordance with the hypotheses in our framework (Table 4.1). We find larger declines in learning for high-scoring students

[68] Figure A4.5 shows heterogeneous learning effects by grade 6 exam score quintile. We find significant negative coefficients between 0.1 and 0.2 s.d. for the three highest scoring quintiles, and positive but statistically insignficant coefficients between 0.05 and 0.1 s.d. for the two lowest scoring quintiles.

[69] We also estimate heterogeneous effects by gender, mother's education, and assets. We find a negative effect of about 0.1 s.d. for the groups with higher grade 6 exam scores: girls, students whose mother completed tertiary education, and students with an above median asset index. We find no effects on the other, lower-scoring groups.

("always access" and "lost access") than improvements in learning for low-scoring students ("never access" and "gained access"). By comparing the effects between groups, we can make two suggestive conclusions on the mechanisms. First, our findings for "always access" and "never access" students suggest that high-scoring students indeed benefit more from higher-scoring peers than low-scoring students do (taking the direct peer effects and instruction level effect together). In Appendix A4.5, we provide additional evidence for this finding by exploiting that some schools started tracking. Learning only declined for "always access" students in mixed classrooms; not for those in classrooms with similar peers as before the zoning policy. "Never access" students in mixed classrooms did not see learning improvements. Second, the larger negative coefficient for "lost access" students than for "always access" students, despite a smaller change in peers, suggests that "lost access" students saw an additional negative effect from the decrease in school quality on top of peer effects. The larger positive coefficient for "gained access" students than for "never access" students also suggests that they benefited from higher school quality. However, the "gained access" students also saw a larger increase in peer scores than "never access" students, so we cannot rule out that the difference in peers drove the difference in effects between those groups.

Table 4.8: Effect on Student Outcomes by Public School Access Subgroup

	(1)	(2)	(3)	(4)	(5)
	Grade 8 score	Finds instruction level difficult	Aspires to go to university	Takes tutoring classes	
				At school	*Private*
Gained Access					
Zoning cohort	0.12	0.04	-0.08***	-0.22***	-0.02
	(0.09)	(0.03)	(0.03)	(0.05)	(0.03)
Merit cohort mean	-0.86	0.27	0.64	0.46	0.23
Observations	1854	1785	1839	1834	1832
Lost Access					
Zoning cohort	-0.23**	-0.07**	-0.02	-0.02	-0.02
	(0.11)	(0.03)	(0.02)	(0.06)	(0.04)
Merit cohort mean	0.28	0.25	0.89	0.43	0.38
Observations	1300	1266	1294	1293	1292
Always Access					
Zoning cohort	-0.13*	-0.08***	-0.03	-0.06	-0.04*
	(0.07)	(0.02)	(0.02)	(0.06)	(0.02)
Merit cohort mean	0.52	0.26	0.89	0.41	0.44
Observations	3409	3305	3389	3377	3370
Never Access					
Zoning cohort	-0.03	-0.01	-0.05*	-0.07	-0.04
	(0.08)	(0.02)	(0.03)	(0.06)	(0.04)
Merit cohort mean	-0.79	0.26	0.72	0.37	0.29
Observations	947	908	942	935	936

Note: Standard errors in parentheses and corrected for clustering at the school level. Each model controls for the standardized UASDA score, gender, age at the time of UASDA exam, an asset index, an indicator for the mother having completed tertiary education and kelurahan (and indicators for missing values in these controls). Numbers are corrected for under-sampling of private schools using sampling weights. * p<0.10 ** p<0.05 *** p<0.01

We find suggestive evidence that public school teachers decreased the difficulty level of their instruction with lower-scoring students, whereas private school teachers did not increase the difficulty level with higher-scoring students. The share of students that find the instruction level difficult declined by 8 percentage points for “always access” students. The teacher survey results corroborate these findings: 23 percent of public school teachers report that they changed the difficulty of their instruction level (Table A4.11). The nearly zero coefficient for “never access” students indicates no change in the instruction level in private schools. In the teacher survey, only 13 percent of teachers in private schools reported to have changed their methods after the zoning policy (Table A4.11). Similarly, “gained access” students found the post-reform instruction level in public schools only somewhat more difficult than the pre-reform instruction level in private schools (insignificant), while “lost access” students found the post-reform private school instruction level significantly easier than the pre-reform public school instruction level. The share of “lost access” students that find the instruction level difficult declined by 7 percentage points. The easier instruction level in both public and private schools may have slowed down the learning progress of high-scoring students.

Even though "gained access" students saw large improvements in school quality and peer scores, their learning gains were modest and statistically insignificant. The framework suggests that a more difficult instruction level would limit their benefits, but we do not find evidence for this mechanism. Therefore, we looked at changes in indicators for these students' own educational investments (see Section 4.2 for a discussion of the literature). First, we find that the share taking tutoring classes at school almost halved. The decline in tutoring could be due to substitution between school resources and parents' own investments, or due to a higher price of tutoring classes in public schools than in private schools. Teacher reports corroborate that fewer students took tutoring classes after the zoning policy (Table A4.12). Second, the share that aspires to go to university declined by 8 percentage points.[70] Students may feel demotivated to aim for university enrollment when they have a lower rank in the classroom. If school-based tutoring classes and aspirations affect learning, these mechanisms could have limited learning benefits for "gained access" students. The policy could have long-term negative effects on "gained access" students if indeed fewer of them enroll in university.

4.6.3 Robustness Checks

We perform several checks to confirm the robustness of our results. Table A4.4 shows that our results are not driven by small differences between cohorts that we found in Table 4.4. First, our results do not depend on conditioning on background characteristics (column 5). Second, applying Inverse Probability Weighting (IPW) for improved balance makes little difference to our impact estimates (column 6). We weight the observations with the inverse of the estimated probability to be in the first zoning cohort based on background characteristics. We estimate the propensity to be in the first zoning cohort using a logit model that includes KMS participation, the standardized mean UASDA score, the kelurahan, the asset index, gender, age at the time of the UASDA exam, and whether the student's mother completed tertiary education.[71]

In addition, we use Least Absolute Shrinkage and Selection Operator (LASSO) machine learning techniques as an alternative method to predict public school enrollment. We estimate the probability of enrollment in public school under each policy using a logit

[70] We checked differences in these effects by gender because generally girls are more likely to get demotivated from negative feedback. However, we find that the effect was concentrated amongst boys (not shown). This may be explained by boys having lower classroom ranks than girls on average. The boys in the "gained access" group scored 8 percentiles lower on the grade 6 exam than girls.

[71] All variables are interacted with an indicator for a missing value, such that all students with a non-missing UASDA score are included in the model.

model, including grade 6 exam scores, age, KMS status, asset index, mother's education, gender, distance to each public school, their kelurahan, indicators for missing values and interaction terms between the grade 6 exam score, KMS status and distance to each public school.[72] We allocate students with the highest predicted probability of public school enrollment under each policy to the available public school seats to determine the same four groups. Note that this predicted probability of public school enrollment depends both on the students' eligibility based on the selection criteria and their school preferences. This may make the predictions of enrollment more accurate than our main simulation results, but we cannot interpret the LASSO results causally as intent-to-treat estimates. Hence, the main purpose of this analysis is to show that our method to predict public school access performs well compared to data-driven methods. Table A4.3 shows that the LASSO method indeed achieves a somewhat larger difference in public school enrollment between the zoning and merit cohort for the "gained access" and "lost access" groups. However, Table A4.4 shows that this method results in lack of balance. Cohorts differ from each other within the groups in terms of their grade 6 exam percentile (column 3). Nevertheless, the impact results are similar to our main results (column 7).

We also show that our results for students whose access remained the same are not driven by a change in schools or assigned teachers by including school and teacher fixed effects. Although the zoning and merit cohort students in the "always access" and "never access" groups enrolled in schools with on average similar school value-added (Table 4.8), the cohorts were not distributed across schools in exactly the same way (Figure A4.4). Moreover, because we find that some schools started tracking their classrooms, one might be concerned that the decline in learning on "always access" students is due to exposure to different teachers. For instance, schools could allocate better teachers to the classrooms with low-scoring students to help them.

We add school fixed effects to make sure that we are comparing students within the same schools, and add teacher fixed effects to compare students who were taught by the same teachers. The model with school fixed effects is specified in Equation 4.4 for student i in school s

$$Y_{i,s}^2 = \gamma_1 T_{i,s} + \gamma_2 Y_{i,s}^1 + \gamma_3 X_{i,s} + \lambda_s + \epsilon_{i,s} \tag{4.4}$$

where λ_s are school indicators that capture cohort-invariant school characteristics. We then replace the school fixed effects with teacher fixed effects and remove students who were taught by teachers that only worked for that school in either 2019 or 2020 (see

[72] The software finds λ_{gmax}, which is the smallest λ that excludes all variables from the model. We then use Cross Validation as the method of variable selection. This method chooses the model λ that minimizes an estimate of the out-of-sample prediction error.

Appendix A4.4 for further discussion of teacher turnover). Indonesian secondary schools have subject teachers, so we estimate the teacher fixed effect model for each subject separately.

Table 4.9 shows that the results are similar to our main results. We still do not find significant effects on test scores for "never access" students, if anything we find a larger decline in learning of 0.08 s.d.. Including school fixed effects increases the coefficient for the zoning cohort from 0.13 to 0.16 s.d. and it is more precisely estimated. Results are also similar when including teacher fixed effects. That "always access" students in the same schools taught by the same teacher saw a decline in learning suggests that this is indeed driven by a change in their peer composition.

Table 4.9: Effect on Test Scores for Those Whose Access Remained the Same and Were Taught by the Same Teacher

	(1)	(2)	(3)
	Grade 8 score		
	Average	Mathematics	Indonesian
Always Access			
Zoning cohort	-0.16**	-0.19**	-0.13*
	(0.07)	(0.09)	(0.07)
Observations	3409	2870	2567
Merit cohort mean	0.52	0.58	0.44
Never Access			
Zoning cohort	-0.08	-0.05	-0.03
	(0.08)	(0.09)	(0.10)
Observations	947	713	735
Merit cohort mean	-0.79	-0.77	-0.63
School fixed effects	Yes	No	No
Teacher fixed effects	No	Yes	Yes

Note: Standard errors in parentheses and corrected for clustering at the school level. Each model controls for the standardized UASDA score, gender, age at the time of UASDA exam, an asset index, an indicator for the mother having completed tertiary education and kelurahan (and indicators for missing values in these controls). Numbers are corrected for under-sampling of private schools using sampling weights. * p<0.10 ** p<0.05 *** p<0.01

4.7 Discussion and Conclusion

This chapter provides empirical evidence that making high-quality schools less selective can affect learning outcomes of students throughout an education system. We show that an equality-enhancing admission policy can generate larger learning losses to high-achieving students than gains to low-achieving students in a context with large variation in school quality and student performance. We provide evidence for substantial learning effects from an exogenous decrease in peer scores on incumbent students. Despite much attention for admission policies in the literature, only few papers study effects of school integration on incumbent students and students displaced to lower-quality schools.

The "zoning policy" expanded access to high-quality public schools for relatively low-achieving students, while displacing relatively high-achieving students to lower-quality private schools. Grade 6 exam scores of incoming students were 0.4 s.d. lower in public schools after the zoning policy, and 0.4 s.d. higher in private schools. Overall, we find that the zoning policy resulted in a small decline in learning. We find moderate learning improvements of 0.12 s.d. for "gained access" students, but large learning losses of 0.23 s.d. for "lost access" students. We also find that the large influx of low-scoring students in public schools generated negative effects of 0.16 s.d. on "always access" students. However, "never access" students in private schools did not benefit from higher-scoring peers.

We explored teacher responses to changes in student composition, and student responses to changes in school quality and peers, as mechanisms that may explain differences in learning effects between groups. First, we find suggestive evidence that teachers decreased the difficulty of their instruction level in public schools after the influx of low-scoring students, while teachers in private schools did not increase the difficulty of their instruction levels. Both "always access" and "lost access" students found the instruction level easier after the zoning policy. In addition, "gained access" students halved their take-up of tutoring classes and fewer of them aspired to go to university.

Our findings have two main implications for policies that seek to promote integration. First, high-achieving students seem to benefit more from enrollment in high-quality schools than low-achieving students, likely because learning improvements of being grouped with higher-achieving peers are larger for this group. Second, the performance of high-quality schools as measured by school value-added partly depends on student composition, especially if the change is large. Students of similar backgrounds in the same schools learned less with lower-achieving peers. Hence, the impacts of school admission policies depend on the magnitude of the change in student composition, and how key actors in the education system may react. This is relevant when generalizing results of studies to impacts of admission policies in other contexts.[73]

Our study shows that, at least in the short-run, there is a trade-off between learning equality and average learning when the change in student composition in large. Policy-makers may accept the learning decline among high-achieving students if they aim for a

[73] For instance, recent studies in Pakistan aim to predict how redistributing students across schools could improve learning outcomes when school value-added differs across student types but is unchanging with student composition (Andrabi et al., 2020; Bau, 2022). The assumption that the peer group does not affect learning is mostly based on a small share of students changing schools, which is unlikely to lead to changes in teaching methods.

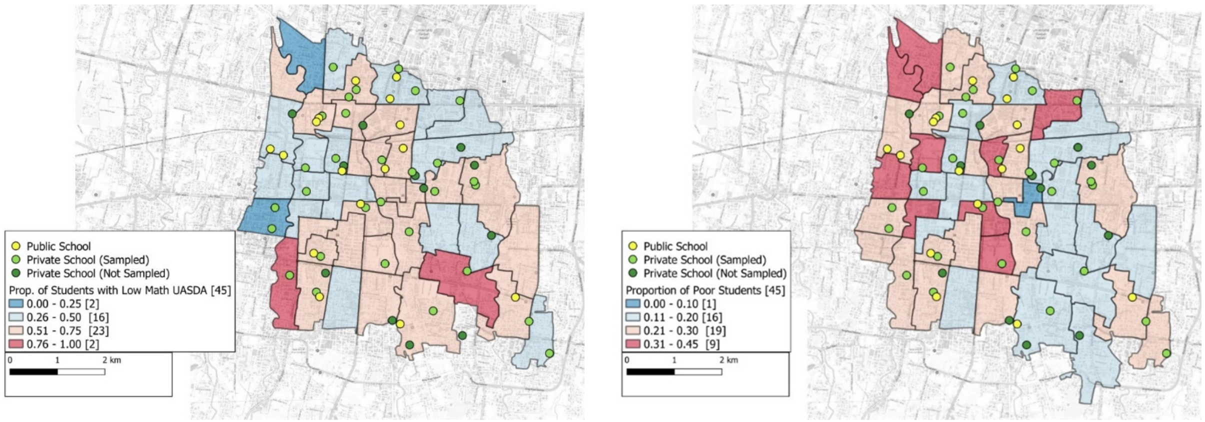
fairer education system. If they aim for improving learning outcomes among low-scoring students while limiting learning losses for high-scoring students, additional investments are likely necessary. They could consider supporting teachers in dealing with heterogeneous classrooms, or improving resources in low-quality schools. The latter would allow for some degree of ability sorting, while avoiding allocation of more and better resources to high-achieving students.

In closing, we stress several caveats when interpreting our results. First, our study timeframe was a brief 18 months. In public schools, the benefits of the "gained access" group could grow over time; and the negative effects for the "always access" group could decline, as teachers become more comfortable with the new student body. Second, we caution against residency-based admissions policies, as they could in the long run encourage greater residential segregation when wealthier parents move closer to the most preferred schools (Abdulkadiroğlu and Andersson, 2022; Black, 1999). Finally, our study examined only one narrow primary outcome (test scores), while benefits of selective schools might materialize for other outcomes, such as non-cognitive skills, university enrollment, wages or other aspects of human capital (Anstreicher et al., 2022; Jackson et al., 2020). Future studies would benefit from considering a wider range of outcomes over a longer timeframe.

Appendix

A4.1 Tables and Figures

Figure A4.1: Student Scores and SES by Kelurahan in Yogyakarta

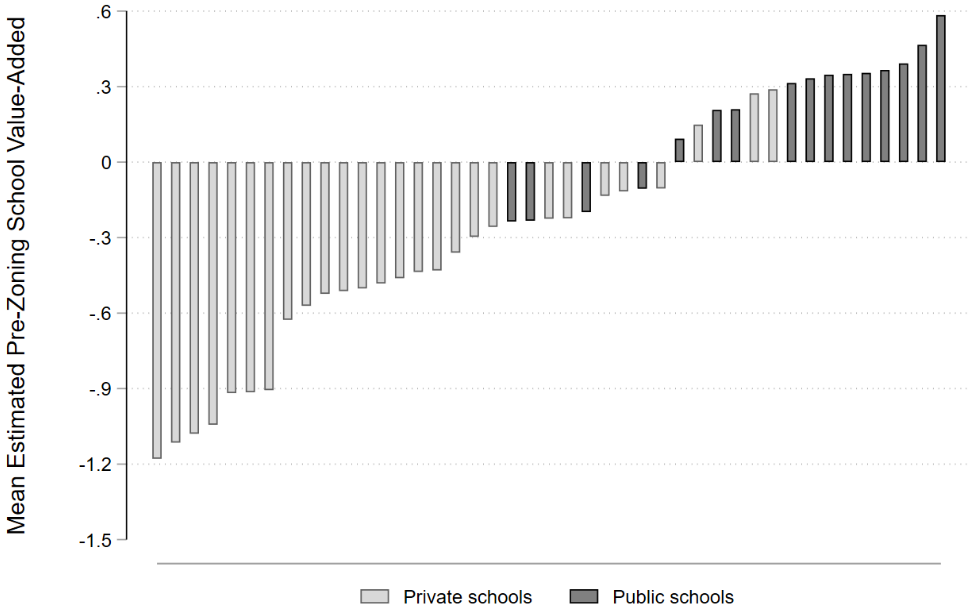


(a) Proportion of Students Scoring Below Median on the Mathematics Grade 6 Exam

(b) Proportion of Students who Participated in KMS (Poor)

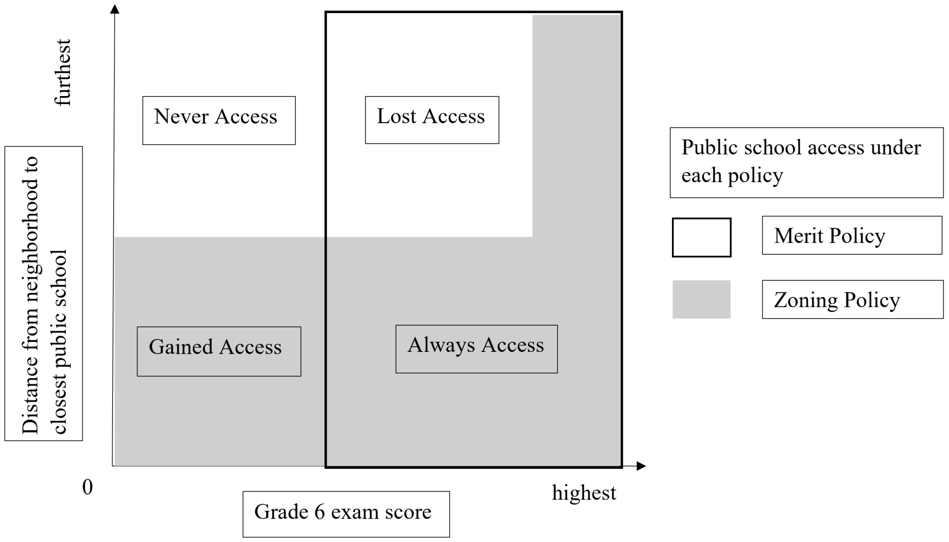
Source: Administrative data from the Yogyakarta education agency 2017.
Note: Sampling details are in Section 3. We surveyed all public schools and a subset of private schools for this study. The background colors indicate the share of students scoring below the median on the grade 6 leaving exam in mathematics and the share who received KMS benefits (i.e. poor) by kelurahan, i.e., the borders are kelurahan, an administrative unit equivalent to a village. Brackets indicate number of kelurahan.

Figure A4.2: School Value-Added Under Merit Policy



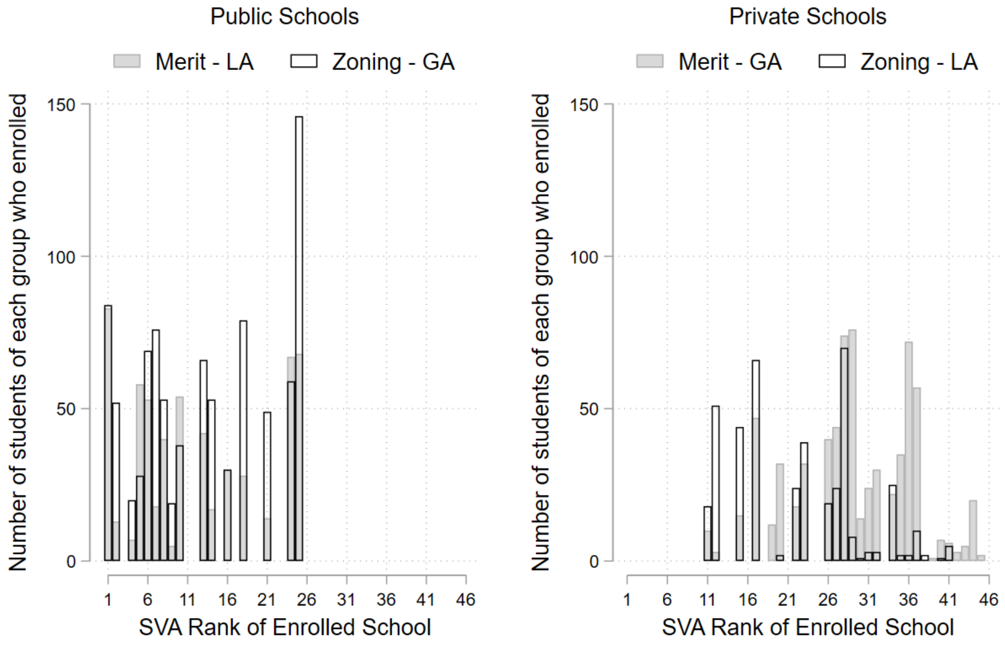
Note: Figure plots ρ_s as specified in Equation 4.2. The figure excludes three private schools which had zero or only one observation with non-missing grade 6 and grade 8 scores. These schools are small with less than 8 merit cohort students.

Figure A4.3: Schematic Illustration of Change in Public School Access by Student Type

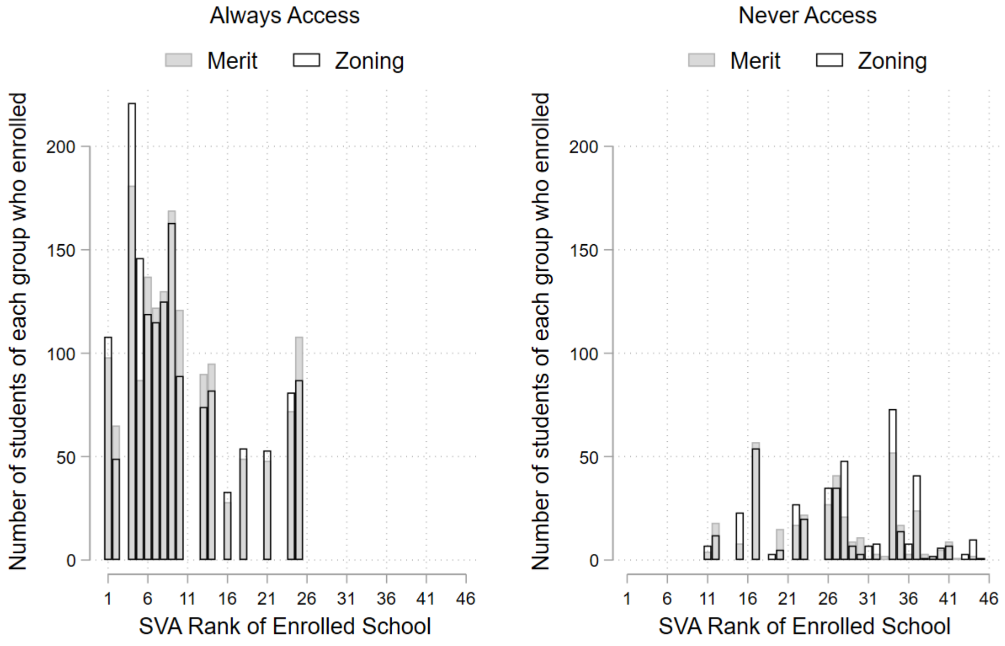


Note: This is a graphic depiction of the simulation to predict public school access by student type as described in Section 4.5.1. For simplicity, the figure ignores admission by KMS status in the merit policy although we take KMS participation into account in the actual group identification. The yellow and green blocks represent students with UASDA scores that qualified them for public school, all of whom who had public school access under the merit policy and a small share of whom maintained this access under the first zoning policy (and remained green) as under zoning there were still 15 percent of students admitted using a UASDA ranking. We refer to this latter group as the "always access" group. Students in the yellow block only had access under the merit policy and are called the "lost access" group. Students in the blue block only had access under the first zoning policy, defined as "gained access," since they lived closest to public schools and had relatively lower UASDA scores. Students with relatively lower UASDA scores who lived farther from public schools are represented in the white the "never access" group.

Figure A4.4: Enrollment in Public and Private Schools by Subgroup and Cohort



(a) Gained and Lost Access



(b) Always and Never Access

Note: LA is short for "lost access" and GA is short for "gained access". SVA stands for school value-added. Two private schools have missing bars in the figure, suggesting that no "gained" or "lost" access students enrolled there. These are small schools with fewer than 10 students per grade.

Table A4.1: Student Characteristics Before and After the Zoning Policy, by Public and Private School

	(1)	(2)	(3)	(4)	(5)	(6)
	Public schools			Private schools		
	Merit	Zoning	Difference	Merit	Zoning	Difference
Standardized grade 6 exam score	0.55	0.10	-0.44***	-0.63	-0.16	0.47***
	(0.76)	(0.96)	(0.08)	(0.88)	(0.90)	(0.07)
Within school grade 6 exam standard deviation	0.50	0.89	0.39***	0.67	0.76	0.08***
	(0.16)	(0.09)	(0.05)	(0.09)	(0.10)	(0.03)
Class size	30.82	31.47	0.65	25.34	26.80	1.46**
	(2.45)	(1.77)	(0.40)	(5.07)	(4.23)	(0.65)
Male	0.45	0.48	0.03*	0.55	0.50	-0.05**
	(0.50)	(0.50)	(0.01)	(0.50)	(0.50)	(0.02)
KMS participant	0.39	0.42	0.04	0.32	0.22	-0.10***
	(0.49)	(0.49)	(0.02)	(0.47)	(0.42)	(0.03)
Standardized asset index	-0.01	-0.12	-0.10*	0.09	0.26	0.18***
	(1.01)	(0.95)	(0.06)	(1.05)	(0.98)	(0.06)
Mother completed tertiary education	0.44	0.43	-0.01	0.49	0.61	0.12***
	(0.50)	(0.50)	(0.02)	(0.50)	(0.49)	(0.04)
Distance from neighborhood to school (km)	1.09	0.36	-0.73***	1.53	1.90	0.37***
	(1.23)	(0.65)	(0.08)	(2.01)	(2.36)	(0.13)
Travel minutes to school	17.58	13.42	-4.16***	19.10	16.09	-3.02***
	(11.20)	(9.52)	(0.32)	(15.64)	(10.86)	(0.75)
Moved house in grade 6	0.11	0.11	-0.01	0.14	0.11	-0.03**
	(0.32)	(0.31)	(0.01)	(0.34)	(0.31)	(0.01)
Observations	2,503	2,592	5,095	1,400	1,062	2,462

Note: Table includes students for whom we have a UASDA and an SLA score. The change in student composition in terms of gender, KMS participation, the asset index and mother's education looks similar when including students for whom we don't have a UASDA score (not shown). Standard deviations are in parentheses and standard errors between brackets. The number of observations is slightly different for mother's education due to students not knowing their mother's education level (n=3,122 in the merit cohort and n=2,968 in the zoning cohort). Stars indicate the significance level of the difference with the mean of the zoning cohort as estimated using a t-test, corrected for clustering at the school level. * p<0.10 ** p<0.05 *** p<0.01

Table A4.2: Correlation Between Grade 6 Exam Score and Distance to Closest Public School by Cohort

	(1)	(2)	(3)
	Grade 6 exam score		
	Average	Mathematics	Indonesian
Distance to closest public school	0.02	0.02	-0.00
	(0.02)	(0.02)	(0.02)
Zoning cohort	-0.03	-0.01	-0.03
	(0.11)	(0.12)	(0.08)
Second zoning cohort	0.09	0.09	0.08
	(0.07)	(0.07)	(0.06)
Zoning cohort × Distance	-0.05	-0.06	-0.03
	(0.04)	(0.05)	(0.04)
Second zoning cohort × Distance	-0.06	-0.08	-0.04
	(0.04)	(0.05)	(0.03)
Observations	11,517	11,517	11,517

Note: Average UASDA score is the unweighted average score in math, Indonesian and science. Standard errors in parentheses and corrected for clustering at the school level. Each model controls for gender, age at time of the UASDA exam, an asset index, an indicator for the mother having completed tertiary education and kelurahan (and indicators for missing values in these controls). Numbers are corrected for under-sampling of private schools using sampling weights. * p<0.10 ** p<0.05 *** p<0.01

Table A4.3: Actual Public School Enrollment by Simulated Public School Access Category and Cohort

	Lost Access		Gained Access		Always Access		Never Access	
	Main	LASSO	Main	LASSO	Main	LASSO	Main	LASSO
Share in public school								
Zoning cohort	27.0	16.9	85.1	82.3	87.1	88.4	16.2	16.9
Merit cohort	71.6	77.4	25.5	19.1	85.3	87.9	16.4	14.6
Share of sample								
Zoning cohort	15.1	11.9	25.5	17.3	45.5	53.7	13.9	17.1
Merit cohort	19.4	12.8	24.1	18.2	45.1	51.4	11.4	17.7

Note: “Merit” indicates the last student cohort admitted under the merit policy, and “zoning” indicates the first cohort under the zoning policy. The table includes students with non-missing UASDA and SLA scores, although the simulation is performed on all students with non-missing UASDA scores. Numbers are corrected for under-sampling of private schools using sampling weights.

Table A4.4: Robustness Checks: No Control Variables, Inverse-Probability Weighting (IPW) Within Cohorts, Group Allocation Using LASSO

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Balance on grade 6 score			Grade 8 score			
	Main	IPW	LASSO	Main	No controls	IPW	LASSO
All							
Zoning cohort	-1.60	-0.43		-0.08	-0.08	-0.07	
	(2.94)	(3.20)		(0.05)	(0.06)	(0.06)	
Merit cohort mean	0.06	0.03		-0.05	-0.05	-0.06	
Observations	7510	7469		7510	7510	7468	
Gained Access							
Zoning cohort	1.08	-1.49	3.78*	0.12	0.14	0.12	0.11
	(1.29)	(1.40)	(2.09)	(0.09)	(0.11)	(0.10)	(0.10)
Merit cohort mean	-1.01	-0.89	-1.03	-0.86	-0.86	-0.77	-0.89
Observations	1854	1843	1326	1843	1854	1843	1320
Lost Access							
Zoning cohort	-3.58*	0.72	-4.14	-0.23**	-0.22*	-0.23*	-0.18*
	(1.85)	(2.06)	(4.52)	(0.11)	(0.12)	(0.12)	(0.10)
Merit cohort mean	0.55	0.42	0.72	0.28	0.28	0.20	0.41
Observations	1300	1291	927	1291	1300	1291	926
Always Access							
Zoning cohort	-1.67	0.26	-4.71**	-0.13*	-0.14	-0.13	-0.12
	(2.16)	(2.39)	(2.18)	(0.07)	(0.09)	(0.08)	(0.08)
Merit cohort mean	0.78	0.72	0.59	0.52	0.52	0.47	0.36
Observations	3409	3392	3950	3391	3409	3391	3924
Never Access							
Zoning cohort	0.11	0.32	-4.80**	-0.03	-0.09	-0.04	-0.03
	(1.15)	(1.27)	(1.83)	(0.08)	(0.08)	(0.10)	(0.06)
Merit cohort mean	-0.91	-0.92	-0.55	-0.79	-0.79	-0.84	-0.48
Observations	947	943	1307	943	947	943	1298

Note: Standard errors in parentheses and corrected for clustering at the school level. Observations in model (2) and (3) are weighted with the inverse of the propensity to be in the first zoning cohort, calculated in each group separately. Models (3) and (5) control for gender, age at the time of UASDA exam, an asset index, an indicator for the mother having completed tertiary education and kelurahan (and indicators for missing values in these controls). Numbers are corrected for under-sampling of private schools using sampling weights. * p<0.10 ** p<0.05 *** p<0.01

A4.2 Data

The students in our sample are graduates from primary schools in Yogyakarta who enrolled in any of the public junior high schools or any of the 30 (out of 41) sampled private schools. Table A4.5 shows how our sample is build up. We have grade 6 exam score data for all graduates from primary schools in Yogyakarta[74] (row 1) and we tested all students enrolled in the sampled secondary schools using the SLA in grade 7 and grade 8. Because the education system in Yogyakarta is open to students from surrounding districts, and because students from Yogyakarta can enroll in secondary schools in surrounding districts, these groups are not the same. Row 4 shows the number of students who graduated from a primary school in Yogyakarta and enrolled in one of the sampled secondary schools. The total number of students enrolled in the sampled schools was smaller in the zoning cohort (row 3) but this seems to be driven by students from outside Yogyakarta since we do not see a similar decline in row 4. Our data confirm that most students enrolled in the sampled schools for whom we do not have a grade 6 exam score lived outside Yogyakarta: about 20 percent of students with a non-missing grade 6 exam score lived outside the city in each cohort, while this was about 80 percent of students with a missing grade 6 exam score (not shown). The mixed policy sample is analysed in Appendix A4.3.

Table A4.5: Sample

	(1)	(2)	(3)
	Merit	Zoning	Mixed
Grade 6 exam takers	7139	7200	7345
Enrolled in unsampled private schools	704	646	794
Enrolled in sampled schools	5943	5590	5874
Grade 6 exam + in sampled school	4080	4002	4403
Grade 6 exam + in sampled school + grade 7 SLA		3834	4130
Grade 6 exam + in sampled school + grade 8 SLA	3870	3640	

Note: The grade 6 exam (UASDA) is administered in May while the 2019 SLA was administered in January and the 2020 SLA in February. Table includes students from all 16 public schools and 30 out of 41 private schools for whom we have a UASDA test score.

We believe that the full population of grade 6 exam takers (row 1) should be comparable over time due to universal primary school enrollment. However, the exam itself has different items across years, potentially creating differences in the difficulty of the test. Later cohorts scored significantly lower on the mathematics test and higher on the

[74] Grade 6 leaving exams are not comparable between districts, so we did not attempt to get access to that data for students from outside the city.

Indonesian test compared to the cohort one year before them.[75] Because the average mathematics score went down and the Indonesian score went up during our study period, we suspect that the tests were of varying difficulty (even if the local government aims to make all UASDA equally difficult) rather than the three subsequent cohorts of students having different underlying skills.[76] Therefore, to be able to compare grade 6 exam scores between cohorts, we standardize the exam score within each cohort, assuming that the exam ranks students in the same ways. Our standardization method improves comparability across cohorts, assuming the tests rank students similarly, even if the tests were of varying difficulty.

The sample we use in our main analysis is shown in row 6. There are two possible sources of selection bias in our analysis. First, selection of primary school graduates from Yogyakarta into the sampled schools could be different under each policy. For instance, high-scoring students may prefer schools outside Yogyakarta over the public schools if they anticipate a low-scoring peer group in the public schools after the zoning policy. In Table 4.4 we found no evidence for such selection as students in our sample are comparable between cohorts on several characteristics, including the grade 6 exam percentile score based on the exam score distribution of all exam takers. Second, the difference in observations between row 4 and row 6 is due to small attrition. The students who refused to participate in the SLA had about 0.3 s.d. lower standardized UASDA scores than students who participated (not shown) and attrition is larger in the zoning cohort than in the merit cohort (the difference between row 4 and 6 is 9.0 percent in the zoning cohort and 5.1 percent for the merit cohort). However, on average, Table A4.6 shows that the attrition makes little difference to the average standardized grade 6 exam score in our sample. Leaving out students for whom we do not have a SLA score increases the grade 6 exam score with 0.02 s.d. in both cohorts, or with 0.5 percentile in the merit cohort and 0.7 percentile in the zoning cohort. In addition, the sample is balanced across cohorts on various student characteristics (Table 4.4). Therefore, we believe that attrition does not affect the comparability of the merit and zoning cohort nor the validity of our analysis.

[75] The first zoning cohort scored on average 2 points lower on the mathematics test and 5 points higher on the Indonesian test than the merit cohort, and the mixed policy cohort scored 9 points lower on the mathematics test and 1 point higher on the Indonesian test than the first zoning cohort.

[76] Grading was also done slightly differently between the merit and zoning 1 cohorts. Test scores were rounded in the merit cohort such that the math UASDA scores were reported in 2.5-point intervals and the Indonesian UASDA scores were reported in 2-point intervals. There was no rounding in the zoning 1 or zoning 2 cohorts. However, if we round the zoning 1 and 2 cohort scores as was done for the merit cohort, this does not explain the difference in mean scores between the cohorts.

Table A4.6: Average Grade 6 Exam Scores for Students Enrolled in the Sampled Schools by SLA Participation

	(1)	(2)	(3)	(4)
	Merit		Zoning	
	All	Non-missing SLA	All	Non-missing SLA
Standardized grade 6 exam score	0.04	0.06	-0.01	0.01
	(1.00)	(1.00)	(0.96)	(0.95)
Grade 6 exam percentile	51.42	51.94	49.63	50.34
	(29.14)	(29.11)	(28.65)	(28.46)
Observations	4080	3870	4002	3640

Note: The grade 6 exam score is standardized using the distribution of the full population of exam takers.

Finally, we make several minor assumptions due to data limitations. First, we are missing neighborhood data for 9 percent of students in the first zoning cohort and 24 percent in the merit cohort.[77] These location data are missing for several schools entirely. If data were missing, we imputed with the predicted neighborhood based on the students' primary school, assuming that these students live nearby their primary school. Public primary schools in Yogyakarta have allocated seats based on house-to-school proximity for decades.[78] Second, we are missing data on whether students participated in the KMS program (i.e. were eligible for school vouchers) for 23 percent of the merit cohort and 33 percent of the first zoning cohort. We assumed students with missing data with an above-mean asset index did not qualify for the KMS program. Among students we know participated in KMS, 82 percent had a below-mean asset index, while 38 percent of students who did not participate in KMS had a below-mean asset index.

[77] We have location data for more students in the first zoning cohort than in the merit cohort because the house location is included in the admissions data for the first zoning cohort, but not for the merit cohort.

[78] We are missing primary school information for zoning students from one private school, so we assumed that those students with missing data lived near this private school. This is unlikely to affect the group allocation as these private school students probably lived too far from public schools to have access.

A4.3 The mixed policy

Many education officials and parents in Yogyakarta resisted the zoning policy. District education leadership expected that zoning could provide greater access to public schools for lower-testing students who lived near public schools. However, they were still concerned that these changes would not adequately help traditionally excluded students, that some students would not have access to any public school because they lived in a part of the city without public schools, and that zoning would bring down average grade 9 leaving exam scores, hurting Yogyakarta's reputation as one of the highest-performing districts in the country. Many parents felt that their high-testing students would be unfairly excluded from public schools. Their resistance took the form of city government making numerous appeals throughout 2017 to the MoE for an exemption to implementing the zoning policy. In late 2018, Yogyakarta's mayor requested that officials from the local education agency meet with MoE lawyers to discuss their appeal (interview with local government officials, November 2018). These appeals failed, and Yogyakarta reluctantly first implemented the zoning policy in the 2018-19 school year.

Ultimately, the city government sided with parents, again making the UASDA a critical component of admissions in 2019. Parents lobbied government representatives, some with the help of community-based organizations, and the local ombudsman, which had the potential to lead to a lawsuit against the city (interview with Ombudsman, August 2018). The city also threatened to file a lawsuit against the Ministry of Education. Therefore, the zoning policy was partly reversed to, what we call, the "mixed policy". The third column in Table A4.7 indicates how the number of seats allocated based on each selection criteria was revised. The policy revision happened after only one year of zoning. The mixed policy in 2019 was not a complete reversion to the merit policy, but increased the UASDA-based share of seats in a school to 40 percent, compared to 15 percent under the first zoning policy and 55 percent under the merit policy. Under the mixed policy, just 30 percent of seats were allocated according to proximity criteria, while this was 75 percent under the first zoning policy. Another 10 percent of seats were reserved for a new category of "talented students;" students in this category were designated by the district as having a special talent related to athletics, arts, or academics. These students were also admitted based on the UASDA score rank. Another 10 percent of seats were reserved for KMS-participating students and ranked by UASDA scores. The remaining 10 percent of seats were allocated to students from outside Yogyakarta and from relocated families, as was the case with the first zoning policy. This 2019 policy continues today.

Table A4.7: Allocation of Seats Within Each School Under the Merit and Zoning Policies

Student Category	Selection Criterion	Share of seats		
		Merit	Zoning	Mixed
Yogyakarta residents	Grade 6 exam score	55	15	40
Poor Yogyakarta residents	Grade 6 exam score	25	0	10
Yogyakarta residents	Proximity to school	0	75	30
Non-Yogyakarta residents	Grade 6 exam score	20	5	5
Relocated students	Grade 6 exam score	0	5	5
"Special talents"	Grade 6 exam score	0	0	10

Note: The students' grade 6 exam score is an unweighted average of a student's mathematics, Indonesian, and science scores. Proximity is measured by the difference between the student's neighborhood and each school. Poverty status is determined by participation in the KMS program.

We conducted an analysis of the mixed policy primarily to support our claim that one can interpret impacts of the policy change as causal. We only have test score data for the mixed policy student cohort in grade 7, 6 months after enrollment. While the six-month time frame over which we have learning data is possibly too short a period to allow schools to readjust to the policy change, we can show how the student composition and initial learning changed as a result of the second policy. Because the mixed policy partially reversed the first policy, reallocating more seats to higher scoring students again, we expected this partial reversal to result in opposite signs for the impact coefficients for the zoning and mixed policy cohorts.

It was complicated to replicate the mixed policy with four public school access groups for students who attended primary school in Yogyakarta because we cannot identify the "special talent" students (10 percent of public school seats) and because location data is missing for 23 percent of students, even after imputations with primary school location. Therefore, we show these results by UASDA quintile instead of the groups specified in Section 4.5.1. We estimated the same value-added model, as specified in Equation 4.3, to compare test score value added between the zoning and the mixed policy cohorts within each quintile. We present a balance test between the zoning and mixed policy cohort in Table A4.8. The cohorts were similar in terms of UASDA scores and wealth. We cannot directly compare the mixed policy cohort to the merit cohort because the merit cohort was only tested in grade 8.

Table A4.8: Balance Between Zoning and Mixed Policy Cohorts

	(1) Grade 6 percentile	(2) Male	(3) Age	(4) Eligible for voucher	(5) Asset Index	(6) Mother completed tertiary education
Mixed policy cohort	0.02	0.02	0.03*	-0.00	0.00	-0.02
	(0.07)	(0.01)	(0.02)	(0.02)	(0.03)	(0.01)
Zoning cohort mean	0.01	0.49	12.50	0.36	0.00	0.49
Observations	7,896	7,896	7,855	7,892	7,889	5,777

Note: Table shows the difference between the zoning and mixed policy cohorts. Students were eligible for school vouchers if their household participated in the KMS program. Missing KMS participation was imputed using the asset index. Age was measured at the time of the grade 6 exam. Standard errors in parentheses and corrected for clustering at the school level. Numbers are corrected for under-sampling of private schools using sampling weights. * p<0.10 ** p<0.05 *** p<0.01

Before showing the impact results in Figure A4.5, we confirmed that public school enrollment indeed reversed with the mixed policy. Table A4.9 shows the share of each UASDA quintile enrolled in public school by cohort. The highest-scoring UASDA quintile saw only a small 12 percentage point decline in public school enrollment, while the lowest-scoring quintile saw a 45 percentage point increase in public school enrollment after the first zoning policy. Table A4.9 shows that the highest-scoring UASDA quintile saw an increase in public school enrollment again with the mixed policy, from 77 to 80 percent enrolled in public school, and the lowest-scoring UASDA quintile saw a 27 percentage point decrease, about half the size of the public school enrollment change between the merit and first zoning policy.

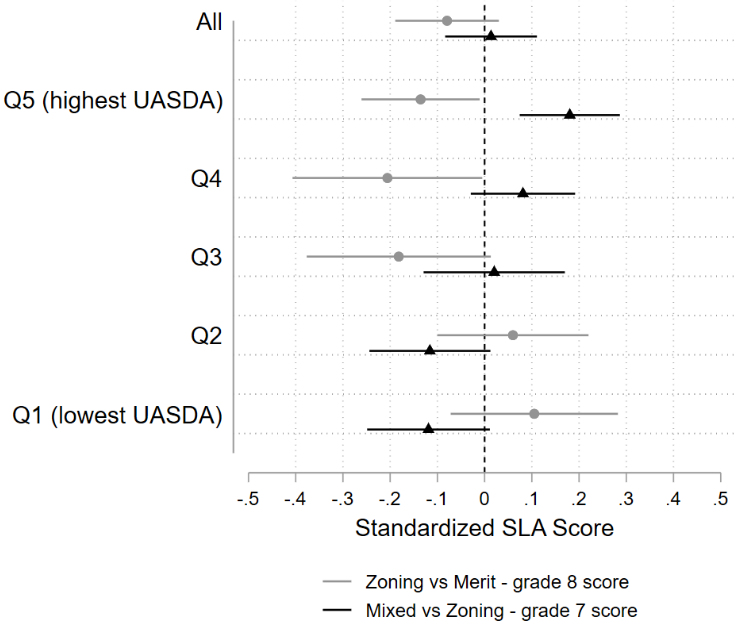
Table A4.9: Percent of Students Enrolled in Public School by Grade 6 Exam Score Quintile and Policy Type

	Percent in public school		
	Merit	Zoning	Mixed
Total	58.5	65.3	60.8
Quintile 1 (lowest)	13.9	58.5	31.1
Quintile 2	43.4	60.0	50.6
Quintile 3	69.1	63.1	63.4
Quintile 4	80.6	68.8	76.2
Quintile 5 (highest)	89.3	76.5	80.2

Note: Table includes UASDA graduates who enrolled in sampled junior secondary schools. This mechanically overestimates the share of students enrolled in public school as all 16 public schools are sampled but 30 out of 41 private schools. Numbers are corrected for under-sampling of private schools using sampling weights.

In Figure A4.5, we show impact estimates for the mixed policy, six months after the second policy was enacted. Comparing the estimates for the zoning and the mixed policy, we find a “bounce back” effect, even though the mixed policy was not a complete reversion to the merit policy and the results for the mixed policy are only after 6 months (compared to after 18 months for the first zoning policy). The negative impacts of the first zoning policy were mostly concentrated in the top three quintiles and more positive effects with the bottom two quintiles, which saw the largest increase in public school enrollment. Yet, between the zoning and mixed policy cohorts, grade 7 SLA scores for the top quintile increased by 0.18 s.d. and grade 7 SLA scores for the bottom quintile decreased by 0.12 s.d. These results demonstrate that the effects are driven by the different student allocation mechanisms.

Figure A4.5: Impact of the Two Zoning Policies on Test Scores by Grade 6 Exam Score Quintile



Note: We cannot compare the three policies directly since the merit cohort was not tested in grade 7 and the mixed policy cohort was not tested in grade 8. Figure presents the estimated coefficients and 95 percent confidence intervals for β_1 in Equation 4.6. Standard errors are corrected for clustering at the school level. Each model controls for gender, age at time of UASDA exam, asset index, an indicator for whether the child's mother completed tertiary education and kelurahan (and indicators for missing values in these controls). Numbers are corrected for under-sampling of private schools using sampling weights.

A4.4 Teacher and principal survey results

Teachers would play a major role in Yogyakarta's response to the zoning policy. While we don't have a comprehensive picture of how school staff responded to changes in student composition, we have some data on these adaptations. We collected survey data from almost all principals (15/16 public school principals and 28/30 private school principals), and from all mathematics and Indonesian teachers, in 2019 and 2020. We interviewed teachers to ask about their background characteristics like salary, years of experience, civil servant status, and tutoring activities; and teaching practice, such as how they might have adjusted lessons or teaching due to the policy changes. We interviewed principals about school facilities and school responses to the policy changes.

We present teacher characteristics in Table A4.10 for public and private schools separately. We also separate between teachers who taught both the zoning and merit cohort (panel teachers), and teachers that left after teaching the merit cohort (only merit), or arrived at the same time as the zoning cohort (only zoning). Looking at panel teachers, the table shows that teachers in private schools were almost 10 years younger and had 10 years less experience than public school teachers. They were also less likely to be certified and registered, and made less than half the salary of public school teachers. Teacher turnover was quite large in private schools, where 18 teachers left and 25 were newly hired. The new teachers were older, had about four additional years of experience, were more likely to be registered as teacher and also made somewhat more salary. In public schools, they hired more teachers than the number that left. The new teacher in public schools were younger, had about 10 years less experience, were less likely to be certified and had lower salary.

Table A4.10: Teacher Characteristics by School Type and Survey Round

	(1)	(2)	(3)	(4)	(5)	(6)
	Panel teachers		Only merit		Only zoning	
	Public	Private	Public	Private	Public	Private
Math teacher	0.56	0.47	0.43	0.56	0.33	0.60
	(0.50)	(0.50)	(0.53)	(0.51)	(0.48)	(0.50)
Indonesian teacher	0.44	0.53	0.57	0.44	0.67	0.40
	(0.50)	(0.50)	(0.53)	(0.51)	(0.48)	(0.50)
Male	0.27	0.36	0.43	0.22	0.05	0.28
	(0.45)	(0.48)	(0.53)	(0.43)	(0.22)	(0.46)
Age	49.08	38.58	48.71	34.83	37.19	39.96
	(9.85)	(11.48)	(15.52)	(12.43)	(12.83)	(14.96)
Years of experience	24.27	13.97	22.71	11.39	12.71	15.96
	(10.30)	(11.20)	(14.31)	(12.10)	(12.53)	(15.06)
University degree	1.00	1.00	1.00	0.94	1.00	0.96
	(0.00)	(0.00)	(0.00)	(0.24)	(0.00)	(0.20)
Certified	0.90	0.44	0.71	0.28	0.48	0.32
	(0.31)	(0.50)	(0.49)	(0.46)	(0.51)	(0.48)
Permanent contract	0.79	0.66	0.71	0.33	0.62	0.52
	(0.41)	(0.48)	(0.49)	(0.49)	(0.50)	(0.51)
Registered	0.90	0.63	0.71	0.44	0.67	0.56
	(0.31)	(0.49)	(0.49)	(0.51)	(0.48)	(0.51)
Monthly salary (1000 IDR)	6042.83	2595.99	5357.29	1651.39	3887.95	2227.56
	(2721.18)	(2507.19)	(4387.22)	(1251.37)	(2872.38)	(2302.02)
Observations	48	59	7	18	21	25

Note: Permanent contracts are civil servant contracts for public school teachers. The monthly salary only includes salary from teaching. Standard deviations between parentheses. * p<0.10 ** p<0.05 *** p<0.01

Because it is hard to quantitatively measure the methods that teachers used in the classroom, we directly asked the teachers if they changed their methods after the zoning policy. We also asked the principals about school policies and teacher methods. The results are presented in Table A4.11. We only included panel teachers in the table to make sure that their answers are due to the zoning policy rather than a change in schools. An overwhelming majority of public school teachers (83 percent) reported they changed some kind of teaching or classroom management in response to the new student composition; this figure was 46 percent in private schools. Almost all public school principals reported school policy changes, and about half of private school principals did. However, most principals and teachers mentioned policies and methods that had less to do with learning, and more with religion and behavior in the classroom, collaborations with parents and marketing for the school to attract more or different students.

Still, we find suggestive evidence that a substantial part of teachers adjusted the difficulty level of their instruction. 23 percent of teachers in public schools and 14 percent of teachers in private schools reported they changed lesson difficulty specifically after the policy change. For example, one public school teacher stated, “We used to have high-performing students. Teachers can just give them lessons and assignments, and they could complete the assignment without any difficulties. But it is different right now. Teachers now have to make more preparations in terms of pedagogical skills as well as how to approach students” (January 2018). Likewise, another public school teacher stated, ”It is challenging to teach zoning students. I need to work harder to make students understand the lessons. Before zoning, teaching was not that hard [but] after zoning, it is difficult to make students pay attention to their lessons, let alone get them to study” (January, 2021). At the same time, a private school teacher reported: “I only provided more exercise questions; there was no change from previous years.” (January 2021). Public school teachers took action to adapt, even if these changes may not have been effective (yet) in markedly improving learning outcomes. It is plausible that a school system could take longer than 18 months to adjust to such a dramatic change in student composition.

In addition, the survey results echo our finding that schools started tracking, i.e. grouping students in classrooms by past performance. 30 percent of public school teachers said that classrooms were tracked compared to 22 percent of private school teachers. Principals of 3 public and 3 private schools mentioned tracking as well. Such grouping could help teachers buffer the impact of greater school heterogeneity by making classrooms more homogeneous, although teachers would still likely have to make some adjustments if they were accustomed to teaching higher-performing students. While tracking was implemented in a minority of schools, it happened enough to result in a smaller change in peers at the classroom level than at the school level after the zoning policy (Table 4.7).

Table A4.11: Self-Reported Teacher Response to Zoning Cohort

	(1)	(2)	(3)	(4)	(5)
	Teachers			Principals	
	Public	Private	Difference	Public	Private
Any school policy	0.63	0.44	0.19*	10	11
	(0.49)	(0.50)	(0.10)		
Tracking	0.30	0.22	0.08	3	3
	(0.47)	(0.42)	(0.10)		
More experienced teacher for zoning students	0.07	0.00	0.07**	1	2
	(0.25)	(0.00)	(0.03)		
Additional classes for zoning students	0.07	0.00	0.07*	2	0
	(0.25)	(0.00)	(0.03)		
Teacher changed methods (any)	0.83	0.46	0.38***	13	11
	(0.38)	(0.50)	(0.09)		
Changed difficulty of instruction	0.23	0.14	0.09	1	1
	(0.42)	(0.35)	(0.07)		
Extra classes for struggling students	0.19	0.10	0.09	4	3
	(0.39)	(0.30)	(0.07)		
Extra tasks	0.21	0.07	0.14**	3	1
	(0.41)	(0.25)	(0.06)		
Observations	48	59	107	15	28

Note: Table only includes teachers who taught both the merit and zoning students. Standard deviations between parentheses. * p<0.10 ** p<0.05 *** p<0.01

Student reports about a decline in tutoring are corroborated by teacher reports. We estimated the change in tutoring by teachers using a simple model for teacher i given by

$$Y_i = \beta_1 T_i + \gamma_i + \epsilon_i \tag{4.5}$$

where Y_i is the outcome variable related to tutoring, T_i is an indicator for the year in which they taught the zoning cohort, γ_i are teacher fixed effects and ϵ_i is the error term. We only use the panel teachers. The results are presented in Table A4.12.

The share of public school teachers reporting that they conducted tutoring over the school year declined by 32 percentage points after zoning, amounting to about 45 fewer minutes of tutoring per week. Among private school teachers, tutoring went down by 10 percentage points or about 11 minutes per week, but these changes were not statistically significant.

Table A4.12: Change in Tutoring by Teachers

	(1)	(2)	(3)	(4)
	Tutoring outside teaching hours		Minutes per week	
	Public	Private	Public	Private
Zoning	-0.32**	-0.10	-44.50	-10.73
	(0.13)	(0.11)	(25.87)	(14.93)
Pre-zoning mean	0.54	0.55	78.75	51.35
Number of teachers	45	55	44	50

Note: Table only includes teachers who taught both the merit and zoning students. Standard deviations between parentheses. * p<0.10 ** p<0.05 *** p<0.01

A4.5 Changes in students composition and correlations with learning

In this appendix, we study correlations between the change in student composition and learning impacts, both at the school and the classroom level. We explore to what extent our results are directly explained by the student composition. We measure peer composition using the leave-me-out mean and the standard deviation of peer grade 6 exam scores.

Note that our peer score measure is incomplete because we do not have grade 6 exam scores for students who graduated from primary schools outside the city. We are therefore missing grade 6 exam scores for a substantial number of peers: about 30 percent of students enrolled in the sampled schools (Table A4.5). Nevertheless, we believe that the grade 6 exam scores are a useful proxy for the actual student composition. Table A4.13 shows that leave-me-out mean scores are strongly correlated between the grade 6 exam and the SLA, and the SLA tested nearly all students.[79] Even though these SLA scores are affected by the policy, they are still informative on the level of sorting. The strong correlations suggest that students from outside Yogyakarta sorted across schools and classrooms in a similar way as students from Yogyakarta.

Table A4.13: Correlation Matrix Between Mean Scores of Peers Using Different Tests

	Merit cohort	Zoning cohort	
	Grade 8 SLA	Grade 8 SLA	Grade 7 SLA
School mean			
Grade 6 UASDA	0.96	0.78	0.80
Grade 7 SLA		0.87	1
Classroom mean			
Grade 6 UASDA	0.91	0.83	0.80
Grade 7 SLA		0.86	1

Note: Table presents Pearson's pairwise correlation coefficients. Merit cohort was not tested in grade 7.

Change in student composition at the school level

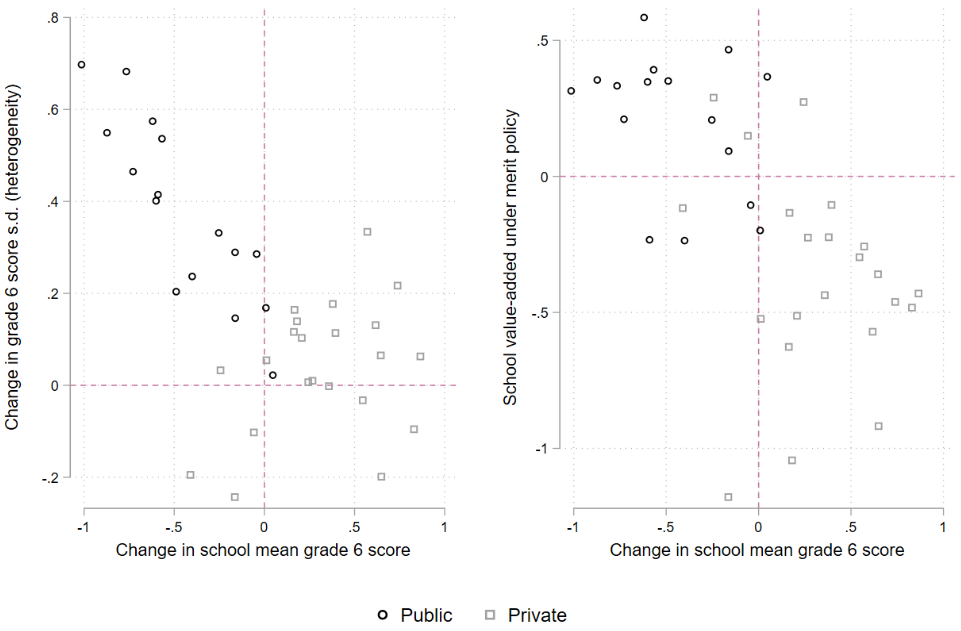
Did schools with a larger change in student composition also see a larger change in learning outcomes? In Figure A4.6, we present the change in student composition for public and private schools (left), and the correlation between the change in student composition and school quality (right). It shows that the change in student composition, and therefore the treatment intensity, varied a lot across schools. Public schools saw the

[79] The SLA includes 91.2 percent of the zoning cohort in grade 7, 94.2 percent in grade 8, and 91.4 percent of the merit cohort in grade 8.

largest increase in student heterogeneity. Private schools mainly saw a large increase in the mean grade 6 score of incoming students.

In the right panel of Figure A4.6, we show what type of schools experienced larger changes in the student composition, specifically in terms of school value-added under the merit policy. Because students of different ability levels were mixed in schools after the zoning policy, the change in student composition was largest for schools with the highest and lowest scoring students under the merit policy. The figure shows that the public schools with the largest change in student composition were relatively high quality. Private schools with the largest change in student composition were relatively low quality.

Figure A4.6: Change in Student Composition at School Level



Note: Each marker represents a school. Figure excludes 8 private schools with less than 10 students.

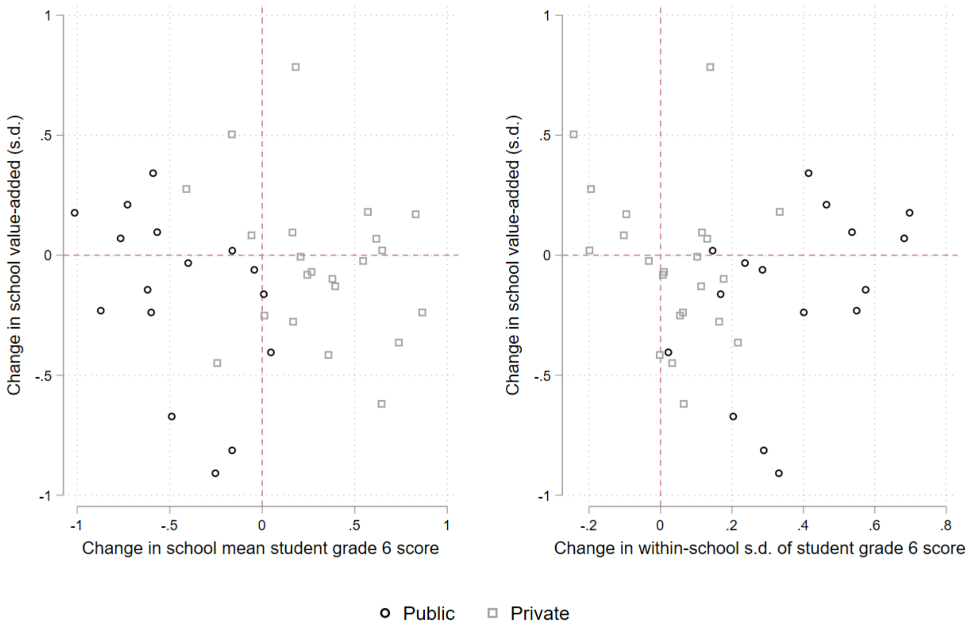
To examine how learning impacts correlate with the change in student composition, we estimate the learning impact in each school separately by including school fixed effects in our main model, and interacting them with the indicator for the zoning cohort. This is specified in Equation 4.6 for student i in school s

$$Y^2_{i,s} = \gamma_1 T_{i,s} + \gamma_2 T_{i,s} \times \lambda_s + \gamma_3 Y^1_{i,s} + \gamma_4 X_{i,s} + \lambda_s + \epsilon_{i,s} \tag{4.6}$$

where λ_s are school indicators.

Figure A4.7 plots the γ_2 for each school against the change in student composition in terms of mean scores of incoming students and the standard deviation of scores. The change in school student composition does not seem to correlate with the change in learning outcomes. Public schools with the largest change in their student composition did not see the largest decline in their test scores. Because these were higher-quality schools, perhaps they were able to limit the learning decline through their school policy responses. It suggests that these public schools performed well irrespective of their student composition.

Figure A4.7: Correlation Between Change in Student Composition and Learning Change at School Level

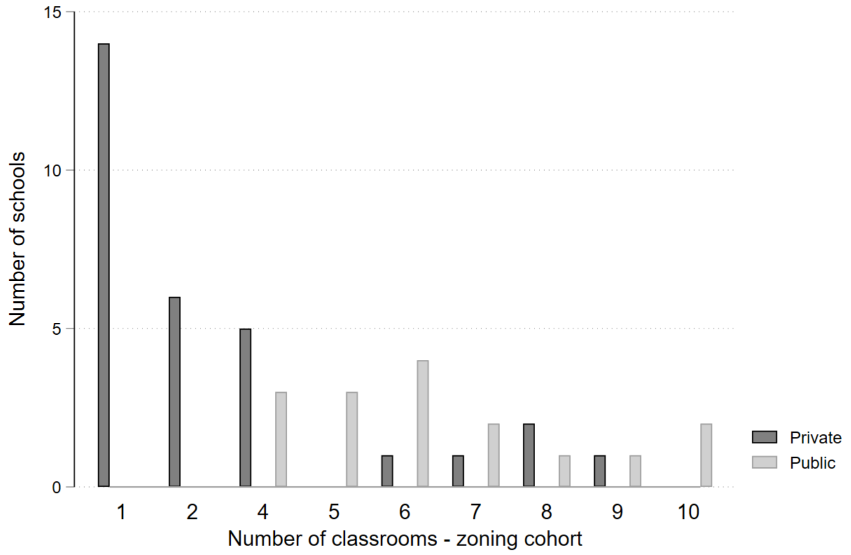


Note: Each marker represents a school. Figure excludes 8 private schools with less than 10 students. The effects on grade 8 scores shown in the figure are estimated by interacting the zoning indicator with school indicators in the model, and is estimated on the full sample.

Tracking into classrooms

Schools could determine their own policy on student allocation across classrooms. Schools can decide on the class size, and on the extent to which they group students with similar characteristics. Most schools had enough students for more than one class and had to decide how to allocate students across the classes, as shown in Figure A4.8. Public schools had between 4 and 10 classrooms. Private schools had between 1 and 9 classrooms. In 13 private schools, student enrollment was so small that they only had one classroom and could not make decisions on classroom composition.

Figure A4.8: Number of Classrooms



We do not find that schools adjusted class size in response to the policy, although the number of classes slightly changed between cohorts for some of the schools. One public school added one class, while another public school removed one class. Amongst private schools, 9 schools removed at least one class (2 schools removed 2 and one even removed 3 classes) and 2 private schools added a class. However, class size barely changed, so this seems to be due to changes in enrollment. Class size only increased from 30.5 to 31.4 students in public schools, and from 23.2 to 24.2 students in private schools.[80]

In terms of grouping students, some principals and teachers reported in the survey that their school started tracking students across classrooms (Appendix A4.4). Tracking was mentioned by 3 out of 15 public school principals and 3 out of 28 private school principals. It was also mentioned by 30 percent of public school teachers, and 22 percent of private school teachers. In addition, we found that the change in classroom peer scores was smaller than the change in school peer scores for all four groups (Table 4.7), pointing towards tracking.

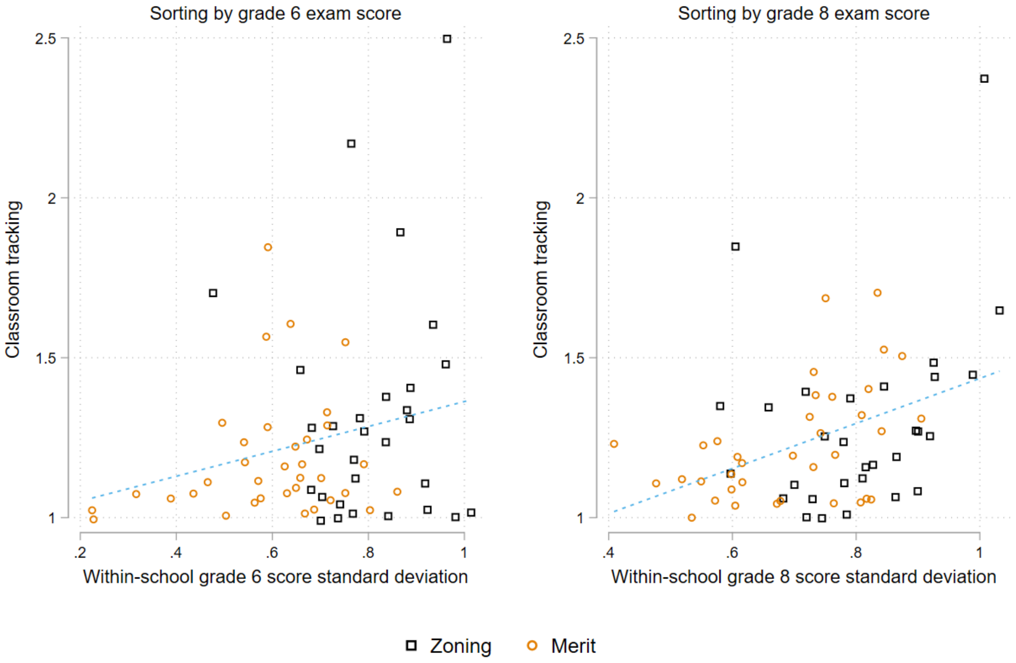
We use a data-driven method to measure the level of tracking in schools. The measure we use is the "absolute" tracking measure by Collins and Gan (2013), which is similar to the measure used in Antonovics et al. (2022) (as explained in their Appendix B). The measure takes the overall standard deviation of test scores within school-grade cohorts

[80] These numbers are slightly different from the ones in Table A4.1 because these numbers are averaged over classrooms while the table averaged over students, which gives classrooms with more students a larger weight.

and divides it by the standard deviation within a class. This way, we get a measure of tracking for each classroom. We then take the average across classrooms to get to a school-level measure of tracking. A value of 1 means that there is no tracking, as the variation within the school and its classrooms is similar, while higher values indicate more tracking.[81]

Figure A4.9 presents the level of tracking by the heterogeneity of the student composition at the school level, by cohort. We show results for sorting by the grade 6 exam score and by the grade 8 SLA score, because the grade 6 exam score has many missing values. The figure shows that schools with larger heterogeneity in their student composition tracked more. On average, the tracking measure increased from 1.2 to 1.4 between the merit and zoning cohort in public schools, while it remained 1.2 in private schools. Results are similar between tests.

Figure A4.9: Tracking by Heterogeneity of Student Composition at School Level



Note: Each marker represents a school. Figure only includes schools with more than 1 classroom.

[81] This absolute measure of tracking is affected by the number of classrooms and the distribution of class sizes. Antonovics et al. (2022) therefore develop a relative tracking measure that conditions on this by capturing the portion of potential tracking that is realized. As shown before, class size barely changes so we do not expect class size to affect our tracking measure. However, public schools have more classrooms and therefore have a larger potential to track.

Correlation between change in peers and learning for “always access” and “never access” students

We exploit the fact that schools started tracking to study whether students with a smaller change in peers also saw a smaller change in learning.[82] We match classrooms in the same schools before and after the policy change based on the classroom student composition in terms of the 25th, 50th and 75th percentile. We do this for “always access” students in public schools first, after which we perform a similar exercise for “never access” students in private schools. If the results for these groups are driven by a change in peers, we would not expect to find any effects on those students in tracking schools whose classroom peers did not change.

For “always access” students, the results are presented in Table A4.14. We matched 17 classes under the zoning policy to 17 classes under the merit policy in the same public schools. The table shows that “always access” students in these classrooms were balanced on their grade 6 exam score between cohorts, and they did not experience a change in their classroom peer composition. “Always access” students in those same schools who were allocated to the other classrooms were also balanced between cohorts, but their classroom peers had lower scores on average and were more heterogeneous. As expected, learning did not significantly decrease for students whose classroom peers stayed the same, while it did decrease by 0.3 s.d. for students in the same schools with similar grade 6 exam scores who were in classrooms with lower-scoring peers. These results suggest that the learning effects for “always access” students were indeed driven by a change in peers. Interestingly, the share of students who found the instruction level difficult decreased in both classroom types, suggesting that direct peer effects were relatively strong compared with indirect peer effects from a change in the instruction level.

We also show results for schools in which we could not match any classrooms.[83] The peer composition for “always access” students changed most in these schools, as they scored 0.6 s.d. lower. We find that learning also declined for “always access” students in these schools. Note that the model in column 5 conditions on the grade 6 score to correct for the imbalance between cohorts (column 1).

Note that learning effects in schools with and without matched classrooms cannot be directly compared. As shown by the merit cohort mean, the schools without matched

[82] We cannot study how tracking at the school level affected the change in learning, because the decision to track can be correlated with other school characteristics. For instance, those schools with the largest shocks to their student composition were more likely to track, but were also of higher average quality.

[83] Note that this does not necessarily mean that these schools did not implement any tracking. Two out of these 6 schools did implement some form of tracking. While one school created a separate class for low-achieving students, the other created 2 separate classes for high-achieving students. However, the classroom compositions were different between cohorts.

classrooms were some of the most selective schools prior to the policy change. Their teachers may have been able to support the high-scoring students in a different way to limit the decline in learning from the change in peers.

Table A4.14: Learning Effect on "Always Access" Students in Public Schools by Classroom Type

	(1)	(2)	(3)	(4)	(5)
	Grade 6 score	Classroom peers		Finds instruction	Grade 8 score
	balance	*mean*	*s.d.*	level difficult	
Matched classrooms in tracking schools					
25 classrooms, mean tracking measure = 1.8					
Zoning	0.05	0.00	0.05	-0.11**	-0.06
	(0.04)	(0.07)	(0.03)	(0.05)	(0.12)
School fixed effects	Yes	Yes	Yes	Yes	Yes
Observations	531	531	531	512	531
Merit cohort mean	0.49	0.36	0.48	0.29	0.24
Other classrooms in tracking schools					
43 classrooms, mean tracking measure = 1.5					
Zoning	0.01	-0.27***	0.16***	-0.11**	-0.29***
	(0.06)	(0.09)	(0.03)	(0.04)	(0.10)
School fixed effects	Yes	Yes	Yes	Yes	Yes
Observations	904	904	904	873	904
Merit cohort mean	0.36	0.20	0.50	0.29	0.23
Schools with no matched classrooms					
93 classrooms, mean tracking measure = 1.2					
Zoning	-0.32***	-0.61***	0.42***	-0.08***	-0.15***
	(0.04)	(0.06)	(0.04)	(0.02)	(0.05)
School fixed effects	Yes	Yes	Yes	Yes	Yes
Observations	1594	1594	1594	1552	1594
Merit cohort mean	1.19	1.12	0.37	0.25	0.84

Note: Mean tracking measure is calculated using the zoning cohort. The number of classrooms and observations include both the merit and the zoning cohort. Standard errors in parentheses and corrected for clustering at the classroom level. Models control for standardized grade 6 exam score (except for model 1), gender, age at the time of UASDA exam, an asset index, an indicator for the mother having completed tertiary education and kelurahan (and indicators for missing values in these controls). * p<0.10 ** p<0.05 *** p<0.01

We do a similar analysis for "never access" students. The results are shown in Table A4.15. There were only 3 schools with 41 students in total in which we could not match classrooms, so we leave these out of the table due to the small number of observations. "Never access" students who were enrolled in the matched classrooms were balanced on their grade 6 exam score and saw only a small increase in the heterogeneity of their classroom, and no change in the mean score of their peers. Other "never access" students in the same schools were together in the same classrooms with the newly enrolled high-scoring students. We find a large decrease in learning of 0.4 s.d. for the students with similar peers, while learning did not decrease for those who were in classrooms with higher-scoring peers. Possibly, teachers shifted their attention to classrooms with higher-scoring students. These results are in contrast with findings of Duflo et al. (2011), who found that low-scoring students benefit from sorting students into classrooms by ability.

Table A4.15: Learning Effect on "Never Access" Students in Private Schools by Classroom Type

	(1)	(2)	(3)	(4)	(5)
	Grade 6 score	Classroom peers		Finds instruction	Grade 8 score
	balance	*mean*	*s.d.*	level difficult	
Matched classrooms					
52 classrooms, mean tracking measure = 1.2					
Zoning	-0.04	-0.02	0.08**	-0.04	-0.38**
	(0.08)	(0.07)	(0.03)	(0.06)	(0.16)
School fixed effects	Yes	Yes	Yes	Yes	Yes
Observations	244	244	244	231	244
Merit cohort mean	-0.87	-0.59	0.59	0.28	-0.48
Other classrooms in same schools					
78 classrooms, mean tracking measure = 1.2					
Zoning	0.13***	0.44***	0.14***	0.00	0.08
	(0.05)	(0.06)	(0.03)	(0.05)	(0.11)
School fixed effects	Yes	Yes	Yes	Yes	Yes
Observations	441	441	441	422	441
Merit cohort mean	-0.97	-0.83	0.54	0.23	-0.96

Note: Mean tracking measure is calculated using the zoning cohort. The number of classrooms and observations include both the merit and the zoning cohort. Standard errors in parentheses and corrected for clustering at the classroom level. Models control for standardized grade 6 exam score (except for model 1), gender, age at the time of UASDA exam, an asset index, an indicator for the mother having completed tertiary education and kelurahan (and indicators for missing values in these controls). * p<0.10 ** p<0.05 *** p<0.01

Due to classroom tracking, there may also be "gained access" and "lost access" students who enrolled in a different school type but with similar classroom peers in terms of ability. Comparing the merit and zoning cohort within that group, we would then only capture an effect from different school resources, such as being taught by a more experienced teacher. We tried to use the simulation results to identify counterfactual private school students for "gained access" students in each of the public schools.[84] Unfortunately, even though our simulation predicts public school enrollment well, it does not predict well which exact public school the students would enroll in. Only 20 percent of "gained access" students actually enrolled in the assigned public school based on the

[84] The simulation does not allocate students to a specific public school when replicating the merit policy. It essentially takes all public schools as one big school to which the highest scoring students are allocated, see Section 5.1. Therefore, we cannot do this school-level analysis for the "lost access" students.

simulation (the closest public school to their neighborhood). Therefore, we refrain from pursuing the analysis for "gained access" students.

We also refrain from estimating correlations between peer scores and learning outcomes directly. These estimates would be hard to interpret in our study setting due to the so-called "reflection problem": it is impossible to distinguish the effect of peers on the individual from the effect of the individual on peers if both are determined simultaneously (Paloyo, 2020). Generally, higher scoring students have higher-scoring peers, and it is difficult to disentangle the selection effects from actual peer effects. We showed that (1) that the change in student composition at the school level is correlated with school quality and (2) that some schools started tracking students into classrooms. Hence, even though the shock to student composition across all schools was exogenous, differences in the change in peers between schools and classrooms were endogenous.

Bibliography

Abdi, A. P. (2019). Kemendikbud temukan kasus kecurangan terstruktur saat unbk smp. *Tirto.id*. Retrieved from https://tirto.id/kemendikbud-temukan-kasus-kecurangan-terstruktur-saat-unbk-smp-d8ipk.

Abdulkadiroğlu, A. and Andersson, T. (2022). School choice. *NBER Working paper 29822*.

Abdulkadiroğlu, A., Pathak, P. A., and Walters, C. R. (2018). Free to choose: Can school choice reduce student achievement? *American Economic Journal: Applied Economics*, 10(1):175–206.

Afkar, R., De Ree, J., and Khairina, N. (2018). *Who learns what in basic education? Evidence from Indonesia*. World Bank, Washington DC.

Akmal, M. and Pritchett, L. (2021). Learning equity requires more than equality: Learning goals and achievement gaps between the rich and the poor in five developing countries. *International Journal of Educational Development*, 82(102350).

Alfons, M. (2019). 126 siswa curang saat unbk 2019, kemendikbud: Otomatis nilai nol. *detikNews*. Retrieved from https://news.detik.com/berita/d-4539834/126-siswa-curang-saat-unbk-2019-kemendikbud-otomatis-nilai-nol.

Anderman, E. M. (2015). India's "cheating mafia" gets to work as school exam season hits. *The Conversation*. Retrieved from https://theconversation.com/why-students-at-prestigious-high-schools-still-cheat-on-exams-91041.

Andrabi, T., Bau, N., Das, J., and Khwaja, A. I. (2020). Private schooling, learning, and civic values in a low-income country. *Unpublished Working Paper*.

Andrabi, T., Das, J., Khwaja, A. I., Vishwanath, T., and Zajonc, T. (2008). *Pakistan Learning and Educational Achievement in Punjab Schools (LEAPS) - Instights to inform the education policy debate*. World Bank, Washington, DC.

Andrabi, T., Das, J., Khwaja, A. I., and Zajonc, T. (2011). Do value-added estimates add value? Accounting for learning dynamics. *American Economic Journal: Applied Economics*, 3(3):29–54.

Angrist, J., Hull, P., Pathak, P. A., and Walters, C. (2020). Simple and Credible Value-Added Estimation Using Centralized School Assignment. *NBER Working Paper 28241.*

Angrist, J., Hull, P., and Walters, C. R. (2022). Methods for measuring school effectiveness. *NBER Working Paper 30803.*

Angrist, J. and Lavy, V. (2009). The effects of high stakes high school achievement awards: Evidence from a randomized trial. *American Economic Review*, 99(4):1384–1414.

Angrist, 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.

Angrist, J. D. and Lang, K. (2004). Does school integration generate peer effects? Evidence from Boston's metco program. *American Economic Review*, 94(5):1613–1634.

Anstreicher, G., Fletcher, J., and Thompson, O. (2022). The long-run impacts of court-ordered desegregation. *NBER Working Paper 29926.*

Antonovics, K., Black, S. E., Berry Cullen, J. J., and Meiselman, A. Y. (2022). Patterns, determinants, and consequences of ability tracking: Evidence from Texas public schools. *IZA Discussion Paper 15528.*

Badan Standar Nasional Pendidikan (2006). *Standar isi untuk satuan pendidikan dasar dan menengah - Standar kompetensi dan kompetensi dasar SD/MI.* Badan Standar Nasional Pendidikan.

Barrow, L., Sartain, L., and de la Torre, M. (2020). Increasing access to selective high schools through place-based affirmative action: Unintended consequences. *American Economic Journal: Applied Economics*, 12(4):135–163.

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.

Bau, N. (2022). Estimating an equilibrium model of horizontal competition in education. *Journal of Political Economy*, 130(7).

Bau, N. and Das, J. (2020). Teacher value added in a low-income country. *American Economic Journal: Economic Policy*, 12(1):62–96.

Bau, N., Das, J., and Yi Chang, A. (2021). New evidence on learning trajectories in a low-income setting. *World Bank Policy Research Working Paper 9597.*

Beatty, A., Berkhout, E., Bima, L., Pradhan, M., and Suryadarma, D. (2021). Schooling progress, learning reversal: Indonesia's learning profiles between 2000 and 2014. *International Journal of Educational Development*, 85(102436).

Berkhout, E., Dharmawan, G., Beatty, A., Suryadarma, D., and Pradhan, M. (2022). Who benefits and loses from large changes to student composition? Assessing impacts of lowering school admissions standards in Indonesia. *RISE Working Paper 22/094.*

Berlinski, S., Busso, M., and Giannola, M. (2022). Helping struggling students and benefiting all: Peer effects in primary education. *CSEF Working Paper 634.*

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.

Biantoro, B. and Arfianti, A. (2019). Issues in the implementation of Computer-based National Exam (CBNE) in Indonesian secondary schools. *Advances in Social Science, Education and Humanities Research*, 353:399–403.

Black, S. (1999). Do better schools matter? Parental valuation of elementary education. *Quarterly Journal of Economics*, 114(2):577–599.

Black, S., Denning, J., and Rothstein, J. (2020). Winners and losers? The effect of gaining and losing access to selective colleges on education and labor market outcomes. *NBER Working Paper 26821.*

Booij, A. S., Leuven, E., and Oosterbeek, H. (2017). Ability peer effects in university: Evidence from a randomized experiment. *The Review of Economic Studies*, 84:547–578.

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.

Bramoullé, Y., Djebbari, H., and Fortin, B. (2020). Peer effects in networks: A survey. *Annual Review of Economics*, 12:603–629.

Burgess, S., Dickson, M., and Macmillan, L. (2020). Do selective schooling systems increase inequality? *Oxford Economic Papers*, 72(1):1–24.

Burke, M. A. and Sass, T. R. (2013). Classroom peer effects and student achievement. *Journal of Labor Economics*, 31(1):51–82.

Callaway, B. and Sant'Anna, P. H. (2020). Difference-in-Differences with multiple time periods. *Journal of Econometrics.*

Chin, S. and Kwon, E. (2019). The effects of ability tracking on the academic performance in the secondary school: Evidence from South Korea. *USC Dornsife Institute for New Economic Thinking Working Paper 19-12.*

Collins, C. A. and Gan, L. (2013). Does sorting students improve scores? An analysis of class composition. *NBER Working Paper 18848.*

Cullen, J. B., Jacob, B. A., and Levitt, S. D. (2006). The effect of school choice on participants: Evidence from randomized lotteries. *Econometrica*, 74(5):1191–1230.

Dalla-Zuanna, A., Liu, K., and Salvanes, K. G. (2022). Pulled-in and crowded-out: Heterogeneous outcomes of merit-based school choice. *CEPR Discussion Paper DP16853*.

Das, J., Dercon, S., Habyarimana, J., Krishnan, P., Muralidharan, K., and Sundararaman, V. (2013). School inputs, household substitution, and test scores. *American Economic Journal: Applied Economics*, 5(2):29–57.

De Chaisemartin, C. and D'Haultfoeuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9):2964–2996.

De Chaisemartin, C. and D'Haultfœuille, X. (2022). Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey. *The Econometrics Journal*, 00:1–30.

De Ree, J. (2016). How much teachers know and how much it matters in class. *World Bank Policy Research Working Paper 7556*.

De Ree, J., Muralidharan, K., Pradhan, M., and Rogers, H. (2018). Double for nothing? Experimental evidence on an unconditional teacher salary increase. *Quarterly Journal of Economics*, pages 993–1039.

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.

Deming, D. J., Hastings, J. S., Kane, T. J., and Staiger, D. O. (2014). School choice, school quality, and postsecondary attainment. *American Economic Review*, 104(3):991–1013.

Denning, J. T., Murphy, R., and Weinhardt, F. (2021). Class rank and long-run outcomes. *NBER Working Paper 27468*.

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

Economist (2011). More cheating, or else! scandals in the classroom. *The Economist*. Retrieved from https://www.economist.com/asia/2011/07/07/more-cheating-or-else.

Feldman, D. C. (1984). The development and enforcement of group norms. *Academy of Management Review*, 9(1):47–53.

Ferraz, C., Finan, F., and Moreira, D. B. (2012). Corrupting learning: Evidence from missing federal education funds in Brazil. *Journal of Public Economics*, 96(9-10):712–

726.

Filmer, D. and Pritchett, L. H. (2001). Estimating wealth effects without expenditure data-or tears: An application to educational enrollments in states of India. *Demography*, 38(1):115–132.

Frankenberg, E., Karoly, L. A., Gertler, P., Achmad, S., Agung, I., Hatmadji, S. H., and Sudharto, P. (1995). *The 1993 Indonesian Family Life Survey: Overview and field report*. RAND and Demographic Institute Indonesia, Santa Monica, CA.

Galbiati, R., Henry, E., Jacquemet, N., and Lobeck, M. (2021). How laws affect the perception of norms: Empirical evidence from the lockdown. *PLoS ONE*, 16(9):e0256624.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*.

Government of Indonesia (1998). *Petunjuk pelaksanaan wajib belajar pendidikan dasar sembilan tahun*. Government of Indonesia.

Gust, S., Hanushek, E. A., and Woessmann, L. (2022). Global universal basic skills: Current deficits and implications for world development. *RISE Working Paper 22/114*.

Guyon, N., Maurin, E., and Mcnally, S. (2012). The effect of tracking students by ability into different schools: A natural experiment. *Journal of Human Resources*, 47(3):684–721.

Hanson, B. A., Harris, D. J., and Brennan, R. L. (1987). A comparison of several statistical methods for examining allegations of copying. *ACT Research Report Series 87-15*.

Hanushek, E. A. (2013). Economic growth in developing countries: The role of human capital. *Economics of Education Review*, 37:204–212.

Hanushek, E. A. (2020). Education production functions. In *The Economics of Education: A Comprehensive Overview*, pages 161–170. Academic Press.

Hanushek, E. A., Kain, J. F., O'Brien, D. M., and Rivkin, S. G. (2005). The market for teacher quality. *NBER Working Paper 11154*.

Hanushek, E. A. and Woessmann, L. (2008). The role of cognitive skills in economic development. *Journal of Economic Literature*, 46(3):607–668.

Hill, D. V., Hughes, R. P., Lenard, M. A., Liebowitz, D. D., and Page, L. C. (2022). New schools and new classmates: The disruption and peer group effects of school reassignment. *NBER Working paper 30085*.

Imberman, S. A., Kugler, A. D., and Sacerdote, B. I. (2012). Katrina's children: Evidence on the structure of peer effects from hurricane evacuees. *American Economic Review*, 102(5):2048–2082.

Jackson, C. K. (2020). Does School Spending matter? A new answer to the old question. In Tach, L., Dunifon, R., and Miller, D. L., editors, *Confronting inequality: How policies and practices shape children's opportunities.*, pages 165–186. American Psychological Association, Washington, DC, US.

Jackson, C. K., Porter, S., Easton, J., and Kiguel, S. (2020). Who benefits from attending effective schools? Examining heterogeneity in high school impacts. *NBER Working paper 28194*.

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, 89(5-6):761–796.

Jacob, B. A. and Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics*, 118(3):843–877.

Jones, S., Schipper, Y., Ruto, S., and Rajani, R. (2014). Can your child read and count? Measuring learning outcomes in East Africa. *Journal of African Economies*, 23(5):643–672.

Jong, H. N. (2015). New exam system fails to prevent cheating, leaks. *The Jakarta Post*. Retrieved from https://www.thejakartapost.com/news/2015/04/16/new-exam-system-fails-prevent-cheating-leaks.html.

Kaffenberger, M. (2019). A typology of learning profiles: Tools for analysing the dynamics of learning. *RISE Insight Note*.

Kaffenberger, M. and Pritchett, L. (2020). Aiming higher: Learning profiles and gender equality in 10 low- and middle-income countries. *International Journal of Educational Development*, 79:102272.

Kemdikbud (2017). Peraturan menteri pendidikan dan kebudayaan Republik Indonesia Nomor 17 tahun 2017.

Kementerian Pendidikan dan Kebudayaan (2013). *Kurikulum 2013 - Kompetensi dasar sekolah dasar (SD)/madrasah ibtidaiyah (MI)*. Kementerian Pendidikan dan Kebudayaan.

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(4):584–616.

Luschei, T. F. (2017). 20 Years of TIMSS: Lessons for Indonesia. *Indonesian Research Journal in Education*, 1(1):6–17.

Martinelli, C., Parker, S. W., Pérez-Gea, A. C., and Rodrigo, R. (2018). Cheating and incentives: Learning from a policy experiment. *American Economic Journal: Eco-*

nomic Policy, 10(1):298–325.

Mendolia, S., Paloyo, A. R., and Walker, I. (2018). Heterogeneous effects of high school peers on educational outcomes. *Oxford Economic Papers*, 70(3):613–634.

Mullis, I., Martin, M., Foy, P., and Arora, A. (2012). *TIMSS 2011 international results in mathematics*. TIMSS & PIRLS International Study Center, Boston College.

Mullis, I., Martin, M., Foy, P., and Hooper, M. (2016). *TIMSS 2015 international results in mathematics*. TIMSS & PIRLS International Study Center, Boston College.

Mullis, I., Martin, M., Foy, P., Olson, J., Preuschoff, C., Erberber, E., Arora, A., and Galia, J. (2008). *TIMSS 2007 international mathematics report: Findings from IEA's trends in international mathematics and science study at the fourth and eighth grades*. TIMSS & PIRLS International Study Center, Boston College.

Mullis, I., Martin, M., Gonzalez, E., and Chrostowski, S. (2004). *Findings from IEA's trends in international mathematics and science study at the fourth and eighth grades*. TIMSS & PIRLS International Study Center, Boston College.

Muralidharan, K. and Sundararaman, V. (2015). The aggregate effect of school choice: Evidence from a two-stage experiment in India. *Quarterly Journal of Economics*, 130(3):1011–1066.

Neal, D. (2013). The consequences of using one assessment system to pursue two objectives. *Journal of Economic Education*, 44(4):339–352.

OECD (2019). Annex B1. Results for countries and economies. In *PISA 2018 Results (Volume I)*. OECD Publishing, Paris.

OECD (2020). Sorting and selecting students between and within schools. In *PISA 2018 Results (Volume V): Effective Policies, Successful Schools*. OECD Publishing, Paris.

Oosterbeek, H., Ruijs, N., and de Wolf, I. (2020). Using admission lotteries to estimate heterogeneous effects of elite schools. *SSRN Electronic Journal*.

Paloyo, A. R. (2020). Peer effects in education: Recent empirical evidence. In *The Economics of Education: A Comprehensive Overview*, pages 291–305. Elsevier Ltd.

Pop-Eleches, C. and Urquiola, M. (2013). Going to a better school: Effects and behavioral responses. *American Economic Review*, 103(4):1289–1324.

Pritchett, L. and Beatty, A. (2015). Slow down, you're going too fast: Matching curricula to student skill levels. *International Journal of Educational Development*, 40(2015):276–288.

Pritchett, L. and Sandefur, J. (2020). Girls' schooling and women's literacy: schooling targets alone won't reach learning goals. *International Journal of Educational Development*, 78(102242).

Pusat Data dan Teknologi Informasi Kemdikbud (2019). Overview of junior secondary schools by status of school. Available at: http://statistik.data.kemdikbud.go.id/index.php/page/smp.

Rahmawati and Asrijanty (2016). Integrity index of national exam: An effort to gain precise information on achievement of curriculum standards. In *Conference Proceedings of the 42nd Conference of the International Association for Educational Assessment.*

Rahmawati, N. (2019). 100 SMP peraih nilai tertinggi nasional UNBK tahun 2019. *Depoedu.com.* Retrieved from https://www.depoedu.com/2019/06/03/edu-talk/100-smp-peraih-nilai-tertinggi-nasional-unbk-tahun-2019/.

Rao, G. (2019). Familiarity does not breed contempt: Generosity, discrimination, and diversity in Delhi schools. *American Economic Review*, 109(3):774–809.

Rarasati, N., Dharmawan, G., Swarnata, A., Zulfa, A. H., and Lim, D. (2020). Comprehensive Reading and Mathematics Assessment Tool (CERMAT). *SMERU Technical Report.*

Reinikka, R. and Svensson, J. (2005). Fighting corruption to improve schooling: Evidence from a newspaper campaign in Uganda. *Journal of the European Economic Association*, 3(2-3):259–267.

Rolleston, C. (2014). Learning profiles and the 'skills gap' in four developing countries: A comparative analysis of schooling and skills development. *Oxford Review of Education*, 40(1):132–150.

Rolleston, C. and James, Z. (2015). After access: Divergent learning profiles in Vietnam and India. *Prospects*, 45(3):285–303.

Roth, A. E. (2008). Deferred acceptance algorithms: History, theory, practice, and open questions. *International Journal of Game Theory*, 36:537–569.

Sacerdote, B. (2011). Peer effects in education: How might they work, how big are they and how much do we know thus far? In *Handbook of the Economics of Education*, volume 3, pages 249–277. Elsevier B.V.

Sacerdote, B. (2014). Experimental and quasi-experimental analysis of peer effects: Two steps forward? *Annual Review of Economics*, 6:253–272.

Safi, M. (2018). Why students at prestigious high schools still cheat on exams. *The Guardian.* Retrieved from https://www.theguardian.com/world/2018/apr/03/india-school-exam-season-cheating-mafia-.

Shi, Y. (2020). Who benefits from selective education? Evidence from elite boarding school admissions. *Economics of Education Review*, 74(101907).

Siddiq, T. (2018). Kemendikbud: Unbk tekan anggaran ujian nasional hingga 70 persen. *Tempo*. Retrieved from https://nasional.tempo.co/read/1069466/kemendikbud-unbk-tekan-anggaran-ujian-nasional-hingga-70-persen/full.

Singh, A. (2015). Private school effects in urban and rural India: Panel estimates at primary and secondary school ages. *Journal of Development Economics*, 113:16–32.

Singh, A. (2020a). Learning more with every year: School year productivity and international learning divergence. *Journal of the European Economic Association*, 18(4):1770–1813.

Singh, A. (2020b). Myths of official measurement: Auditing and improving administrative data in developing countries. *RISE Working Paper 20/042*.

Spaull, N. and Kotze, J. (2015). Starting behind and staying behind in South Africa: The case of insurmountable learning deficits in mathematics. *International Journal of Educational Development*, 41:13–24.

StataCorp (2017). *Stata statistical software* (Release 15). StataCorp LLC.

Statistics Indonesia (2020). Educational indicators, 1994-2019. Retrieved August, 2020, from https://www.bps.go.id/statictable/2010/03/19/1525/indikator-pendidikan-1994-2019.html.

Strauss, J., Beegle, K., Sikoki, B., Dwiyanto, A., Herawati, Y., and Witoelar, F. (2004). *The third wave of the Indonesia Family Life Survey: Overview and field report*. RAND, Labor and Population.

Strauss, J., Witoelar, F., , and Sikoki, B. (2016). *The fifth wave of the Indonesia Family Life Survey: Overview and field report*. RAND, Labor and Population.

Strauss, J., Witoelar, F., Sikoki, B., and Wattie, A. M. (2009). *The fourth wave of the Indonesia Family Life Survey: Overview and field report*. RAND, Labor and Population.

Sugiarti, T. (2014). *Struktur Kurikulum 2013*. Kementerian Pendidikan dan Kebudayaan.

Sun, L. and Abraham, S. (2020). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*.

Sundaryani, F. S. (2015). Students get high scores by cheating. *The Jakarta Post*. Retrieved from https://www.thejakartapost.com/news/2015/06/15/students-get-high-scores-cheating.html.

Suryadarma, D., Suryahadi, A., Sumarto, S., and Rogers, F. H. (2006). Improving student performance in public primary schools in developing countries: Evidence from Indonesia. *Education Economics*, 14(4):401–429.

UNESCO (2021). VIEW - Global education monitoring report completion rates. Retrieved in 2022, from https://www.education-estimates.org.

UNESCO (2022). New estimation confirms out-of-school population is growing in sub-Saharan Africa. *Factsheet 62/Policy Paper 48.*

UNESCO Institute for Statistics (2018). Metadata for the global and thematic indicators for the follow-up and review of SDG 4 and Education 2030. Retrieved from http://tcg.uis.unesco.org/wp-content/uploads/sites/4/2019/04/sdg4-metadata-global-thematic-indicators-en.pdf.

United Nations General Assembly (2015). Transforming our world: The 2030 agenda for sustainable development. *Resolution 70/1.*

UU No. 14 (2005). Art. IX, tentang Guru dan Dosen (Law No. 14 on Teachers and Lecturers).

UU No. 20 (2003). Art. X, tentang Sistem Pendidikan Nasional (Law No. 20 on National Education System).

UU No. 22 (1999). Art. XI. paragraph 2, tentang Pemerintahan Daerah (Law No. 22 on Local Government).

Van Der Linden, W. J. and Sotaridona, L. (2006). Detecting answer copying when the regular response process follows a known response model. *Journal of Educational and Behavioral Statistics*, 31(3):283–304.

Vegas, E. and Coffin, C. (2015). When education expenditure matters: An empirical analysis of recent international data. *Comparative Education Review*, 59(2):289–304.

Wang, S., Jiao, H., Young, M. J., Brooks, T., and Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K-12 reading assessments: A meta-analysis of testing mode effects. *Educational and Psychological Measurement*, 68(1):5–24.

Widiatmo, H. (2006). Metode untuk mendeteksi penyontekan jawaban pada tes pilihan ganda: studi kasus smp di kabupaten garut. *Pusat Penelitian Pendidikan, Balitbang Diknas*, pages 219–226.

World Bank (2013). *Indonesia - Spending more or spending better: Improving education financing in Indonesia.* World Bank.

World Bank (2018a). *Learning more, growing faster.* World Bank.

World Bank (2018b). *World development report 2018: Learning to realize education's promise.* International Bank for Reconstruction and Development and The World Bank.

Summary

Developing countries have made massive progress in school enrollment in the past decades. However, many children do not acquire foundational skills in reading and mathematics during their schooling. This thesis examines this problem in the context of Indonesia. Indonesia has achieved near universal enrollment up to grade 9, but not even a third of grade 9 students is proficient in reading and mathematics. The thesis contains three studies that analyze effects on student academic performance of different policies that aimed to improve the quality and fairness of the education system.

Chapter 2 describes the main changes to the Indonesian education system between 2000 and 2014, and studies the development of numeracy skills of children during that period. The Indonesian Government doubled their education expenses, decentralized education management, and increased teacher standards. However, the chapter finds that numeracy skills declined using a nearly nationally representative household survey. The study rules out that this decline is due to selection of lower-performing students into schools. Learning profiles show that many students fell behind curriculum expectations as they progressed through their schooling. The findings suggest that a slower curriculum pace in early grades may help students to learn the necessary foundational skills for learning in later grades.

Chapter 3 estimates the impacts of computer-based testing on exam scores, which was implemented to prevent cheating. The national grade 9 exam had been contaminated with widespread cheating for years. An algorithm that detects suspicious answer patterns found evidence for cheating in a third of all junior secondary schools. Computer-based testing was expected to prevent cheating by generating almost unique test versions for each student. To measure the impact of computer-based testing, the study exploits the staggered implementation of the policy across schools between 2015 and 2019. It finds a massive drop in exam scores once schools switch to computers, which was concentrated among schools that had suspicious answer patterns in previous years. Cheating also decreased among schools that still took the exam on paper, possibly because cheating became less accepted. Hence, computer-based testing was successful at preventing cheating. Removing the opportunity to cheat on the exam may increase the effort that students

and teachers put into learning in order to achieve high grades. However, computer-based exam scores did not improve within three years of implementation, despite high stakes on the exam.

Chapter 4 evaluates the learning effects of making top school less selective. The primary admission criterion for high-quality public schools in Yogyakarta changed from an exam score ranking to a neighborhood-to-school distance ranking. The policy gave many low-achieving students access to public schools, and displaced many high-achieving students to lower-quality private schools. The study compares test score value-added of two otherwise similar student cohorts admitted before and after the policy change. Average learning slightly declined. Using the admissions criteria, the study then identifies students whose access changed and students whose access stayed the same by predicting public school access under both policies for each student. Learning of students who gained access to public schools improved moderately, but learning of students displaced to private schools declined twice as much. Incumbent students who remained in public schools learned less with their new lower-scoring peer group. Yet, students who remained in private schools did not benefit from a higher-scoring peer group. Survey results suggest that teachers simplified their instructions to accommodate for the lower-performing students, which may have slowed down the learning progress of high-achieving students. Hence, the policy decreased learning inequality, but mostly at the expense of high-achieving students.

The results indicate that achieving equitable quality education is a challenge for Indonesia, despite massive investments in its education system. To help overcome this challenge, future research is necessary to better understand how education policies translate into classroom practices.

Nederlandse Samenvatting (Summary in Dutch)

Hoewel de meeste kinderen in lage- en middeninkomenslanden tegenwoordig naar school gaan, leren veel van hen nog te weinig. Dit probleem wordt in dit proefschrift onderzocht in de context van Indonesië. Bijna alle Indonesische kinderen gaan tot hun vijftiende naar school, maar nog geen derde van hen behaalt de benodigde basisvaardigheden in lezen, schrijven en rekenen. Het proefschrift bevat drie studies naar de leereffecten van verschillende beleidswijzigingen, die als doel hadden om de kwaliteit en kansengelijkheid van het onderwijssysteem te verbeteren.

Hoofdstuk 2 bestudeert hoe de rekenvaardigheden van 7 tot 18 jaar oude kinderen zich hebben ontwikkeld tussen 2000 en 2014, een periode waarin belangrijke beleidsveranderingen hebben plaatsgevonden in het Indonesische onderwijssysteem. De Indonesische overheid heeft in die periode de onderwijsuitgaven verdubbeld, de beleidsvoering van onderwijs gedecentraliseerd en de bekwaamheidseisen voor leraren verhoogd. Echter, deze studie vindt dat rekenvaardigheden in die periode zijn verslechterd, gebruik makende van bijna nationaal representatieve onderzoeksgegevens. De studie bevestigt dat deze verslechtering niet komt door een grotere instroom van laagscorende kinderen in de scholen. Zogenoemde "leerprofielen" laten zien dat veel kinderen steeds verder achterliepen op het curriculum naarmate ze overgingen naar hogere klassen. De bevindingen suggereren dat een minder ambitieus curriculum in de eerste klassen wellicht kan helpen om kinderen de benodigde basisvaardigheden bij te brengen voordat ze overgaan naar hogere klassen.

Hoofdstuk 3 evalueert de impact op examencijfers wanneer examens afgenomen worden op computers in plaats van op papier. Dit beleid werd toegepast om examenfraude tegen te gaan. Fraude op het centrale examen aan het eind van de lagere middelbare school (derde klas) was een groot probleem in Indonesië. Er werden bijvoorbeeld antwoordmodellen verspreid. Hoe groot dit probleem daadwerkelijk was werd aangetoond door middel van een algoritme dat verdachte antwoordpatronen detecteert: ongeveer een derde van alle lagere middelbare scholen werd verdacht van fraude. De Indonesische overheid verwachtte deze grootschalige fraude te bestrijden door het examen af te nemen op computers, omdat ze zo een bijna unieke versie van het examen konden genereren voor iedere student. De studie maakt gebruik van de gefaseerde invoering van computerexamens tussen 2015 en 2019 om de impact te meten van dit beleid.

De resultaten laten zien dat computers inderdaad examenfraude konden voorkomen. Er was een grote daling in examencijfers wanneer scholen naar computers overstapten. Om aan te tonen dat dit effect door een afname in fraude komt en niet door een gebrek aan computervaardigheden, laat de studie zien dat de daling in examencijfers was geconcentreerd onder scholen met verdachte antwoordpatronen in voorgaande jaren. Daarnaast was het effect ook vergelijkbaar tussen scholen die al computers hadden voor het nieuwe beleid en scholen die computers moesten aanschaffen om het beleid uit te kunnen voeren. Fraude nam ook af onder scholen die het examen nog op papier deden, zoals gemeten door het algoritme, waarschijnlijk omdat fraude minder werd geaccepteerd. Wanneer fraude niet meer mogelijk is, zou het kunnen dat leerlingen en leraren meer moeite steken in het verbeteren van hun vaardigheden om hoge cijfers te behalen. De studie test deze hypothese door het effect op examencijfers te meten tot drie jaar na invoering van het nieuwe beleid. De studie vindt geen toename in examencijfers, ondanks het grote belang bij hoge cijfers voor zowel leerlingen als leraren.

Hoofdstuk 4 onderzoekt hoe een verandering in de verdeling van leerlingen over lagere middelbare scholen hun leeruitkomsten beïnvloed. De publieke scholen in Yogyakarta, die bekend staan om hun hoge kwaliteit, namen voor de beleidswijziging alleen leerlingen aan met de hoogste examencijfers. Na de beleidswijziging moesten ze de meeste plekken toewijzen aan leerlingen uit de dichtstbijzijnde woonbuurten. Dit beleid had als doel de kansengelijkheid te verbeteren in het onderwijs door laagscorende, dichtbij wonende leerlingen toegang te geven tot de goed aangeschreven publieke scholen. Omdat plekken in publieke scholen beperkt waren, moesten hierdoor veel hoogscorende leerlingen die verder weg woonden van de publieke scholen naar private scholen van lagere kwaliteit. De studie vergelijkt de toename in vaardigheden gedurende anderhalf jaar tussen leerlingcohorten die toegelaten zijn voor en na de beleidswijziging. Vervolgens simuleert de studie beide selectiemechanismen om subgroepen van leerlingen definiëren die wel of geen toegang zouden hebben tot de publieke scholen onder ieder beleid, en meet de effecten apart voor elk van deze groepen.

De resultaten laten zien dat leerlingen gemiddeld iets minder leerden onder het nieuwe beleid dan onder het oude beleid. De nieuwe laagscorende leerlingen in de publieke scholen leerden daar iets meer dan in de private scholen, maar hoogscorende leerlingen die niet langer toegang hadden tot publieke scholen leerden twee keer zo veel minder in private scholen. Ook leerlingen die in de publieke scholen bleven leerden minder nu zij in de klas zaten met lager scorende klasgenoten. Tegelijkertijd presteerden leerlingen die in private scholen bleven niet beter met hoger scorende klasgenoten. Aan de hand van vragenlijsten vindt de studie aanwijzingen dat leraren in publieke scholen hun lessen iets makkelijker hadden gemaakt om de laagscorende leerlingen tegemoet te komen. Ook vonden de leerlingen lessen in private scholen makkelijker dan in publieke scholen. Dit zou kunnen verklaren waarom hoogscorende leerlingen minder leerden onder het nieuwe beleid. De resultaten tonen aan dat het beleid ongelijkheid in vaardigheden heeft verminderd, maar vooral ten koste van de hoogscorende leerlingen.

Dit proefschrift laat zien dat het verbeteren van de kwaliteit en kansengelijkheid van het Indonesische onderwijssysteem nog een uitdaging is, ondanks de grote investeringen die de overheid gemaakt heeft in de afgelopen decennia. Om deze uitdaging te overwinnen is meer onderzoek nodig naar hoe onderwijsbeleid zich vertaald naar leeruitkomsten in de praktijk.

List of Authors

Schooling Progress, Learning Reversal: Indonesia's Learning Profiles Between 2000 and 2014

Main authors: Amanda Beatty[*] and Emilie Berkhout
Luhur Bima[‡] provided research assistance, and Menno Pradhan[†] and Daniel Suryadarma[‡§] mainly supervised. This chapter was published in the *International Journal of Educational Development* (volume 85, 102436) in 2021.

Using Technology to Prevent Fraud in High Stakes National School Examinations: Evidence from Indonesia

Main author: Emilie Berkhout
Menno Pradhan and Daniel Suryadarma supervised and helped writing, Rahmawati[**] developed the integrity index and shared the data, Arya Swarnata[‡] provided research assistance.

Who Benefits and Loses from Making Top Schools Less Selective? Evidence From a Large Change in Student Composition in Indonesian Schools

Main author: Emilie Berkhout
Goldy Dharmawan[‡**] was in charge of the data collection and provided research assistance, Amanda Beatty supervised and helped writing, Daniel Suryadarma and Menno Pradhan mainly supervised.

[*] Mathematica

[†] University of Amsterdam, Vrije Universiteit, AIGHD

[‡] SMERU Research Institute

[§] Asian Development Bank Institute. The views expressed are those of the authors and do not necessarily reflect the views of ADBI, ADB, its Board of Directors, or the governments they represent.

[**] Government of Indonesia